СОДЕРЖАНИЕ

Том 62, номер 5, 2022 год

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

МАТЕМАТИЧЕСКАЯ ФИЗИКА

EDN: WWAJNF
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 723–741

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 517.983.3+512.643.8

ОБ АЛГОРИТМЕ НАИЛУЧШЕГО ПРИБЛИЖЕНИЯ МАТРИЦАМИ МАЛОГО РАНГА В НОРМЕ ЧЕБЫШЁВА[1)]

© 2022 г. Н. Л. Замарашкин[1,*], С. В. Морозов[1,], Е. Е. Тыртышников[1,***]**

[1]119333 Москва, ул. Губкина, 8, Институт вычислительной математики им. Г.И. Марчука РАН, Россия

**e-mail: nikolai.zamarashkin@gmail.com*

***e-mail: stanis-morozov@yandex.ru*

****e-mail: eugene.tyrtyshnikov@gmail.com*

Поступила в редакцию 18.11.2021 г.
Переработанный вариант 18.11.2021 г.
Принята к публикации 16.12.2021 г.

Задача приближения матрицами малого ранга встречается в вычислительной математике повсеместно. Традиционно эта задача решается в спектральной или фробениусовой нормах, где эффективность приближения связана со скоростью убывания сингулярных чисел матрицы. Однако недавние результаты показывают, что в других нормах это требование не является необходимым. В данной работе предлагается метод решения задачи о приближении матрицами малого ранга в чебышёвской норме, который способен за приемлемое время строить эффективные приближения для матриц без убывания сингулярных чисел. Библ. 12. Фиг. 3.

Ключевые слова: приближение матрицами малого ранга, алгоритм Ремеза, чебышёвское приближение.

DOI: 10.31857/S0044466922050143

1. ВВЕДЕНИЕ

Матрицы малого ранга встречаются в науке повсеместно. Они находят многочисленные применения в вычислительной математике [1], вычислительной гидродинамике [2], рекомендательных системах [3], машинном обучении [4] и многих других задачах, как инструмент малопараметрического приближения матриц.

Однако в большинстве известных случаев для приближаемых матриц делается разной степени обоснованности дополнительное предположение о быстром убывании их сингулярных чисел, удобство которого в первую очередь связано с наличием эффективных алгоритмов построения оптимальных или близких к оптимальным приближений в унитарно инвариантных нормах [5], [6], [7].

С другой стороны, в современных приложениях, в особенности относящихся к области больших данных, часто более естественно использовать другие матричные нормы. Например, в классической схеме рекомендательных систем рассматривается матрица рейтингов, строки которой соответствуют фильмам, музыкальным произведениям или некоторым товарам, а столбцы — пользователям. Значения элементов матрицы определяют рейтинги, которые выставили пользователи. Для восстановления отсутствующих рейтингов и построения последующих рекомендаций строится малоранговое приближение матрицы по известным элементам. В этом случае более естественным представляется приближать матрицу не в спектральной или фробениусовой нормах, а поэлементно, стараясь наилучшим образом приблизить значения всех рейтингов. Более того, как следует из статьи [8], статистические модели описания рейтинговых матриц приводят, вообще говоря, к матрицам с медленным убыванием сингулярных чисел, но допускающим поэлементное приближение матрицами малого ранга. Последнее означает, что использование алгоритмов малоранговой аппроксимации матриц на основе сингулярных разложений в рекомендательных системах едва ли можно считать обоснованным.

[1)]Работа выполнена при финансовой поддержке РНФ (проект 21-71-10072).

Введем в рассмотрение норму (такую норму естественно называть чебышёвской)

$$\|X\|_C = \max_{i,j} |x_{ij}|,$$

для которой рассмотрим задачу построения наилучшего малорангового приближения. А именно, пусть задана матрица $A \in \mathbb{C}^{m\times n}$, целое r и требуется найти $U \in \mathbb{C}^{m\times r}$ и $V \in \mathbb{C}^{n\times r}$ такие, что

$$\mu = \inf_{U\in\mathbb{C}^{m\times r}, V\in\mathbb{C}^{n\times r}} \|A - UV^{\mathrm{T}}\|_C. \tag{1}$$

При этом матрицы $\widehat{U}$ и $\widehat{V}$, удовлетворяющие

$$\|A - \widehat{U}\widehat{V}^{\mathrm{T}}\|_C = \mu,$$

будем называть матрицами наилучшего приближения A ранга r.

Несмотря на естественность постановки ((1)), на сегодняшний день эта задача мало изучена. Известны асимптотические оценки на точность приближения ((1)) (см. [8]) и метод построения локальных минимумов задачи ((1)) в случае ранга 1 (см. [9]).

В настоящей работе предлагается и обосновывается алгоритм решения задачи

$$\mu = \inf_{U\in\mathbb{R}^{m\times r}} \|A - UV^{\mathrm{T}}\|_C$$

для произвольного ранга. На основе этого алгоритма развивается метод нахождения локальных минимумов задачи ((1)) для произвольного ранга. Большое количество численных экспериментов показывают, что асимптотические оценки, доказанные в [8], в общем случае не оптимальны.

Оставшаяся часть статьи организована следующим образом. В разд. 2 приведены известные в литературе результаты о рассматриваемой задаче. В разд. 3 приведены базовые результаты о свойствах решения задачи, включая вопросы существования, единственности и непрерывности решения. Кроме того, обсуждается вопрос о существовании и свойствах характеристических множеств, приводятся известные результаты о методах решения задачи с матрицей размера $(r+1)\times r$. Наконец, мы рассматриваем критерии определения оптимальности решений, которые будут полезны в дальнейшем. В разд. 4 приводится и обосновывается комбинаторная формула решения задачи, а также предлагается обобщенный алгоритм Ремеза, позволяющий на практике находить решения за полиномиальное число операций. В разд. 5 приводится алгоритм решения задачи в случае, когда обе матрицы U и V считаются неизвестными. Численные эксперименты из разд. 6 демонстрируют эффективность предложенного метода, а также открывают ряд новых вопросов об асимптотической точности приближений матриц в чебышёвской норме. Разд. 7 завершает работу.

2. СУЩЕСТВУЮЩИЕ РЕЗУЛЬТАТЫ

Насколько нам известно, задача построения и анализа малоранговых приближений матриц в чебышёвской норме исследована мало. Мы будем опираться на две работы [8], [9]. Первая из них содержит результаты об асимптотических свойствах чебышёвских приближений матриц (в отсутствие предположения об убывании сингулярных чисел), а вторая – метод нахождения локальных оптимумов в задаче (1) для ранга 1.

Остановимся на этих работах подробнее. Одним из результатов, доказанных в [8], является

Теорема 1. *Пусть $X \in \mathbb{R}^{m\times n}$, где $m \geq n$ и $0 < \varepsilon < 1$. Тогда при*

$$r = \lceil 72\log(2n+1)/\varepsilon^2 \rceil$$

имеем

$$\inf_{\mathrm{rank}\, Y \leq r} \|X - Y\|_C \leq \varepsilon \|X\|_2.$$

Из этой теоремы видно, что чебышёвские приближения малого ранга обладают большим потенциалом. Так, для любой последовательности матриц с ограниченной спектральной нормой, при фиксированной точности приближения ε, ранг чебышёвского приближения растет не более чем логарифмически. Например, при приближении единичной матрицы меньшим рангом в спектральной или фробениусовой норме, точность наилучшего приближения матрицы $n\times n$ рангом $n-1$ равна 1. В то же время в чебышёвской норме единичная матрица может быть при-

ближена с любой фиксированной точностью $\varepsilon > 0$ с рангом, который логарифмически возрастает с порядком матрицы n. В [8] это свойство чебышёвской нормы называется одной из основных причин, по которой матрицы, возникающие в анализе данных, могут быть эффективно приближены малым рангом.

Насколько нам известно, единственной статьей, в которой исследуется задача (1), является работа [9], где рассматривается случай приближения ранга 1. Для этого в [9] сначала решается задача вида

$$\mu = \inf_{u \in \mathbb{R}^m} \|A - uv^{\mathrm{T}}\|_C. \tag{2}$$

Легко видеть, что для каждой строки матрицы A задача (2) может быть решена независимо и, следовательно, сводится к задаче

$$\mu = \inf_{u \in \mathbb{R}} \|a - uv\|_\infty.$$

Отсюда вытекает простой алгоритм решения задачи (2). Для того чтобы получить локальный минимум решения задачи (1), авторы применили метод альтернанса. Пусть задан некоторый начальный вектор $v^{(0)}$, решая задачу (2) при фиксированном $v = v^{(0)}$, найдем решение $u^{(1)}$. Далее при фиксированном $u = u^{(1)}$ решим задачу

$$\mu = \inf_{v^{\mathrm{T}} \in \mathbb{R}^n} \|A - uv^{\mathrm{T}}\|_C$$

и найдем решение $v^{(1)}$. Продолжая по такой схеме, мы сойдемся к некоторому решению, которое, однако, не всегда является глобальным решением задачи (1). Кроме того, в [9] приведено необходимое условие оптимальности решения U, V задачи (1) для ранга 1. Для простоты формулировок предположим, что все элементы U и V ненулевые.

Утверждение 1. *Пусть все элементы векторов u и v ненулевые и они являются решением задачи* (1). *Пусть $R = A - uv^{\mathrm{T}}$. Тогда в матрице R существует цикл, т.е. набор индексов $(i_1, j_1), (i_1, j_2), (i_2, j_2), \ldots, (i_k, j_k), (i_k, j_1)$ такой, что*

1) *индексы $i_1, \ldots, i_k$ различны*;

2) *индексы $j_1, \ldots, j_k$ различны*;

3) *в каждой из этих позиций в матрице R достигается максимальное по модулю значение*;

4) *пусть (i_t, j_p) и (i_g, j_h) являются соседними в цикле, т.е. различными парами индексов такими, что $t = g$ или $p = h$. Тогда знаки величин $u_{i_t} v_{j_p} r_{i_t j_p}$ и $u_{i_g} v_{j_h} r_{i_g j_h}$ различны.*

Далее мы докажем обобщения этих результатов на случай произвольного ранга.

3. ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ

Приведем некоторые базовые результаты, которые пригодятся нам в дальнейшем. Значительная их часть является переложением известных результатов теории чебышёвских приближений функций на матричный случай [10], [11]. В этом разделе нас будет интересовать задача

$$\mu = \inf_{U \in \mathbb{C}^{m \times r}} \|A - UV^{\mathrm{T}}\|_C. \tag{3}$$

Матрицу $\widehat{U}$, удовлетворяющую

$$\|A - \widehat{U}V^{\mathrm{T}}\|_C = \mu,$$

будем называть матрицей *наилучшего приближения A* по системе векторов V. Задача состоит в том, чтобы по заданной матрице $A \in \mathbb{C}^{m \times n}$ и матрице $V \in \mathbb{C}^{n \times r}$ найти матрицу $\widehat{U} \in \mathbb{C}^{m \times r}$ наилучшего приближения. Легко понять, что эта задача разбивается на m независимых подзадач для каждой строки матрицы A, для которой требуется найти соответствующую строку матрицы $\widehat{U}$. Поэтому далее будем решать следующую задачу. Пусть заданы матрица $V \in \mathbb{C}^{n \times r}$ и вектор $a \in \mathbb{C}^n$. Требуется найти

$$\mu = \inf_{u \in \mathbb{C}^r} \|a - Vu\|_\infty \tag{4}$$

и вектор $\hat{u} \in \mathbb{C}^r$, удовлетворяющий

$$\left\|a - V\hat{u}\right\|_\infty = \mu.$$

3.1. Существование, единственность, непрерывность

Приведем результаты о существовании, единственности и непрерывности решения задачи (4). Существование решения, т.е. такого $\hat{u} \in \mathbb{C}^r$, что

$$\left\|a - V\hat{u}\right\|_\infty = \mu,$$

очевидно, поэтому перейдем сразу к вопросу о единственности решения. Здесь и далее будем обозначать нижним индексом столбец матрицы, а верхним индексом строку. Введем понятие *чебышёвской системы векторов*.

Определение 1. Столбцы матрицы $V \in \mathbb{C}^{n\times r}$ образуют чебышёвскую систему векторов, если любые r строк матрицы V линейно независимы.

Это понятие тесно связано с единственностью решения задачи наилучшего приближения. А именно, верна следующая

Теорема 2 (Хаар [10]). *Пусть матрица $V \in \mathbb{C}^{n\times r}$ и $n > r$. Тогда для того, чтобы для любого вектора $a \in \mathbb{C}^n$ существовал единственный вектор $\hat{u} \in \mathbb{C}^r$ ее наилучшего равномерного приближения необходимо и достаточно, чтобы столбцы матрицы V образовывали чебышёвскую систему.*

Также верно

Утверждение 2. *Пусть столбцы матрицы $V \in \mathbb{C}^{n\times r}$, где $n > r$, образуют чебышёвскую систему и $V\hat{u}$ — вектор наилучшего равномерного приближения. Тогда, по крайней мере, в $r+1$ точке достигается максимальное значение, т.е. существуют $i_1,\ldots,i_{r+1}$, в которых выполняется равенство*

$$|a_{i_j} - (V\hat{u})_{i_j}| = \|a - V\hat{u}\|_\infty, \quad j = 1,\ldots,r+1.$$

Доказательство. Пусть точек, в которых достигается максимальное по модулю значение $r_1 < r+1$. Тогда решив систему с r_1 уравнениями и r неизвестными, строки которой линейно независимы, мы можем получить вектор $p \in \mathbb{C}^r$ такой, что

$$(Vp)_{i_j} = a_{i_j} - (V\hat{u})_{i_j}, \quad j = 1,\ldots,r_1.$$

Но тогда вектор $V(\hat{u} + \delta p)$ при достаточно малом δ отклоняется от a меньше, чем $V\hat{u}$. Пришли к противоречию.

Рассмотрим вопрос о непрерывности решения.

Теорема 3 (Никольский [10]). *Пусть система столбцов матрицы $V \in \mathbb{C}^{n\times r}$, где $n > r$, является чебышёвской. Тогда коэффициенты вектора наилучшего равномерного приближения $\hat{u}$ непрерывно зависят от приближаемого вектора a, и системы столбцов V, т.е. $\forall\varepsilon > 0\ \exists\delta = \delta(a,V,\varepsilon) > 0$ такое, что если $\|a-b\|_\infty + \|V-W\| < \delta$, то $\|\hat{u}(a,V) - \hat{u}(b,W)\|_\infty < \varepsilon$, где через $\hat{u}(a,V)$ и $\hat{u}(b,W)$ обозначены коэффициенты оптимального решения для векторов a по системе V и b по системе W соответственно.*

3.2. Характеристические множества

Пусть J обозначает набор индексов $J = \{1,2,\ldots,n\}$, а J' и J'' — некоторые подмножества J. Пусть

$$\mu(J') = \inf_{u\in\mathbb{C}^r} \left\|a(J') - V(J')u\right\|_\infty,$$

где через $V(J')$ обозначена подматрица матрицы V, содержащая строки с номерами из множества J', а через $a(J')$ обозначен подвектор вектора a, содержащий элементы с номерами из J'.

Определение 2. Множество J' называется *характеристическим множеством*, если $\mu(J) = \mu(J')$ и для любого подмножества $J'' \subsetneq J'$ $\mu(J'') < \mu(J)$.

Далее мы докажем, что если система столбцов матрицы V линейно независима, то существует по крайней мере одно характеристическое множество, содержащее не более $2r+1$ точек в комплексном и не более $r+1$ точек в вещественном случае.

Для дальнейшего нам понадобятся следующие обозначения [10], [11]. Пусть $\lambda \geq 0$ и пусть

$$F(j,u) = \left|a_j - u^{\mathrm{T}} v^j\right|,$$

$$K(j,\lambda) = \left\{u \in \mathbb{C}^r \mid F(j,u) \leq \lambda\right\},$$

$$K(J',\lambda) = \bigcap_{j \in J'} K(j,\lambda) = \left\{u \in \mathbb{C}^r \mid F(j,u) \leq \lambda, \forall j \in J'\right\}.$$

Утверждение 3. *Верны следующие вложения*

$$K(j,\lambda') \subset K(j,\lambda''), \quad K(J',\lambda') \subset K(J',\lambda''), \quad 0 \leq \lambda' < \lambda'',$$

$$K(J'',\lambda) \subset K(J',\lambda), \quad J' \subset J''.$$

Лемма 1. *Пусть столбцы матрицы V линейно независимы. Тогда множество $K(j,\lambda)$ выпукло и замкнуто, а множество $K(J,\lambda)$ ограничено для любых $\lambda \geq 0$.*

Доказательство. Докажем замкнутость множеств $K(j,\lambda)$. Пусть u_1 — предельная точка множества $K(j,\lambda)$. Тогда в любой ее окрестности существуют точки $u \in K(j,\lambda)$

$$F(j,u_1) \leq |F(j,u_1) - F(j,u)| + F(j,u).$$

Так как $u \in K(j,\lambda)$, то $F(j,u) \leq \lambda$. Функция

$$F(j,u) = |a_j - u^{\mathrm{T}} v^j|$$

очевидно непрерывна по u при любом фиксированном j. Тогда для любого $\varepsilon > 0$ существует $\delta > 0$ такая, что если $|u - u_1| < \delta$, то $|F(j,u_1) - F(j,u)| < \varepsilon$. Таким образом, имеем

$$F(j,u_1) < \varepsilon + \lambda$$

для любого $\varepsilon > 0$, откуда $u_1 \in K(j,\lambda)$. Замкнутость доказана.

Докажем ограниченность $K(J,\lambda)$. Рассмотрим вектор Vu при $\|u\|_1 = 1$. Величина $\|Vu\|_\infty$ является непрерывной по u функцией на компактном множестве, поэтому достигает в некоторой точке своего минимального значения

$$M = \|V\hat{u}\|_\infty \leq \|Vu\|_\infty.$$

Поскольку столбцы V линейно независимы, любая их нетривиальная линейная комбинация не равна 0 и $M > 0$.

Пусть $\|u\|_1 \geq \dfrac{C+1}{M}$, где $C > 0$ — некоторая константа. Тогда

$$\|a - Vu\|_\infty \geq \|Vu\|_\infty - \|a\|_\infty \geq \|u\|_1 M - \|a\|_\infty \geq C + 1 - \|a\|_\infty.$$

Тогда при $\|u\|_1 \geq \dfrac{C+1}{M}$ не может быть выполнено условие $F(J,u) \leq \lambda = C - \|a\|_\infty$, т.е. $u \notin K(j,\lambda)$ при $\lambda \leq C - \|a\|_\infty$. В силу произвольности C ограниченность доказана.

Докажем выпуклость множества $K(j,\lambda)$. Пусть $u_1, u_2 \in K(j,\lambda)$, т.е. $F(j,u_1) \leq \lambda$ и $F(j,u_2) \leq \lambda$. Нужно доказать, что $F(j,\tau u_1 + (1-\tau)u_2) \leq \lambda$ для любого $\tau \in (0,1)$:

$$F(j,\tau u_1 + (1-\tau)u_2) = |a_j - (\tau u_1 + (1-\tau)u_2)^{\mathrm{T}} v^j| =$$

$$= |\tau(a_j - u_1^{\mathrm{T}} v^j) + (1-\tau)(a_j - u_2^{\mathrm{T}} v^j)| \leq$$

$$\leq \tau F(j,u_1) + (1-\tau)F(j,u_2) \leq \lambda.$$

Обозначим через M_k множество всевозможных упорядоченных подмножеств J_k, состоящих из k элементов $i_1,\ldots,i_k$, взятых из множества J. Обозначим через $\mu_k(J)$ точную верхнюю грань наименьших отклонений от нуля функции $F(j,u)$ на всевозможных подмножествах $J_k \in M_k$:

$$\mu_k(J) = \max_{J_k \in M_k} \mu(J_k) = \max_{J_k \in M_k} \min_{u \in \mathbb{C}^r} \max_{j \in J_k} F(j,u).$$

Легко доказать

Утверждение 4. *Верно неравенство* $\mu_k(J) \le \mu_{k+1}(J) \le \mu(J) \quad \forall k$.

Для дальнейшего анализа понадобится следующая

Теорема 4 (Хелли). *Если множество K замкнутых выпуклых множеств точек $x \in \mathbb{R}^r$ содержит не менее $r+1$ множеств* (*среди которых могут быть одинаковые*), *пересечение любых $r+1$ множеств из K не пусто и пересечение некоторого конечного числа множеств из K ограничено*, *то пересечение всех множеств из K не пусто.*

Основываясь на теореме Хелли, докажем следующий результат.

Теорема 5 (Шнирельман [10]). *Если существует такое $\lambda_0 > \mu_{2r+1}$ ($\lambda_0 > \mu_{r+1}$ в вещественном случае*), *что для любого j и любом λ, $\mu_{2r+1} < \lambda < \lambda_0$ ($\mu_{r+1} < \lambda < \lambda_0$ в вещественном случае*), *множество $K(j,\lambda)$ замкнуто и выпукло*, *и пересечение некоторого конечного числа множеств $K(j,\lambda)$ ограничено*, *то $\mu_{2r+1}(J) = \mu(J)$ ($\mu_{r+1}(J) = \mu(J)$ в вещественном случае*).

Доказательство. Пусть $k = r+1$ в вещественном случае и $k = 2r+1$ в комплексном. В силу конечности множеств J_k и M_k имеем, что максимумы и минимумы достигаются, поэтому при любом $\lambda > \mu_k(J)$ множество $K(J_k,\lambda)$ не пусто. Кроме того, так как $K(j,\lambda)$ по условию выпукло и замкнуто, то

$$K(J_k,\lambda) = \bigcap_{j \in J_k} K(j,\lambda)$$

не пусто и выпукло для любого $J_k \in M_k$.

Убедимся, что выполнены все условия теоремы Хелли. В качестве совокупности множеств K возьмем множества $K(j,\lambda)$, $j \in J$. По условию теоремы они замкнуты и выпуклы, и пересечение некоторого конечного числа этих множеств ограничено. В вещественном случае то, что пересечение любых $j+1$ множеств не пусто, эквивалентно тому, что $K(J_{r+1},\lambda)$ не пусто, что было показано выше. В комплексном случае нам нужно работать с пространством $\mathbb{C}^r$, которое мы отождествим с $\mathbb{R}^{2r}$, поэтому нам требуется, чтобы пересечение любых $2r+1$ множеств было не пусто, что также было показано выше. Итак, все условия теоремы Хелли выполнены и мы имеем, что

$$K(J,\lambda) = \bigcap_{j \in J} K(j,\lambda)$$

не пусто, замкнуто, выпукло и ограничено.

Пусть последовательность $\{\lambda_t\}$ убывающая, $\mu_k < \lambda_t < \lambda_0$ и стремится к μ_k. При этом $K(J,\lambda_{t+1}) \subset K(J,\lambda_t)$ и пересечение

$$K = \bigcap_{t=1}^{\infty} K(J,\lambda_t)$$

не пусто, замкнуто и ограничено. Пусть $u_0 \in K$. Это значит, что $u_0 \in K(J,\lambda_t)$, что значит, что $F(j,u_0) \le \lambda_t$ для любых $j \in J$ и любого t. В пределе при $t \to \infty$ получаем, что $F(j,u_0) \le \mu_k(J)$ для любого $j \in J$, откуда

$$\mu(J) \le \max_{j \in J} F(j,u_0) \le \mu_k(J).$$

Но, как было отмечено выше, $\mu_k(J) \le \mu(J)$.

Эта теорема позволяет сформулировать следующий результат.

Теорема 6. *Пусть столбцы матрицы $V \in \mathbb{C}^{n\times r}$, где $n \ge r$ линейно независимы и вектор a не принадлежит образу матрицы V. Тогда существует по крайней мере одно характеристическое множество*, *состоящее не более чем из $2r+1$ точек в комплексном случае и $r+1$ точек в вещественном. Кроме то-*

го, если система столбцов матрицы V является чебышёвской, то любое характеристическое множество состоит не менее чем из $r+1$ точек.

Доказательство. Результат для произвольной системы сразу следует из леммы 0 и теоремы 5. В случае чебышёвской системы на любом множестве из r и менее точек можно решить систему и точно приблизить вектор в этих точках, а поскольку множество является характеристическим, то это противоречит условию, что a не принадлежит образу V.

3.3. О задаче поиска равноудаленных точек

Введем понятие равноудаленной точки системы.

Определение 3. Пусть $V \in \mathbb{R}^{(r+1)\times r}$ и $a \in \mathbb{R}^{r+1}$. Пусть система

$$Vu = a$$

несовместна. Будем называть точку u *равноудаленной точкой системы*, если

$$\rho(u) = |(v^1, u) - a_1| = |(v^2, u) - a_2| = \ldots = |(v^{r+1}, u) - a_{r+1}|.$$

Будем называть точку u *наилучшей равноудаленной точкой системы*, если она является равноудаленной и величина $\rho(u)$ минимальна.

Приведем результаты [12] о том, как устроено множество всех равноудаленных точек системы в вещественном случае.

Пусть

$$\hat{V}_j = \begin{bmatrix} v_1^1 & v_2^1 & \ldots & v_r^1 \\ v_1^2 & v_2^2 & \ldots & v_r^2 \\ \vdots & \vdots & \vdots & \vdots \\ v_1^{j-1} & v_2^{j-1} & \ldots & v_r^{j-1} \\ v_1^{j+1} & v_2^{j+1} & \ldots & v_r^{j+1} \\ \vdots & \vdots & \vdots & \vdots \\ v_1^{r+1} & v_2^{r+1} & \ldots & v_r^{r+1} \end{bmatrix}$$

и $\hat{a}_j = (a_1, a_2, \ldots, a_{j-1}, a_{j+1}, \ldots, a_{r+1})^{\mathrm{T}}$. Тогда обозначим через $D_j = \det \hat{V}_j$ и $\hat{u}^j$ решение системы $\hat{V}_j u = \hat{a}_j$. Верны следующие теоремы о множестве равноудаленных точек системы.

Теорема 7 (Дзядык [12]). *Пусть задана несовместная система уравнений $Vu = a$, где $V \in \mathbb{R}^{(r+1)\times r}$ и $a \in \mathbb{R}^{r+1}$. Тогда верно следующее.*

1) *При каждом $j = 1, \ldots, r+1$ имеет место равенство*

$$(v^j, \hat{u}^j) - a_j = \frac{(-1)^{j+1}}{D_j} \sum_{\nu=1}^{r+1} (-1)^{\nu} a_{\nu} D_{\nu}.$$

2) *Для любых действительных k_j, $j = 1, \ldots, r+1$, таких, что*

$$\sum_{j=1}^{n+1} |D_j| e^{ik_j} \neq 0$$

точка u, определяемая по формуле

$$u = \rho \sum_{j=1}^{r+1} \frac{\hat{u}^j e^{ik_j}}{|(v^j, \hat{u}^j) - a_j|} = \rho \frac{\sum\limits_{j=1}^{r+1} |D_j| \hat{u}^j e^{ik_j}}{\left| \sum\limits_{j=1}^{r+1} (-1)^j D_j a_j \right|},$$

где

$$\rho = \left(\sum_{j=1}^{r+1}\frac{e^{ik_j}}{|(v^j,\hat{u}^j)-a_j|}\right)^{-1} = \frac{\left|\sum_{j=1}^{r+1}(-1)^j D_j a_j\right|}{\sum_{j=1}^{r+1}|D_j|e^{ik_j}}$$

является равноудаленной точкой системы $Vu = a$, при этом $|\rho|$ выражает V-расстояние от точки и до a.

3. *Всякая V-равноудаленная точка и системы $Vu = a$ может быть при некоторых действительных k_j представлена по формуле выше. При этом k_j могут, в частности, быть выражены по формуле*

$$k_j = \arg((v^j,u)-a_j) - \arg((v^j,\hat{u}^j)-a_j).$$

Теорема 8 (Дзядык [12]). *Пусть задана несовместная система уравнений $Vu = a$, где $V \in \mathbb{R}^{(r+1)\times r}$ и $a \in \mathbb{R}^{r+1}$. Тогда наилучшая равноудаленная от системы точка и определяется по формуле*

$$u^* = \rho^*\sum_{j=1}^{r+1}\frac{\hat{u}^j}{|(v^j,\hat{u}^j)-a_j|} = \frac{\sum_{j=1}^{r+1}|D_j|\hat{u}^j}{\sum_{j=1}^{r+1}|D_j|},$$

где

$$\rho^* = \left(\sum_{j=1}^{r+1}\frac{1}{|(v^j,\hat{u}^j)-a_j|}\right)^{-1} = \frac{\left|\sum_{j=1}^{r+1}(-1)^j D_j a_j\right|}{\sum_{j=1}^{r+1}|D_j|}.$$

Доказательство. Достаточно заметить, что величина

$$\rho = \left(\sum_{j=1}^{r+1}\frac{e^{ik_j}}{|(v^j,\hat{u}^j)-a_j|}\right)^{-1}$$

принимает наименьшее значение, когда все слагаемые сонаправлены, т.е. $e^{ik_1} = e^{ik_2} = \ldots = e^{ik_{r+1}}$.

3.4. Критерии оптимальности

Приведем несколько критериев оптимальности приближений. Эти критерии интересны как сами по себе, так и позволяют получить новые важные сведения о задаче.

Теорема 9 (Колмогоров). *Пусть заданы система векторов с матрицей $V \in \mathbb{C}^{n\times r}$ и вектор $a \in \mathbb{C}^n$, который следует приблизить линейной комбинацией столбцов V. Для того чтобы вектор $V\hat{u}$ был для a вектором наилучшего равномерного приближения, необходимо и достаточно, чтобы на множестве $E = E(V\hat{u})$ всех точек, на которых для вектора $V\hat{u}$ достигается максимальное по модулю значение разности, для всех векторов вида Vu выполнялось*

$$\min_{j\in E}\operatorname{Re}((Vu)_j\overline{(a_j-(V\hat{u})_j)}) \le 0.$$

Доказательство. Необходимость. Пусть $V\hat{u}$ является вектором наилучшего равномерного приближения для a. От противного, пусть

$$\min_{j\in E}\operatorname{Re}((Vu)_j\overline{(a_j-(V\hat{u})_j)}) > c > 0$$

для некоторого вектора $u \in \mathbb{C}^r$. Пусть

$$G = \max_{j\in E} |a_j - (V\hat{u})_j|, \quad G' = \max_{j\notin E} |a_j - (V\hat{u})_j|,$$

$$h = G - G' > 0, \quad M = \max_j |(Vu)_j|, \quad \lambda = \max\left\{\frac{c}{M^2}, \frac{h}{2M}\right\} > 0.$$

Докажем, что тогда вектор $V(\hat{u} + \lambda u)$ приближает вектор a лучше.

1. Пусть $j \in E$. Тогда

$$|a_j - (\hat{u} + \lambda u)^{\mathrm{T}} v^j|^2 = (a_j - \hat{u}^{\mathrm{T}} v^j - \lambda u^{\mathrm{T}} v^j) \cdot \overline{((a_j - \hat{u}^{\mathrm{T}} v^j) - \lambda(u^{\mathrm{T}} v^j))} =$$

$$= |a_j - \hat{u}^{\mathrm{T}} v^j|^2 + \lambda^2 |u^{\mathrm{T}} v^j|^2 - 2\lambda \operatorname{Re}((Vu)_j \overline{(a_j - (V\hat{u})_j)}) \le$$

$$\le G^2 + \lambda^2 M^2 - 2\lambda \operatorname{Re}((Vu)_j \overline{(a_j - (V\hat{u})_j)}) <$$

$$< G^2 + \lambda^2 M^2 - 2\lambda c \le G^2 + \lambda \frac{c}{M^2} M^2 - 2\lambda c =$$

$$= G^2 - \lambda c < G^2.$$

2. Пусть $j \notin E$. Тогда имеем

$$|a_j - (\hat{u} + \lambda u)^{\mathrm{T}} v^j| \le |a_j - \hat{u}^{\mathrm{T}} v^j| + \lambda |u^{\mathrm{T}} v^j| \le G' + \lambda M \le G - h + \frac{h}{2M} M = G - h/2 < G.$$

Отсюда следует, что вектор $V(\hat{u} + \lambda u)$ приближает вектор a лучше. Получили противоречие.

Достаточность. Пусть выполнено условие критерия Колмогорова с вектором коэффициентов $\hat{u}$ и пусть $u \in \mathbb{C}^r$ произвольный. Рассмотрим вектор $w = V(u - \hat{u})$. Выберем индекс j_0, для которого выполняется неравенство для вектора w:

$$\operatorname{Re}((V(u - \hat{u}))_{j_0} \overline{(a_{j_0} - (V\hat{u})_{j_0})}) \le 0.$$

Тогда имеем

$$|a_{j_0} - (Vu)_{j_0}|^2 = |a_{j_0} - (V\hat{u})_{j_0} - ((Vu)_{j_0} - (V\hat{u})_{j_0})|^2 =$$

$$= |a_{j_0} - (V\hat{u})_{j_0}|^2 + |(Vu)_{j_0} - (V\hat{u})_{j_0}|^2 - 2\operatorname{Re}((V(u - \hat{u}))_{j_0} \overline{(a_{j_0} - (V\hat{u})_{j_0})}) \ge$$

$$\ge |a_{j_0} - (V\hat{u})_{j_0}|^2.$$

Отсюда видно, что для любого вектора $u \in \mathbb{C}^r$, вектор $\hat{u}$ дает приближение не хуже, т.е. является оптимальным.

Приведем другой, в некоторых ситуациях более удобный критерий оптимальности. В некотором смысле он является переформулировкой критерия Колмогорова с использованием следующей леммы.

Лемма 2. *Пусть u_{ij}, $i = 1,\ldots,m$, $j = 1,\ldots,n$ – некоторые числа. Для того чтобы существовали числа $\delta_i \ge 0$, $i = 1,\ldots,m$, не все равные нулю и такие, что*

$$\sum_{i=1}^{m} \delta_i u_{ij} = 0, \quad j = 1,\ldots,n,$$

необходимо и достаточно, чтобы для любой системы чисел c_j, $j = 1,\ldots,n$, неравенства

$$\operatorname{Re} \sum_{j=1}^{n} c_j u_{ij} > 0, \quad i = 1,\ldots,m,$$

не выполнялись одновременно.

Доказательство. Необходимость. Пусть при $\delta_i \geq 0$, $i = 1,\ldots,m$, выполнено

$$\sum_{i=1}^{m}\delta_i u_{ij} = 0, \quad j = 1,\ldots,n.$$

Тогда имеем

$$\sum_{i=1}^{m}\delta_i \operatorname{Re}\sum_{j=1}^{n}c_j u_{ij} = \operatorname{Re}\sum_{j=1}^{n}c_j\sum_{i=1}^{m}\delta_i u_{ij} = 0,$$

откуда следует, что для любой системы c_j условия

$$\operatorname{Re}\sum_{j=1}^{n}c_j u_{ij} > 0, \quad i = 1,\ldots,m,$$

не могут быть выполнены одновременно.

Достаточность. Введем величину

$$v = v(\delta_1,\ldots,\delta_m) = \sum_{j=1}^{n}\left|\sum_{i=1}^{m}\delta_i u_{ij}\right|, \quad \delta_i \geq 0, \quad \sum_{i=1}^{m}\delta_i = 1.$$

Так как функция v непрерывна на компактном множестве, то она принимает минимальное значение v_0 при $\delta_i = \delta_i^0$. Покажем, что условие леммы эквивалентно тому, что если неравенства $\operatorname{Re}\sum_{j=1}^{n}c_j u_{ij} > 0$ не могут быть выполнены одновременно, то $v_0 = 0$.

Докажем это от противного. Пусть неравенства не могут быть выполнены одновременно ни при каком выборе c_j, но $v_0 > 0$. Возьмем

$$c_j = \sum_{i=1}^{m}\delta_i^0\overline{u}_{ij}$$

и пусть при таком выборе c_j, не ограничивая общности, не выполнено неравенство при $i = m$:

$$\operatorname{Re}\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\delta_i^0\overline{u}_{ij}\right)u_{mj} \leq 0.$$

Пусть

$$v_* = \sum_{j=1}^{m}|u_{mj}|^2, \quad \lambda = \frac{v_*}{v_* + v_0} < 1.$$

Выберем δ_i следующим образом:

$$\delta_i = \begin{cases}\lambda\delta_i^0, & i = 1,2,\ldots,m-1,\\ (1-\lambda) + \lambda\delta_m^0, & i = m,\end{cases}$$

и покажем, что $v(\delta_1,\ldots,\delta_m) < v_0$. Действительно,

$$v = \sum_{j=1}^{n}\left|\sum_{i=1}^{m}\delta_i u_{ij}\right| = \sum_{j=1}^{n}\left|(1-\lambda)u_{mj} + \lambda\sum_{i=1}^{m}\delta_i^0 u_{ij}\right| =$$

$$= (1-\lambda)^2\sum_{j}|u_{mj}|^2 + \lambda^2\sum_{j=1}^{n}\left|\sum_{i=1}^{m}\delta_i^0 u_{ij}\right| + 2\lambda(1-\lambda)\operatorname{Re}\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\delta_i^0\overline{u}_{ij}u_{mj}\right).$$

Как было отмечено выше,

$$\operatorname{Re}\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\delta_i^0\overline{u}_{ij}u_{mj}\right) \leq 0,$$

но $\lambda \geq 0, 1-\lambda > 0$, откуда

$$v \leq (1-\lambda)^2 v_* + \lambda^2 v_0 = \frac{v_0^2}{(v_* + v_0)^2} v_* + \frac{v_*^2}{(v_* + v_0)^2} v_0 = v_0 v_* \frac{v_* + v_0}{(v_* + v_0)^2} = \frac{v_*}{v_* + v_0} v_0 = \lambda v_0 < v_0.$$

Пришли к противоречию с оптимальностью v_0, следовательно, $v_0 = 0$.

Замечание 1. Если все $u_{ij} \in \mathbb{R}$, то c_j достаточно выбирать вещественными.

Используя эту лемму и критерий Колмогорова, докажем другой критерий оптимальности поэлементного приближения

Теорема 10 (Ремез). *Пусть заданы система векторов с матрицей $V \in \mathbb{C}^{n\times r}$ и вектор $a \in \mathbb{C}^n$, который следует приблизить линейной комбинацией столбцов V. Пусть вектор $V\hat{u}$ достигает максимальных по модулю значений разности в позициях $E = \{i_1, \ldots, i_t\}$. Тогда $V\hat{u}$ является вектором наилучшего равномерного приближения для a, тогда и только тогда, когда существуют $\delta_k \geq 0$, $k = 1, \ldots, t$, не все из которых равны нулю, такие, что*

$$\sum_{k\in E} \delta_k \overline{(a_k - \hat{u}^{\mathrm{T}} v^k)} v_j^k = 0, \quad j = 1, \ldots, r.$$

Доказательство. Достаточность. Пусть выполнены условия

$$\sum_{k\in E} \delta_k \overline{(a_k - \hat{u}^{\mathrm{T}} v^k)} v_j^k = 0, \quad j = 1, \ldots, r.$$

Пусть $u_{ij} = \overline{(a_k - \hat{u}^{\mathrm{T}} v^k)} v_j^k$ в терминах предыдущей леммы. Тогда, согласно лемме, условия

$$\operatorname{Re} \sum_{j\in E} c_j v_j^k \overline{(a_k - \hat{u}^{\mathrm{T}} v^k)} > 0$$

не выполнены одновременно для любого выбора c_j. Принимая во внимание, что $\sum_{j\in E} c_j v_j^k$ задает произвольный вектор вида Vc на строках из E, получаем, что выполнен критерий Колмогорова и $V\hat{u}$ является оптимальным приближением.

Необходимость. Если $V\hat{u}$ – оптимальный, то выполнен критерий Колмогорова:

$$\min_{k\in E} \operatorname{Re} \sum_{j\in E} c_j v_j^k \overline{(a_k - \hat{u}^{\mathrm{T}} v^k)} \leq 0,$$

следовательно,

$$\operatorname{Re} \sum_{j\in E} c_j v_j^k \overline{(a_k - \hat{u}^{\mathrm{T}} v^k)} > 0$$

не выполнены одновременно и тогда по лемме существуют $\delta_k \geq 0$, удовлетворяющие условиям Ремеза.

Замечание 2. Всегда существует оптимальное решение, в котором максимальные значения достигаются в t точках, где $1 \leq t \leq 2r+1$ в комплексном случае и $1 \leq t \leq r+1$ в вещественном, причем условие критерия Ремеза выполняется с $\delta_k > 0$.

Замечание 3. Заметим, что условие Ремеза может быть переписано в виде

$$\sum_{k\in E} \delta_k \operatorname{sign}\{\overline{(a_k - \hat{u}^{\mathrm{T}} v^k)}\} v_j^k = 0, \quad j = 1, \ldots, r.$$

Пусть мы каким-либо образом нашли множество E. Тогда, используя критерий Ремеза, легко найти решение. Решим систему

$$\sum_{k\in E} v_j^k s_k = 0, \quad j = 1, \ldots, r.$$

Это система с $r+1$ переменной и r уравнениями, которая имеет нетривиальное решение. Найдем это решение. Тогда

$$\delta_k = |s_k|, \quad \operatorname{sign}\{\overline{(a_k - \hat{u}^{\mathrm{T}} v^k)}\} = \operatorname{sign} s_k. \tag{5}$$

Поскольку чебышёвская система векторов всегда имеет характеристическое множество, состоящее ровно из $r+1$ элементов, для чебышёвской системы из $r+1$ уравнений и r неизвестных, уравнение (5) задает знаки величин $\overline{(a_k - \hat{u}^{\mathrm{T}} v^k)}$ для наилучшей равноудаленной точки. Кроме того, модули $|a_k - \hat{u}^{\mathrm{T}} v^k|$ равны, что позволяет в вещественном случае выписать систему из r с r неизвестными, решением которой будет наилучшая равноудаленная точка. Таким образом, наилучшая равноудаленная точка чебышёвской системы $(r+1)\times r$ может быть найдена за $O(r^3)$ операций с помощью двух решений систем линейных уравнений.

4. О ЗАДАЧЕ ПОИСКА ХАРАКТЕРИСТИЧЕСКИХ МНОЖЕСТВ В ВЕЩЕСТВЕННОМ СЛУЧАЕ

Перейдем к вопросу о методах решения задачи (4).

4.1. Комбинаторная формула решения

Пусть решается задача приближения вектора $a \in \mathbb{R}^n$ по системе векторов $V \in \mathbb{R}^{n\times r}$. Пусть $\widetilde{V} \in \mathbb{R}^{(r+1)\times r}$ и $\tilde{a} \in \mathbb{R}^{r+1}$. Тогда через $\left[\widetilde{V}\ \ \tilde{a}\right] \in \mathbb{R}^{(r+1)\times(r+1)}$ мы обозначим матрицу, первые r столбцов которой являются столбцами матрицы $\widetilde{V}$, а последний столбец $\tilde{a}$. Обозначим через $V(i_1,\ldots,i_k)$ подматрицу матрицы V, содержащую строки $i_1,\ldots,i_k$. Аналогично обозначим через $a(i_1,\ldots,i_k)$ подвектор вектора a, содержащий элементы $i_1,\ldots,i_k$. Кроме того, обозначим через $\widetilde{V}_{\setminus k}$ подматрицу матрицы $\widetilde{V}$, в которой удалена k-я строка.

Теорема 11. *Пусть решается задача наилучшего равномерного приближения вектора $a \in \mathbb{R}^n$ по чебышёвской системе векторов $V \in \mathbb{R}^{n\times r}$. Пусть*

$$\mu = \inf_{u\in\mathbb{R}^r} \|a - Vu\|_\infty .$$

Тогда получаем

$$\mu = \max_{i_1,i_2,\ldots,i_{r+1}} \frac{\left|\det\left[V(i_1,i_2,\ldots,i_{r+1})\ \ a(i_1,i_2,\ldots,i_{r+1})\right]\right|}{\sum\limits_{k=1}^{r+1}\left|\det\left(V(i_1,i_2,\ldots,i_{r+1})_{\setminus k}\right)\right|}.$$

Доказательство. По теореме 5 имеем, что

$$\mu(J) = \mu_{r+1}(J),$$

а согласно определению верно

$$\mu_{r+1}(J) = \max_{J_{r+1}\in M_{r+1}} \mu(J_{r+1}).$$

Остается применить теорему 8 для получения явного вида $\mu(J_{r+1})$.

4.2. Аналог теоремы о чебышёвском альтернансе

В утверждении 2 было показано, что для оптимального решения в векторе невязки достигаются максимальные по модулю значения по крайней мере в $r+1$ позициях. Широко известен результат о чебышёвском альтернансе для непрерывных функций. В этом случае помимо того, что в векторе невязки достигаются максимальные по модулю значения, имеет место чередование знаков. Аналогичный результат может быть доказан и в матричном случае.

Лемма 3. *Пусть решается задача наилучшего равномерного приближения вектора $a \in \mathbb{R}^{n+1}$ по чебышёвской системе векторов $V \in \mathbb{R}^{(n+1)\times n}$ и вектор z^* является вектором наилучшего равномерного*

приближения (*наилучшей равноудаленной точкой системы*). *Обозначим через* $w = a - Vz^*$ *вектор невязки. Тогда знаки величин*

$$w_1 D_1, w_2 D_2, \ldots, w_{n+1} D_{n+1} \tag{6}$$

чередуются.

Доказательство. В обозначениях теорем 7 и 8 имеем

$$z^* = \sum_{j=1}^{n+1} \frac{|D_j|}{\sum\limits_{\nu=1}^{n+1} |D_\nu|} z^j.$$

Кроме того, из теоремы 7 имеем

$$(v^j, z^j) - a_j = \frac{(-1)^{j+1}}{D_j} \sum_{\nu=1}^{n+1} (-1)^\nu a_\nu D_\nu = \frac{(-1)^{j+1}}{D_j} X,$$

где

$$X = \sum_{\nu=1}^{n+1} (-1)^\nu a_\nu D_\nu$$

не зависит от j. Согласно определению z^j получаем

$$Vz_j = \left[a_1 \;\; a_2 \;\; \ldots \;\; a_{j-1} \;\; \tilde{a}_j \;\; a_{j+1} \;\; \ldots \;\; a_{n+1} \right]^{\mathrm{T}},$$

где $\tilde{a}_j = (v^j, z^j)$. Тогда имеем

$$Vz^* = \sum_{j=1}^{n+1} \frac{|D_j|}{\sum\limits_{\nu=1}^{n+1} |D_\nu|} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{j-1} \\ \tilde{a}_j \\ a_{j+1} \\ \vdots \\ a_{n+1} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{j-1} \\ a_j \\ a_{j+1} \\ \vdots \\ a_{n+1} \end{bmatrix} - \frac{1}{\sum\limits_{\nu=1}^{n+1} |D_\nu|} \left(\begin{bmatrix} |D_1|\tilde{a}_1 \\ |D_2|\tilde{a}_2 \\ \vdots \\ |D_{j-1}|\tilde{a}_{j-1} \\ |D_j|\tilde{a}_j \\ |D_{j+1}|\tilde{a}_{j+1} \\ \vdots \\ |D_{n+1}|\tilde{a}_{n+1} \end{bmatrix} - \begin{bmatrix} |D_1|a_1 \\ |D_2|a_2 \\ \vdots \\ |D_{j-1}|a_{j-1} \\ |D_j|a_j \\ |D_{j+1}|a_{j+1} \\ \vdots \\ |D_{n+1}|a_{n+1} \end{bmatrix} \right).$$

Отсюда

$$a - Vz^* = \frac{1}{\sum\limits_{\nu=1}^{n+1} |D_\nu|} \begin{bmatrix} |D_1|(a_1 - \tilde{a}_1) \\ |D_2|(a_2 - \tilde{a}_2) \\ \vdots \\ |D_{n+1}|(a_{n+1} - \tilde{a}_{n+1}) \end{bmatrix} = \frac{1}{\sum\limits_{\nu=1}^{n+1} |D_\nu|} \cdot \begin{bmatrix} |D_1| \dfrac{(-1)^1}{D_1} X \\ |D_2| \dfrac{(-1)^2}{D_2} X \\ \vdots \\ |D_{n+1}| \dfrac{(-1)^{n+1}}{D_{n+1}} X \end{bmatrix},$$

поскольку

$$a_j - \tilde{a}_j = a_j - (v^j, z^j) = \frac{(-1)^j}{D_j} X.$$

Обозначая через

$$C = \frac{X}{\sum\limits_{\nu=1}^{n+1} |D_\nu|},$$

получаем, что

$$w = a - Vz^* = C\begin{bmatrix} (-1)^1 \operatorname{sign} D_1 \\ (-1)^2 \operatorname{sign} D_2 \\ \vdots \\ (-1)^{n+1} \operatorname{sign} D_{n+1} \end{bmatrix}.$$

А тогда верно

$$w_j D_j = C(-1)^j |D_j|,$$

и последовательность

$$w_1 D_1, w_2 D_2, \ldots, w_{n+1} D_{n+1}$$

имеет чередующиеся знаки.

4.3. Обобщенный алгоритм Ремеза для матриц

Заметим, что лемма 3 показывает, что матричная задача о наилучшем равномерном приближении является более общей, чем аналогичная задача для непрерывных функций. Действительно, для непрерывных функций, также как и для матриц, известен результат, что существует характеристическое множество из $r+1$ элементов при приближении по системе из r функций (т.е. полиномами степени $r-1$). При известном характеристическом множестве задача сводится к решению матричной задачи с матрицей Вандермонда. Заметим, что определитель матрицы Вандермонда может быть вычислен по формуле

$$W(x_1, x_2, \ldots, x_r) = \prod_{j<i} (x_i - x_j).$$

Откуда легко видеть, что знак определителя матрицы Вандермонда зависит только от порядка, в котором берутся точки. Если точки каждый раз берутся в возрастающем порядке, то все определители имеют одинаковый знак и в формуле (6) остаются только знаки элементов невязки. Приведенные рассуждения позволяют обобщить известный для непрерывных функций алгоритм Ремеза для построения наилучшего чебышёвского приближения на матричный случай.

Заметим, что теорема 11 уже позволяет решать задачу (4) за конечное число операций. Для этого достаточно перебрать всевозможные варианты характеристических множеств (все $r+1$-элементные подмножества строк) и решить для каждого из них задачу о поиске наилучшей равноудаленной точки. Однако находить характеристическое множество можно намного быстрее.

Приведем алгоритм решения задачи о наилучшем равномерном приближении в вещественном случае. Пусть имеются матрица $V \in \mathbb{R}^{n\times r}$ и вектор $a \in \mathbb{R}^n$.

1. Выберем произвольное множество из $r+1$ индексов строк матрицы V. Обозначим это множество через I_1 и возьмем $t = 1$.

2. Решим задачу о наилучшем равномерном приближении для матрицы $V(I_t)$ и вектора $a(I_t)$. Эта задача может быть решена за $O(r^3)$ операций для чебышёвской системы векторов. Получим решение u_t.

3. Вычислим невязку $w_t = Vu_t - a$ и найдем в векторе w_t максимальный по модулю элемент. Эта операция требует $O(nr)$ операций. Обозначим позицию максимального по модулю элемента j_t. Если $j_t \in I_t$, то согласно замечанию к критерию Ремеза, множество I_t является характеристическим и u_t является решением задачи наилучшего равномерного приближения.

4. Если $j_t \notin I_t$, то попробуем заменить каждый из элементов множества I_t на j_t. Пусть $I_t = \{i_1^t, i_2^t, \ldots, i_{r+1}^t\}$ и пусть $I_t^k = I_t \backslash \{i_k^t\} \cup j_t$. Решим задачу с матрицей $V(I_t^k)$ и вектором $a(I_t^k)$ и найдем максимум модуля невязки на множестве I_t^k, $w_t^k = V(I_t^k)u_t^k - a(I_t^k)$. Пусть $l = \arg\max_k \|w_t^k\|_\infty$. Это шаг требует $O(r^4)$ операций.

5. $I_{t+1} = I_t^l$, $t = t+1$, и перейдем к шагу 2.

Теорема 12. *Пусть система векторов $V \in \mathbb{R}^{n\times r}$ является чебышёвской и вектор $a \in \mathbb{R}^n$. Тогда обобщенный алгоритм Ремеза находит решение задачи о наилучшем равномерном приближении за конечное число операций.*

Доказательство. Пусть I_t – текущее множество индексов и $w_t = Vu_t - a$. Пусть

$$E_t = \|w_t(I_t)\|_\infty.$$

Пусть в векторе w_t максимальный по модулю элемент достигается в позиции j_t. Тогда рассмотрим задачу о наилучшем равномерном приближении для подматрицы, взятой на множестве строк с номерами $I_t \cup \{j_t\}$. Для этой задачи существует характеристическое множество из $r+1$ элементов. Заметим, что оно не может целиком состоять из элементов множества I_t, поскольку для оптимального решения на этом множестве в позиции j_t достигается строго большее значение элемента невязки. Это означает, что характеристическое множество содержит j_t и r элементов из множества I_t, т.е. получается заменой в I_t одного из элементов на j_t. Обозначим это множество через $\hat{I}_t$. Покажем, что при этом ошибка оптимального приближения на новом множестве $\hat{I}_t$ строго больше, чем на множестве I_t. Действительно, пусть при оптимальном приближении на множестве I_t получается ошибка δ, а на элементе в позиции j_t соответствующее решение дает ошибку ε. Заметим, что $\varepsilon > \delta$. Аналогично, пусть при оптимальном приближении на множестве $\hat{I}_t$ на самом множестве получается ошибка δ_1, а на удаленном из I_t элементе ошибка ε_1. Заметим, что поскольку $\hat{I}_t$ — характеристическое множество, то $\varepsilon \le \delta_1$. Тогда получаем

$$\delta < \varepsilon \le \delta_1,$$

откуда следует, что ошибка оптимального приближения на новом множестве $\hat{I}_t$ строго больше, чем ошибка оптимального приближения на множестве I_t. Откуда следует, что

$$E_{t+1} > E_t.$$

Но поскольку существует конечное число $r+1$–элементных подмножеств, последовательность $\{E_t\}$ не может быть бесконечной и достигает своего максимального значения на некотором множестве, что согласно рассуждениям теоремы 11 говорит о том, что найденное множество является характеристическим и было построено оптимальное решение.

5. ЗАДАЧА О ЧЕБЫШЁВСКОМ ПРИБЛИЖЕНИИ МАТРИЦ

Построив метод решения задачи (3), мы можем перейти к задаче

$$\mu = \inf_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}} \|A - UV^{\mathrm{T}}\|_C. \tag{7}$$

5.1. Необходимое условие оптимальности

В [9] было доказано необходимое условие оптимальности решения задачи (7). В свете полученных выше результатов мы можем легко получить это условие, а также обобщить его на случай произвольного ранга. Пусть пара $(\widehat{U}, \widehat{V})$ является решением задачи (7). Тогда матрица $\widehat{U}$ является решением задачи

$$\mu = \inf_{U\in\mathbb{R}^{m\times r}} \left\|A - U\widehat{V}^{\mathrm{T}}\right\|_C.$$

В матрице $A - \widehat{U}\widehat{V}^{\mathrm{T}}$ в некоторой позиции (i, j) достигается максимальный по модулю элемент. Рассмотрим задачу для i-й строки матрицы

$$\mu = \inf_{u\in\mathbb{R}^{m\times r}} \|a^i - u\widehat{V}^{\mathrm{T}}\|_C. \tag{8}$$

Ясно, что $\hat{u}^i$ является оптимальным решением задачи (8), в противном случае можно было бы заменить i-ю строку матрицы $\widehat{U}$ на оптимальное решение и получить лучший результат в задаче (7).

В силу оптимальности решения мы имеем, что в векторе $a^i - \hat{u}^i \widehat{V}^{\mathrm{T}}$ максимальное по модулю значение достигается в $r+1$ позиции и знаки невязки и определителей матрицы $\widehat{V}$ в этих позициях чередуются (см. лемму 3). Для получения необходимого условия из [9] достаточно заметить, что при $r = 1$ определителями являются элементы вектора $\widehat{V}$, а знак $\hat{u}^i$, очевидно, не меняется в пределах одного столбца, откуда следует утверждение 1. В случае произвольного ранга в каждом столбце и каждой строке, в которых достигается максимальный по модулю элемент, максимальное по модулю значение достигается в $r+1$ позиции и знаки невязки и определителей чередуются согласно лемме 3.

5.2. Метод решения

Построим итерационный процесс решения задачи (7). Пусть задана некоторая матрица $A \in \mathbb{R}^{m\times n}$. Пусть также задана некоторая чебышёвская матрица U_0. Найдем оптимальное чебышёвское приближение $A = U_0 V^{\mathrm{T}}$ и обозначим результат V_1. Предположим, что система векторов V_1 чебышёвская. Затем найдем оптимальное чебышёвское приближение $A = UV_1^{\mathrm{T}}$ и обозначим результат через U_1, снова предполагая, что система U_1 чебышёвская. Найдем оптимальное чебышёвское приближение $A = U_1 V^{\mathrm{T}}$ и обозначим результат через V_2. Продолжая по описанной схеме, заметим, что при этом величина $\rho_k = \|A - U_k V_k^{\mathrm{T}}\|_C$ не возрастает и ограничена снизу и, следовательно, сходится.

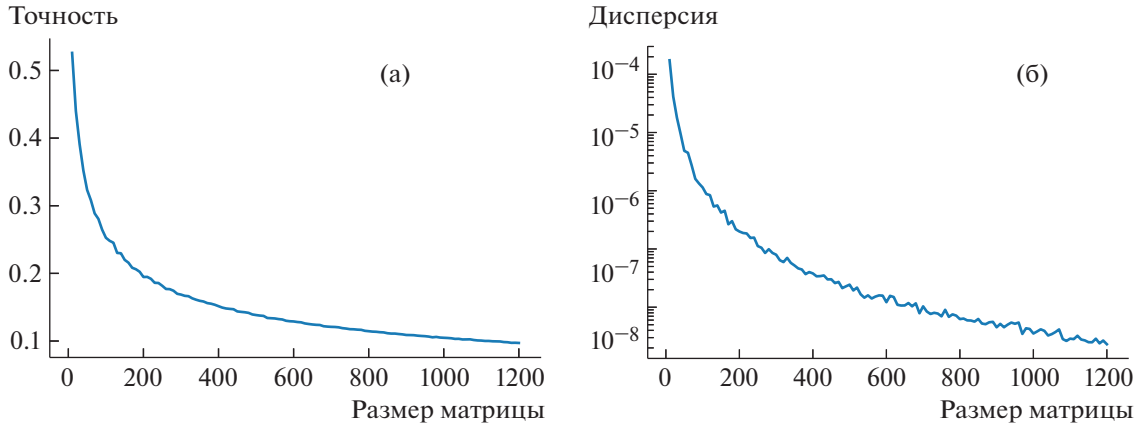
6. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

В этом разделе мы приведем ряд численных экспериментов по применению метода, описанного в п. 5.2 , для построения малоранговых чебышёвских приближений для матриц, сингулярные числа которых не убывают. Для проведения экспериментов был реализован алгоритм из п. 5.2 на C++. В эксперименте генерировались случайные матрицы, сингулярные числа которых распределены равномерно на отрезке [1, 2]. Для этого генерировались две случайные матрицы из стандартного нормального распределения, для них строилось QR-разложение и факторы Q выбирались в качестве левых U и правых V сингулярных векторов. В качестве матрицы сингулярных чисел Σ генерировалась диагональная матрица, диагональные элементы которой равномерно распределены на отрезке [1, 2]. После этого строилась матрица $U\Sigma V$. Размеры матриц варьировались от 10 до 1400 с шагом 10, а ранг приближения выбирался как $r = \sqrt{n}$, где n — размер матрицы. Для каждого размера генерировалось 10 случайных матриц и для каждой из них запускался метод альтернанса из 20 случайных точек. Таким образом, для каждого размера выполнялось 200 запусков метода альтернанса. Для каждой матрицы вычислялись среднее значение точности μ_n^i и дисперсия σ_n^i чебышёвского приближения по 20 стартовым точкам. Далее для каждого размера эти величины усреднялись по 10 матрицам

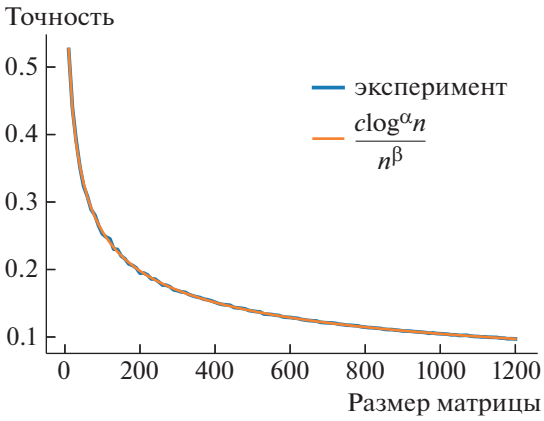
$$\mu_n = \frac{1}{10}\sum_{i=1}^{10}\mu_n^i, \quad \sigma_n = \frac{1}{10}\sum_{i=1}^{10}\sigma_n^i.$$

На фиг. 1а приведен график зависимости μ_n от размера матрицы, а на фиг. 1б график зависимости σ_n от размера матрицы. График для дисперсии приведен в логарифмическом масштабе. Одной из интересных особенностей, которые видны из этого графика, является то, что дисперсия падает с ростом размера задачи. Так, например, для матрицы размера 1400×1400 при приближении рангом 37, при запуске с 20 случайных точек, получаются следующие значения ошибки приближения:

$$\begin{bmatrix} 0.09111796 & 0.09098914 & 0.09103979 & 0.09101653 & 0.09097955 \\ 0.09112523 & 0.09102086 & 0.09097652 & 0.09099676 & 0.0909908 \\ 0.09106326 & 0.0911168 & 0.09108753 & 0.09101277 & 0.09098213 \\ 0.09103401 & 0.09106984 & 0.09097417 & 0.09094869 & 0.09092307 \end{bmatrix}.$$



Фиг. 1. Усредненная ошибка приближения и дисперсия по 20 стартовым точкам и 10 случайным матрицам для различных размеров матриц.



Фиг. 2. Усредненная ошибка приближения по 20 стартовым точкам и 10 случайным матрицам для различных размеров матриц и ее аналитическое приближение.

Была оценена асимптотика зависимости точности приближения от размера матрицы. Точность приближения искалась в виде $\frac{c\log^{\alpha} n}{n^{\beta}}$. Оптимальными оказались следующие значения:

$$c = 0.995139,$$
$$\alpha = 0.604346,$$
$$\beta = 0.495001.$$

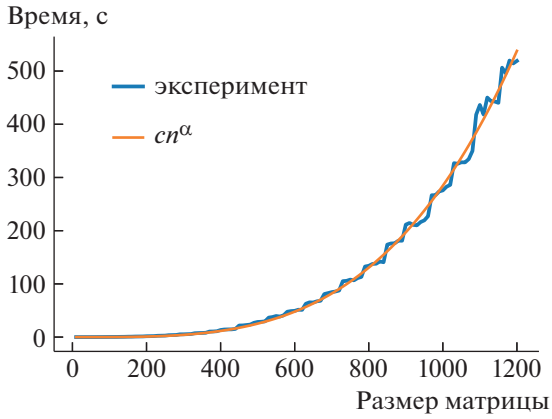
Кривая $\frac{c\log^{\alpha} n}{n^{\beta}}$ с оптимальными значениями параметров также изображена на фиг. 2. Заметим, что эта оценка соответствует

$$\varepsilon \approx \frac{\log^{0.6} n}{n^{0.5}},$$

в то время как известная теоретическая оценка (см. [8, теорема 1]) при подстановке $r = \sqrt{n}$ дает

$$\varepsilon \le \frac{6\sqrt{2}\log^{0.5}(2n+1)}{n^{0.25}},$$

откуда следует неоптимальность последней оценки.



Фиг. 3. Усредненное время работы по 200 запускам для различных размеров матриц и его аналитическое приближение.

Кроме того, в процессе эксперимента измерялось время работы программы. Для каждого размера матрицы время усреднялось по 200 запускам. На фиг. 3 приведен график зависимости времени работы от размера матрицы. Для времени работы также была оценена асимптотика в виде cn^{α}. Оптимальными оказались следующие значения:

$$c = 9.13302\text{e} - 09,$$

$$\alpha = 3.49745.$$

Это означает, что при $r = \sqrt{n}$ на практике сложность работы алгоритма составляет $O(n^{3.5})$.

7. ЗАКЛЮЧЕНИЕ

В работе был предложен метод решения задачи о наилучшем малоранговом приближении матрицы в чебышёвской норме в случае, когда известен один из факторов разложения. С использованием предложенного метода и метода альтернанса был построен алгоритм построения малорангового приближения в чебышёвской норме для произвольного ранга. Описанный метод существенно обобщает все известные на сегодняшний день методы решения задачи о чебышёвских приближениях матриц и улучшает известные теоретические оценки приближения, описанные в [8]. Численные эксперименты показывают, что метод способен приближать матрицы с хорошей точностью даже при отсутствии убывания сингулярных чисел и имеет приемлемую асимптотическую сложность.

СПИСОК ЛИТЕРАТУРЫ

1. *Bebendorf M.* A means to efficiently solve elliptic boundary value problems // Hierarchical Matrices. LNCS. 2008. V. 63. P. 49–98.
2. *Son S.W., Chen Z., Hendrix W., Agrawal A., Liao W.K., Choudhary A.* Data compression for the exascale computing era-survey // Supercomputing Frontiers and Innovations. 2014. V. 1. N. 2. P. 76–88.
3. *He X., Zhang H., Kan M.Y., Chua T.S.* Fast matrix factorization for online recommendation with implicit feedback // Proc. of the 39th Internat. ACM SIGIR conference on Research and Development in Information Retrieval. 2016. P. 549–558.
4. *Yang C., Akimoto Y., Kim D., Udell M.* Oboe: Collaborative filtering for automl initialization // arXiv preprint arXiv:1808.03233. 2018.
5. *Goreinov S., Tyrtyshnikov E., Zamarashkin N.* A theory of pseudoskeleton approximations // Linear Algebra and its Appl. 1997. V. 261. N. 1. P. 1–21.
6. *Osinsky A., Zamarashkin N.* Pseudo-skeleton approximations with better accuracy estimates // Linear Algebra and its Appl. 2018. V. 537. P. 221–249.

7. *Halko P.M.N., Tropp J.* Finding Structures with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions // SIAM Review. 2011. V. 53. N. 2. P. 217–288.
8. *Udell M., Townsend A.* Why are big data matrices approximately low rank? // SIAM Journal on Math. of Data Sci. 2019. V. 1. N. 1. P. 144–160.
9. *Даугавет В.* О равномерном приближении функции двух переменных, заданной таблично, произведением функций одной переменной // Ж. вычисл. матем. и матем. физ. 1971. Т. 11. N. 2. С. 289–303.
10. *Дзядык В.К.* Введение в теорию равномерного приближения функций полиномами. М.: Наука, 1977.
11. *Смирнов В.И., Лебедев Н.А.* Конструктивная теория функций комплексного переменного // М.–Л.: Наука, 1964.
12. *Дзядык В.К.* О приближении функций на множествах, состоящих из конечного числа точек // Сб. “Теория приближения функций и ее приложения”, Киев. 1974. Р. 69–80.

EDN: RPZLUI
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 742–756

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.613

АЛГОРИТМ РАЗДЕЛЕНИЯ МАТРИЧНОГО СПЕКТРА ОТНОСИТЕЛЬНО УГЛА[1)]

© 2022 г. Э. А. Бибердорф

630090 Новосибирск, пр-т Акад. Коптюга, 4, Институт математики им. С.Л. Соболева СО РАН, Россия

e-mail: biberdorf@ngs.ru

Поступила в редакцию 21.09.2021 г.
Переработанный вариант 12.11.2021 г.
Принята к публикации 14.01.2022 г.

Представлен алгоритм разделения спектра относительно кусочно-гладкой кривой, а именно относительно угла. В работе дано полное описание задачи и ее решения, начиная от постановки и краткого изложения теоретических основ базовых алгоритмов дихотомии и заканчивая численными примерами использования алгоритма дихотомии матричного спектра относительно угла. Библ. 29. Фиг. 4. Табл. 1.

Ключевые слова: матричный спектр, угловая область, секториальный оператор, дихотомия спектра.

DOI: 10.31857/S0044466922050027

ВВЕДЕНИЕ

В данной работе предлагается алгоритм для решения задачи о расположении спектра относительно некоторого угла. Эта проблема связана с исследованиями так называемых секториальных операторов, спектр которых находится внутри определенного сектора комплексной плоскости. В частности, к секториальным относятся многие дифференциальные операторы (см. [1]). Параметрами сектора определяется скорость роста или убывания в зависимости от времени решений уравнений с оператором такого типа [2].

Структура работы следующая. В первом разделе обсуждаются особенности различных постановок спектральных задач для несимметричных матриц, связанные с чувствительностью собственных значений к возмущению матрицы. В разд. 2 приведен ряд фактов, на которых основаны простейшие алгоритмы дихотомии. Далее, в разд. 3 описаны способы решения вспомогательных задач и введены необходимые обозначения. В разд. 4 дано обоснование алгоритма дихотомии относительно угла. Затем (разд. 5) приведены примеры его использования. В частности, показано применение алгоритма к спектральной задаче для оператора Орра-Зоммерфельда для плоскопараллельного течения Пуазейля.

1. ПОСТАНОВКИ СПЕКТРАЛЬНЫХ ЗАДАЧ

1.1. Роль псевдоспектра при решении прикладных задач

Информация о спектрах линейных операторов необходима для решения большого круга прикладных задач. Постановки многих из них (например, задач устойчивости) обладают общими специфическими свойствами. В частности, для решения таких задач не требуется определения каждого собственного значения в отдельности. Вместо этого нужна информация о расположении групп собственных значений в определенных областях комплексной плоскости (например, областях устойчивости и неустойчивости).

Другое общее свойство прикладных задач связано с тем, что модуль непрерывности собственных значений несимметричных операторов ведет себя непредсказуемым образом. Для его исследования вводится понятие ε-спектра или псевдоспектра $\Lambda_\varepsilon(A)$ (см. [1], [3]). По определению

[1)]Работа выполнена в рамках государственного задания ИМ СО РАН (проект № FWNF-2022-0008).

к ε-спектру $n \times n$ матрицы A относятся те комплексные числа λ, для которых матрица $A - \lambda I_n$ “почти вырождена”

$$\sigma_{\min}(A - \lambda I_n) \le \varepsilon,$$

где I_n – единичная матрица, $\sigma_{\min}$ – минимальное сингулярное число.

Визуализировать структуру ε-спектра можно, изображая график функции

$$f(\lambda) = \log_{10}\sigma_{\min}(A - \lambda I_n), \tag{1.1}$$

например, с помощью линий уровня (см. примеры в разд. 5).

Размеры пятен ε-спектра определяют, насколько могут отличаться собственные значения $|\lambda_j(A) - \lambda_j(A + \Delta A)|$ исходной и возмущенной матриц, $\|A\| \le \varepsilon$. Расположением пятен объясняются на первый взгляд парадоксальные примеры, когда решения системы $y' = Ay$ устойчивы, а возмущенной системы $y' = (A + \Delta A)y$ – неустойчивы. Этим же объясняется тот факт, что решения устойчивых систем могут на начальных временных интервалах показывать большой рост (“практическая” неустойчивость), что на практике может приводить к разрушениям инженерных конструкций, развитию турбулентности течений и т.д. [4].

Таким образом, для решения прикладных задач знания расположения точек спектра недостаточно, необходима информация о расположении и параметрах спектральных пятен. Нетрудно заметить, что спектральная задача в классической постановке ориентирована именно на вычисление отдельных собственных значений. То есть с точки зрения приложений она с одной стороны избыточна, а с другой стороны – недостаточна.

1.2. Исследования псевдоспектра и критерии расположения спектра в заданной области

Движение в сторону альтернативных постановок задач можно видеть во многих работах. Так, целый ряд статей посвящен методам изображения псевдоспектра. С точки зрения вычислений это весьма затратная процедура. Поэтому предпринимаются попытки создать более быстрые алгоритмы для вычисления некоторого приближения к функции (1.1) (см., например, [5]). Авторы других работ концентрируются на конкретных операторах и численно исследуют, какие из собственных значений наименее устойчивы к возмущению матрицы [6]. Заметим, что изображение псевдоспектра матрицы не гарантирует, что на нем будут видны все собственные значения, лежащие в данной области. Некоторые пятна ε-спектра могут быть настолько небольшими, что их легко пропустить при вычислении значений функции (1.1) даже на сетке с довольно мелкой ячейкой.

Другой подход может быть связан с подбором некоторого численного критерия (или критериев), который определенным образом характеризует часть спектра, находящуюся в фиксированной области в целом. В качестве известного примера можно привести теорему Раусса-Гурвица о том, что все корни полинома с вещественными коэффициентами лежат строго в левой полуплоскости тогда и только тогда, когда положительны все главные миноры матрицы Гурвица [7]. Этот факт является одним из многих результатов, связывающих расположение корней полиномов и свойства определенных матриц и знакопеременных квадратичных форм, собранных в подробном обзоре [8]. В частности, там упоминается работа [9] о расположении всех корней полинома в угловой области $-\theta < \arg x < \theta$.

Для конструирования других критериев можно использовать свойства некоторых интегралов по контуру γ, ограничивающему заданную область. Например, нетрудно установить, что матрица

$$P = \frac{1}{2\pi i}\oint_{\gamma}(\lambda I_n - A)^{-1}d\lambda \tag{1.2}$$

является проектором на инвариантное подпространство матрицы A, соответствующее собственным значениям $\lambda_j(A)$, лежащим внутри контура γ. Причем след проектора равен числу этих собственных значений с учетом кратности [1]. Можно сказать, что этот факт является матричным вариантом принципа аргумента.

Однако при вычислении перечисленных критериев также могут возникать проблемы вплоть до ложного результата, связанные с расположением пятен псевдоспектра вблизи границы области.

1.3. Задача дихотомии матричного спектра

К конструктивному решению проблемы новой постановки спектральной задачи подошла научная группа под руководством С.К. Годунова. Ими была сформулирована задача о дихотомии матричного спектра. Суть ее состоит в том, чтобы определить

1) есть ли на заданной кривой точки спектра матрицы,

2) если на кривой отсутствуют точки спектра, определить базисы инвариантных подпространств матрицы, соответствующие собственным значениям, находящимся по разные стороны от кривой, и привести матрицу к клеточно-диагональному виду.

Заметим, что сходство названия задачи с ε-дихотомией пространства решений дифференциальных уравнений, описанной в [10], неслучайно, так как и в том и в другом случае производится разделение пространства на прямую сумму подпространств.

При этом критерием дихотомии и ответом на первый вопрос задачи является (возможно, с дополнительной нормировкой) интеграл

$$H = \oint_{\gamma} (\lambda I_n - A)^{-1} \left(\bar{\lambda} I_n - A^*\right)^{-1} d\lambda, \tag{1.3}$$

который сходится только при отсутствии собственных значений матрицы на γ, а норма которого зависит от расположения пятен псевдоспектра вблизи кривой γ. Можно видеть, что в отличие от матричных критериев [7], [8], матрица H является самосопряженной и положительно-определенной. Для ответа на второй вопрос задачи дихотомии используется интеграл (1.2). Заметим, что в случае разделения спектра относительно единичной окружности или мнимой оси матрицы H и P являются решениями уравнений Ляпунова и их обобщений [1], [11].

Несмотря на то что в основе метода лежат контурные интегралы, для их определения не используется численное интегрирование. Вместо этого алгоритм представляет собой последовательность итераций, состоящих из QR-разложения и других матричных операций. Сходимость итераций связана с величиной H и зависит от тех спектральных пятен, которые пересекают границу области. А значит, в процессе выполнения алгоритма диагностируется ситуация, когда псевдоспектр может критически повлиять на результат. Подробнее см. в следующем разделе.

В данный момент разработаны алгоритмы дихотомии относительно окружности и мнимой оси, которые можно назвать базовыми, так как они являются основной составляющей частью всех других алгоритмов дихотомии [12]–[17]. Кроме того, существуют алгоритмы дихотомии относительно кривых второго порядка [18]–[20] и алгоритм дихотомии корней полинома относительно окружности [21]. Более подробный обзор алгоритмов дихотомии и их приложений представлен в статье [22].

2. БАЗОВЫЕ АЛГОРИТМЫ

2.1. Математическая основа

В данном разделе мы представим обоснование алгоритма дихотомии относительно единичной окружности, опираясь исключительно на свойства решений краевых задач для линейных разностных уравнений

$$\begin{gathered} U_{j+1} = AU_j + f_j, \quad \|f_j\| \le C < \infty, \\ \|U_j\| \le C < \infty, \quad -\infty < j < \infty \end{gathered} \tag{2.1}$$

(здесь и далее обозначение нормы $\|\cdot\|$, примененное к вектору, означает евклидову норму, примененное к матрице – операторную норму). Основную роль при этом играет тот факт, что существование и единственность решения задачи (2.1) обусловлены расположением спектра матрицы A относительно единичной окружности. При этом разрешимость такой задачи тесно связана с разрешимостью задачи для матричной разностной функции Грина

$$\begin{gathered} G_{k+1} = AG_k, \quad k \ne 0, \\ G_1 = AG_0 + I_n, \\ \|G_k\| \le C \le \infty, \quad -\infty < k < \infty. \end{gathered} \tag{2.2}$$

Заметим, что утверждения, которые будут приведены ниже, в известной степени аналогичны утверждениям о разрешимости краевой задачи для системы линейных обыкновенных диффе-

ренциальных уравнений, заданной на бесконечном интервале (см. [14], [23]–[25]), правда, в последнем случае спектр разделяется относительно мнимой оси.

В первую очередь поставим вопрос об условиях единственности решений (2.1) и (2.2). Для ответа на него используется следующая лемма о разрешимости однородной задачи:

Лемма 1. *Однородная краевая задача*

$$\begin{gathered} U_{j+1}^{[\text{од}]} = AU_j^{[\text{од}]}, \\ \left\|U_j^{[\text{од}]}\right\| \le C < \infty, \quad -\infty < j < \infty, \end{gathered} \tag{2.3}$$

имеет нетривиальное решение тогда и только тогда, когда среди собственных значений матрицы A есть такие, модуль которых равен единице.

Доказательство следует из представления общего решения однородных разностных уравнений в виде $U_j^{[\text{од}]} = A^j U_0^{[\text{од}]}$. Решение уравнения из (2.1) будет нетривиальном в случае $\left\|U_0^{[\text{од}]}\right\| \ne 0$ и ограниченным, если вектор $U_0^{[\text{од}]}$ является линейной комбинацией собственных векторов, соответствующих собственным значениям $|\lambda(A)| = 1$. В противном случае нетривиальных решений однородной задачи нет.

Следующее утверждение определяет условия единственности решений задач (2.1) и (2.2).

Лемма 2. *Решения краевых задач* (2.1) *и* (2.2) *единственны тогда и только тогда, когда собственные значения матрицы A не лежат на единичной окружности*

$$|\lambda_i(A)| \ne 1, \quad i = 1, 2, \ldots, n. \tag{2.4}$$

Теперь рассмотрим вопрос о существовании решений. Заметим, что если спектр матрицы A делится единичной окружностью на два непересекающихся множества, то с помощью подобного преобразования она может быть приведена к блочно-диагональной форме (см., например, [1]):

$$A = T\begin{bmatrix} \Lambda_0 & 0 \\ 0 & \Lambda_\infty \end{bmatrix} T^{-1}, \quad |\lambda_i(\Lambda_0)| < 1, \quad |\lambda_i(\Lambda_\infty)| > 1. \tag{2.5}$$

Лемма 3. *Если выполнено условие* (2.4)*, то решение задачи* (2.2) *существует и может быть представлено в явном виде*

$$G_k = \begin{cases} T\begin{bmatrix} \Lambda_0^{k-1} & 0 \\ 0 & 0 \end{bmatrix} T^{-1}, & k > 0, \\ T\begin{bmatrix} 0 & 0 \\ 0 & -\Lambda_\infty^k \end{bmatrix} T^{-1}, & k \le 0. \end{cases} \tag{2.6}$$

То, что заданная в (2.6) функция G_k является решением задачи (2.2), проверяется непосредственной подстановкой.

Замечание. Если у матрицы A есть собственные значения, лежащие на единичной окружности, то представления, аналогичные (2.5) и (2.6), отличающиеся только нестрогими неравенствами для спектров подматриц Λ_0 или Λ_∞, также имеют место. Однако условия ограниченности решения (2.2) и/или его единственности при этом нарушаются.

Лемма 4. *Если выполнено условие* (2.4)*, то решение задачи* (2.2) *существует и может быть представлено в виде*

$$U_j = \sum_{k=-\infty}^{+\infty} G_{j-k} f_k. \tag{2.7}$$

Сходимость ряда в формуле (2.7) для любой ограниченной разностной функции правой части f_k является следствием леммы 3. Справедливость представления (2.7) для решения задачи (2.1) может быть проверена непосредственной подстановкой.

Приведенные выше леммы 1–4 представляют собой отдельные этапы доказательства следующей теоремы.

Теорема 1. *Для существования и единственности решения задачи* (2.1) *необходимо и достаточно выполнение условия* (2.4).

Рассмотрим величину

$$\|H\| = \left\| \sum_j G_j^* G_j \right\| \tag{2.8}$$

и заметим, что этот ряд также как и (2.7) сходится только при условии (2.4). А значит, можно сформулировать следующий критерий разделения спектра матрицы A единичной окружностью как следствие теоремы (1).

Утверждение 1. *Условие* (2.4) *выполняется тогда и только тогда, когда* $\|H\| < \infty$.

Обратим внимание, что матрица $P_{\text{in}} = G_1$ является проектором на инвариантное подпространство матрицы A, соответствующее собственным значениям, лежащим внутри единичной окружности. Проектор на подпространство, соответствующее собственным значениям, лежащим вне окружности, имеет вид $P_{\text{out}} = I_n - P_{\text{in}}$. Размерность инвариантных подпространств вычисляется по формулам $n_{\text{in}} = \operatorname{tr} P_{\text{in}}$, $n_{\text{out}} = \operatorname{tr} P_{\text{out}}$. С помощью сингулярного разложения (см., в частности, [1])

$$P_{\text{in}} = [U_1, W_1] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V_1, \quad P_{\text{out}} = [U_2, W_2] \begin{bmatrix} \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} V_2$$

вычисляются ортогональные базисы U_1, U_2 инвариантных подпространств и матрица перехода $T = [U_1, U_2]$, которая приводит матрицу A к клеточно-диагональному виду (2.5).

Обобщая вышесказанное на случай матричных пучков $A - \lambda B$, можно сформулировать следующее

Утверждение 2. *На единичной окружности отсутствуют собственные значения пучка* $A - \lambda B$ *и возможно представление*

$$A - \lambda B = T \begin{bmatrix} \Lambda - \lambda I_\Lambda & 0 \\ 0 & I_M - \lambda M \end{bmatrix} S,$$
$$\det T \neq 0, \quad \det S \neq 0,$$
$$|\lambda_i(\Lambda)| < 1, \quad |\lambda_i(M)| < 1,$$

где I_Λ, I_M *– единичные матрицы, тогда и только тогда, когда существует единственное решение краевой задачи*

$$AG_{k+1} - BG_k = 0, \quad k \neq 0,$$
$$G_1 - G_0 = I_n, \tag{2.9}$$
$$\|G_k\| \leq C < \infty, \quad -\infty < k < \infty.$$

Причем численным критерием качества дихотомии спектра $A - \lambda B$ *является величина* (2.8).

Таким образом, задача о разделении матричного спектра единичной окружностью сводится к решению краевой задачи (2.9), которая имеет постоянные матричные коэффициенты и может быть эффективно решена с использованием метода удвоений и ортогональных исключений.

2.2. Дихотомия единичной окружностью и мнимой осью

Приведенные выше факты лежат в основе алгоритма дихотомии единичной окружностью, который, в свою очередь, является ключевой частью алгоритмов дихотомии относительно остальных кривых. Существует несколько вариантов этого алгоритма. Однако основной цикл каждого из них представляет собой метод удвоений, т.е. переход от уравнений $A_k U_j - B_k U_{j+2^k} = 0$ к уравнениям $A_{k+1} U_j - B_{k+1} U_{j+2^{k+1}} = 0$ путем исключения промежуточного слагаемого при помощи QR-разложения соответствующей матрицы.

Заметим также, что условия сходимости основного цикла могут быть сформулированы по-разному. Можно следить за абсолютной сходимостью, как это сделано ниже, или за относительной. В условие можно также включать сходимость проектора P_k. Кроме того, существуют оценки, позволяющие априорно оценить число итераций, необходимых для сходимости алгоритма при

условии $\|H\| \le \omega_{\max}$. Эту оценку также можно использовать в качестве верхней границы числа итераций. При этом нужно принимать во внимание, что она, как правило, очень завышена, так как в обычной ситуации для сходимости с высокой точностью достаточно нескольких итераций.

Алгоритм дихотомии матричного спектра относительно единичной окружности

Дано: матричный пучок $A_0 - \lambda B_0$, ε_{it} – требуемая точность итерационного процесса, $\omega_{\max}$, $\mu_{\max}$ – максимальные значения критерия дихотомии и числа обусловленности матрицы.

Если $\operatorname{cond}(A_0 - B_0) > \mu_{\max}$,

то дихотомия невозможна, конец расчетов.

иначе вычисляется начальное приближение матричного критерия дихотомии H

$$H_0 = (A_0 - B_0)^{-1}\left(A_0A_0^* + B_0B_0^*\right)\left(A_0^* - B_0^*\right)^{-1}$$

Цикл пока $\|H_k - H_{k-1}\| > \varepsilon_{it}$

Если $\|H_k\| \ge \omega_{\max}$ или $\operatorname{cond}(A_k + B_k) > \mu_{\max}$,

то дихотомия невозможна, конец расчетов.

иначе вычисляются вспомогательные матрицы $V_{k+1} = (A_k + B_k)^{-1}A_k$, $U_{k+1} = I_n - V_{k+1}$,

приближение матричного критерия дихотомии H $H_{k+1} = U_{k+1}H_kU_{k+1}^* + V_{k+1}H_kV_{k+1}^*$,

преобразование матричного пучка $qr\left(\begin{bmatrix} -B_k & A_k & 0 \\ A_k & 0 & -B_k \end{bmatrix}\right) = \begin{bmatrix} * & * & * \\ 0 & A_{k+1} & -B_{k+1} \end{bmatrix}$,

приближение матрицы проектора P_{in} $P_k = -(A_{k+1} - B_{k+1})^{-1}B_{k+1}$.

Конец цикла.

Алгоритм дихотомии спектра матрицы A относительно мнимой оси основан на свойствах экспоненциального отображения, которое переводит левую полуплоскость в единичный круг. Таким образом, его отличие от дихотомии относительно окружности заключается в том, что в качестве начального матричного пучка используется $A_0 - \lambda B_0 = e^{\tau A} - \lambda I_n$. Параметр τ выбирается так, чтобы обеспечить быструю сходимость при вычислении матричной экспоненты, например, $\tau \approx 1/2\|A\|$. Однако в случае матриц с большой нормой, например, аппроксимирующих дифференциальные операторы, такой выбор τ может приводить к ложным результатам. Поэтому для матриц с большой нормой используется прием, описанный в [17], заключающийся в выборе τ в виде степени двойки $\tau = 2^{-K}$ и последующем применении K дополнительных итераций.

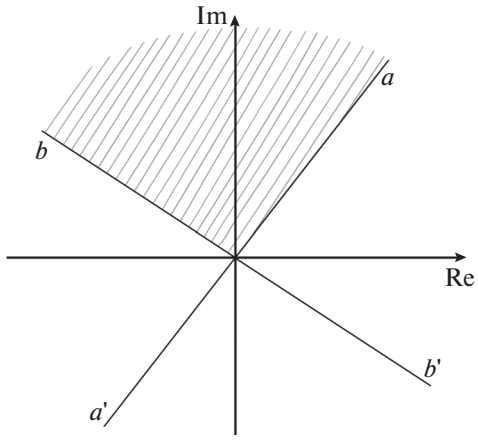
3. ОБОЗНАЧЕНИЯ И ВСПОМОГАТЕЛЬНЫЕ АЛГОРИТМЫ

В данном разделе приводятся алгоритмы, которые войдут в качестве составной части в алгоритм дихотомии относительно угла. Кроме этого, здесь будет определен ряд обозначений, которые будут использоваться в основном алгоритме.

3.1. Используемые обозначения

Основными математическими объектами, которые будут рассматриваться далее, являются лучи (или полупрямые) на комплексной плоскости $\mathbb{C}$, начало которых будет находиться в точке 0. Обозначать лучи будем малыми буквами a, b, c, … . Угол, который образуют два луча a и b, будем обозначать ab или $\angle ab$ (см. фиг. 1). Положительный отсчет меры угла будет направлен против часовой стрелки. Тогда, очевидно, верно равенство для сопряженных углов $\angle ab = 2\pi - \angle ba$.

Прямые, проходящие через точку (0, 0), будем обозначать двумя буквами aa', где a и a' – лучи с началом в точке (0, 0), составляющие вместе данную прямую. Такое двойное обозначение позволяет отождествлять прямую aa' и развернутый угол aa'. Кроме того, по аналогии с вещественной и мнимой осью для каждой прямой будем задавать направление. То есть один из лучей будем называть положительным, другой – отрицательным. Эти направления можно задавать или полностью произвольно, или исходя из удобства выкладок. При обозначениях мы будем учитывать



Фиг. 1. Угловая область.

направление прямой следующим образом: у прямой aa' полупрямая a будет положительной, а полупрямая a' – отрицательной. Так как основным инструментом для решения основной задачи будет являться дихотомия матричного спектра мнимой осью, которая делит плоскость на правую и левую полуплоскости, то по аналогии мы определим правые и левые полуплоскости для всех прямых, у которых мы определим направление. *Правой* назовем полуплоскость, в которой находится угол $a'a$ (от отрицательного направления a' прямой aa' к положительному a), и обозначим через $\mathbb{C}^r_{aa'}$. Соответственно вторую полуплоскость, в которой находится угол aa' (от положительного направления к отрицательному), будем называть *левой* и обозначать через $\mathbb{C}^l_{aa'}$. В этих обозначениях часть плоскости внутри угла ab является пересечением левой полуплоскости относительно прямой aa' и правой полуплоскости относительно прямой bb': $\mathbb{C}^l_{aa'} \bigcap \mathbb{C}^r_{bb'}$.

3.2. Дихотомия спектра матрицы произвольной прямой

Для того, чтобы исследовать расположение спектра матрицы A относительно произвольной прямой aa', проходящей через начало координат, необходимо совершить поворот комплексной плоскости на угол φ между прямой aa' и мнимой осью $\xi = e^{i\varphi}\lambda$. При таком преобразовании прямая aa' перейдет в мнимую ось, а исходная спектральная задача $Av = \lambda v$ в задачу для нового спектрального параметра $A_\varphi v = \xi v$, где $A_\varphi = e^{i\varphi}A$. Из этих равенств видно, что собственные векторы, а значит, и инвариантные подпространства у матриц A и A_φ совпадают, а собственные значения отличаются на угол φ. А значит, дихотомия спектра A_φ относительно мнимой оси позволит определить проекторы P^r и P^l на инвариантные подпространства матрицы A, соответствующие собственным значениям, лежащим в правой $\mathbb{C}^r_{aa'}$ и левой $\mathbb{C}^l_{aa'}$ полуплоскостях. С помощью сингулярного разложения получаем ортогональные базисы U^r, U^l в инвариантных подпространствах, после чего можем представить матрицу в клеточно-диагональном виде:

$$A = T\begin{bmatrix} A^r & \\ & B^l \end{bmatrix}T^{-1}, \quad T = \left[U^r, U^l\right].$$

3.3. Проверка наличия собственных значений матрицы на луче

Стороны угла являются лучами, поэтому проверка отсутствия собственных значений заданной матрицы A на сторонах угла сводится к вопросу о нахождении собственных значений на луче.

Обозначим через a луч с началом в точке (0, 0), а угол между ним и вещественной положительной полуосью α. Причем отсчет угла производим от луча a к полуоси против часовой стрелки.

Тогда задача о локализации спектра $n \times n$ матрицы A на заданном луче a при помощи поворота сводится к вопросу о расположении спектра матрицы $A_\alpha = e^{i\alpha}A$ на вещественной положительной полуоси.

Далее заметим, что среди собственных значений матрицы A_α отсутствуют положительные вещественные числа тогда и только тогда, когда среди собственных значений матричного пучка $A_\alpha - \xi^2 I_n$, квадратичного относительно спектрального параметра ξ, нет вещественных чисел. Из равенства определителей

$$\det\left(A_\alpha - \xi^2 I_n\right) = (-1)^n \det(\mathbf{A} - \xi I_{2n}),$$

где

$$\mathbf{A} = \begin{bmatrix} & I_n \\ A_\alpha & \end{bmatrix},$$

следует, что спектр квадратичного матричного пучка $A_\alpha - \xi^2 I_n$ совпадает со спектром матрицы $\mathbf{A}$ удвоенного размера.

И, наконец, очевидно, что наличие собственных значений матрицы $\mathbf{A}$ на вещественной оси эквивалентно наличию собственных значений матрицы $i\mathbf{A}$ на мнимой оси.

Итак, для того чтобы определить наличие или отсутствие точек спектра матрицы A на луче a, нужно вычислить критерий дихотомии относительно мнимой оси для матрицы $i\mathbf{A}$. Заметим, что вычисления проекторов в данном случае не происходит, так как луч не разделяет плоскость на две непересекающиеся области.

4. ЗАДАЧА О ДИХОТОМИИ ОТНОСИТЕЛЬНО УГЛА

4.1. Формулировка задачи

На комплексной плоскости задан угол с вершиной в начале координат, образованный лучами a, b (см. фиг. 1). Задача заключается в том, чтобы разделить спектр заданной матрицы A на две части в зависимости от расположения относительно угла ab. Для этого нужно будет определить, лежат ли на сторонах угла ab собственные значения матрицы A. В случае, если стороны угла свободны от точек спектра, нужно вычислить проекторы на инвариантные подпространства матрицы, соответствующие собственным значениям, лежащим, соответственно, внутри и вне угла ab, определить базисы этих инвариантных подпространств и привести матрицу A к клеточно-диагональному виду.

Так как угол состоит из двух лучей, то для того, чтобы определить критерий дихотомии спектра углом, нужно определить численный критерий отсутствия точек спектра на каждом луче отдельно (см. разд. 3). Если определены значения $\omega(a)$ и $\omega(b)$ для обеих сторон угла, то в качестве критерия дихотомии углом можно взять их сумму:

$$\omega(ab) = \omega(a) + \omega(b).$$

Далее нужно вычислить проекторы P_{in} и $P_{\text{out}} = I_n - P_{\text{in}}$ на инвариантные подпространства, соответствующие собственным значениям, лежащим внутри и вне угла ab соответственно, а также привести подобным преобразованием матрицу A к клеточно-диагональному виду

$$A = T\begin{bmatrix} A_{\text{in}} & \\ & B_{\text{out}} \end{bmatrix}T^{-1}. \tag{4.1}$$

Здесь собственные значения подматрицы A_{in} лежат внутри угла ab, а собственные значения B_{out} – вне угла.

4.2. Дихотомия прямыми aa' и bb'

Для выделения части спектра, лежащей внутри угла ab, можно использовать алгоритм дихотомии прямыми, являющимися продолжением сторон угла. Обозначим соответствующие прямые через aa' и bb'.

Случай 1. На обеих прямых aa' и bb' отсутствуют собственные значения матрицы A. В этой ситуации алгоритм решения задачи дихотомии максимально прост, но, очевидно, неприменим в общем случае. Решение сводится к последовательному выполнению дихотомии спектра A относительно обеих прямых с вычислением проекторов $P^l_{bb'}$ и $P^r_{aa'}$. Тогда проектор на инвариантное подпространство, соответствующее собственным значениям, находящимся внутри угла ab, представляет собой произведение $P_{\text{in}} = P^l_{bb'}P^r_{aa'}$. Сингулярное разложение проекторов P_{in} и $I_n - P_{\text{in}}$ позволяет получить базисы инвариантных подпространств и привести матрицу A к клеточно-диагональному виду.

Случай 2. На прямой aa' нет собственных значений матрицы A. Произведем дихотомию спектра относительно этой прямой (см. разд. 3). В результате чего множество всех собственных значений разделится на два непересекающихся подмножества, лежащих в полуплоскостях $\mathbb{C}^l_{aa'}$ и $\mathbb{C}^r_{aa'}$, разделенных прямой aa'. При этом будут вычислены проекторы

$$P_1 = P^l_{aa'}, \quad I_n - P_1 = P^r_{aa'} \tag{4.2}$$

на инвариантные подпространства матрицы A размерности $n_1 = \operatorname{tr} P_1$ и $n - n_1$, соответствующих собственным значениям, лежащим в полуплоскостях, разделенных прямой aa', базисы в данных инвариантных подпространствах $U^l_{aa'}$, $U^r_{aa'}$ и матрицу перехода

$$T_1 = \left[U^l_{aa'}, U^r_{aa'}\right], \tag{4.3}$$

приводящую матрицу A к клеточно-диагональному виду:

$$A = T_1 \begin{bmatrix} A_1 & \\ & B_1 \end{bmatrix} T_1^{-1}. \tag{4.4}$$

Здесь спектр подматрицы A_1 полностью лежит в левой полуплоскости $\mathbb{C}^l_{aa'}$, а спектр B_1 – в правой $\mathbb{C}^r_{aa'}$

$$\lambda_j(A_1) \in \mathbb{C}^l_{aa'}, \quad \lambda_j(B_1) \in \mathbb{C}^r_{aa'}.$$

Луч b лежит в левой полуплоскости $\mathbb{C}^l_{aa'}$ прямой aa', тогда как его продолжение b' лежит в правой полуплоскости. А внутренняя часть угла ab представляет собой пересечение полуплоскостей $\mathbb{C}^l_{aa'} \bigcap \mathbb{C}^r_{bb'}$. Следовательно, для завершения решения задачи необходимо разделить спектр матрицы A_1 относительно прямой bb'. В результате применения метода дихотомии матричного спектра прямой будут получены проекторы

$$P_2 = P^r_{bb'}, \quad I_{n_1} - P_2 = P^l_{bb'}, \tag{4.5}$$

базисы $U^r_{bb'}$, $U^l_{bb'}$ и матрица перехода

$$T_2 = \left[U^r_{bb'}, U^l_{bb'}\right], \tag{4.6}$$

приводящую матрицу A_1 к клеточно-диагональному виду:

$$A_1 = T_2 \begin{bmatrix} A_2 & \\ & B_2 \end{bmatrix} T_2^{-1}, \tag{4.7}$$

где

$$\lambda_j(A_2) \in \mathbb{C}^r_{bb'}, \quad \lambda_j(B_2) \in \mathbb{C}^l_{bb'}.$$

Случай 3. На прямой bb' отсутствуют собственные значения матрицы A. Если при этом критерий дихотомии спектра прямой aa' бесконечен, то порядок действий поменяется. Сначала нужно провести дихотомию прямой bb'. Результат этого действия: проекторы

$$P_1 = P^r_{bb'}, \quad I_n - P_1 = P^l_{bb'}, \tag{4.8}$$

на инвариантные подпространства матрицы A, соответствующих собственным значениям, лежащим в полуплоскостях, разделенных прямой bb', базисы в данных инвариантных подпространствах $U^r_{bb'}$, $U^l_{bb'}$ и матрицу перехода

$$T_1 = \left[U^r_{bb'}, U^l_{bb'}\right], \tag{4.9}$$

приводящую матрицу A к клеточно-диагональному виду (4.4), причем

$$\lambda_j(A_1) \in \mathbb{C}^r_{bb'}, \quad \lambda_j(B_1) \in \mathbb{C}^l_{bb'}.$$

Далее производится дихотомия спектра подматрицы A_1 относительно прямой aa'. Будут получены: проекторы

$$P_2 = P^l_{aa'}, \quad I_{n_1} - P_2 = P^r_{aa'}, \tag{4.10}$$

базис $U^l_{aa'}$, $U^r_{aa'}$ и матрица перехода

$$T_2 = \left[U^l_{aa'}, U^r_{aa'}\right] \tag{4.11}$$

такая, что имеет место представление (4.7), где

$$\lambda_j(A_2) \in \mathbb{C}^l_{aa'}, \quad \lambda_j(B_2) \in \mathbb{C}^r_{aa'}.$$

Итоговые матрицы проекторов и матричные разложения будут иметь следующий вид:

$$P_{\text{in}} = P_1 T_1 \begin{bmatrix} P_2 & \\ & 0_{n-n_1} \end{bmatrix} T_1^{-1} = T_1 \begin{bmatrix} P_2 & \\ & 0_{n-n_1} \end{bmatrix} T_1^{-1} P_1, \quad P_{\text{out}} = I_n - P_{\text{in}}, \tag{4.12}$$

$$T = T_1 \begin{bmatrix} T_2 & \\ & I_{n-n_1} \end{bmatrix}, \quad A_{\text{in}} = A_2, \quad B_{\text{out}} = \begin{bmatrix} B_2 & \\ & B_1 \end{bmatrix}. \tag{4.13}$$

Здесь символом 0_{n-n_1} обозначена квадратная нулевая матрица размера $(n-n_1)\times(n-n_1)$.

Упрощенный алгоритм дихотомии углом

Дано: матрица A, угол ab на комплексной плоскости с вершиной в точке (0, 0), максимально допустимое значение критерия дихотомии $\omega_{\max}$.

Шаг 1. Проверка наличия собственных значений матрицы на лучах a и b.

Если $\omega(ab) < \omega_{\max}$, то выполняются шаги 2, 3.

Шаг 2. Если $\omega(aa') < \omega_{\max}$, то вычисляются матрицы (4.2), (4.3), (4.5), (4.6). Если $\omega(aa') \geq \omega_{\max}$, но $\omega(bb') < \omega_{\max}$, то вычисляются матрицы (4.8), (4.9), (4.10), (4.11).

Шаг 3. Построение матриц (4.12), (4.13), участвующих в разложении (4.1).

Результат. Значение критерия дихотомии ω, проекторы P_{in}, P_{out} и разложение (4.1), если на сторонах угла ab собственных значений матрицы A нет, $\omega = \omega_{\max}$ в случае пересечения сторонами угла ab пятен ε-спектра матрицы при малых значениях ε.

4.3. Случай невозможности дихотомии прямыми aa' и bb'

Более сложной является ситуация, когда на обоих лучах a' и b', продолжающих стороны угла ab, находятся собственные значения матрицы A. В этом случае в алгоритм вводится дополнительный этап, целью которого является приведение исходной матрицы к такому клеточно-диагональному виду

$$A = T_0 \begin{pmatrix} A_0 & \\ & B_0 \end{pmatrix} T_0^{-1}, \tag{4.14}$$

где у матрицы A_0 нет собственных значений на прямых aa' и bb'. Описанный ниже алгоритм является одним из возможных методов построения разложения (4.14).

Разделим угол $b'a$ на равные части лучами $c_1, c_2, \ldots, c_{n-1}$. Так как у матрицы A всего n собственных значений с учетом кратности, и на прямых aa' и bb' по предположению находятся собственные значения, то найдется прямая $c_k c_k'$, на которой точки спектра матрицы A будут отсутствовать. Здесь c_k' – луч, продолжающий луч c_k, при этом внутренность угла ab лежит в левой полуплоскости $\mathbb{C}^l_{c_k c_k'}$. Таким образом, дополнительным этапом алгоритма является дихотомия спектра матрицы A прямой $c_k c_k'$. Результатом этого этапа являются матрицы проекторов

$$P_0 = P^l_{c_k c_k'}, \quad I_n - P_0 = P^l_{c_k c_k'},$$

базисы $U^r_{c_k c_k'}$, $U^l_{c_k c_k'}$ и матрица перехода

$$T_0 = \left[U^l_{c_k c_k'}, U^r_{c_k c_k'}\right], \tag{4.15}$$

которая позволяет представить исходную матрицу в виде (4.14).

Далее к матрице A_0 применяется упрощенный алгоритм. В итоге будут получены следующие матрицы:

$$P_{\text{in}} = P_0 \begin{bmatrix} P_1 & \\ & 0_{n-n_0} \end{bmatrix} \begin{bmatrix} P_2 & \\ & 0_{n-n_1} \end{bmatrix}, \quad P_{\text{out}} = I_n - P_{\text{in}}, \tag{4.16}$$

$$T = T_0 \begin{bmatrix} T_1 & \\ & I_{n-n_0} \end{bmatrix} \begin{bmatrix} T_2 & \\ & I_{n-n_1} \end{bmatrix}, \quad A_{\text{in}} = A_2, \quad B_{\text{out}} = \begin{bmatrix} B_2 & & \\ & B_1 & \\ & & B_0 \end{bmatrix}, \tag{4.17}$$

где $n_0 = \operatorname{tr} P_0$, $n_1 = \operatorname{tr} P_1$.

Алгоритм дихотомии углом

Дано: матрица A, угол ab на комплексной плоскости с вершиной в точке (0, 0), максимально допустимое значение критерия дихотомии $\omega_{\max}$.

Шаг 1. Проверка наличия собственных значений матрицы на лучах a и b. Если $\omega(ab) < \omega_{\max}$, то выполняются шаги 2, 3.

Шаг 2. Проверка наличия собственных значений матрицы A на прямых aa', bb'.

Если $\omega(aa') < \omega_{\max}$ или $\omega(bb') < \omega_{\max}$, то $A_0 = A$, $T_0 = I_n$ и выполняется шаг 4.

Если $\omega(aa') \geq \omega_{\max}$ и $\omega(bb') \geq \omega_{\max}$, то выполняется шаг 3.

Шаг 3. Строятся лучи $c_1, \ldots c_{n-1}$, делящие угол $b'a$ на равные части.

В цикле находится прямая $c_k c_k'$, для которой $\omega(c_k c_k') < \omega_{\max}$. Производится дихотомия спектра матрицы A этой прямой и вычисляется разложение (4.14) с приводящей матрицей (4.15).

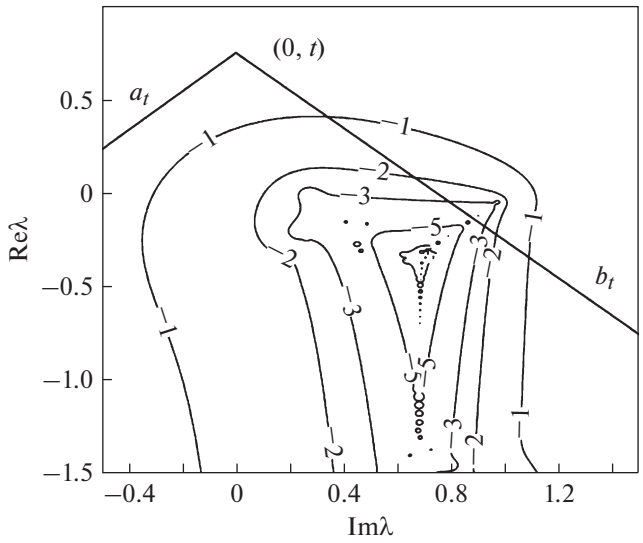
Шаг 4. Применение упрощенного алгоритма к матрице A_0.

Результат. Значение критерия дихотомии ω, проекторы P_{in}, P_{out} (см. формулы (4.16), (4.17)) и разложение (4.1), если на сторонах угла ab собственных значений матрицы A нет, $\omega = \omega_{\max}$ в случае пересечения сторонами угла ab пятен ε-спектра матрицы при малых значениях ε.

5. ПРИМЕРЫ

5.1. Определение сектора дифференциального оператора

В качестве первого примера мы рассмотрим спектральную задачу для уравнения Орра–Зоммерфельда для плоскопараллельного течения Пуазейля вязкой несжимаемой жидкости. Изучению спектральных характеристик оператора Орра–Зоммерфельда посвящено множество работ, например, уже упоминавшиеся [4], [6], [17]. В статье [26] реализована идея разложения собственных функций в ряды в граничных точках, а обширный список литературы, приведенный в этой работе, содержит ссылки на разносторонние исследования оператора Орра–Зоммерфельда. Среди них можно отдельно отметить [27]. Изложенный в данной работе подход перекли-



Фиг. 2. Спектральный портрет матрицы $\mathcal{A}$, аппроксимирующей оператор Орра–Зоммерфельда.

кается с задачей дихотомии, так как алгоритм применяется к точкам границы области с некоторым шагом, а позволяет сделать вывод о числе собственных значений, лежащих внутри области.

С одной стороны, спектральная задача для уравнения Орра–Зоммерфельда имеет важное прикладное значение с точки зрения исследования развития турбулентности. С другой – это содержательный пример несамосопряженного дифференциального оператора, и исследование его спектра представляет собой нетривиальную, но при этом хорошо изученную задачу. Таким образом, оператор Орра–Зоммерфельда является идеальным объектом для апробации новых вычислительных методов.

Рассмотрим матричный пучок $A - \lambda B$, являющийся дискретной аппроксимацией оператора Орра–Зоммефельда [28], [29]

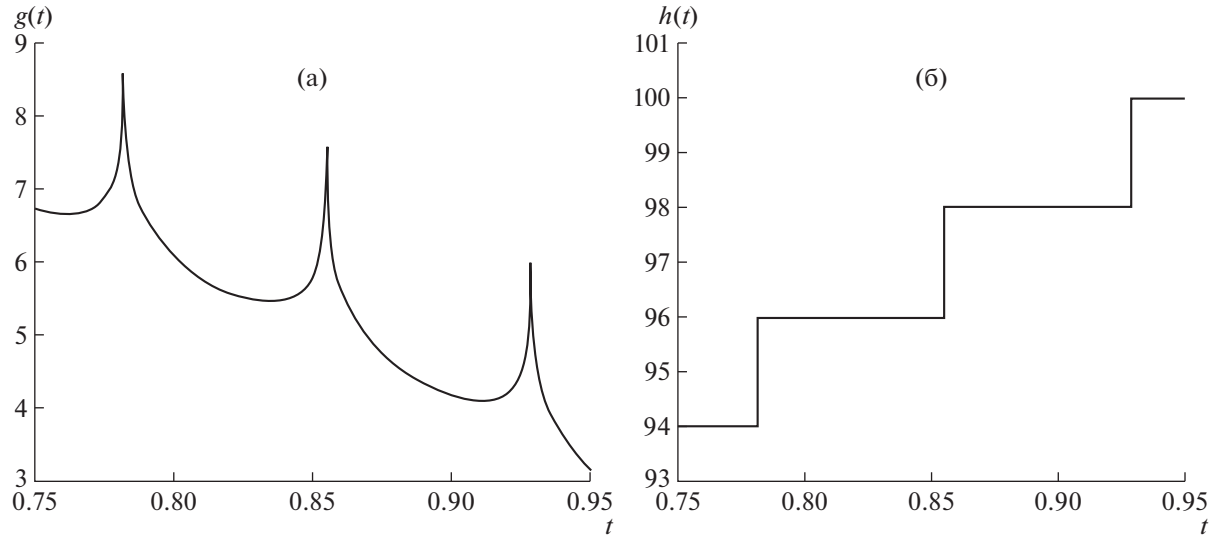
$$A = \alpha\ \mathrm{diag}(U)\left(D_2 - \kappa^2 I\right) + 2\alpha I + i\left(D_4 - 2\kappa^2 D_2 + \kappa^4 I\right)/\mathrm{Re}, \quad B = D_2 - \kappa^2 I.$$

Здесь D_2, D_4 – матрицы коллокационных производных второго и четвертого порядков, учитывающие однородные условия Дирихле и Неймана, I – единичная матрица, α и $\kappa^2 = \alpha^2 + \beta^2$ – спектральные параметры, Re – число Рейнольдса, $U(y) = 1 - y^2$, $-1 \le y \le 1$ – профиль основного течения Пуазейля, $\mathrm{diag}(U)$ – диагональная матрица, на диагонали которой стоят значения функции U в точках Гаусса-Лобатто $y_j = \cos \pi/n$. Далее будем полагать $n = 100$, $\mathrm{Re} = 6000$, $\alpha = 1.02$, $\beta = 0$.

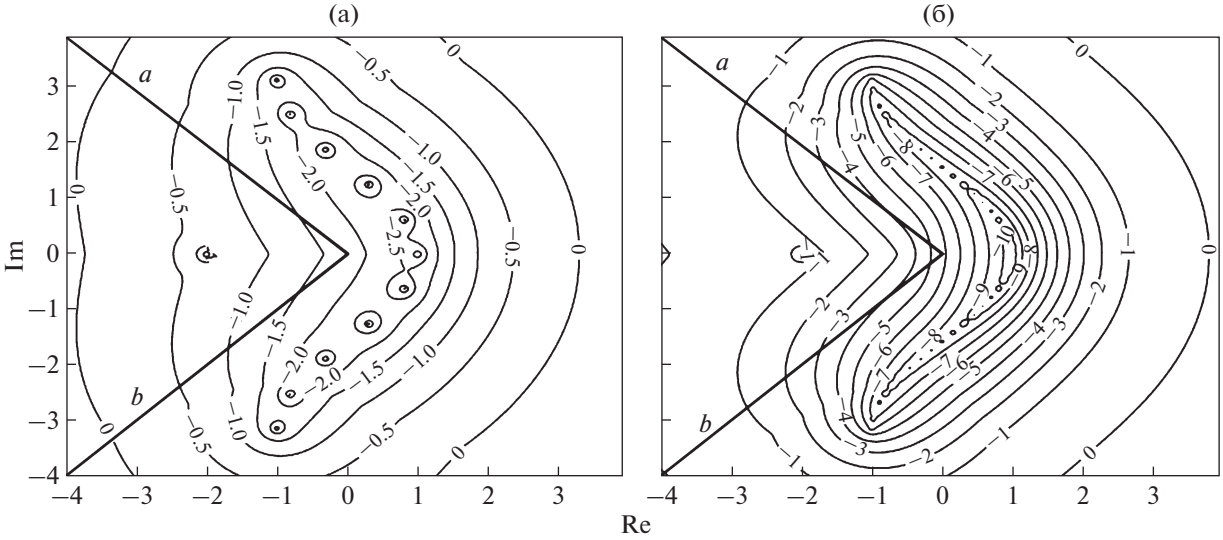
Часть ε-спектра матрицы $\mathcal{A} = B^{-1}A$ изображена на фиг. 2 в виде линий уровня функции $f(\lambda) = \log_{10}\sigma_{\min}(\mathcal{A} - \lambda I)$. Цифрами обозначены значения, которые принимает функция $f(\lambda)$ на данном контуре.

Пусть стороны угла ab образуют с положительной вещественной полуосью углы $5\pi/4$ и $7\pi/4$. Рассмотрим семейство ломаных $a_t b_t$, полученных из угла ab переносом его вершины вдоль мнимой оси в точку $(0, t)$ (см. фиг. 2). Очевидно, что дихотомия спектра матрицы $\mathcal{A}$ относительно угла $a_t b_t$ равносильна дихотомии спектра матрицы $\mathcal{A} - itI$ относительно угла ab. Для каждого t из заданного интервала вычислим критерий дихотомии $\omega(a_t b_t)$ и проектор $P_{\mathrm{in}}(a_t b_t)$ на инвариантное подпространство, соответствующее собственным значениям матрицы $\mathcal{A}$, находящимся внутри угла $a_t b_t$.

На фиг. 3а изображен график функции $g(t) = \log_{10}\omega(a_t b_t)$, значения которой конечны, если на сторонах угла $a_t b_t$ нет точек спектра матрицы $\mathcal{A}$, и бесконечны в обратном случае. То есть “пики”



Фиг. 3. График критерия дихотомии относительно угла $a_t b_t$ (а), след проектора на инвариантное подпространство матрицы $\mathcal{A}$, соответствующее собственным значениям внутри угла $a_t b_t$ (б).



Фиг. 4. Спектральные портреты матриц с дугообразным спектром при $n = 10$ (а) и при $n = 40$ (б).

графика $g(t)$ указывают на значения t, при которых собственные значения $\mathcal{A}$ оказываются на стороне угла $a_t b_t$.

На фиг. 3б изображен график кусочно-постоянной функции $h(t) = \mathrm{tr}\ P_{\mathrm{in}}(a_t b_t)$. Значения этой функции совпадают с числом собственных значений матрицы $\mathcal{A}$, находящихся внутри угла $a_t b_t$. При прохождении стороны угла $a_t b_t$ через точки спектра функция меняет значение на величину, равную числу собственных значений, оказавшихся в этот момент на стороне угла. На представленном графике величина каждого такого изменения равна двум. Это согласуется с тем фактом (см., например, [28]), что кратность собственных значений на ветви P (группа собственных значений, для которых $\mathrm{Re}\lambda \to 1$) равна двум, каждому из них соответствуют две собственные функции – четная и нечетная. Кроме того, при $t > t_0 = 0.9284$ функция $h(t)$ тождественно равна размерности матрицы $n = 100$. Это означает, что сектор с вершиной в точке $(0, t)$, $t > t_0$, и углом полураствора $\pi/4$ содержит все собственные значения матрицы $\mathcal{A}$.

Таблица 1

	Размерность n	Критерий дихотомии углом $\log \omega(ab)$	Критерий дополнительной дихотомии	Точность проектора $\log \|P^2 - P\|$	Точность проектора $\log \|PA - AP\|$
1	10	2.5	4.0	−13.9	−21.9
2	20	5.2	8.5	−11.4	−20.6
3	30	7.9	13.2	−8.7	−17.4
4	35	9.3	15.6	−7.6	−15.7
5	40	10.6	>16	–	–
6	40	10.6	13.3	−9.6	−12.3

5.2. Случай дугообразного спектрального пятна

Описанный в предыдущем разделе алгоритм даст ложноотрицательный результат, если пятно ε-спектра матрицы для достаточно малого ε имеет вид дуги, охватывающей вершину угла. В этом случае может оказаться, что при конечном значении критерия дихотомии $\omega(ab)$ предварительное разделение спектра любой прямой cc' невозможно, а значит, невозможно выполнение алгоритма в целом.

Для того, чтобы решить эту проблему, можно произвести предварительное разделение спектра (шаг 3 алгоритма) относительно какой-то выпуклой кривой, охватывающей вершину угла, но полностью лежащей внутри спектральной дуги.

Рассмотрим пример двухдиагональной матрицы A размера $n \times n$, на главной диагонали которой стоят числа

$$A(j, j) = \cos\xi_j + i\xi_j, \quad \text{где} \quad \xi_j = \pi\left(2\frac{j}{n} - n\right), \quad 1 \le j \le n,$$

и $A(n+1, n+1) = -2$, а на побочной диагонали $A(j, j+1) = 2$. Изображение ε-спектра этой матрицы при $n = 10$ и $n = 40$ приведено на фиг. 4.

Зафиксируем угол ab, стороны которого образуют с положительной вещественной полуосью углы $3\pi/4$ и $5\pi/4$, и применим алгоритм из разд. 4 дихотомии относительно угла ab к этим матрицам. При этом положим $\omega_{\max} = 10^{16}$. Результаты вычислений представлены в строках 1–5 табл. 1.

Значения, приведенные в последних двух столбцах, говорят о том, что проекторы вычисляются с большой точностью.

В строке 5 табл. 1 мы видим, что при $n = 40$ $\omega(ab) < \omega_{\max}$, т.е. спектральные пятна достаточно хорошо разделены сторонами угла. Однако вспомогательная дихотомия относительно одной из прямых $c_k c_k'$ невозможна, так как $\omega(c_k c_k') \ge \omega_{\max}$. Пробелы в последних двух ячейках этой строки говорят о том, что при $\omega \ge \omega_{\max}$ дальнейшие вычисления не проводились.

Однако дихотомия спектра при $n = 40$ может быть осуществлена, если для дополнительной дихотоми вместо прямой выбрать окружность. В данном конкретном случае это окружность радиуса $R = 3$ с центром в точке $(-3, 0)$. Результаты вычислений см. в табл. 1, строка 6.

6. ЗАКЛЮЧЕНИЕ

Данная работа продолжает серию публикаций, посвященных новой постановке спектральной задачи – дихотомии матричного спектра относительно заданной кривой. Впервые поставлена задача разделения спектра относительно кусочно-гладкой кривой, а именно относительно угла, и представлен алгоритм для ее решения. Очевидно, что идея этого метода может быть использована при решении широкого круга задач о разделении спектра матрицы относительно кусочно-гладких кривых, состоящих из дуг кривых второго порядка и отрезков прямых.

СПИСОК ЛИТЕРАТУРЫ

1. *Годунов С.К.* Современные аспекты линейной алгебры. Новосибирск: Научная книга, 1997. С. 388.
2. *Годунов С.К.* Дихотомия спектра и критерий устойчивости для секториальных операторов // Сиб. мат. журн. 1995. Т. 36. № 6. С. 1328–1335.
3. *Trefethen L.N., Embree M.* Spectra and Pseudospectra. Princeton University Press, 2005. P. 606.
4. *Trefethen L.N., Trefethen A.E., Satish C.R., Tobin A.* Hydrodynamic Stability without Eigenvalues // Science, New Series. 1993. V. 261. № 5121. P. 578–584.
5. *Toh K.-Ch., Trefethen L.* Calculation of Pseudospectra by the Arnoldi Iteration // SIAM J. Sci. Comput. 1999. V. 17. № 1. P. 1–15.
6. *Reddy S.C., Schmid P.J., Henningson D.S.* Pseudospectra of the Orr–Sommerfeld Operator // SIAM J. on Applied Mathematics. 1993. V. 53. № 1. P. 15–47.
7. *Постников М.М.* Устойчивые многочлены. М.: Наука, 1981. С. 176.
8. *Крейн М.Г., Неймарк М.А.* Метод симметрических и эрмитовых форм в теории отделения корней алгебраических уравнений. Харьков: ГНТИ Украины, 1936. С. 40.
9. *Fujiwara M.* Über die Nullstellen der ganzen Funktionen vom Geschlecht Null und Eins // Tôhoku Math. J. 1925. V. 25. P. 29.
10. *Далецкий Ю.Г., Крейн М.Г.* Устойчивость решений дифференциальных уравнений в банаховом пространстве. М.: Наука, 1970. С. 534.
11. *Бибердорф Э.А.* Гарантированная точность в прикладных задачах линейной алгебры. Новосибирск: РИЦ НГУ, 2008. С. 145.
12. *Годунов С.К.* Круговая дихотомия матричного спектра // Сиб. матем. журн. 1986. Т. 27. № 5. С. 24–37.
13. *Булгаков А.Я., Годунов С.К.* Круговая дихотомия матричного спектра // Сиб. матем. журн. 1988. Т. 29. № 5. С. 59–70.
14. *Малышев А.Н.* Введение в вычислительную линейную алгебру. Новосибирск: Наука, Сибирское отделение. 1991. С. 228.
15. *Godunov S.K., Sadkane M.* Spectral Analysis of Symplectic Matrices with Application to the Theory of Parametric Resonance // SIAM J. on Matrix Analysis and Applications. 2006. V. 28. № 4. С. 1083–1096.
16. *Буньков В.Г., Годунов С.К., Курзин В.Б., Садкане М.* Применение нового математического аппарата “Одномерные спектральные портреты матрицы” к решению проблемы аэроупругих колебаний решеток лопастей // Ученые записки ЦАГИ. 2009. Т. 40. № 6. С. 3–13.
17. *Бибердорф Э.А., Блинова М.А., Попова Н.И.* Модификации метода дихотомии матричного спектра и их применение к задачам устойчивости // Сиб. ж. вычисл. матем. 2018. Т. 21. № 2. С. 139–153.
18. *Godunov S.K., Sadkane M.* Elliptic dichotomy of a matrix spectrum // Linear Algebra and its Applications. 1996. V. 248. P. 205–232.
19. *Malyshev A.N., Sadkane M.* On parabolic and elliptic spectral dichotomy // SIAM Journal on Matrix Analysis and its Applications. 1997. V. 18. P. 265–278.
20. *Блинова М.А., Попова Н.И., Бибердорф Э.А.* Приложение дихотомии матричного спектра к исследованию устойчивости течений // Марчуковские научные чтения – 2017. Труды Международной научной конференции. 2017. С. 106–112.
21. *Бибердорф Э.А.* Критерий дихотомии корней полинома единичной окружностью // Сиб. журн. индустр. матем. 2000. Т. 3. № 1. С. 16–32.
22. *Biberdorf E.* Development of the matrix spectrum dichotomy method // Continuum mechanics, applied mathematics and scientific computing: Godunov’s legacy – A liber amicorum to Professor Godunov; Book series: Advanced Structured Materials. 2020. P. 37–43.
23. *Демиденко Г.В., Матвеева И.И.* Обыкновенные дифференциальные уравнения в задачах. Новосибирск: ИПЦ НГУ, 2021. С. 246.
24. *Годунов С.К.* Обыкновенные дифференциальные уравнения с постоянными коэффициентами. Новосибирск: Изд. НГУ, 1994. С. 263.
25. *Фадеев С.И., Когай В.В.* Линейные и нелинейные краевые задачи для систем обыкновенных дифференциальных уравнений. Новосибирск: ИПЦ НГУ, 2018. С. 290.
26. *Скороходов С.Л.* Численный анализ спектра задачи Орра–Зоммерфельда // Ж. вычисл. матем. и матем. физ. 2007. Т. 47. № 10. С. 1672–1691.
27. *Курочкин С.В.* Метод выявления неустойчивости и поиска неустойчивых собственных значений в задаче Орра–Зоммерфельда // Ж. вычисл. матем. и матем. физ. 2001. Т. 41. № 1. С. 86–94.
28. *Бойко А.А., Грек Г.Р., Довгаль А.В., Козлов В.В.* Физические механизмы перехода к турбулентности в открытых течениях. Ижевск: НИЦ “Регулярная и хаотическая динамика”, 2006. С. 301.
29. *Trefethen L.N.* Spectral Methods in MATLAB. SIAM. Philadelphia. 2000. P. 163.

EDN: AUOYFN
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 757–767

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.613

О ПАРАХ СИММЕТРИЧНЫХ ТЁПЛИЦЕВЫХ МАТРИЦ, КВАДРАТЫ КОТОРЫХ СОВПАДАЮТ[1)]

© 2022 г. В. Н. Чугунов

119333 Москва, ул. Губкина, 8, Институт вычислительной математики им. Г.И. Марчука, Россия

e-mail: chugunov.vadim@gmail.com

Поступила в редакцию 10.09.2021 г.
Переработанный вариант 10.09.2021 г.
Принята к публикации 14.01.2022 г.

Дано полное описание пар симметричных тёплицевых матриц, квадраты которых совпадают. Библ. 3.

Ключевые слова: тёплицева матрица, циркулянт, косой циркулянт, инволютивная матрица.

DOI: 10.31857/S0044466922050039

1. ПОСТАНОВКА ЗАДАЧИ

Тёплицевой называется комплексная $n \times n$-матрица T, имеющая вид

$$T = \begin{pmatrix} t_0 & t_1 & t_2 & \dots & t_{n-1} \\ t_{-1} & t_0 & t_1 & \dots & t_{n-2} \\ t_{-2} & t_{-1} & t_0 & \dots & t_{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ t_{-n+1} & t_{-n+2} & t_{-n+3} & \dots & t_0 \end{pmatrix}. \tag{1}$$

Хорошо известными частными случаями тёплицевых матриц являются циркулянты и косые циркулянты. Тёплицева матрица (1) называется *циркулянтом*, если

$$t_{-j} = t_{n-j}, \quad j = 1, 2, \dots, n-1,$$

и *косым циркулянтом* при

$$t_{-j} = -t_{n-j}, \quad j = 1, 2, \dots, n-1.$$

Обобщением циркулянтов и косых циркулянтов служат φ-*циркулянты* – тёплицевы матрицы, для которых

$$t_{-j} = \varphi t_{n-j}, \quad j = 1, 2, \dots, n-1,$$

где φ – некоторое число.

Рассмотрим следующую задачу: описать пары симметричных тёплицевых матриц (T_1, T_2), удовлетворяющих условиям

$$T_1^2 = T_2^2, \quad T_1 \neq \pm T_2.$$

В предлагаемой работе дается полное решение этой задачи в виде списка множеств требуемых пар матриц. В разд. 2 формулируется теорема, являющаяся главным результатом статьи, доказательство которой проводится в разд. 4. В разд. 3 приводятся вспомогательные утверждения.

Напомним вначале некоторые определения и факты. Согласно [1], если C – циркулянт, то для него справедливо спектральное разложение

$$C = F_n^* D F_n, \tag{2}$$

[1)]Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках реализации программы Московского центра фундаментальной и прикладной математики (соглашение № 075-15-2019-1624).

где $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ – диагональная матрица, F_n – (нормированная) матрица дискретного преобразования Фурье

$$F_n = \frac{1}{\sqrt{n}}\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & \epsilon & \epsilon^2 & \ldots & \epsilon^{n-1} \\ 1 & \epsilon^2 & \epsilon^4 & \ldots & \epsilon^{2(n-1)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & \epsilon^{n-1} & \epsilon^{2(n-1)} & \ldots & \epsilon^{(n-1)^2} \end{pmatrix}$$

и $\epsilon = \exp\left(\frac{2\pi i}{n}\right)$ – первообразный корень n-й степени из единицы.

Если S – косой циркулянт, то вместо (2) имеем

$$S = G_{-1}F_n^* D F_n G_{-1}^*, \tag{3}$$

где

$$G_{-1} = \mathrm{diag}\left(1, \psi, \psi^2, \ldots, \psi^{n-1}\right),$$

$\psi = e^{\frac{i\pi}{n}}$ есть корень n-й степени из (-1).

В дальнейшем мы будем использовать матрицу-перестановку

$$\mathscr{P}_n = \begin{pmatrix} & & & 1 \\ & & 1 & \\ & \ldots & & \\ & 1 & & \\ 1 & & & \end{pmatrix}, \tag{4}$$

называемую иногда перъединичной матрицей.

2. ГЛАВНЫЙ РЕЗУЛЬТАТ

Теорема. *Ненулевые симметричные тёплицевы матрицы T_1 и T_2 удовлетворяют условиям*

$$T_1^2 = T_2^2, \quad T_1 \neq \pm T_2, \tag{5}$$

тогда и только тогда, когда они входят хотя бы в один из описываемых ниже классов:

Класс 1. Матрицы T_1 и T_2 являются циркулянтами, связанными соотношением

$$T_2 = T_1 C_0,$$

где C_0 – симметричный нескалярный инволютивный циркулянт.

Класс 2. Матрицы T_1 и T_2 суть косые циркулянты, для которых выполнятся равенство

$$T_2 = T_1 S_0.$$

Здесь S_0 – симметричный нескалярный инволютивный косой циркулянт.

Класс 3. Матрицы T_1 и T_2 имеют вид

$$T_1 = \alpha C_0 + \beta S_0,$$
$$T_2 = \alpha S_0 + \beta C_0.$$

При этом C_0 и S_0 – симметричные инволютивные циркулянт и косой циркулянт соответственно, не являющиеся одновременно скалярными матрицами, α, β – некоторые числа, $\alpha \neq \pm\beta$.

3. ВСПОМОГАТЕЛЬНЫЕ УТВЕРЖДЕНИЯ

Для доказательства главного результата нам понадобятся некоторые дополнительные утверждения. Начнем с результата, который принадлежит к тёплицеву фольклору. Все знают о нем, но никто не знает первоисточника.

Лемма 1. *Произведение нескалярных тёплицевых матриц T_1 и T_2 тогда и только тогда является тёплицевой матрицей, когда T_1 и T_2 принадлежат хотя бы одному из следующих классов*:

Класс 1'. Матрицы T_1 и T_2 суть φ-циркулянты для одного и того же числа $\varphi \neq 0$.

Класс 2'. Обе матрицы T_1 и T_2 – верхнетреугольные или же обе – нижнетреугольные.

Другим результатом, нужным в дальнейшем, является следующий факт.

Лемма 2. *Матрица T является нескалярным симметричным φ-циркулянтом тогда и только тогда, когда T – симметричный циркулянт или косой циркулянт.*

Доказательство леммы 2. Так как достаточность очевидна, то установим лишь необходимость.

Пусть T – φ-циркулянт с первой строкой $t_0, t_1, \ldots, t_{n-1}$. В силу нескалярности T найдется число $j > 0$ такое, что $t_j \neq 0$. Из определения φ-циркулянта и его симметричности имеем

$$t_j = t_{-j} = \varphi t_{n-j} = \varphi t_{-(n-j)} = \varphi^2 t_j,$$

или

$$\left(1 - \varphi^2\right) t_j = 0.$$

Так как $t_j \neq 0$, то $\varphi = \pm 1$. Лемма 2 доказана.

Из лемм 1 и 2 следует

Лемма 3. *Квадрат симметричной нескалярной тёплицевой матрицы T тогда и только тогда является тёплицевой матрицей, когда T – симметричный циркулянт или косой циркулянт.*

Также нам потребуются критерии симметричности циркулянта и косого циркулянта.

Лемма 4. *Циркулянт C со спектральным разложением* (2) *является симметричной матрицей тогда и только тогда, когда*

$$d_j = d_{n+2-j}, \quad j = 2, 3, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor.$$

Доказательство леммы 4. Запишем условие симметричности циркулянта C, используя спектральное разложение (2):

$$F_n^* D F_n = F_n D F_n^*.$$

После умножения слева и справа на F_n приходим к соотношению

$$D F_n^2 = F_n^2 D.$$

Так как $F_n^2 = \mathcal{P}_1 \oplus \mathcal{P}_{n-1}$ (см. [2, лемма 1.2.17]), получаем утверждение леммы. Лемма 4 доказана.

Лемма 5. *Пусть S – косой циркулянт, для которого записано спектральное разложение* (3). *Матрица S является симметричной тогда и только тогда, когда*

$$d_1 = d_2, \quad d_j = d_{n+3-j}, \quad j = 3, \ldots, \left\lfloor \frac{n}{2} \right\rfloor + 1.$$

Доказательство леммы 5. Запишем условие симметричности косого циркулянта S, используя спектральное разложение (3):

$$G_{-1} F_n^* D F_n G_{-1}^* = G_{-1}^* F_n D F_n^* G_{-1}.$$

Умножение слева на $F_n G_{-1}^*$, а справа на $G_{-1}^* F_n$ приводит к равенству

$$D F_n \left(G_{-1}^*\right)^2 F_n = F_n \left(G_{-1}^*\right)^2 F_n D.$$

Учитывая лемму 1.2.20 из [2], имеем

$$F_n\left(G_{-1}^*\right)^2 F_n = \mathcal{P}_2 \oplus \mathcal{P}_{n-2}.$$

Вместе с предыдущим равенством это дает утверждение леммы. Лемма 5 доказана.

4. ДОКАЗАТЕЛЬСТВО ГЛАВНОГО РЕЗУЛЬТАТА

В этом разделе приведем обоснование теоремы, являющейся основным результатом.

Хорошо известно, что всякую тёплицеву матрицу можно однозначно представить в виде суммы скалярной матрицы, циркулянта и косого циркулянта с нулевыми диагоналями, поэтому запишем матрицы T_1 и T_2 в виде

$$T_1 = t_0^{(1)} I_n + C^{(1)} + S^{(1)}, \quad T_2 = t_0^{(2)} I_n + C^{(2)} + S^{(2)}, \tag{6}$$

где $C^{(1)}$, $C^{(2)}$ – циркулянты, $S^{(1)}$, $S^{(2)}$ – косые циркулянты с нулевыми диагоналями.

Обозначим элементы первых строк циркулянтов $C^{(1)}$ и $C^{(2)}$ через $0, c_1^{(1)}, \ldots, c_{n-1}^{(1)}$ и $0, c_1^{(2)}, \ldots, c_{n-1}^{(2)}$ соответственно. Аналогично, элементы первых строк косых циркулянтов $S^{(1)}$ и $S^{(2)}$ обозначим как $0, s_1^{(1)}, \ldots, s_{n-1}^{(1)}$ и $0, s_1^{(2)}, \ldots, s_{n-1}^{(2)}$.

Подставим представления (6) в (5):

$$\begin{aligned}&\left(t_0^{(1)}\right)^2 I_n + 2t_0^{(1)}C^{(1)} + 2t_0^{(1)}S^{(1)} + \left(C^{(1)}\right)^2 + \left(S^{(1)}\right)^2 + C^{(1)}S^{(1)} + S^{(1)}C^{(1)} = \\ &= \left(t_0^{(2)}\right)^2 I_n + 2t_0^{(2)}C^{(2)} + 2t_0^{(2)}S^{(2)} + \left(C^{(2)}\right)^2 + \left(S^{(2)}\right)^2 + C^{(2)}S^{(2)} + S^{(2)}C^{(2)},\end{aligned} \tag{7}$$

или

$$\begin{aligned}&C^{(1)}S^{(1)} + S^{(1)}C^{(1)} - C^{(2)}S^{(2)} - S^{(2)}C^{(2)} = \\ &= -\left(t_0^{(1)}\right)^2 I_n - 2t_0^{(1)}C^{(1)} - 2t_0^{(1)}S^{(1)} - \left(C^{(1)}\right)^2 - \left(S^{(1)}\right)^2 + \\ &+ \left(t_0^{(2)}\right)^2 I_n + 2t_0^{(2)}C^{(2)} + 2t_0^{(2)}S^{(2)} + \left(C^{(2)}\right)^2 + \left(S^{(2)}\right)^2.\end{aligned}$$

Матрица в правой части, как сумма циркулянтов и косых циркулянтов, тёплицева, значит, и матрица в левой части должна быть тёплицевой:

$$\begin{aligned}&\left\{C^{(1)}S^{(1)} + S^{(1)}C^{(1)} - C^{(2)}S^{(2)} - S^{(2)}C^{(2)}\right\}_{k,m} = \\ &= \left\{C^{(1)}S^{(1)} + S^{(1)}C^{(1)} - C^{(2)}S^{(2)} - S^{(2)}C^{(2)}\right\}_{k+1,m+1},\end{aligned}$$

$$k, m = 1, \ldots, n-1.$$

Подробная запись последнего равенства

$$\begin{aligned}&\sum_{l=1}^{n}\left\{C^{(1)}\right\}_{k,l}\left\{S^{(1)}\right\}_{l,m} + \sum_{l=1}^{n}\left\{S^{(1)}\right\}_{k,l}\left\{C^{(1)}\right\}_{l,m} - \\ &- \sum_{l=1}^{n}\left\{C^{(2)}\right\}_{k,l}\left\{S^{(2)}\right\}_{l,m} - \sum_{l=1}^{n}\left\{S^{(2)}\right\}_{k,l}\left\{C^{(2)}\right\}_{l,m} - \\ &- \sum_{l=1}^{n}\left\{C^{(1)}\right\}_{k+1,l}\left\{S^{(1)}\right\}_{l,m+1} - \sum_{l=1}^{n}\left\{S^{(1)}\right\}_{k+1,l}\left\{C^{(1)}\right\}_{l,m+1} + \\ &+ \sum_{l=1}^{n}\left\{C^{(2)}\right\}_{k+1,l}\left\{S^{(2)}\right\}_{l,m+1} + \sum_{l=1}^{n}\left\{S^{(2)}\right\}_{k+1,l}\left\{C^{(2)}\right\}_{l,m+1} = 0\end{aligned}$$

в силу тёплицевости $C^{(1)}$, $S^{(1)}$, $C^{(2)}$ и $S^{(2)}$ приобретает вид

$$\sum_{l=1}^{n} c_{l-k}^{(1)} s_{m-l}^{(1)} + \sum_{l=1}^{n} s_{l-k}^{(1)} c_{m-l}^{(1)} - \sum_{l=1}^{n} c_{l-k}^{(2)} s_{m-l}^{(2)} - \sum_{l=1}^{n} s_{l-k}^{(2)} c_{m-l}^{(2)} -$$
$$- \sum_{l=1}^{n} c_{l-k-1}^{(1)} s_{m+1-l}^{(1)} - \sum_{l=1}^{n} s_{l-k-1}^{(1)} c_{m+1-l}^{(1)} + \sum_{l=1}^{n} c_{l-k-1}^{(2)} s_{m+1-l}^{(2)} + \sum_{l=1}^{n} s_{l-k-1}^{(2)} c_{m+1-l}^{(2)} = 0.$$

Заменим индекс суммирования l на p, полагая $p = l$ в первых четырех суммах и $p = l-1$ в остальных:

$$\sum_{p=1}^{n} c_{p-k}^{(1)} s_{m-p}^{(1)} + \sum_{p=1}^{n} s_{p-k}^{(1)} c_{m-p}^{(1)} - \sum_{p=1}^{n} c_{p-k}^{(2)} s_{m-p}^{(2)} - \sum_{p=1}^{n} s_{p-k}^{(2)} c_{m-p}^{(2)} -$$

$$- \sum_{p=0}^{n-1} c_{p-k}^{(1)} s_{m-p}^{(1)} - \sum_{p=0}^{n-1} s_{p-k}^{(1)} c_{m-p}^{(1)} + \sum_{p=0}^{n-1} c_{p-k}^{(2)} s_{m-p}^{(2)} + \sum_{p=0}^{n-1} s_{p-k}^{(2)} c_{m-p}^{(2)} = 0.$$

Выполняя элементарные преобразования, приходим к равенству

$$c_{n-k}^{(1)} s_{-(n-m)}^{(1)} - c_{-k}^{(1)} s_{m}^{(1)} + s_{n-k}^{(1)} c_{-(n-m)}^{(1)} - s_{-k}^{(1)} c_{m}^{(1)} - c_{n-k}^{(2)} s_{-(n-m)}^{(2)} + c_{-k}^{(2)} s_{m}^{(2)} - s_{n-k}^{(2)} c_{-(n-m)}^{(2)} + s_{-k}^{(2)} c_{m}^{(2)} = 0.$$

Так как $C^{(1)}$, $C^{(2)}$ – циркулянты, $S^{(1)}$, $S^{(2)}$ – косые циркулянты, то можем записать

$$-c_{n-k}^{(1)} s_{m}^{(1)} - c_{n-k}^{(1)} s_{m}^{(1)} + s_{n-k}^{(1)} c_{m}^{(1)} + s_{n-k}^{(1)} c_{m}^{(1)} + c_{n-k}^{(2)} s_{m}^{(2)} + c_{n-k}^{(2)} s_{m}^{(2)} - s_{n-k}^{(2)} c_{m}^{(2)} - s_{n-k}^{(2)} c_{m}^{(2)} = 0,$$

или

$$c_{n-k}^{(1)} s_{m}^{(1)} - s_{n-k}^{(1)} c_{m}^{(1)} - c_{n-k}^{(2)} s_{m}^{(2)} + s_{n-k}^{(2)} c_{m}^{(2)} = 0.$$

Заменяя k на $n-k$, получаем

$$c_{k}^{(1)} s_{m}^{(1)} - c_{m}^{(1)} s_{k}^{(1)} = c_{k}^{(2)} s_{m}^{(2)} - c_{m}^{(2)} s_{k}^{(2)}. \tag{8}$$

Введем в рассмотрение две вспомогательные $(n-1)\times 2$-матрицы $\mathcal{F}$ и $\mathcal{G}$, задавая их формулами

$$\mathcal{F} = \begin{bmatrix} c_1^{(1)} & s_1^{(1)} \\ c_2^{(1)} & s_2^{(1)} \\ \vdots & \vdots \\ c_{n-1}^{(1)} & s_{n-1}^{(1)} \end{bmatrix}, \quad \mathcal{G} = \begin{bmatrix} c_1^{(2)} & s_1^{(2)} \\ c_2^{(2)} & s_2^{(2)} \\ \vdots & \vdots \\ c_{n-1}^{(2)} & s_{n-1}^{(2)} \end{bmatrix},$$

и векторы $c^{(1)} = \left(c_1^{(1)}, c_2^{(1)}, \ldots, c_{n-1}^{(1)}\right)^{\mathrm{T}}$, $c^{(2)} = \left(c_1^{(2)}, c_2^{(2)}, \ldots, c_{n-1}^{(2)}\right)^{\mathrm{T}}$, $s^{(1)} = \left(s_1^{(1)}, s_2^{(1)}, \ldots, s_{n-1}^{(1)}\right)^{\mathrm{T}}$, $s^{(2)} = \left(s_1^{(2)}, s_2^{(2)}, \ldots, s_{n-1}^{(2)}\right)^{\mathrm{T}}$.

Из условия симметричности матриц T_1 и T_2 следует симметричность $C^{(1)}$, $C^{(2)}$, $S^{(1)}$ и $S^{(2)}$, поэтому имеем соотношения

$$\begin{aligned} \mathcal{P}_{n-1} c^{(1)} &= c^{(1)}, \\ \mathcal{P}_{n-1} s^{(1)} &= -s^{(1)}, \\ \mathcal{P}_{n-1} c^{(2)} &= c^{(2)}, \\ \mathcal{P}_{n-1} s^{(2)} &= -s^{(2)}. \end{aligned} \tag{9}$$

Определим величины

$$\Delta_{km}^{\mathcal{F}} = \det\begin{pmatrix} c_k^{(1)} & s_k^{(1)} \\ c_m^{(1)} & s_m^{(1)} \end{pmatrix} = c_k^{(1)} s_m^{(1)} - c_m^{(1)} s_k^{(1)},$$

$$\Delta_{km}^{\mathcal{G}} = \det\begin{pmatrix} c_k^{(2)} & s_k^{(2)} \\ c_m^{(2)} & s_m^{(2)} \end{pmatrix} = c_k^{(2)} s_m^{(2)} - c_m^{(2)} s_k^{(2)}.$$

Теперь (8) принимает вид

$$\Delta_{km}^{\mathcal{F}} = \Delta_{km}^{\mathcal{G}}, \quad k,m = 1, \ldots, n-1. \tag{10}$$

На основании равенства (10) рассмотрим несколько взаимоисключающих случаев, определяемых значениями рангов матриц $\mathcal{F}$ и $\mathcal{G}$, и в каждом из них найдем решение уравнения (5).

I. Матрицы $\mathcal{F}$ и $\mathcal{G}$ нулевые, тогда T_1 и T_2 являются скалярными матрицами и условия (5) не выполнены.

II. Матрица $\mathcal{G}$ нулевая, а $\mathcal{F}$ ненулевая. В этом случае матрица T_2 будет скалярной. Из уравнения (5) получаем, что T_1 – скалярное кратное инволютивной матрицы и, кроме того, квадрат T_1 является тёплицевой матрицей. По лемме 3 заключаем, что либо T_1 – симметричный циркулянт и пара (T_1, T_2) принадлежит классу 3 с $\beta = 0$ (скалярная матрица является косым циркулянтом), либо T_1 является симметричным косым циркулянтом и пара (T_1, T_2) принадлежит классу 3 с $\alpha = 0$.

III. Матрица $\mathcal{F}$ нулевая, $\mathcal{G}$ ненулевая. Повторяя рассуждения предыдущего случая, снова получаем, что пара (T_1, T_2) принадлежит классу 3.

IV. Матрицы $\mathcal{F}$ и $\mathcal{G}$ ненулевые и $\operatorname{rank}\mathcal{F} = 1$. В равенствах (10) все миноры $\Delta_{km}^{\mathcal{F}} = 0$, а потому и все миноры $\Delta_{km}^{\mathcal{G}} = 0$. Поскольку $\mathcal{G}$ – ненулевая матрица, то $\operatorname{rank}\mathcal{G} = 1$.

Так как $\operatorname{rank}\mathcal{F} = 1$, найдется ненулевой вектор $z^{(1)}$ такой, что $c^{(1)}$ и $s^{(1)}$ можно представить в виде $c^{(1)} = \gamma_1 z^{(1)}$ и $s^{(1)} = \delta_1 z^{(1)}$. Числа γ_1 и δ_1 удовлетворяют условию

$$|\gamma_1| + |\delta_1| \neq 0.$$

Соотношения (9) запишем в виде

$$\gamma_1 \mathcal{P}_{n-1} z^{(1)} = \gamma_1 z^{(1)}, \quad \delta_1 \mathcal{P}_{n-1} z^{(1)} = -\delta_1 z^{(1)},$$

или

$$\gamma_1 \left(\mathcal{P}_{n-1} z^{(1)} - z^{(1)}\right) = 0, \quad \delta_1 \left(\mathcal{P}_{n-1} z^{(1)} + z^{(1)}\right) = 0.$$

Если $\gamma_1\delta_1 \neq 0$, то из условий $\mathcal{P}_{n-1} z^{(1)} = z^{(1)}$ и $\mathcal{P}_{n-1} z^{(1)} = -z^{(1)}$ получаем, что $z^{(1)}$ – нулевой вектор. Однако $z^{(1)} \neq 0$, поэтому $\gamma_1\delta_1 = 0$.

Так как $\operatorname{rank}\mathcal{G} = 1$, найдется ненулевой вектор $z^{(2)}$ такой, что $c^{(2)}$ и $s^{(2)}$ можно представить в виде $c^{(2)} = \gamma_2 z^{(2)}$ и $s^{(2)} = \delta_2 z^{(2)}$. Числа γ_2 и δ_2 удовлетворяют условию

$$|\gamma_2| + |\delta_2| \neq 0.$$

Из (9) имеем

$$\gamma_2 \mathcal{P}_{n-1} z^{(2)} = \gamma_2 z^{(2)}, \quad \delta_2 \mathcal{P}_{n-1} z^{(2)} = -\delta_2 z^{(2)},$$

или

$$\gamma_2 \left(\mathcal{P}_{n-1} z^{(2)} - z^{(2)}\right) = 0, \quad \delta_2 \left(\mathcal{P}_{n-1} z^{(2)} + z^{(2)}\right) = 0.$$

Если $\gamma_2\delta_2 \neq 0$, то из равенств $\mathcal{P}_{n-1}z^{(2)} = z^{(2)}$ и $\mathcal{P}_{n-1}z^{(2)} = -z^{(2)}$ следует, что $z^{(2)}$ – нулевой вектор. Однако $z^{(2)} \neq 0$, поэтому $\gamma_2\delta_2 = 0$.

Приходим к совокупности соотношений

$$c^{(1)} = \gamma_1 z^{(1)}, \quad s^{(1)} = \delta_1 z^{(1)}, \quad \gamma_1\delta_1 = 0, \quad |\gamma_1| + |\delta_1| \neq 0,$$
$$c^{(2)} = \gamma_2 z^{(2)}, \quad s^{(2)} = \delta_2 z^{(2)}, \quad \gamma_2\delta_2 = 0, \quad |\gamma_2| + |\delta_2| \neq 0,$$

из которых заключаем, что возможны четыре взаимоисключающих случая, определяемых равенством нулю или отличием от нуля чисел γ_1, δ_1, γ_2 и δ_2.

Если $\delta_1 = \delta_2 = 0$, то T_1 и T_2 – циркулянты, которые запишем как

$$T_1 = F_n^* D_1 F_n, \quad T_2 = F_n^* D_2 F_n,$$

где $D_1 = \operatorname{diag}\left(d_1^{(1)}, d_2^{(1)}, \ldots, d_n^{(1)}\right)$ и $D_2 = \operatorname{diag}\left(d_1^{(2)}, d_2^{(2)}, \ldots, d_n^{(2)}\right)$ – диагональные матрицы.

Решаемое уравнение приобретает вид

$$D_1^2 = D_2^2, \quad D_1 \neq \pm D_2,$$

из которого имеем, что

$$D_2 = D_1 D_0,$$

где $D_0 = \operatorname{diag}\left(d_1^{(0)}, d_2^{(0)}, \ldots, d_n^{(0)}\right)$ – нескалярная инволютивная диагональная матрица, для которой

$$d_j^{(0)} = d_{n+2-j}^{(0)}, \quad j = 2, 3, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor.$$

Последние условия нужны для обеспечения симметричности T_2.

В результате получаем соотношение

$$T_2 = F_n^* D_2 F_n = F_n^* D_1 F_n F_n^* D_0 F_n = T_1 C_0,$$

где C_0 – симметричный нескалярный инволютивный циркулянт. Пара (T_1, T_2) принадлежит классу 1.

Если $\gamma_1 = \gamma_2 = 0$, то T_1 и T_2 суть косые циркулянты вида

$$T_1 = G_{-1} F_n^* D_1 F_n G_{-1}^*, \quad T_2 = G_{-1} F_n^* D_2 F_n G_{-1}^*,$$

где $D_1 = \operatorname{diag}\left(d_1^{(1)}, d_2^{(1)}, \ldots, d_n^{(1)}\right)$ и $D_2 = \operatorname{diag}\left(d_1^{(2)}, d_2^{(2)}, \ldots, d_n^{(2)}\right)$ – диагональные матрицы.

Подставляя в уравнение (5), снова имеем соотношение

$$D_2 = D_1 D_0,$$

где $D_0 = \operatorname{diag}\left(d_1^{(0)}, d_2^{(0)}, \ldots, d_n^{(0)}\right)$ – нескалярная инволютивная диагональная матрица, для которой

$$d_1^{(0)} = d_2^{(0)},$$
$$d_j^{(0)} = d_{n+3-j}^{(0)}, \quad j = 3, \ldots, \left\lfloor \frac{n}{2} \right\rfloor + 1,$$

что обеспечивает симметричность T_2.

В этом случае можем записать

$$T_2 = G_{-1} F_n^* D_2 F_n G_{-1}^* = G_{-1} F_n^* D_1 F_n G_{-1}^* G_{-1} F_n^* D_0 F_n G_{-1}^* = T_1 S_0,$$

где S_0 – симметричный нескалярный инволютивный косой циркулянт. Пара (T_1, T_2) принадлежит классу 2.

Если $\delta_1 = \gamma_2 = 0$, то T_1 – циркулянт, T_2 – косой циркулянт. Равенство $T_1^2 = T_2^2$ можно рассматривать как систему

$$T_1^2 = \xi I_n,$$
$$T_2^2 = \xi I_n,$$

из которой следует, что T_1 и T_2 – скалярные кратные инволютивных циркулянта и косого циркулянта соответственно (класс 3 с $\beta = 0$).

Если $\delta_2 = \gamma_1 = 0$, то, рассуждая как в случае выше, приходим к ситуации, когда T_1 и T_2 – скалярные кратные инволютивных косого циркулянта и циркулянта соответственно. Получаем пару из класса 3 для $\alpha = 0$.

V. Пусть теперь матрицы $\mathcal{F}$ и $\mathcal{G}$ ненулевые и $\operatorname{rank}\mathcal{F} > 1$. Так как $\mathcal{F}$ – матрица размера $(n-1)\times 2$, то $\operatorname{rank}\mathcal{F} = 2$. Поэтому в равенствах (10) найдется ненулевой минор $\Delta_{km}^{\mathcal{F}}$, а значит, и ненулевой минор $\Delta_{km}^{\mathcal{G}}$. Тем самым $\operatorname{rank}\mathcal{G} = 2$.

Применяя лемму из [3], можем написать

$$\mathcal{G} = \mathcal{F}W. \tag{11}$$

Представим матрицу W в виде

$$W = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}.$$

Определитель этой матрицы равен единице:

$$w_{11}w_{22} - w_{12}w_{21} = 1. \tag{12}$$

Используя соотношения (9), можем написать

$$\mathcal{P}_{n-1}\mathcal{F} = \mathcal{F}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathcal{P}_{n-1}\mathcal{G} = \mathcal{G}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Умножая равенство (11) слева на $\mathcal{P}_{n-1}$, получаем соотношение

$$\mathcal{G}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \mathcal{F}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}W,$$

или

$$\mathcal{G} = \mathcal{F}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}W\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Почленно вычитая из последнего равенства соотношение (11), имеем

$$\mathcal{F}\left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}W\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - W\right] = 0.$$

Так как матрица $\mathcal{F}$ имеет полный ранг, можем написать

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}W\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - W = 0,$$

или с учетом вида матрицы W:

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = 0.$$

Приходим к равенству

$$\begin{pmatrix} w_{11} & -w_{12} \\ -w_{21} & w_{22} \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix},$$

из которого следует, что $w_{12} = w_{21} = 0$. Пусть $w_{11} = \lambda$, тогда, так как матрица W имеет определитель 1, то $w_{22} = 1/\lambda$ и W – матрица вида

$$W = \begin{pmatrix} \lambda & 0 \\ 0 & \dfrac{1}{\lambda} \end{pmatrix}.$$

Поэтому верны соотношения

$$C^{(2)} = \lambda C^{(1)}, \quad S^{(2)} = \frac{1}{\lambda} S^{(1)}. \tag{13}$$

При этом из условия $\operatorname{rank} \mathscr{T} = 2$ следует, что $C^{(1)}$ и $S^{(1)}$ – нескалярные матрицы.

Рассмотрим сначала два особых случая. Если $\lambda = 1$, то $C^{(2)} = C^{(1)}$, $S^{(2)} = S^{(1)}$ и, используя (6), можем записать

$$T_1 = t_0^{(1)} I_n + \hat{T}, \quad T_2 = t_0^{(2)} I_n + \hat{T}$$

для некоторой нескалярной матрицы $\hat{T}$. Подстановка в уравнение (5) дает условие

$$\left(t_0^{(1)} - t_0^{(2)}\right)\left(\left(t_0^{(1)} + t_0^{(2)}\right) I_n + 2\hat{T}\right) = 0,$$

из которого следует, что $t_0^{(1)} = t_0^{(2)}$ и, поэтому, $T_1 = T_2$. Этот случай не дает новых классов. Аналогично для $\lambda = -1$.

Пусть теперь $\lambda \neq \pm 1$. Подстановка представлений (13) в (7) дает соотношение

$$\left(t_0^{(1)}\right)^2 I_n + 2t_0^{(1)} C^{(1)} + 2t_0^{(1)} S^{(1)} + \left(C^{(1)}\right)^2 + \left(S^{(1)}\right)^2 =$$
$$= \left(t_0^{(2)}\right)^2 I_n + 2t_0^{(2)} \lambda C^{(1)} + 2\frac{t_0^{(2)}}{\lambda} S^{(1)} + \lambda^2 \left(C^{(1)}\right)^2 + \frac{1}{\lambda^2}\left(S^{(1)}\right)^2,$$

или

$$\left(1 - \lambda^2\right)\left(C^{(1)}\right)^2 + 2\left(t_0^{(1)} - \lambda t_0^{(2)}\right) C^{(1)} =$$
$$= \left(\left(t_0^{(2)}\right)^2 - \left(t_0^{(1)}\right)^2\right) I_n + \left(\frac{1}{\lambda^2} - 1\right)\left(S^{(1)}\right)^2 + 2\left(\frac{t_0^{(2)}}{\lambda} - t_0^{(1)}\right) S^{(1)}.$$

Так как в последнем равенстве слева стоит циркулянт, справа косой циркулянт, то это соотношение эквивалентно системе

$$\begin{gathered} \left(1 - \lambda^2\right)\left(C^{(1)}\right)^2 + 2\left(t_0^{(1)} - \lambda t_0^{(2)}\right) C^{(1)} = \xi I_n, \\ \left(\left(t_0^{(2)}\right)^2 - \left(t_0^{(1)}\right)^2\right) I_n + \left(\frac{1}{\lambda^2} - 1\right)\left(S^{(1)}\right)^2 + 2\left(\frac{t_0^{(2)}}{\lambda} - t_0^{(1)}\right) S^{(1)} = \xi I_n \end{gathered} \tag{14}$$

для некоторого числа ξ.

Запишем циркулянт $C^{(1)}$ в виде $C^{(1)} = F_n^* D_1 F_n$ и подставим в первое уравнение (14)

$$\left(1 - \lambda^2\right) F_n^* D_1^2 F_n + 2\left(t_0^{(1)} - \lambda t_0^{(2)}\right) F_n^* D_1 F_n = \xi I_n.$$

После домножения слева на F_n, а справа на F_n^* имеем

$$\left(1 - \lambda^2\right) D_1^2 + 2\left(t_0^{(1)} - \lambda t_0^{(2)}\right) D_1 - \xi I_n = 0.$$

Получаем, что каждый диагональный элемент матрицы D_1 должен удовлетворять одному и тому же квадратному уравнению

$$\left(1 - \lambda^2\right) x^2 + 2\left(t_0^{(1)} - \lambda t_0^{(2)}\right) x - \xi = 0 \tag{15}$$

относительно x. Для квадратного уравнения $ax^2 + bx + c = 0$ условимся записывать корни как

$$\frac{-b \pm \left(b^2 - 4ac\right)^{1/2}}{2a} = \gamma \pm \delta.$$

Тогда, если обозначить корни уравнения (15) как $\gamma_1 \pm \delta_1$, то диагональную матрицу D_1 можно представить в виде

$$D_1 = \gamma_1 I_n + \delta_1 D_0^{(1)},$$

где $D_0^{(1)}$ – нескалярная инволютивная диагональная матрица, подчиненная условиям

$$\left\{D_0^{(2)}\right\}_{jj} = \left\{D_0^{(2)}\right\}_{n+2-j,n+2-j}, \quad j = 2, 3, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor.$$

Тогда для самой матрицы C_1 справедливо представление

$$C^{(1)} = F_n^* D_1 F_n = \gamma_1 I_n + \delta_1 C_0, \tag{16}$$

где C_0 – нескалярный симметричный инволютивный циркулянт.

Проведем аналогичные рассуждения для косого циркулянта $S^{(1)}$. А именно, подставим представление $S^{(1)} = G_{-1} F_n^* D_2 F_n G_{-1}^*$ во второе уравнение (14)

$$\left(\frac{1}{\lambda^2} - 1\right) G_{-1} F_n^* D_2^2 F_n G_{-1}^* + 2\left(\frac{t_0^{(2)}}{\lambda} - t_0^{(1)}\right) G_{-1} F_n^* D_2 F_n G_{-1}^* =$$
$$= \left(\xi - \left(t_0^{(2)}\right)^2 + \left(t_0^{(1)}\right)^2\right) I_n.$$

После умножения слева на $F_n G_{-1}^*$, а справа на $G_{-1} F_n^*$ имеем

$$\left(\frac{1}{\lambda^2} - 1\right) D_2^2 + 2\left(\frac{t_0^{(2)}}{\lambda} - t_0^{(1)}\right) D_2 = \left(\xi - \left(t_0^{(2)}\right)^2 + \left(t_0^{(1)}\right)^2\right) I_n.$$

Получаем, что каждый диагональный элемент матрицы D_2 должен удовлетворять одному и тому же квадратному уравнению

$$\left(\frac{1}{\lambda^2} - 1\right) x^2 + 2\left(\frac{t_0^{(2)}}{\lambda} - t_0^{(1)}\right) x - \left(\xi - \left(t_0^{(2)}\right)^2 + \left(t_0^{(1)}\right)^2\right) = 0 \tag{17}$$

относительно x. Если предположить, что уравнение (17) имеет корни $\gamma_2 \pm \delta_2$, то диагональную матрицу D_2 можно записать в виде

$$D_2 = \gamma_2 I_n + \delta_2 D_0^{(2)},$$

где $D_0^{(2)}$ – нескалярная инволютивная диагональная матрица, подчиненная условиям

$$\left\{D_0^{(2)}\right\}_{11} = \left\{D_0^{(2)}\right\}_{22},$$
$$\left\{D_0^{(1)}\right\}_{jj} = \left\{D_0^{(1)}\right\}_{n+3-j,n+3-j}, \quad j = 3, \ldots, \left\lfloor \frac{n}{2} \right\rfloor + 1.$$

Тогда для самой матрицы $S^{(1)}$ справедливо представление

$$S^{(1)} = G_{-1} F_n^* D_2 F_n G_{-1}^* = \gamma_2 I_n + \delta_2 S_0, \tag{18}$$

в котором S_0 – нескалярный симметричный инволютивный косой циркулянт.

Заметим, что из нескалярности $C^{(1)}$ и $S^{(1)}$ следуют условия

$$\delta_1 \neq 0, \qquad \delta_2 \neq 0. \tag{19}$$

Используя представления (6), (13), (16) и (18), можем записать

$$\begin{gathered} T_1 = t_0^{(1)} I_n + C^{(1)} + S^{(1)} = t_0^{(1)} I_n + \gamma_1 I_n + \delta_1 C_0 + \gamma_2 I_n + \delta_2 S_0 = \\ = \left(t_0^{(1)} + \gamma_1 + \gamma_2\right) I_n + \delta_1 C_0 + \delta_2 S_0 = a I_n + \delta_1 C_0 + \delta_2 S_0, \\ T_2 = t_0^{(2)} I_n + C^{(2)} + S^{(2)} = t_0^{(2)} I_n + \lambda C^{(1)} + \frac{1}{\lambda} S^{(1)} = \\ = t_0^{(2)} I_n + \lambda\gamma_1 I_n + \lambda\delta_1 C_0 + \frac{1}{\lambda}\gamma_2 I_n + \frac{1}{\lambda}\delta_2 S_0 = \\ = \left(t_0^{(2)} + \lambda\gamma_1 + \frac{1}{\lambda}\gamma_2\right) I_n + \lambda\delta_1 C_0 + \frac{1}{\lambda}\delta_2 S_0 = b I_n + \lambda\delta_1 C_0 + \frac{1}{\lambda}\delta_2 S_0, \end{gathered} \tag{20}$$

где $a = t_0^{(1)} + \gamma_1 + \gamma_2$, $b = t_0^{(2)} + \lambda\gamma_1 + \frac{1}{\lambda}\gamma_2$.

Подставим выражения (20) в (5)

$$\left(a^2 + \delta_1^2 + \delta_2^2\right) I_n + 2a\delta_1 C_0 + 2a\delta_2 S_0 = \left(b^2 + \lambda^2\delta_1^2 + \frac{1}{\lambda^2}\delta_2^2\right) I_n + 2b\lambda\delta_1 C_0 + \frac{2b}{\lambda}\delta_2 S_0,$$

или

$$2(a - \lambda b)\delta_1 C_0 = \left(b^2 + \lambda^2\delta_1^2 + \frac{1}{\lambda^2}\delta_2^2 - a^2 - \delta_1^2 - \delta_2^2\right) I_n - 2\left(a - \frac{b}{\lambda}\right)\delta_2 S_0.$$

Матрица в левой части является нескалярным циркулянтом, в правой – косым циркулянтом. Это возможно лишь в случае, если

$$a = \lambda b \tag{21}$$

и

$$2\left(a - \frac{b}{\lambda}\right)\delta_2 S_0 = \left(b^2 + \lambda^2\delta_1^2 + \frac{1}{\lambda^2}\delta_2^2 - a^2 - \delta_1^2 - \delta_2^2\right) I_n. \tag{22}$$

В последнем равенстве матрица в левой части является нескалярной, в правой – скалярной. Чтобы это равенство было верным, должно выполняться условие

$$a = \frac{b}{\lambda}. \tag{23}$$

Так как $\lambda \neq \pm 1$, то из (21) и (23) получаем, что $a = b = 0$ и (22) превращается в условие

$$\delta_2 = \xi\lambda\delta_1, \quad \xi = \pm 1.$$

Подстановка в (20) дает представление для T_1 и T_2

$$T_1 = \delta_1 C_0 + \xi\lambda\delta_1 S_0 = \delta_1 C_0 + \lambda\delta_1(\xi S_0),$$

$$T_2 = \lambda\delta_1 C_0 + \delta_1(\xi S_0).$$

Приходим к классу 3 с $\alpha = \delta_1$, $\beta = \lambda\delta_1$ и нескалярными инволютивными циркулянтом C_0 и косым циркулянтом ξS_0. Из условия $\lambda \neq \pm 1$ и (19) имеем, что $\alpha \neq \pm\beta$. Теорема доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Воеводин В.В., Тыртышников Е.Е.* Вычислительные процессы с тёплицевыми матрицами. М: Наука, 1987.
2. *Чугунов В.Н.* Нормальные и перестановочные тёплицевы и ганкелевы матрицы. М: Наука, 2017.
3. *Ефимов Н.В., Розендорн Е.Р.* Линейная алгебра и многомерная геометрия. М: Наука, 1975.

EDN: RGBVJH
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 768–776

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 517.977

РЕКОНСТРУКЦИЯ ВХОДНОГО ВОЗДЕЙСТВИЯ В ПАРАБОЛИЧЕСКОМ ВКЛЮЧЕНИИ, НЕРАЗРЕШЕННОМ ОТНОСИТЕЛЬНО ПРОИЗВОДНОЙ

© 2022 г. В. И. Максимов

620990 Екатеринбург, ул. С. Ковалевской, 16, Институт математики и механики УрО РАН, Россия

e-mail: maksimov@imm.uran.ru

Поступила в редакцию 03.06.2021 г.
Переработанный вариант 03.06.2021 г.
Принята к публикации 16.12.2021 г.

Рассматривается задача реконструкции распределенных входных воздействий в параболических включениях, неразрешенных относительно производной. Указывается алгоритм решения задачи, который является устойчивым к информационным помехам и погрешностям вычислений. Алгоритм основан на комбинации методов теории некорректных задач и теории позиционного управления. Он позволяет осуществить процесс реконструкции неизвестных входных воздействий на основе неточных измерений решений включений в дискретные достаточно частые моменты времени. Библ. 18.

Ключевые слова: динамическое восстановление, метод управляемых моделей.

DOI: 10.31857/S0044466922040093

1. ВВЕДЕНИЕ

Обсуждается проблема реконструкции распределенных входных воздействий в параболических включениях, неразрешенных относительно производной. Суть проблемы такова. Имеется параболическое включение. Эволюция его фазового состояния, т.е. решение включения, порождается неизвестным входным воздействием. Само решение априори не задано. Требуется организовать процесс реконструкции (восстановления) входа при условии, что в дискретные (достаточно частые) моменты неточно измеряется решение. Указанная выше проблема относится к классу обратных задач динамики и в более общем контексте вкладывается в проблематику теории некорректных задач [1]–[5]. Один из методов исследования подобных задач, основанный на идеях теории позиционного управления [6] и теории некорректных задач [2], был развит в работах [7]–[14]. Суть этого метода состоит в том, что алгоритм реконструкции представляется в виде алгоритма управления некоторой вспомогательной динамической системой – моделью. Управление в модели конструируется на основе текущих измерений решения таким образом, что его реализация во времени приближает неизвестное входное воздействие. В данной статье, продолжающей [7]–[14], указывается алгоритм решения указанной выше проблемы, являющийся устойчивым к информационным и вычислительным помехам. При этом рассматривается случай отсутствия “мгновенных” ограничений на входные воздействия. Другие алгоритмы динамического восстановления “неограниченных” управлений в распределенных системах, основанные на принципе обратной связи, см. в работах [12]–[14], в которых приведена достаточно обширная библиография.

2. ПОСТАНОВКА ЗАДАЧИ. МЕТОД РЕШЕНИЯ

Рассматривается включение следующего вида:

$$\begin{gathered}\beta(y(t,\eta))_t - \Delta y(t,\eta) \ni u(t,\eta), \quad (t,\eta) \in T \times \Omega, \\ y(t,\sigma) = 0, \quad (t,\sigma) \in T \times \Gamma, \\ y(0,\eta) = y_0(\eta) \quad \eta \in \Omega.\end{gathered} \tag{2.1}$$

Здесь $T = [0, \vartheta]$ – промежуток времени, $\vartheta = \text{const} \in (0, +\infty)$, Ω – область в пространстве $\mathbb{R}^n$ с достаточно гладкой границей Γ, Δ – оператор Лапласа, $\beta(\cdot) : R \to R$ – максимально монотонный граф со свойствами: $0 \in \beta(0)$, для некоторого $\omega > 0$ верно неравенство

$$(\beta(r) - \beta(s))(r - s) \geq \omega |r - s|^2 \quad \forall r, s \in R, \tag{2.2}$$

отображение β переводит ограниченные множества в ограниченные множества.

Заметим, что в виде включения (2.1) может быть формализована двухфазная задача Стефана (см., например, [15]–[17]). При этом

$$\beta(r) = \begin{cases} c_1 r, & r < 0, \\ [0, c_2], & r = 0, \\ c_3 r + c_2, & r > 0. \end{cases}$$

Следуя [17], [18], пару функций $\{y(\cdot), v(\cdot)\}$, $y(\cdot) = y(\cdot; v_0, u(\cdot))$, $v(\cdot) = v(\cdot; v_0, u(\cdot))$, удовлетворяющую соотношениям

$$\begin{gathered} v_t(t, \eta) - \Delta y(t, \eta) = u(t, \eta), \quad (t, \eta) \in T \times \Omega, \\ y(t, \sigma) = 0, \quad (t, \sigma) \in T \times \Gamma, \\ v(t, \eta) \in \beta(y(t, \eta)), \quad (t, \eta) \in T \times \Omega, \\ v(0, \eta) = v_0(\eta) \in \beta(y_0(\eta)), \quad \eta \in \Omega, \end{gathered} \tag{2.3}$$

назовем *решением включения* (2.1), если эти функции таковы

$$\begin{gathered} y(\cdot) \in W^{1,2}(T; H) \cap L_2(T; V), \\ v(\cdot) \in W^{1,2}(T; V^*) \cap L_\infty(T; H). \end{gathered}$$

Здесь $H = L_2(\Omega)$, $V = H_0^1(\Omega)$, $V^* = H^{-1}(\Omega)$. Скалярные произведения в последних двух пространствах определяются следующим образом:

$$(x, y)_V = \int_\Omega \{x(\eta) y(\eta) + \nabla x(\eta) \nabla y(\eta)\} d\eta \quad \forall x, y \in V,$$

$$(x, y)_{V^*} = -\left\langle \Delta^{-1} x, y \right\rangle_{V \times V^*} \quad \forall x, y \in V^*,$$

где символ $\langle \cdot, \cdot \rangle_{V \times V^*}$ означает двойственность пространств V и V^*, а оператор Δ действует из $H_0^1(\Omega)$ в $H^{-1}(\Omega)$ (канонический изоморфизм $H_0^1(\Omega)$ на $H^{-1}(\Omega)$).

На протяжении всей статьи мы будем пользоваться известным свойством троек Гельфанда, которое состоит в следующем: двойственность между пространствами V и V^* эквивалентна скалярному произведению в пространстве H:

$$(u, v)_H = \langle v, u \rangle_{V \times V^*} \quad \forall u \in H \subset V^*, \quad v \in V \subset H.$$

Имеет место

Теорема 1 (см. [16, с. 152]). *Пусть* $\eta \to v_0(\eta) \in L_2(\Omega)$, $y_0(\eta) = \beta^{-1}(v_0(\eta)) \in H_0^1(\Omega)$. *Тогда для любого* $u(\cdot) \in L_2(T; H)$ *существует единственное решение включения* (2.1).

Заметим, что при выполнении условия (2.2) отображение β^{-1} однозначно и липшицево. Поэтому будем полагать, что функция $y_0(\eta)$ удовлетворяет условию теоремы 1.

Обсуждаемая в данной статье задача такова. На промежутке времени T реализуется решение включения (2.1), порождаемое неизвестным входным воздействием $u(\cdot) \in L_2(T; H)$. Промежуток T разбит на конечное число полуинтервалов $[\tau_i, \tau_{i+1})$, $i \in [0 : m-1]$, $\tau_{i+1} = \tau_i + \delta$, $\tau_0 = 0$, $\tau_m = \vartheta$. В узлах разбиения $\tau_i \in \Delta = \{\tau_i\}_{i=0}^m$, измеряются (приближенно) величины $v(\tau_i)$, т.е. находятся элементы $\xi_i^h \in V^*$ со свойствами:

$$\left| v(\tau_i) - \xi_i^h \right|_{V^*} \leq h. \tag{2.4}$$

Здесь $h \in (0, 1)$ – величина погрешности измерения. Решение включения (2.1) неизвестно. Задача состоит в приближенном восстановлении $u(\cdot)$ на основе неточного измерения $v(\tau_i)$.

Для решения сформулированной выше задачи воспользуемся методом вспомогательных позиционно-управляемых моделей [6]–[10]. Согласно этому методу задача реконструкции неизвестного входного воздействия по результатам измерения решения заменяется другой задачей, а именно задачей позиционного управления вспомогательной системой, называемой *моделью*. Таким образом, задача восстановления функции $u(\cdot)$ сводится к следующим двум задачам:

1) задаче выбора вспомогательной модели;

2) задаче формирования управления моделью по принципу обратной связи.

В работе (см. [10, гл. I]) было отмечено, что для достаточно широкого класса систем с распределенными параметрами в качестве моделей удобно брать “копии” реальных систем. Оказывается, что и в нашей ситуации в качестве модели можно брать “копию” включения (2.1), которая имеет следующий вид:

$$\begin{gathered} w_t^h(t,\eta) - \Delta z^h(t,\eta) = u^h(t,\eta), \quad (t,\eta) \in T \times \Omega, \\ z^h(t,\sigma) = 0, \quad (t,\sigma) \in T \times \Gamma, \\ w^h(t,\eta) \in \beta(z^h(t,\eta)), \quad (t,\eta) \in T \times \Omega, \\ w^h(0,\eta) = v_0(\eta) \in \beta(y_0(\eta)), \quad \eta \in \Omega. \end{gathered} \tag{2.5}$$

Здесь $u^h(\cdot)$ – управление, закон формирования которого требуется сконструировать. Решение модели (2.5) определяется аналогично решению уравнения (2.1) (см. (2.3)).

Закон формирования управления в модели $u^h(\cdot)$ (при каждом $h \in (0, 1)$) отождествим с парой $S_h = (\Delta_h, \mathcal{U}_h)$, где

$$\Delta_h = \{\tau_{h,i}\}_{i=0}^{m_h}$$

есть разбиение отрезка T на полуинтервалы $[\tau_{h,i}, \tau_{h,i+1})$, $\tau_{h,i+1} = \tau_{h,i} + \delta$, $\delta = \delta(h)$, $\tau_{h,0} = 0$, $\tau_{h,m_h} = \vartheta$, $\mathcal{U}_h$ – отображение, ставящее в соответствие каждой тройке $p_i^h = \{\tau_i, \xi_i^h, w^h(\tau_i)\}$ элемент

$$v_i^h = \mathcal{U}_h(p_i^h) \in H. \tag{2.6}$$

При этом управление $u^h(\cdot)$ в правой части включения (2.5) определяется по правилу

$$u^h(t) = v_i^h, \quad t \in [\tau_i, \tau_{i+1}). \tag{2.7}$$

3. АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ

Опишем алгоритм решения рассматриваемой задачи. Фиксируем некоторую функцию $\alpha = \alpha(h) : (0, 1) \to (0, 1)$, $\alpha(h) \to 0$ при $h \to 0$.

Отображение $\mathcal{U}_h$ зададим следующим образом:

$$\mathcal{U}_h(p_i^h) = \arg\min\left\{2(\Delta^{-1}(w^h(\tau_i) - \xi_i^h), u)_H + \alpha|u|_H^2 : u \in H\right\} = \alpha^{-1}\Delta^{-1}(\xi_i^h - w^h(\tau_i)), \tag{3.1}$$

где $\alpha = \alpha(h)$.

После того как модель M и семейство $S_h = (\Delta_h, \mathcal{U}_h)$ выбраны, работу алгоритма восстановления $u(\cdot)$ осуществляем по следующей схеме. До начального момента фиксируем погрешность h, число $\alpha = \alpha(h)$ и разбиение $\Delta = \Delta_h = \{\tau_i\}_{i=0}^m$, $(\tau_i = \tau_{h,i}, m = m_h)$ отрезка T с шагом $\delta = \delta(h)$. На i-м шаге алгоритма, осуществляемом на промежутке времени $[\tau_i, \tau_{i+1})$, выполняем следующие операции. Сначала измеряем (с ошибкой) фазовое состояние $v(\tau_i)$, т.е. находим элемент $\xi_i^h \in V^*$ со свойством (2.4). Затем, зная $w^h(\tau_i)$ и ξ_i^h, по правилу (2.6), (2.7), (3.1) определяем управление в модели (2.5). После этого вместо траектории $w^h(t)$, $z^h(t)$, $t \in [0, \tau_i]$, формируем фазовую траекторию $w^h(t)$, $z^h(t)$, $t \in (\tau_i, \tau_{i+1}]$, т.е. осуществляем корректировку памяти.

Прежде, чем перейти к доказательству основных утверждений работы, приведем одну теорему, которая нам понадобится в дальнейшем.

Теорема 2 (см. [17, предложение 1.3, с. 125]). *Пусть выполнены условия теоремы* 1. *Тогда*

$$y(\cdot; y_0, v_0, u_j(\cdot)) \to y(\cdot; y_0, v_0, u(\cdot)) \quad в \quad C(T; H),$$

$$v(\cdot; y_0, v_0, u_j(\cdot)) \to v(\cdot; y_0, v_0, u(\cdot)) \quad в \quad C(T; V^*),$$

если $u_j(\cdot) \to u(\cdot)$ *слабо в* $L_2(T; H)$ *при* $j \to \infty$. *Кроме того, для всех* $u(\cdot) \in L_2(T; H)$ *справедливы оценки*

$$\left|y_t(\cdot; y_0, v_0, u(\cdot))\right|_{L_2(T;H)} + \left|y(\cdot; y_0, v_0, u(\cdot))\right|_{L_\infty(T;V)} \le b_1\left(1 + \left|u(\cdot)\right|_{L_2(T;H)}\right),$$

$$\left|v_t(\cdot; y_0, v_0, u(\cdot))\right|_{L_2(T;V^*)} + \left|v(\cdot; y_0, v_0, u(\cdot))\right|_{L_\infty(T;H)} \le b_2\left(1 + \left|u(\cdot)\right|_{L_2(T;H)}\right),$$

где $b_k = b_k\left(|y_0|_V, |v_0|_H\right)$, $k = 1, 2$.

Справедлива

Теорема 3. *Пусть* $\alpha(h) \to 0$, $\delta(h)\alpha^{-1}(h) \to 0$, $h^2\delta^{-1}(h)\alpha^{-1}(h) \to 0$ *при* $h \to 0$. *Тогда найдется такое* $h_* \in (0, 1)$, *что при всех* $h \in (0, h_*)$ *справедливы неравенства*

$$\varepsilon(t) + 2\omega\int_0^t \left|z^h(\tau) - y(\tau)\right|_H^2 d\tau \le d_1\left(h^2\delta^{-1}(h) + \delta(h) + \alpha(h)\right) \quad при\ п.в. \quad t \in T, \tag{3.2}$$

$$\int_0^\vartheta \left|u^h(t)\right|_H^2 dt \le \rho_1(\alpha(h), \delta(h))\int_0^\vartheta |u(t)|_H^2 dt + \rho(h, \alpha(h), \delta(h)), \tag{3.3}$$

где d_1 *– постоянная, не зависящая от* h, α, δ,

$$\varepsilon(t) = \left|w^h(t) - v(t)\right|_{V^*}^2, \quad \rho_1(\alpha(h), \delta(h)) \to 1, \quad \rho(h, \alpha(h), \delta(h)) \to 0 \quad при \quad h \to 0.$$

Доказательство. Вычтем (2.3) из (2.5) и полученное выражение умножим (скалярно в V^*) на разность $w^h(t) - v(t)$. Будем иметь

$$\left(w_t^h(t) - v_t(t), w^h(t) - v(t)\right)_{V^*} + J_t = J_1(t), \tag{3.4}$$

где

$$J_t = \left(\Delta(z^h(t) - y(t)), w^h(t) - v(t)\right)_{V^*},$$

$$J_1(t) = \left(u^h(t) - u(t), w^h(t) - v(t)\right)_{V^*}.$$

Заметим, что справедливо равенство

$$J_t = \left(z^h(t) - y(t), w^h(t) - v(t)\right)_H.$$

Поэтому в силу (2.2) имеем неравенство

$$\omega\left|z^h(t) - y(t)\right|_H^2 \le J_t. \tag{3.5}$$

Нетрудно видеть, что при п.в. $t \in \delta_i = [\tau_i, \tau_{i+1}]$ справедливо неравенство

$$J_1(t) \le \left(w^h(\tau_i) - \xi_i^h, u^h(t) - u(t)\right)_{V^*} + \rho_i(t, h). \tag{3.6}$$

Здесь

$$\rho_i(t, h) = c_0\left(\left|u^h(t)\right|_{V^*} + |u(t)|_{V^*}\right)\left(h + \int_{\tau_i}^t \left\{\left|w_\tau^h(\tau)\right|_{V^*} + |v_\tau(\tau)|_{V^*}\right\} d\tau\right).$$

В таком случае, в силу (3.5), (3.6), при п.в. $t \in \delta_i$ имеем

$$2\omega\left|z^h(t) - y(t)\right|_H^2 + \frac{d\varepsilon(t)}{dt} \le 2\left(w^h(\tau_i) - \xi_i^h, u^h(t) - u(t)\right)_{V^*} + 2\rho_i(t,h). \tag{3.7}$$

Если $x \in V^*$, $y \in H$, то в силу известного свойства троек Гельфанда получаем

$$\left\langle \Delta^{-1}x, y\right\rangle_{V\times V^*} = \left(\Delta^{-1}x, y\right)_H.$$

Значит, при п.в. $t \in \delta_i$

$$\left(w^h(\tau_i) - \xi_i^h, u^h(t) - u(t)\right)_{V^*} = \left(\Delta^{-1}(w^h(\tau_i) - \xi_i^h), u^h(t) - u(t)\right)_H.$$

Неравенство (3.7) влечет оценку

$$\begin{aligned} &2\omega\left|z^h(t) - y(t)\right|_H^2 + \frac{d\varepsilon(t)}{dt} + \alpha\left\{\left|u^h(t)\right|_H^2 - \left|u(t)_H^2\right|\right\} \le 2\left(u^h(t), \Delta^{-1}\left(w^h(\tau_i) - \xi_i^h\right)\right)_H + \\ &+ \alpha\left|u^h(t)\right|_H^2 - 2\left(u(t), \Delta^{-1}\left(w^h(\tau_i) - \xi_i^h\right)\right)_H - \alpha\left|u(t)\right|_H^2 + 2\rho_i(t,h) \quad \text{при п.в} \quad t \in \delta_i. \end{aligned} \tag{3.8}$$

Пусть

$$\varepsilon^h(t) = \varepsilon(t) + 2\omega\int_0^t \left|z^h(\tau) - y(\tau)\right|_H^2 d\tau + \alpha\int_0^t \left\{\left|u^h(\tau)\right|_H^2 - \left|u(\tau)\right|_H^2\right\} d\tau.$$

Учитывая правило выбора управления $u^h(\cdot)$ (см. (2.6), (2.7), (3.1)), а также непрерывность вложения H в V^*, из (3.8) получаем при $t \in \delta_i$, $i \in [0 : m-1]$,

$$\begin{aligned} &\varepsilon^h(t) \le \varepsilon^h(\tau_i) + c_1\int_{\tau_i}^t \left\{\left|u^h(\tau)\right|_H + \left|u(\tau)\right|_H\right\} d\tau \times \left(h + \int_{\tau_i}^t \left\{\left|w_\tau^h(\tau)\right|_{V^*} + \left|v_\tau(\tau)\right|_{V^*}\right\} d\tau\right) \le \\ &\le \varepsilon^h(\tau_i) + c_2h^2 + c_3\delta\int_{\tau_i}^t \left\{\left|u^h(\tau)\right|_H^2 + \left|u(\tau)\right|_H^2\right\} d\tau + c_4\delta\int_{\tau_i}^t \left\{\left|w_\tau^h(\tau)\right|_{V^*}^2 + \left|v_\tau(\tau)\right|_{V^*}^2\right\} d\tau. \end{aligned} \tag{3.9}$$

Суммируя правую и левую части (3.9) по i и учитывая теорему 2, при $t \in T$ получаем

$$\begin{aligned} &\varepsilon(t) + 2\omega\int_0^t \left|z^h(\tau) - y(\tau)\right|^2 d\tau + \alpha\int_0^t \left|u^h(\tau)\right|_H^2 d\tau \le \alpha\int_0^t \left|u(\tau)\right|_H^2 d\tau + c_5h^2\delta^{-1} + \\ &+ c_6\delta\left\{1 + \int_0^t \left\{\left|u^h(\tau)\right|_H^2 + \left|u(\tau)\right|_H^2\right\} d\tau\right\}. \end{aligned} \tag{3.10}$$

В таком случае в силу (3.10),

$$\varepsilon(t) + 2\omega\int_0^t \left|z^h(\tau) - y(\tau)\right|_H^2 d\tau + (\alpha - c_6\delta)\int_0^t \left|u^h(\tau)\right|_H^2 d\tau \le (\alpha + c_6\delta)\int_0^t \left|u(\tau)\right|_H^2 d\tau + c_5h^2\delta^{-1} + c_6\delta. \tag{3.11}$$

В свою очередь, из (3.11) следует неравенство

$$\int_0^t \left|u^h(\tau)\right|_H^2 d\tau \le \rho_1(\alpha,\delta)\int_0^t \left|u(\tau)\right|_H^2 d\tau + \rho_2(h,\delta,\alpha) + \rho_3(\alpha,\delta), \tag{3.12}$$

где

$$\rho_1(\alpha,\delta) = (\alpha + c_6\delta)(\alpha - c_6\delta)^{-1},$$

$$\rho_2(h,\delta,\alpha) = c_5h^2(\delta(\alpha - c_6\delta))^{-1}, \quad \rho_3(\alpha,\delta) = c_6\delta(\alpha - c_6\delta)^{-1}.$$

Учитывая сходимости $\delta(h)\alpha^{-1}(h) \to 0$, $h^2\delta^{-1}(h)\alpha^{-1}(h) \to 0$ при $h \to 0$, заключаем, что при $h \to 0$ справедливы соотношения

$$\rho_1(\alpha(h),\delta(h)) \to 1, \quad \rho_2(h,\delta(h),\alpha(h)) \to 0, \quad \rho_3(\alpha(h),\delta(h)) \to 0.$$

Из (3.11), учитывая (3.12), получаем : найдется такое $h_1 \in (0, 1)$, что при всех $h \in (0, h_1)$ справедливо (при п.в. $t \in T$) неравенство

$$\varepsilon(t) + 2\omega\int_0^t \left|z^h(\tau) - y(\tau)\right|_H^2 d\tau \le c_7\left(\alpha(h) + \delta(h) + h^2\delta^{-1}(h)\right). \tag{3.13}$$

Из (3.12), (3.13) следуют оценки (3.2), (3.3). Теорема доказана.

Теорема 4. *Пусть выполнены условия теоремы* 3. *Тогда имеет место сходимость*

$$u^h(\cdot) \to u(\cdot) \quad в \quad L_2 = L_2(T; H) \quad при \quad h \to 0.$$

Доказательство. Покажем, что, каковы бы ни были последовательность чисел $h_j \to 0+$ при $j \to \infty$, а также последовательность элементов $\xi_i^{h_j}$ со свойствами (2.4) (в (2.4) полагаем $h = h_j$) , имеет место сходимость

$$u^{h_j}(\cdot) \to u(\cdot) \quad \text{в} \quad L_2 \quad \text{где} \quad j \to \infty.$$

Здесь и ниже управления $u^{h_j}(\cdot)$ определяются согласно (2.6), (2.7), (3.1), где полагается $h = h_j$. Предполагая противное, заключаем: найдется подпоследовательность последовательности $u^{h_j}(\cdot)$ (для простоты обозначаем ее тем же символом, т.е. $u^{h_j}(\cdot)$) такая, что

$$u^{h_j}(\cdot) \to u_0(\cdot) \quad \text{слабо в} \quad L_2 \quad \text{при} \quad j \to \infty, \tag{3.14}$$

$$u_0(\cdot) \ne u(\cdot). \tag{3.15}$$

Пусть $q^{h_j}(t) = z^{h_j}(t) - y^0(t)$, $p^{h_j}(t) = w^{h_j}(t) - v^0(t)$, где $\left\{z^{h_j}(\cdot), w^{h_j}(\cdot)\right\}$ – решение системы (2.5) при $h = h_j$, а $\left\{v^0(\cdot), y^0(\cdot)\right\}$ – решение системы

$$\begin{gathered} v_t^0(t,\eta) - \Delta y^0(t,\eta) = u_0(t,\eta), \quad (t,\eta) \in T \times \Omega, \\ y^0(t,\sigma) = 0, \quad (t,\sigma) \in T \times \Gamma, \\ v^0(t,\eta) \in \beta(y^0(t,\eta)), \quad (t,\eta) \in T \times \Omega, \\ v^0(0,\eta) = v_0(\eta) \in \beta(y_0(\eta)), \quad y^0(0,\eta) = y_0(\eta), \quad \eta \in \Omega.. \end{gathered} \tag{3.16}$$

Вычтем (3.16) из (2.5) (в (2.5) мы полагаем $h = h_j$). После этого умножим (скалярно в V^*) полученную разность на $p^{h_j}(t)$. Аналогично (3.4) устанавливаем равенство

$$d\tilde{\varepsilon}^{h_j}(t)/dt + \tilde{I}_{1t}^{h_j} = \tilde{I}_{2t}^{h_j} \quad \text{при п.в.} \quad t \in T, \tag{3.17}$$

где

$$\tilde{\varepsilon}^{h_j}(t) = 1/2\left|p^{h_j}(t)\right|_{V^*}^2, \quad \tilde{I}_{1t}^{h_j} = \int_0^t \left(q^{h_j}(\tau), p^{h_j}(\tau)\right)_H d\tau,$$

$$\tilde{I}_{2t}^{h_j} = \int_0^t \left(w^{h_j}(\tau) - v^0(\tau), u^{h_j}(\tau) - u_0(\tau)\right)_{V^*} d\tau.$$

Учитывая монотонность функции $\beta(\cdot)$, заключаем, что верно неравенство

$$\tilde{I}_{1t}^{h_j} \ge 0 \quad \text{при п.в.} \quad t \in T. \tag{3.18}$$

Рассмотрим $\tilde{I}_{2t}^{h_j}$. Имеем

$$\tilde{I}_{2t}^{h_j} = \int_0^t \left(w^{h_j}(\tau) - v(\tau), u^{h_j}(\tau) - u_0(\tau)\right)_{V^*} d\tau + \int_0^t \left(v(\tau) - v^0(\tau), u^{h_j}(\tau) - u_0(\tau)\right)_{V^*} d\tau.$$

Тогда

$$\sup_{t\in T}\left|I_{2t}^{h_j}\right| \to 0 \quad \text{при} \quad j \to \infty. \tag{3.19}$$

Этот факт вытекает из теоремы 3 и слабой сходимости последовательности функций $u^{h_j}(\cdot)$ к $u_0(\cdot)$ (см. (3.14)). В таком случае из (3.17)–(3.19) получаем

$$\overline{\lim_{j\to\infty}} \sup_{t\in T} \tilde{\varepsilon}^{h_j}(t) \to 0.$$

Учитывая это соотношение, а также теорему 2, устанавливаем справедливость равенств

$$\operatorname{vrai}\sup_{t\in T}\left|v^0(t) - v(t)\right|_{V^*} = 0. \tag{3.20}$$

Ввиду свойств графа β,

$$0 = \left(v^0(t) - v(t), y^0(t) - y(t)\right)_H \geq \omega\left|y^0(t) - y(t)\right|_H^2 \quad \text{при п.в.} \quad t \in T.$$

Отсюда получаем

$$y^0(\cdot) = y(\cdot).$$

Кроме того, в силу (3.20)

$$v^0(\cdot) = v(\cdot).$$

Поэтому

$$u_0(\cdot) = u(\cdot).$$

Последнее противоречит (3.14), (3.15). Следовательно,

$$u^{h_j}(\cdot) \to u(\cdot) \quad \text{слабо в} \quad L_2 \quad \text{при} \quad j \to \infty. \tag{3.21}$$

Ввиду известного свойства слабого предела, из (3.21) вытекает неравенство

$$\underline{\lim_{j\to\infty}}\left|u^{h_j}(\cdot)\right|_{L_2} \geq \left|u(\cdot)\right|_{L_2}. \tag{3.22}$$

Кроме того, в силу (3.3) имеет место оценка

$$\left|u^{h_j}(\cdot)\right|_{L_2}^2 \leq \rho_1(\alpha(h),\delta(h))\left|u(\cdot)\right|_{L_2}^2 + \rho(h,\alpha(h),\delta(h)).$$

В таком случае,

$$\overline{\lim_{j\to\infty}}\left|u^{h_j}(\cdot)\right|_{L_2} \leq \left|u(\cdot)\right|_{L_2}. \tag{3.23}$$

Значит (см. (3.22), (3.23)),

$$\overline{\lim_{j\to\infty}}\left|u^{h_j}(\cdot)\right|_{L_2} \leq \left|u(\cdot)\right|_{L_2} \leq \underline{\lim_{j\to\infty}}\left|u^{h_j}(\cdot)\right|_{L_2}.$$

Отсюда следует сходимость

$$\left|u^{h_j}(\cdot)\right|_{L_2} \to \left|u(\cdot)\right|_{L_2} \quad \text{при} \quad j \to \infty. \tag{3.24}$$

Учитывая (3.21) и (3.24), заключаем

$$u^{h_j}(\cdot) \to u(\cdot) \ \text{ в } \ L_2 \quad \text{при} \quad j \to \infty.$$

Теорема доказана.

4. ОЦЕНКА СКОРОСТИ СХОДИМОСТИ АЛГОРИТМА

Установим оценку скорости сходимости алгоритма. В дальнейшем нам понадобится

Лемма 1 (см. [12]). *Пусть заданы две функции*: $t \to a(t) \in L_2(T;W^*)$ *и* $t \to b(t) \in W$, $t \in T$, *причем* $b(\cdot)$ *является функцией с ограниченной вариацией. Если верны неравенства*

$$\left|\int_0^t a(\tau)d\tau\right|_{W^*} \le \varepsilon, \quad |b(t)|_W \le d, \quad t \in T,$$

то справедлива оценка

$$\int_0^\vartheta \langle b(t), a(t)\rangle_{W\times W^*} d\tau \le \varepsilon(\operatorname{var}_T b(t) + d).$$

Здесь W – банахово пространство с нормой $|\cdot|_W$; символ $\operatorname{var}_T b(t)$ означает полную вариацию функции $b(t)$ на промежутке T, а символ $\langle\cdot,\cdot\rangle_{W\times W^*}$ – двойственность между W и W^*.

Теорема 5. *Пусть выполнены условия теоремы* 3, *и функция* $t \to u(t) \in V$ *при* $t \in T$ *является функцией с ограниченной вариацией. Тогда справедлива следующая оценка скорости сходимости алгоритма*:

$$\int_0^\vartheta |u(t) - u^h(t)|_{V^*}^2 dt \le C\left\{\left(\delta + \alpha + h^2\delta^{-1}\right)^{1/2} + \rho(h,\alpha,\delta) + |1 - \rho_1(\alpha,\delta)|\right\}, \tag{4.1}$$

где C – *положительная постоянная, не зависящая от* h, δ, α.

Доказательство. Заметим, что каково бы ни было $v \in V$, справедливо равенство

$$\left(\Delta(y(t) - z^h(t)), v\right)_{V^*} = \left(y(t) - z^h(t), v\right)_H \quad (t \in T).$$

Поэтому

$$\begin{aligned}\left|\int_0^t \left(u(\tau) - u^h(\tau)\right)d\tau\right|_{V^*} &= \sup_{v\in V, |v|_V\le 1}\left\langle v, \int_0^t \left\{v_\tau(\tau) - w_\tau^h(\tau) - \Delta(y(\tau) - z^h(\tau))\right\}d\tau\right\rangle_{V\times V^*} \le \\ &\le |v(t) - w^h(t)|_{V^*} + \sup_{v\in V, |v|_V\le 1}\left(\int_0^t \left(y(\tau) - z^h(\tau)\right)d\tau, v\right)_H.\end{aligned} \tag{4.2}$$

В силу непрерывности вложения пространства V в пространство H найдется такое число $d_* > 0$, что при всех $x \in V$ получим

$$|x|_H \le d_* |x|_V.$$

Значит, $d_*^{-1}|x|_H \le |x|_V$. В таком случае,

$$\sup_{v\in V, |v|_V\le 1}\left(\int_0^t \left(y(\tau) - z^h(\tau)\right)d\tau, v\right)_H \le \sup_{v\in H, |v|_H\le d_*}\left(\int_0^t \left(y(\tau) - z^h(\tau)\right)d\tau, v\right)_H = d_*\left|\int_0^t \left(y(\tau) - z^h(\tau)\right)d\tau\right|_H. \tag{4.3}$$

Из (4.2), учитывая (4.3), получаем

$$\left|\int_0^t \left(u(\tau) - u^h(\tau)\right)d\tau\right|_{V^*} \le |v(t) - w^h(t)|_{V^*} + d_*\int_0^t |y(\tau) - z^h(\tau)|_H d\tau. \tag{4.4}$$

В свою очередь, из (4.4), в силу (3.2), выводим ($\alpha = \alpha(h)$, $\delta = \delta(h)$)

$$\left|\int_0^t \left(u(\tau) - u^h(\tau)\right)d\tau\right|_{V^*} \le C_1\left(\delta + h^2\delta^{-1} + \alpha\right)^{1/2}, \quad t \in T.$$

Воспользовавшись неравенством (3.3), получим

$$\begin{aligned}\left|u(\cdot)-u^h(\cdot)\right|^2_{L_2(T;H)} &\le (1+\rho_1(\alpha,\delta))\left|u(\cdot)\right|^2_{L_2(T;H)} - 2\int_0^\vartheta \left(u(\tau),u^h(\tau)\right)_H d\tau + \rho(h,\alpha,\delta) = \\ &= 2\int_0^\vartheta \left\langle u(\tau),u(\tau))-u^h(\tau)\right\rangle_{V\times V^*} d\tau + \rho(h,\alpha,\delta) + \left|1-\rho_1(\alpha,\delta)\right|\int_0^\vartheta \left|u(\tau)\right|^2_H d\tau.\end{aligned} \tag{4.5}$$

В силу леммы 1 из (4.5) получаем (4.1). Теорема доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М.: Наука, 1978.
2. *Лаврентьев М.М., Романов В.Г., Шишатский С.П.* Некорректные задачи математической физики и анализа. Новосибирск: Наука, 1980.
3. *Banks H.T., Kunisch K.* Estimation techniques for distributed parameter systems. Boston: Birkhäuser, 1989.
4. *Иванов В.К., Васин В.В., Танана В.П.* Теория линейных некорректных задач и ее приложения. М.: Наука, 1978.
5. *Kabanikhin S.I.* Inverse and ill-posed problems. Theory and Applications. In: Inverse and Ill-Posed Problems Series, 55. Berlin: De Gruyter, 2011.
6. *Красовский Н.Н., Субботин А.И.* Позиционные дифференциальные игры. М.: Наука, 1974. 456 с.
7. *Osipov Yu.S., Kryazhimskii A.V.* Inverse problems for ordinary differential equations: dynamical solutions. Amsterdam: Gordon and Breach, 1995.
8. *Осипов Ю.С., Кряжимский А.В., Максимов В.И.* Методы динамического восстановления входов управляемых систем. Екатеринбург: УрО РАН, 2011.
9. *Осипов Ю.С., Васильев Ф.П., Потапов М.М.* Основы метода динамической регуляризации. М.: МГУ, 1999. 238 с.
10. *Maksimov V.I.* Dynamical inverse problems of distributed systems. Utrecht: VSP, 2002.
11. *Осипов Ю.С., Кряжимский А.В., Максимов В.И.* Динамические обратные задачи для параболических систем // Дифференц. ур-ния. 2000. Т. 36. № 5. С. 579–597.
12. *Васильева Е.В., Максимов В.И.* О динамической реконструкции неограниченных управлений в параболическом уравнении // Дифференц. ур-ния. 2003. Т. 39. № 1. С. 23–29.
13. *Favini A., Maksimov V., Pandolfi L.* A deconvolution problem related to a singular system // J. of Mathematical Analysis and Applications. 2004. V. 292. № 1. P. 60–72.
14. *Maksimov V.I.* On dynamical reconstruction of boundary and distributed inputs in a Schlogl equation // J. of Inverse and Ill-Posed Problems. 2019. V. 27. № 6. P. 877–889.
15. *Brezis H.* Problemes unilateraux // J. Math. Pures Appl. 1972. V. 51. P. 1–168.
16. *Barbu V.* Optimal Control of Variational Inequalities. Pitman: London, 1984.
17. *Tiba D.* Optimal Control of Nonsmooth Distributed Parameter Systems. Berlin: Springer Verlag, 1991.
18. *Neittaanmaki P., Tiba D.* Optimal Control of Nonlinear Parabolic Systems. New York: Marcel Dekker, 1994.

EDN: LMNXKR
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 777–789

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 519.85

НЕПРЕРЫВНЫЙ ПРОЕКЦИОННЫЙ ОБОБЩЕННЫЙ ЭКСТРАГРАДИЕНТНЫЙ КВАЗИНЬЮТОНОВСКИЙ МЕТОД ВТОРОГО ПОРЯДКА ДЛЯ РЕШЕНИЯ СЕДЛОВЫХ ЗАДАЧ

© 2022 г. В. Г. Малинов

432000 Ульяновск, ул. Толстого, 42, УлГУ, Россия

e-mail: vgmalinov@mail.ru

Поступила в редакцию 16.09.2020 г.
Переработанный вариант 04.10.2021 г.
Принята к публикации 14.01.2022 г.

Исследуется указанный метод решения седловых задач для выпукло-вогнутых гладких функций с липшицевыми частными градиентами на выпуклом замкнутом подмножестве конечномерного евклидова пространства. Средствами выпуклого анализа доказаны сходимость и экспоненциальная скорость сходимости метода. Библ. 11.

Ключевые слова: выпукло-вогнутая функция, седловая задача, непрерывный проекционный обобщенный экстраградиентный квазиньютоновский метод.

DOI: 10.31857/S0044466922050088

1. ВВЕДЕНИЕ

1.1. Известны множество методов оптимизации и меньшее число методов для решения седловых задач (см. [1]–[11]); последние предназначены для отыскания седловых точек или точек равновесия**.** Напомним, точку $(\mathbf{x}^*, \mathbf{u}^*) \in Q \times U \subset E^n \times E^m$, называют *седловой точкой* всякой функции $\varphi(\mathbf{x}, \mathbf{u})$, $\mathbf{x} \in Q \subset E^n$, $\mathbf{u} \in U \subset E^m$, с непустыми множествами $Q \subset E^n$, $U \subset E^m$, в евклидовых пространствах E^n и E^m, если эта точка есть решение системы неравенств:

$$\varphi(\mathbf{x}^*, \mathbf{u}) \le \varphi(\mathbf{x}^*, \mathbf{u}^*) \le \varphi(\mathbf{x}, \mathbf{u}^*) \quad \forall \mathbf{x} \in Q, \quad \mathbf{u} \in U. \tag{1.1}$$

Существуют непрерывные проекционные методы отыскания седловых точек (НПМОСТ) и минимизации (НПММ), и итеративные для соответствующих математических моделей (ММ) исследуемых процессов. К разработке новых методов решения этих задач приводят экстремальные задачи теории игр, математической экономики, математической физики, оптимального управления [1]–[10]. В связи с наличием все более сложных ММ, приводящих к седловым задачам, и небогатым разнообразием методов их решения, актуальна задача исследования новых НПМОСТ [1]–[3], [7], [10], [11].

В работах [3], [10] детально рассмотрены применения седловых методов, в [10] седловые задачи охарактеризованы “как мостик, через который можно попытаться перенести развитую технику решения задач оптимизации для решения игровых задач”, исследованы управляемые (обратными связями по производной, по невязке и смешанными) НПМОСТ второго порядка. Разработана методика преобразования седловых задач и методов к равновесным. В [11] построены и изучены не исследованные в [10] равновесные методы второго порядка на основе управляемого (дифференциального) НПМОСТ второго порядка. В [3] исследовано, наряду с другими, несколько перспективных игровых равновесных методов. Здесь предлагается и исследуется базовый НПМОСТ второго порядка с переменной метрикой, на основе которого можно построить частные методы решения конкретных случаев седловых и равновесных задач в перечисленных науках.

Рассматриваемые НПМОСТ соответствуют задаче Коши для обыкновенных дифференциальных уравнений (ОДУ), описывающих динамические процессы. Существенно одно из благоприятных свойств дифференциальных моделей — возможность использования численных мето-

дов вычислительной математики, для решения ОДУ в алгоритмах численной реализации НПМОСТ и НПММ [3]–[6].

Седловые задачи для конкретных математических моделей решаются при своих требованиях (к пространствам, множествам и функциям), выражающихся в постановке задачи и влияющих на метод ее решения. В этой работе предлагается и исследуется НПМОСТ для решения задачи в следующей постановке.

1.2. Постановка задачи

Требуется решить задачу об отыскании седловой точки $(\mathbf{x}^*, \mathbf{u}^*) \in Q \times U \subset E^n \times E^m$ функции $\varphi(\mathbf{x}, \mathbf{u})$.

Предполагаем следующее: а) множества $Q \subset E^n$, $U \subset E^m$, $Q \times U \subset E^n \times E^m$ непустые выпуклые замкнутые; б) выпукло-вогнутая функция $\varphi(\mathbf{x}, \mathbf{u})$ с овражными гиперповерхностями уровней определена в окрестности подмножества $W \subset Q \times U \subset E^n \times E^m$, выпукла по $\mathbf{x} \in Q \subset E^n$ и вогнута по $\mathbf{u} \in U \subset E^m$, т.е. для всех фиксированных $\mathbf{u} \in U$ функция $g(\mathbf{x}) = \varphi(\mathbf{x}, \mathbf{u})$ выпукла на $Q \subset E^n$, а $\forall\ \mathbf{x} \in Q$ фиксированного функция $h(\mathbf{u}) = \varphi(\mathbf{x}, \mathbf{u})$ вогнута на $U \subset E^m$; в) множество седловых точек $(\mathbf{x}^*, \mathbf{u}^*)$ функции $\varphi(\mathbf{x}, \mathbf{u})$ на $W \subset E^n \times E^m$ непустое, $W_* = Q_* \times U^* \neq \varnothing$; г) частные градиенты функции $\varphi(\mathbf{x}, \mathbf{u})$ липшицевы на $Q \times U$,

$$\begin{aligned} \|\nabla\varphi_x(\mathbf{x}, \mathbf{u}) - \nabla\varphi_x(\mathbf{x}^\wedge, \mathbf{u}^\wedge)\| &\le L\left(\|\mathbf{x} - \mathbf{x}^\wedge\|^2 + \|\mathbf{u} - \mathbf{u}^\wedge\|^2\right)^{1/2}, \\ \|\nabla\varphi_u(\mathbf{x}, \mathbf{u}) - \nabla\varphi_u(\mathbf{x}^\wedge, \mathbf{u}^\wedge)\| &\le L^0\left(\|\mathbf{x} - \mathbf{x}^\wedge\|^2 + \|\mathbf{u} - \mathbf{u}^\wedge\|^2\right)^{1/2}, \end{aligned} \tag{1.2}$$

где $L > 0$, $L^0 > 0$ – константы Липшица**.** В терминах оператора проектирования седловая точка $(\mathbf{x}^*, \mathbf{u}^*) \in W_*$ задачи (1.1) характеризуется равенствами

$$\mathbf{x}^* = P_Q\left[\mathbf{x}^* - \tau\nabla\varphi_x(\mathbf{x}^*, \mathbf{u}^*)\right], \quad \mathbf{u}^* = P_U\left[\mathbf{u}^* + \tau\nabla\varphi_u(\mathbf{x}^*, \mathbf{u}^*)\right], \quad \tau > 0, \tag{1.3}$$

где $P_Q(\cdot)$ и $P_U(\cdot)$ – операторы проектирования на множества Q и U (см. [3], [4]).

1.3. Траектории не всех НПМОСТ сходятся к седловой точке. Например, сходимость на седловых задачах простейшего итеративного метода проекции градиента седлового [1] доказана лишь при весьма ограничительных предположениях сильной выпукло-вогнутости, что не выполняется для многих нужных классов седловых задач [2]. Поэтому предложено несколько способов устранения этого недостатка. Таковым является и изменение самого НПМОСТ: построением экстраградиентного метода (ЭГМ) [2]; включением в ОДУ управления с помощью прогноза и обратных связей [3], приводящим к ЭГМ. Успешный способ улучшения сходимости воплощен в НПМОСТ, использующих прогноз и, аналогично методам минимизации [4]–[6], операторы переменной метрики.

Цель данной работы в широком смысле — распространение подхода к построению методов минимизации из работ [4]–[6] на НПМОСТ; точнее, обоснование НПМОСТ второго порядка с лучшими свойствами, построенного на основе синтеза идеи и теории: НПММ переменной метрики [4], проекционного обобщенного двухточечного экстраградиентного метода минимизации квазиньютоновского (ПОДЭМК) [5], непрерывного проекционного обобщенного экстраградиентного квазиньютоновского метода минимизации (НПОЭКМ) второго порядка [6], итеративного ПОДЭМК седлового из [7]. Наша цель в узком смысле — исследование предложенного НПОЭКМ *седлового* (НПОЭКМС) *второго порядка*.

Последнее включает обоснование вспомогательных утверждений, доказательство сходимости НПОЭКМС и оценки скорости его сходимости для выпукло-вогнутых функций.

2. ПРЕДЛАГАЕМЫЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ

2.1. Предлагается для решения задачи (1.1)–(1.3) *НПОЭКМС второго порядка*, обычно записываемый в виде задачи Коши для системы ОДУ:

$$\begin{gathered}\alpha(t)\mathbf{x}''(t)+\beta(t)\mathbf{x}'(t)+\mathbf{x}(t)=P_Q[\mathbf{y}(t)-\gamma(t)\mathbf{B}^{-1}(\mathbf{y}(t))\nabla\varphi_{\mathbf{x}}(\mathbf{y}(t),\mathbf{u}(t))],\\ \alpha(t)\mathbf{u}''(t)+\beta(t)\mathbf{u}'(t)+\mathbf{u}(t)=P_U[\mathbf{v}(t)+\lambda(t)\mathbf{G}^{-1}(\mathbf{v}(t))\nabla\varphi_u(\mathbf{y}(t),\mathbf{v}(t))],\\ \mathbf{y}(t)=\mathbf{x}(t)-\sigma(t)\mathbf{x}'(t),\quad \mathbf{v}(t)=\mathbf{u}(t)-\theta(t)\mathbf{u}'(t),\quad t\geq 0,\\ \mathbf{x}(t_0)=\mathbf{x}^0,\quad \mathbf{u}(t_0)=\mathbf{u}^0,\quad \mathbf{x}'(t_0)=\mathbf{x}^1,\quad \mathbf{u}'(t_0)=\mathbf{u}^1,\end{gathered}\tag{2.1}$$

где функции $\alpha(t)\in C^2[0,\infty),\beta(t),\gamma(t),\sigma(t),\theta(t)\in C^1[0,\infty)$, — положительные параметры метода, таковы, что $\lim\limits_{t\to\infty}\alpha(t)=\alpha_0$, $\lim\limits_{t\to\infty}\beta(t)=\beta_0$, $\lim\limits_{t\to\infty}\gamma(t)=\gamma_0>0$; $\lim\limits_{t\to\infty}\sigma(t)=\sigma_0>0$, $\lim\limits_{t\to\infty}\theta(t)=\theta_0>0$, $\alpha(t)\geq\alpha_0>0$, $1>\beta(t)>0, 1>\sigma(t)>0, 1>\theta(t)>\theta_0>0$; $\alpha'(t)<0$, $\beta'(t)<0$, $\sigma'(t)<0$, $\gamma'(t)<0$, $\theta'(t)<0$, $\alpha''(t)>0$.

В частности, этим условиям удовлетворяют такие параметры–функции НПОЭКМС (2.1): $\alpha(t)=\alpha_0+\dfrac{1}{1+t}$, $\beta(t)=\beta_0+\dfrac{1}{2+t}$, $\gamma(t)=\gamma_0+\dfrac{1}{3+t}$, $\lambda(t)=\lambda_0+\dfrac{1}{3+t}$, $\sigma(t)=\sigma_0+\dfrac{1}{t+1}$, $\theta(t)=\theta_0+\dfrac{1}{t+1}$.

Операторы $\mathbf{B}(\mathbf{x}):E^n\to E^n\ \forall\mathbf{x}\in E^n$ фиксированного и $\mathbf{G}(\mathbf{u}):E^m\to E^m\ \forall\mathbf{u}\in E^m$ фиксированного — положительно-определенные самосопряженные линейные, изменяющие метрику пространства; они в (2.1) таковы, что

$$m\|\mathbf{v}\|^2\leq(\mathbf{B}(\mathbf{x})\mathbf{v},\mathbf{v})\leq M\|\mathbf{v}\|^2,\quad 0<m\leq M,\quad \mathbf{v},\mathbf{x}\in E^n,\tag{2.2}$$

$$m\|\mathbf{v}\|^2\leq(\mathbf{G}(\mathbf{u})\mathbf{v},\mathbf{v})\leq M\|\mathbf{v}\|^2,\quad 0<m\leq M,\quad \mathbf{v},\mathbf{u}\in E^m.\tag{2.3}$$

Обратные операторы таковы, что

$$\begin{gathered}\|\mathbf{v}\|^2/M\leq(\mathbf{B}^{-1}(\mathbf{x})\mathbf{v},\mathbf{v})\leq\|\mathbf{v}\|^2/m,\quad \mathbf{v},\mathbf{x}\in E^n,\\ \|\mathbf{v}\|^2/M\leq(\mathbf{G}^{-1}(\mathbf{u})\mathbf{v},\mathbf{v})\leq\|\mathbf{v}\|^2/m,\quad \mathbf{v},\mathbf{u}\in E^m.\end{gathered}\tag{2.4}$$

2.2. Для метода (2.1)–(2.4) характеристики вида (1.3) седловой точки $(\mathbf{x}^*,\mathbf{u}^*)$ имеют следующий вид:

$$\mathbf{x}^*=P_Q\left[\mathbf{x}^*-\gamma(t)\mathbf{B}^{-1}(\mathbf{x}^*)\nabla\varphi_x(\mathbf{x}^*,\mathbf{u}^*)\right],\quad \gamma>0,\tag{2.5}$$

$$\mathbf{u}^*=P_U\left[\mathbf{u}^*+\lambda(t)\mathbf{G}^{-1}(\mathbf{u}^*)\nabla\varphi_u(\mathbf{x}^*,\mathbf{u}^*)\right],\quad \lambda>0.\tag{2.6}$$

2.3. **Замечание.** Отметим следующее.

1. В этой работе для простоты обозначены через $\nabla\varphi_x$ – частный градиент по первому аргументу, а $\nabla\varphi_u$ – по второму.

2. Поскольку в (2.1) операторы проектирования в исходной метрике, критерии проекций $\mathbf{w}=P_Q(\mathbf{v})$ по первой и $\mathbf{s}=P_U(\mathbf{v})$ по второй переменным будут по исходной метрике (см. [8, с. 189]):

$$(\mathbf{w}-\mathbf{v},\mathbf{x}-\mathbf{w})\geq 0,\quad \mathbf{x}\in Q,\quad (\mathbf{s}-\mathbf{v},\mathbf{u}-\mathbf{s})\geq 0,\quad \mathbf{u}\in U.\tag{2.7}$$

(Критерии проекций в новой метрике есть неравенства

$$(\mathbf{B}(\mathbf{x})(\mathbf{w}-\mathbf{v},\mathbf{x}-\mathbf{w})\geq 0,\quad \mathbf{x}\in Q,\quad (\mathbf{G}(\mathbf{u})(\mathbf{s}-\mathbf{v},\mathbf{u}-\mathbf{s})\geq 0,\quad \mathbf{u}\in U,$$

но здесь мы пользуемся (2.7), ибо в (2.1) оператор проектирования в исходной метрике.)

3. Имеются и непрерывные, и итеративные, проекционные методы с операторами проектирования: $P_Q(.)$, $P_U(.)$ – в исходной метрике; $P_Q^{\mathbf{G}(\mathbf{v}(t))}$, $P_U^{\mathbf{G}(\mathbf{u}(t))}$ – в новой метрике.

3. ВСПОМОГАТЕЛЬНЫЕ УТВЕРЖДЕНИЯ

Приведем неравенства, дополняющие необходимый для обоснования сходимости и оценки скорости сходимости методов математический аппарат.

3.1. **Лемма 1.** *Если для* (2.1) *при любом фиксированном* $\mathbf{u} \in U \in E^m$ *выпуклая функция* $g_1(\mathbf{x}) \in C^{1,1}(Q)$ *такова, что* $\nabla g_1(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x})\nabla\varphi_x(\mathbf{x},\mathbf{u})$,

$$\|\nabla g_1(\mathbf{w}) - \nabla g_1(\mathbf{v})\| \le K\|\mathbf{w} - \mathbf{v}\|, \quad K = \frac{L}{m} > 0,$$

то

$$(\nabla g_1(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*) \ge 0, \quad \mathbf{x} \in Q. \tag{3.1}$$

Доказательство. Из характеристического свойства (2.5) седловой точки, критерия проекции $\mathbf{w} = P_Q(\mathbf{v})$, $\mathbf{v} \in H_Q$ из (2.7), и равенства из условия леммы 1, имеем

$$(\mathbf{x}^* - \mathbf{x}^* - \gamma\mathbf{B}^{-1}(\mathbf{x}^*)\nabla\varphi_x(\mathbf{x}^*,\mathbf{u}^*), \mathbf{x} - \mathbf{x}^*) = \gamma(\nabla g_1(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*) \ge 0, \tag{3.1a}$$

$\gamma > 0$, $\mathbf{x} \in Q$. Из правого неравенства (3.1а) получим (3.1).

3.2. **Лемма 2.** *Если для* (2.1) *при любом фиксированном* $\mathbf{x} \in Q \subset E^n$ *вогнутая функция* $h_1(\mathbf{u}) \in C^{1,1}(U)$ *такова, что* $\nabla h_1(\mathbf{u}) = \mathbf{G}^{-1}(\mathbf{u})\nabla\varphi_u(\mathbf{x},\mathbf{u})$,

$$\|\nabla h_1(\mathbf{u}) - \nabla h_1(\mathbf{v})\| \le R\|\mathbf{u} - \mathbf{v}\|, \quad R = \frac{L^0}{2m} > 0,$$

то

$$(\nabla h_1(\mathbf{u}^*), \mathbf{u}^* - \mathbf{u}) \ge 0, \quad \mathbf{u} \in U. \tag{3.2}$$

Доказательство проведено для работы [7].

4. ОБОСНОВАНИЕ СХОДИМОСТИ НПОЭКМС (2.1)

Теорема 1. *Пусть выполнены предположения* а)–г) *о задаче и функции* $\varphi(\mathbf{x},\mathbf{u})$ *из разд.*1 *и функциях-параметрах из* п. 2.1; *неравенства* (2.2)–(2.7); *леммы* 1 *и* 2; *параметры метода* (2.1), *функции* $\alpha(t), \beta(t), \gamma(t), \sigma(t), \theta(t)$ *таковы, что*

$$\begin{gathered} 0 < \alpha(t) < \min\{(\beta+\sigma)(3\beta-2\sigma)/4; \beta(\beta+\theta)/2\}, \\ 0 < \gamma(t) < (\beta+2\sigma)/[K(\beta+\sigma)], \quad 0 < \sigma < 3\beta/2, \\ 0 < \lambda < \min\left\{\frac{2\beta(\beta+\theta)-2\alpha}{R(\beta+\theta)(\beta+\theta)}; \frac{3-2\beta-2\theta}{2R}\right\}, \quad \beta+\sigma < 3/2, \\ K(\alpha(t)(\beta(t)+\sigma(t))\gamma(t))' > (4\alpha\beta + 2\alpha\sigma(t))', \quad t \ge 0. \end{gathered} \tag{4.1}$$

Тогда НПОЭКМС (2.1), (4.1) $\forall(\mathbf{x}^0,\mathbf{u}^0) \in E^n \times E^m$ *по норме сходится к седловой точке* $(\mathbf{x}^*,\mathbf{u}^*) \in Q_* \times U^*$ *функции* $\varphi(\mathbf{x},\mathbf{u})$,

$$\begin{gathered} \int_0^{+\infty} \left[\|\mathbf{x}(s) - \mathbf{x}^*\|^2 + \|\mathbf{x}'(s)\|^2 + \|\mathbf{x}''(s)\|^2\right] ds < +\infty, \\ \|\mathbf{x}(t) - \mathbf{x}^*\| + \|\mathbf{x}'(t)\| + \|\mathbf{x}''(t)\| \to 0, \quad t \to \infty, \end{gathered} \tag{4.2}$$

$$\begin{gathered} \int_0^{+\infty} \left[\|\mathbf{u}(s) - \mathbf{u}^*\|^2 + \|\mathbf{u}'(s)\|^2 + \|\mathbf{u}''(s)\|^2\right] ds < +\infty, \\ \|\mathbf{u}(t) - \mathbf{u}^*\| + \|\mathbf{u}'(t)\| + \|\mathbf{u}''(t)\| \to 0, \quad t \to \infty, \end{gathered} \tag{4.3}$$

т.е. $\mathbf{x}^k \to \mathbf{x}^* \in Q_*$, $\mathbf{u}^k \to \mathbf{u}^* \in U^*$ *при* $k \to \infty$.

Доказательство. Представим первые два уравнения НПОЭКМС (2.1), пользуясь (2.7), в виде вариационных неравенств

$$\begin{aligned}\big(\mathbf{w}(t)-\mathbf{y}(t)+\gamma(t)\mathbf{B}^{-1}(\mathbf{y}(t))\nabla\varphi_x\big(\mathbf{y}(t),\mathbf{u}(t)\big),a-\mathbf{w}(t)\big)\ge 0,\quad a\in Q,\\ \big(\mathbf{s}(t)-\mathbf{v}(t)-\lambda(t)\mathbf{G}^{-1}(\mathbf{v}(t))\nabla\varphi_u\big(\mathbf{y}(t),\mathbf{v}(t)\big),b-\mathbf{s}(t)\big)\ge 0,\quad b\in U,\end{aligned}\tag{4.4}$$

где $\mathbf{w}(t)=\alpha(t)\mathbf{x}''(t)+\beta(t)\mathbf{x}'(t)+\mathbf{x}(t)\in Q$, $\mathbf{s}(t)=\alpha(t)\mathbf{u}''+\beta(t)\mathbf{u}'+\mathbf{u}\in U$.

Неравенства (4.4) преобразуем, пользуясь свойствами скалярного произведения, неравенством Коши–Буняковского, нерасширяющим свойством оператора проектирования (см. [8, с. 190]), а также (2.1), (4.1), леммами 1 и 2.

В первом неравенстве (4.4) положим $a=\mathbf{x}^*$, пользуясь леммой 1, неравенство (3.1) умножим на $\gamma>0$, примем $\mathbf{x}=\mathbf{w}(t)$, полученные неравенства сложим и получим

$$\big(\mathbf{w}(t)-\mathbf{y}(t),\mathbf{w}(t)-\mathbf{x}^*\big)\le\gamma(t)\big(\nabla g_1(\mathbf{y}(t))-\nabla g_1(\mathbf{x}^*),\mathbf{x}^*-\mathbf{w}(t)\big),\quad t\ge 0.\tag{4.5}$$

Преобразуем (4.5). Для левой части, пользуясь (2.1), последовательно получаем

$$\begin{aligned}\big(\mathbf{w}(t)-\mathbf{y}(t),\mathbf{w}(t)-\mathbf{x}^*\big)&=\big(\alpha\mathbf{x}''+\beta\mathbf{x}'+\sigma\mathbf{x}',\alpha\mathbf{x}''+\beta\mathbf{x}'+\mathbf{x}-\mathbf{x}^*\big)=\\ &=\big(\alpha\mathbf{x}''+\beta\mathbf{x}',\mathbf{x}-\mathbf{x}^*\big)+\big\|\alpha\mathbf{x}''+\beta\mathbf{x}'\big\|^2+\sigma\big(\mathbf{x}',\mathbf{x}-\mathbf{x}^*\big)+\alpha\sigma\big(\mathbf{x}',\mathbf{x}''\big)+\beta\sigma\big\|\mathbf{x}'\big\|^2=\\ &=\alpha\big(\mathbf{x}'',\mathbf{x}-\mathbf{x}^*\big)+(\beta+\sigma)\big(\mathbf{x}',\mathbf{x}-\mathbf{x}^*\big)+\alpha^2\big\|\mathbf{x}''\big\|^2+(\beta^2+\beta\sigma)\big\|\mathbf{x}'\big\|^2+\alpha(2\beta+\sigma)\big(\mathbf{x}',\mathbf{x}''\big).\end{aligned}\tag{4.6}$$

Правую часть (4.5) оценим с помощью следующего неравенства (см. [8, с. 175])

$$\gamma(t)(\nabla g_1(\mathbf{y})-\nabla g_1(\mathbf{x}^*),\mathbf{x}^*-\mathbf{w})\le K\gamma\big\|\mathbf{y}-\mathbf{w}\big\|^2/4,$$

где $K=\dfrac{L}{m}$ из леммы 1,

$$\begin{aligned}K\gamma(t)\big\|\mathbf{w}(t)-\mathbf{y}(t)\big\|^2/4&=\frac{K\gamma}{4}\big\|\alpha\mathbf{x}''(t)+(\beta+\sigma)\mathbf{x}'(t)\big\|^2=\\ &=\frac{K\gamma}{4}\Big(\alpha^2\big\|\mathbf{x}''\big\|^2+(\beta+\sigma)^2\big\|\mathbf{x}'\big\|^2+2\alpha(\beta+\sigma)\big(\mathbf{x}'',\mathbf{x}'\big)\Big).\end{aligned}$$

Подставив эту оценку и (4.6) в (4.5), получим

$$\alpha(t)\big(\mathbf{x}'',\mathbf{x}-\mathbf{x}^*\big)+a_1(t)\big\|\mathbf{x}''\big\|^2+a_2(t)\big\|\mathbf{x}'\big\|^2+a_3(t)\big(\mathbf{x}',\mathbf{x}''\big)+a_4(t)\big(\mathbf{x}',\mathbf{x}-\mathbf{x}^*\big)\le 0,\quad t\ge 0,\quad \mathbf{x}^*\in Q_*,\tag{4.7}$$

где $a_1(t)=\alpha^2\Big(1-\dfrac{K\gamma}{4}\Big)$; $a_2(t)=\beta^2+\beta\sigma-\dfrac{K\gamma}{4}(\beta+\sigma)^2$; $a_3=\alpha\Big[2\beta+\sigma-\dfrac{K\gamma}{2}(\beta+\sigma)\Big]$; $a_4(t)=\beta(t)+\sigma(t)$; $a_j(t)>0\ \forall j\in[1:4]$, $0<\gamma(t)<\dfrac{4\beta}{K(\beta+\sigma)}$.

Неравенство (4.7) преобразуем с помощью тождеств

$$\begin{aligned}2(\mathbf{x}''(t),\mathbf{x}'(t))=\frac{d}{dt}\big\|\mathbf{x}'(t)\big\|^2;\quad 2(\mathbf{x}'(t),\mathbf{x}(t)-\mathbf{x}^*)=\frac{d}{dt}\big\|\mathbf{x}(t)-\mathbf{x}^*\big\|^2;\\ (\mathbf{x}'',\mathbf{x}-\mathbf{x}^*)=\frac{d^2}{dt^2}\big\|\mathbf{x}-\mathbf{x}^*\big\|^2/2-\big\|\mathbf{x}'\big\|^2,\quad t\ge 0,\end{aligned}\tag{4.8}$$

тогда (обозначив $a_{21}(t)=a_2-\alpha>0$, $a_{31}(t)=a_3(t)/2$, $a_{41}(t)=a_4(t)/2$) получим

$$a_1(t)\big\|\mathbf{x}''\big\|^2+a_{21}(t)\big\|\mathbf{x}'\big\|^2+a_{31}(t)\frac{d}{dt}\big\|\mathbf{x}'\big\|^2+\alpha(t)\frac{d^2}{dt^2}\big\|\mathbf{x}-\mathbf{x}^*\big\|^2/2+a_{41}(t)\frac{d}{dt}\big\|\mathbf{x}-\mathbf{x}^*\big\|^2\le 0,\quad t\ge 0,\tag{4.9}$$

где $0<\alpha<\beta(\beta+\sigma)$, $0<\beta<1$, $0<\gamma<4(\beta^2+\beta\sigma-\alpha)/[K(\beta+\sigma)^2]=\gamma^{11}$.

Проинтегрировав (4.9) на отрезке $[\xi, t]$, $t > \xi \geq 0$, придем к неравенству

$$\int_\xi^t [a_1(s)\|\mathbf{x}''\|^2 + a_{22}(s)\|\mathbf{x}'\|^2 + a_{42}(s)\|\mathbf{x} - \mathbf{x}^*\|^2]ds +$$

$$+ a_{23}(t)\|\mathbf{x}'(t)\|^2 + a_{43}(t)\|\mathbf{x} - \mathbf{x}^*\|^2/2 + \alpha(t)\frac{d}{dt}\|\mathbf{x} - \mathbf{x}^*\|^2/2 \leq C_1(\xi, \mathbf{x}^*), \tag{4.10}$$

$$t > \xi \geq 0, \quad \mathbf{x}^* \in Q_*,$$

где $a_{22}(s) = a_{21}(s) - a_{31}'(s) > 0$ при $a_{31}'(s) < 0$, что выполняется для

$$K(\alpha(t)(\beta(t) + \sigma(t))\gamma(t))' > (4\alpha\beta + 2\alpha\sigma(t))', \quad a_{23}(t) = a_{31}(t),$$

$$C_1(\xi, \mathbf{x}^*) = a_{31}(\xi)\|\mathbf{x}'(\xi)\|^2 + \alpha(\xi)(\mathbf{x}'(\xi), \mathbf{x}(\xi) - \mathbf{x}^*) + \left(a_{41}(\xi) - \frac{1}{2}\alpha'(\xi)\right)\|\mathbf{x} - \mathbf{x}^*\|^2,$$

$$a_{42}(s) = \frac{1}{2}\alpha''(s) - a_{41}'(s) = \frac{1}{2}(\alpha'' - \beta' - \sigma'), \quad a_{43}/2 = a_{41}(t) - \alpha'/2 = \frac{\beta}{2} + \frac{\sigma}{2} - \frac{1}{2}\alpha',$$

коэффициенты положительны при условиях (4.1) и интеграл положителен. Из (4.10) без положительных слагаемых следует

$$\frac{\alpha(t)}{a_{43}(t)}\frac{d}{dt}\|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{x} - \mathbf{x}^*\|^2 \leq 2C_1(\xi, \mathbf{x}^*)/a_{43}(t), \quad t > \xi \geq 0, \quad \mathbf{x}^* \in Q_*.$$

Умножим это неравенство на

$$a_{43}(t)\frac{e(t)}{\alpha(t)}, \quad e(t) = \exp\left(\int_0^t a_{43}(s)\alpha^{-1}(s)ds\right) > 0,$$

$$e(t)\frac{d}{dt}\|\mathbf{x} - \mathbf{x}^*\|^2 + a_{43}(t)e(t)\alpha^{-1}(t)\|\mathbf{x} - \mathbf{x}^*\|^2 \leq 2C_1(\xi, \mathbf{x}^*)e(t)/\alpha, \quad t > \xi \geq 0.$$

Отсюда с учетом производной $e'(t) = a_{43}(t)e(t)/\alpha(t)$, имеем

$$\frac{d}{dt}[e(t)\|\mathbf{x} - \mathbf{x}^*\|^2] \leq 2C_1(\xi, \mathbf{x}^*)e(t)/\alpha, \quad t > \xi \geq 0.$$

Проинтегрировав его на $[\xi, t]$ и умножив полученное неравенство на $e^{-1}(t)$, получим

$$\|\mathbf{x}(t) - \mathbf{x}^*\|^2 \leq \frac{2C_1(\xi, \mathbf{x}^*)}{\alpha e(t)}\int_\xi^t e(s)ds + \frac{e(\xi)\|\mathbf{x}(\xi) - \mathbf{x}^*\|^2}{e(t)} \leq 2C_1(\xi, \mathbf{x}^*)p(t), \quad t > \xi \geq 0,$$

$$p(t) = \frac{1}{\alpha_0 e(t)}\int_\xi^t e(s)ds + \frac{e(\xi)\|\mathbf{x}(\xi) - \mathbf{x}^*\|^2}{2e(t)}C_1^{-1}(\xi, \mathbf{x}^*), \quad \alpha(t) \geq \alpha_0 > 0, \quad \lim_{t\to\infty} p(t) = \alpha_0^{-1}. \tag{4.11}$$

Из (4.11) следует

$$\overline{\lim_{t\to\infty}}\|\mathbf{x}(t) - \mathbf{x}^*\|^2 \leq 2C_1(\xi, \mathbf{x}^*)/\alpha_0, \quad t > \xi \geq 0, \quad \mathbf{x}^* \in Q_*. \tag{4.12}$$

Оценим второе и третье слагаемые для второго соотношения из (4.2) с помощью неравенств (4.10) и (4.7). Учитывая (4.1) и (4.8), существование чисел $r > 0$ (пусть $r = \beta_0\gamma_0\sigma_0\theta_0$) и $\eta \geq 0$ таких, что для $s > \eta \geq \xi \geq 0$ коэффициенты подынтегральных слагаемых в (4.10) $a_1(s) \geq r > 0$, $a_{22}(s) \geq r > 0$, $a_{42}(s) \geq r > 0$, из (4.10) получим

$$r\int_\xi^t \left\{\|\mathbf{x}''(s)\|^2 + \|\mathbf{x}'(s)\|^2 + \|\mathbf{x}(s) - \mathbf{x}^*\|^2\right\}ds + a_{43}(t)\|\mathbf{x}(t) - \mathbf{x}^*\|^2/2 +$$

$$+ \alpha(t)(\mathbf{x}'(t), \mathbf{x}(t) - \mathbf{x}^*) + a_{23}(t)\|\mathbf{x}'(t)\|^2 \leq C_1(\xi, \mathbf{x}^*), \quad t > \xi \geq \eta \geq 0. \tag{4.13}$$

В (4.13) третье слагаемое оценим с помощью известного неравенства

$$2|ab| \leq \varepsilon a^2 + \varepsilon^{-1}b^2 \quad \forall a, b, \varepsilon > 0 \tag{4.14}$$

при $\varepsilon = \beta(t)$, $a = \mathbf{x}'(t)$, $b = \mathbf{x}(t) - \mathbf{x}^*$, т.е.

$$\alpha(t)(\mathbf{x}'(t), \mathbf{x}(t) - \mathbf{x}^*) \geq -\alpha\beta\|\mathbf{x}'(t)\|^2/2 - \alpha\|\mathbf{x}(t) - \mathbf{x}^*\|^2/(2\beta).$$

Тогда из (4.13) следует

$$r\int_\xi^t \left\{\|\mathbf{x}''(s)\|^2 + \|\mathbf{x}'(s)\|^2 + \|\mathbf{x}(s) - \mathbf{x}^*\|^2\right\} ds + (a_{23} - \alpha\beta/2)\|\mathbf{x}'(t)\|^2 +$$

$$+ [a_{41}(t) - \alpha'(t)/2 - \alpha(t)/(2\beta)]\|\mathbf{x}(t) - \mathbf{x}^*\|^2 \leq C_1(\xi, \mathbf{x}^*), \tag{4.15}$$

$$t > \eta \geq \xi \geq 0,$$

где коэффициенты при втором и третьем слагаемых положительны при условиях (4.1), и $a_{23} - \dfrac{\alpha\beta}{2} = \dfrac{\alpha\beta}{2} + \dfrac{\alpha\sigma}{2} - \dfrac{K\alpha\gamma}{4}(\beta + \sigma) - \beta] > \dfrac{\alpha\beta}{4}$, $0 < \alpha < (\beta + \sigma)(3\beta - 2\sigma)/4 < \beta(\beta + \sigma)$ в неравенстве $0 < \gamma < \dfrac{\beta + 2\sigma}{K(\beta + \sigma)} < \gamma^{11}$, $0 < \sigma < \dfrac{3}{2}\beta$; $a_{41} - \dfrac{\alpha}{2\beta} - \alpha'/2 > 0$ в силу (4.1). Из (4.15) получим

$$\int_\xi^t \left[\|\mathbf{x}''(s)\|^2 + \|\mathbf{x}'(s)\|^2 + \|\mathbf{x}(s) - \mathbf{x}^*\|^2\right] ds \leq C_1(\xi, \mathbf{x}^*)/r, \tag{4.15а}$$

$$\|\mathbf{x}'(t)\|^2 \leq 4C_1(\xi, \mathbf{x}^*)(\alpha\beta)^{-1}, \quad t > \eta \geq \xi \geq 0. \tag{4.15б}$$

Из (4.15а) и (4.15б) с учетом условий (4.1) при $t \to \infty$ следует

$$\int_0^\infty \left(\|\mathbf{x}(s) - \mathbf{x}^*\|^2 + \|\mathbf{x}'(s)\|^2 + \|\mathbf{x}''(s)\|^2\right) ds < +\infty \quad \forall \mathbf{x}^* \in Q_*, \tag{4.16}$$

$$\overline{\lim_{t\to\infty}}\|\mathbf{x}'(t)\|^2 \leq 4C_1(\xi, \mathbf{x}^*)(\alpha_0\beta_0)^{-1}, \quad t > \eta \geq \xi \geq 0. \tag{4.17}$$

Далее оценим $\|\mathbf{x}''(t)\|$, исходя из (4.7) и пользуясь (4.14). Неравенства

$$a_3(\mathbf{x}'(t), \mathbf{x}''(t)) \geq -(2a_3/\alpha)\|\mathbf{x}'(t)\|^2 - (a_3\alpha/8)\|\mathbf{x}''(t)\|^2,$$

$$\alpha(\mathbf{x}''(t), \mathbf{x}(t) - \mathbf{x}^*) \geq -(\alpha^2/8)\|\mathbf{x}''(t)\|^2 - 2\|\mathbf{x}(t) - \mathbf{x}^*\|^2,$$

$$(\beta + \sigma)(\mathbf{x}'(t), \mathbf{x}(t) - \mathbf{x}^*) \geq -(\beta + \sigma)\|\mathbf{x}'(t)\|^2/2 - (\beta + \sigma)\|\mathbf{x}(t) - \mathbf{x}^*\|^2/2,$$

следуют из (4.14) соответственно при $\varepsilon = \dfrac{4}{\alpha}$, $\varepsilon = \dfrac{\alpha}{4}$, $\varepsilon = 1$. С их учетом из (4.7) имеем

$$a_{11}^1(t)\|\mathbf{x}''\|^2 \leq a_{24}(t)\|\mathbf{x}'\|^2 + a_{44}(t)\|\mathbf{x} - \mathbf{x}^*\|^2, \quad t \geq 0, \quad \mathbf{x}^* \in Q_*. \tag{4.18а}$$

В (4.18а) оценим коэффициенты при квадратах норм при условиях (4.1):

$$a_{11}^1 = a_1(t) - a_3\alpha/8 - \alpha^2/8 = \alpha^2\left[7 - \frac{K\gamma}{2}(4 - \beta - \sigma) - 2\beta - \sigma\right]/8 \geq \alpha^2/8 = a_{11}(t)$$

$$\text{при} \quad 0 < \gamma < 2(6 - 2\beta - \sigma)/[K(4 - \beta - \sigma)], \quad 2\beta + \sigma < 6, \quad \beta + \sigma < 4;$$

$$a_{24} = 2a_3/\alpha + (\beta + \sigma)/4 - a_2 = \frac{9\beta}{2} + \frac{5\sigma}{2} - \frac{K\gamma}{2}(\beta + \sigma)\left(1 + \frac{\beta + \sigma}{4}\right) - \beta^2 - \beta\sigma \leq$$

$$\leq 2a_3/\alpha + (\beta + \sigma)/4 = a_{25};$$

$$a_{44}(t) \leq (\beta + \sigma)/2 + 2 = (\beta + \sigma + 4)/2 = a_{45}(t).$$

Учитывая эти оценки, из (4.18а) получаем

$$a_{11}(t)\|\mathbf{x}''\|^2 \leq a_{25}(t)\|\mathbf{x}'\|^2 + a_{45}(t)\|\mathbf{x} - \mathbf{x}^*\|^2, \quad t \geq 0, \quad \mathbf{x}^* \in Q_*. \tag{4.18б}$$

Из (4.18б) с учетом оценок (4.11), (4.15б) имеем

$$\|\mathbf{x}''\|^2 \leq [a_{11}(t)]^{-1}\left(2C_1(\xi, \mathbf{x}^*)[2a_{25}(t)(\alpha\beta)^{-1} + a_{45}(t)p(t)]\right). \tag{4.18}$$

С учетом условий (4.1), а также соотношений (4.12), (4.17), из (4.18) следует

$$\overline{\lim_{t\to\infty}}\left\|\mathbf{x}''(t)\right\|^2 \le C_2(\xi,\mathbf{x}^*), \quad t>\eta\ge\xi\ge 0, \quad \mathbf{x}^*\in Q_*, \tag{4.19}$$

где

$$C_2(\xi,\mathbf{x}^*)=2[2a_{25}^0(\beta_0)^{-1}+a_{45}^0][\alpha_0 a_{11}^0]^{-1}C_1(\xi,\mathbf{x}^*), \quad a_{11}^0=\alpha_0^2/8,$$

$$a_{25}^0=\lim_{t\to\infty}a_{25}(t), \quad a_{45}^0=(\beta_0+\sigma_0+4)/2.$$

Асимптотическую устойчивость траектории системы (2.1) и единственность предельной точки траектории можно показать по аналогии с работами [3]–[6].

Из (4.11), (4.12) следует, что траектория $\mathbf{x}(t)$ ограничена, а в силу (4.16) имеем

$$\lim_{t\to\infty}\left[\left\|\mathbf{x}''(t)\right\|^2+\left\|\mathbf{x}'(t)\right\|^2+\left\|\mathbf{x}(t)-\mathbf{x}^*\right\|^2\right]=0 \quad \forall\mathbf{x}^*\in Q_* \tag{4.20а}$$

и существует подпоследовательность $\{t_i\}$, такая, что

$$\left\|\mathbf{x}''(t_i)\right\|\to 0, \quad \left\|\mathbf{x}'(t_i)\right\|\to 0, \quad \left\|\mathbf{x}(t_i)-\mathbf{x}^*\right\|\to 0, \quad i\to\infty, \quad \mathbf{x}^*\in Q_*. \tag{4.20б}$$

Если в $C_1(\xi,\mathbf{x}^*)$ из (4.10) положим $t=t_i$, учтем (4.9) и для $t_i\ge t_1$ обозначим

$$C_1(t_i,\mathbf{x}^*)=a_{31}(t_i)\left\|\mathbf{x}'(t_i)\right\|^2+\alpha(t_i)(\mathbf{x}'(t_i),\mathbf{x}(t_i)-\mathbf{x}^*)+$$

$$+[a_{41}(t_i)-0.5\alpha'(t_i)]\left\|\mathbf{x}(t_i)-\mathbf{x}^*\right\|^2,$$

то в пределе с учетом (4.12), (4.17), (4.20а), (4.20б) получим

$$C_1(t_i,\mathbf{x}^*)\to 0, \quad i\to\infty, \quad \mathbf{x}^*\in Q_*. \tag{4.21}$$

Тогда с учетом (4.21) из (4.15а), (4.16) следует первое доказываемое соотношение из (4.2), а из (4.12), (4.17), (4.19), (4.20б), (4.21) следует второе соотношение из (4.2).

Из второго неравенства (4.4) при $b=\mathbf{u}^*$, с учетом (3.2) из леммы 2, имеем

$$\begin{gathered}(\mathbf{s}(t)-\mathbf{v}(t),\mathbf{s}-\mathbf{u}^*)\le-\lambda(t)(\nabla h_1(\mathbf{v}(t)),\mathbf{u}^*-\mathbf{s}(t)), \quad \mathbf{u}^*\in U^*\subset U,\\ \mathbf{s}(t)=\alpha(t)\mathbf{u}''(t)+\beta(t)\mathbf{u}'(t)+\mathbf{u}(t), \quad \mathbf{v}(t)=\mathbf{u}(t)-\theta(t)\mathbf{u}'(t), \quad t\ge 0,\end{gathered} \tag{4.22}$$

где скалярное произведение в правой части представим в виде суммы двух слагаемых:

$$-(\nabla h_1(\mathbf{v}(t)),\mathbf{u}^*-\mathbf{s}(t))=-(\nabla h_1(\mathbf{v}(t)),\mathbf{u}^*-\mathbf{v}(t))-(\nabla h_1(\mathbf{v}(t)),\mathbf{v}(t-\mathbf{s}(t)). \tag{4.23}$$

Слагаемые в правой части (4.23) преобразуем с помощью неравенства для вогнутой функции (см. [9, § 2.4, с. 44]),

$$(\nabla h_1(\mathbf{v}),\mathbf{u}^*-\mathbf{v}(t))\ge h_1(\mathbf{u}^*)-h_1(\mathbf{v}) \quad \forall\mathbf{v}(t)\in H_U, \quad \mathbf{u}^*\in U^*,$$

и другого неравенства

$$(\nabla h_1(\mathbf{v}),\mathbf{v}-\mathbf{s})\ge h_1(\mathbf{v})-h_1(\mathbf{s})-\frac{R}{2}\left\|\mathbf{v}-\mathbf{s}\right\|^2 \quad \forall\mathbf{v},\mathbf{s}\in H_U$$

(см. [8, гл. 2, § 3, с. 93; R из леммы 2]). Подставим их в (4.23) и преобразованное (4.23) будет иметь вид

$$-(\nabla h_1(\mathbf{v}),\mathbf{u}^*-\mathbf{s})\le h_1(\mathbf{s})-h_1(\mathbf{u}^*)+\frac{R}{2}\left\|\mathbf{v}-\mathbf{s}\right\|^2\le\frac{R}{2}\left\|\mathbf{v}-\mathbf{s}\right\|^2, \tag{4.23а}$$

где учтено, что $h_1(\mathbf{s})-h_1(\mathbf{u}^*)\le 0$ ввиду вогнутости функции $h_1(\mathbf{u})$. Пользуясь (4.23а) в (4.22), получаем

$$(\mathbf{s}(t)-\mathbf{v}(t),\mathbf{s}-\mathbf{u}^*)\le\frac{R\lambda}{2}\left\|\mathbf{v}(t)-\mathbf{s}(t)\right\|^2. \tag{4.24}$$

В (4.24) подставим разложения для левой части и квадрата нормы в правой части

$$(\mathbf{s}(t)-\mathbf{v}(t),\mathbf{s}-\mathbf{u}^*)=(\alpha(t)\mathbf{u}''(t)+(\beta(t)+\theta(t))\mathbf{u}'(t),\mathbf{u}(t)-\mathbf{u}^*+\beta\mathbf{u}'(t)+\alpha\mathbf{u}''(t))=$$

$$= (\alpha(t)\mathbf{u}'', \mathbf{u} - \mathbf{u}^*) + (\beta + \theta)(\mathbf{u}', \mathbf{u} - \mathbf{u}^*) + \alpha^2 \|\mathbf{u}''\|^2 + \beta(\beta + \theta)\|\mathbf{u}'\|^2 + \alpha(2\beta + \theta)(\mathbf{u}', \mathbf{u}''),$$

$$\|\mathbf{v} - \mathbf{s}\|^2 = \|\mathbf{s} - \mathbf{v}\|^2 = \|\alpha(t)\mathbf{u}''(t) + (\beta + \theta)\mathbf{u}'(t)\|^2 =$$

$$= \alpha^2 \|\mathbf{u}''(t)\|^2 + (\beta + \theta)^2 \|\mathbf{u}'(t)\|^2 + 2\alpha(\beta + \theta)(\mathbf{u}'', \mathbf{u}').$$

После их подстановки из (4.24) следует

$$\begin{gathered} b_1(t)\|\mathbf{u}''\|^2 + b_{21}^a(t)\|\mathbf{u}'\|^2 + b_{31}(t)(\mathbf{u}', \mathbf{u}'') + \\ + \alpha(\mathbf{u}'', \mathbf{u} - \mathbf{u}^*) + (\beta + \theta)(\mathbf{u}', \mathbf{u} - \mathbf{u}^*) \le 0, \quad t \ge 0, \end{gathered} \tag{4.25}$$

где

$$0 < \lambda(t) < \frac{2\beta}{R(\beta + \theta)}; \quad b_1(t) = \alpha^2\left(1 - \frac{R\lambda}{2}\right) > 0, \quad b_{21}^a(t) = (\beta + \theta)\left[\beta - \frac{R\lambda}{2}(\beta + \theta)\right] > 0,$$

$$b_{31}(t) = \alpha(2\beta + \theta + 2\beta + 2\theta) = \alpha(t)(4\beta + 3\theta) > 0.$$

Справедливы аналогичные (4.8) тождества

$$\begin{gathered} 2(\mathbf{u}''(t), \mathbf{u}'(t)) = \frac{d}{dt}\|\mathbf{u}'(t)\|^2; \quad 2(\mathbf{u}'(t), \mathbf{u}(t) - \mathbf{u}^*) = \frac{d}{dt}\|\mathbf{u}(t) - \mathbf{u}^*\|^2; \\ (\mathbf{u}'', \mathbf{u} - \mathbf{u}^*) = \frac{d^2}{dt^2}\|\mathbf{u} - \mathbf{u}^*\|^2/2 - \|\mathbf{u}'\|^2, \quad t \ge 0. \end{gathered} \tag{4.8a}$$

Преобразуем (4.25) с помощью тождеств (4.8а), тогда получим

$$b_1(t)\|\mathbf{u}''\|^2 + b_{22}\|\mathbf{u}'\|^2 + b_3(t)\frac{d}{dt}\|\mathbf{u}'\|^2 + \alpha(t)\frac{d^2}{dt^2}\|\mathbf{u} - \mathbf{u}^*\|^2/2 + b_{41}\frac{d}{dt}\|\mathbf{u} - \mathbf{u}^*\|^2 \le 0, \quad t \ge 0, \tag{4.26}$$

где

$$b_{22}(t) = b_{21}^a(t) - \alpha(t), \quad 0 < \lambda < 2[\beta(\beta + \theta) - \alpha]/[R(\beta + \theta)^2] = \lambda^{11}, \quad \alpha < \beta(\beta + \theta),$$

$$b_3 = \frac{1}{2}b_{31} = 2\alpha\beta + 3\alpha\theta/2, \quad b_{41} = \frac{1}{2}(\beta + \theta).$$

Проинтегрировав (4.26), на отрезке $[\xi, t]$, $t > \xi \ge 0$, получим

$$\begin{gathered} \int_\xi^t [b_1(s)\|\mathbf{u}''\|^2 + b_{21}(s)\|\mathbf{u}'\|^2 + b_{42}(s)\|\mathbf{u} - \mathbf{u}^*\|^2]ds + \\ + b_3(t)\|\mathbf{u}'(t)\|^2 + b_{43}(t)\|\mathbf{u} - \mathbf{u}^*\|^2 + \alpha(t)\frac{d}{dt}\|\mathbf{u} - \mathbf{u}^*\|^2 \le C_3(\xi, \mathbf{u}^*), \\ t > \xi \ge 0, \quad \mathbf{u}^* \in U^*, \end{gathered} \tag{4.27}$$

где

$$b_{21}(s) = b_{22}(s) - b_3'(s) > 0, \quad b_{22} > 0,\ b_3' < 0, \quad b_{42}(s) = \alpha''(s) - \frac{1}{2}(\beta'(s) + \theta'(s)) > 0,$$

с учетом (4.8а) имеем

$$C_3(\xi, \mathbf{u}^*) = b_3(\xi)\|\mathbf{u}'(\xi)\|^2 + 2\alpha(\xi)(\mathbf{u}'(\xi), \mathbf{u}(\xi) - \mathbf{u}^*) + b_{43}(\xi)\|\mathbf{u} - \mathbf{u}^*\|^2,$$

$$b_{43}(t) = (\beta + \theta)/2 - \alpha' > 0;$$

коэффициенты и интеграл положительны. Из (4.27) без положительных слагаемых следует

$$\frac{\alpha(t)}{b_{43}(t)}\frac{d}{dt}\|\mathbf{u} - \mathbf{u}^*\|^2 + \|\mathbf{u} - \mathbf{u}^*\|^2 \le C_3(\xi, \mathbf{u}^*)/b_{43}(t), \quad t > \xi \ge 0, \quad \mathbf{u}^* \in U^*. \tag{4.28}$$

Умножим (4.28) на

$$b_{43}(t)\frac{g(t)}{\alpha(t)}, \quad g(t) = \exp\left(\int_0^t b_{43}(s)\alpha^{-1}(s)ds\right) > 0,$$

$$g(t)\frac{d}{dt}\|\mathbf{u}-\mathbf{u}^*\|^2 + b_{43}(t)g(t)\alpha^{-1}(t)\|\mathbf{u}-\mathbf{u}^*\|^2 \le C_3(\xi,\mathbf{u}^*)g(t)\alpha^{-1}(t), \quad t > \xi \ge 0.$$

Отсюда с учетом производной $g'(t) = b_{43}(t)g(t)/\alpha(t)$ имеем

$$\frac{d}{dt}[g(t)\|\mathbf{u}-\mathbf{u}^*\|^2] \le C_3(\xi,\mathbf{u}^*)g(t)/\alpha, \quad t > \xi \ge 0. \tag{4.29}$$

Проинтегрировав (4.29) на $[\xi,t]$ и умножив полученное неравенство на $g^{-1}(t)$, получим

$$\begin{gathered}\|\mathbf{u}(t)-\mathbf{u}^*\|^2 \le \frac{C_3(\xi,\mathbf{u}^*)}{\alpha g(t)}\int_\xi^t g(s)ds + \frac{g(\xi)\|\mathbf{u}(\xi)-\mathbf{u}^*\|^2}{g(t)} \le C_3(\xi,\mathbf{u}^*)q(t), \quad t > \xi \ge 0,\\ q(t) = \frac{1}{\alpha_0 g(t)}\int_\xi^t g(s)ds + \frac{g(\xi)\|\mathbf{u}(\xi)-\mathbf{u}^*\|^2}{C_3(\xi,\mathbf{u}^*)g(t)}, \quad \alpha(t) \ge \alpha_0 > 0, \quad \lim_{t\to\infty} q(t) = \alpha_0^{-1}.\end{gathered} \tag{4.30a}$$

Отсюда имеем

$$\overline{\lim_{t\to\infty}}\|\mathbf{u}(t)-\mathbf{u}^*\|^2 \le C_3(\xi,\mathbf{u}^*)/\alpha_0, \quad t > \xi \ge 0, \quad \mathbf{u}^* \in U^*. \tag{4.30}$$

Оценим второе и третье слагаемые во втором соотношении из (4.3) с помощью неравенств (4.8а) и (4.14). Учитывая (4.1) и то, что существуют числа $r > 0$ и $\eta \ge 0$ такие, что для $s > \eta \ge \xi \ge 0$ коэффициенты подынтегральных слагаемых в (4.27) $b_1(s) \ge r > 0$, $b_{21}(s) \ge r > 0$, $b_{42}(s) \ge r > 0$, и применяя (4.8а), из (4.27) получаем

$$\begin{gathered}r\int_\xi^t\left\{\|\mathbf{u}''(s)\|^2 + \|\mathbf{u}'(s)\|^2 + \|\mathbf{u}(s)-\mathbf{u}^*\|^2\right\}ds + b_{43}(t)\|\mathbf{u}(t)-\mathbf{u}^*\|^2 +\\ +2\alpha(t)(\mathbf{u}'(t),\mathbf{u}(t)-\mathbf{u}^*) + b_3(t)\|\mathbf{u}'(t)\|^2 \le C_3(\xi,\mathbf{u}^*),\\ t > \eta \ge \xi \ge 0.\end{gathered} \tag{4.31}$$

В (4.31) третье слагаемое оценим с помощью неравенства (4.14) при $\varepsilon = \beta$, $a = \mathbf{u}'(t)$, $b = \mathbf{u}(t) - \mathbf{u}^*$, т.е.

$$2\alpha(t)(\mathbf{u}'(t),\mathbf{u}(t)-\mathbf{u}^*) \ge -\alpha\beta\|\mathbf{u}'(t)\|^2 - \alpha\|\mathbf{u}(t)-\mathbf{u}^*\|^2/\beta.$$

Тогда верно

$$\begin{gathered}r\int_\xi^t\left\{\|\mathbf{u}''(s)\|^2 + \|\mathbf{u}'(s)\|^2 + \|\mathbf{u}(s)-\mathbf{u}^*\|^2\right\}ds + b_{32}(t)\|\mathbf{u}'(t)\|^2 + b_{44}(t)\|\mathbf{u}(s)-\mathbf{u}^*\|^2 \le C_3(\xi,\mathbf{u}^*),\\ t > \eta \ge \xi \ge 0,\end{gathered} \tag{4.32}$$

где коэффициенты при втором и третьем слагаемых неотрицательны при условиях (4.1),

$$b_{32}(t) = b_3 - \alpha\beta = 2\alpha\beta + \frac{3\alpha\theta}{2} - \alpha\beta = \alpha\beta + \frac{3\alpha\theta}{2},$$

$$b_{44}(t) = b_{43}(t) - \frac{\alpha}{\beta} > 0, \quad 0 < \alpha \le \frac{1}{2}\beta(\beta+\theta).$$

Неравенство (4.32) эквивалентно системе

$$\int_\xi^t\left[\|\mathbf{u}''(s)\|^2 + \|\mathbf{u}'(s)\|^2 + \|\mathbf{u}(s)-\mathbf{u}^*\|^2\right]ds \le C_3(\xi,\mathbf{u}^*)/r, \tag{4.33a}$$

$$\|\mathbf{u}'(t)\|^2 \le C_3(\xi,\mathbf{u}^*)/b_{32}(t), \quad t > \eta \ge \xi \ge 0. \tag{4.34a}$$

Из (4.33а), (4.34а), с учетом $\lim b_{32}(t) = \alpha_0\beta_0 + \frac{3}{2}\alpha_0\theta_0 = b_{32}^0$ при $t \to \infty$, и условий (4.1) для параметров метода, следует

$$\int_0^\infty \left(\|\mathbf{u}(s) - \mathbf{u}^*\|^2 + \|\mathbf{u}'(s)\|^2 + \|\mathbf{u}''(s)\|^2\right) ds < +\infty \quad \forall \mathbf{u}^* \in U^*, \tag{4.33}$$

$$\overline{\lim_{t\to\infty}} \|\mathbf{u}'(t)\|^2 \le C_3(\xi, \mathbf{u}^*)/b_{32}^0, \quad t > \eta \ge \xi \ge 0. \tag{4.34}$$

Оценку для $\|\mathbf{u}''(t)\|$ получим из (4.25) с помощью (4.8а) и (4.14). С учетом оценок

$$b_{31}(\mathbf{u}'(t), \mathbf{u}''(t)) \ge -2(4\beta + 3\theta)\|\mathbf{u}'(t)\|^2 - \alpha^2(4\beta + 3\theta)\|\mathbf{u}''(t)\|^2/8,$$

$$(\beta + \theta)(\mathbf{u}'(t), \mathbf{u}(t) - \mathbf{u}^*) \ge -(\beta + \theta)\|\mathbf{u}'(t)\|^2/2 - (\beta + \theta)\|\mathbf{u}(t) - \mathbf{u}^*\|^2/2,$$

$$\alpha(\mathbf{u}''(t), \mathbf{u}(t) - \mathbf{u}^*) \ge -\alpha^2\|\mathbf{u}''(t)\|^2/8 - 2\|\mathbf{u}(t) - \mathbf{u}^*\|^2,$$

получаемых из (4.14) соответственно при $\varepsilon = 4/\alpha$, $\varepsilon = 1$, $\varepsilon = \alpha/4$, из (4.25) следует

$$b_{11}^1(t)\|\mathbf{u}''\|^2 \le b_{24}(t)\|\mathbf{u}'\|^2 + b_{44}(t)\|\mathbf{u} - \mathbf{u}^*\|^2, \quad t \ge 0, \quad \mathbf{u}^* \in Q_*, \tag{4.35}$$

где следующие оценки коэффициентов верны при условиях (4.1):

$$b_{11}^1(t) = b_1(t) - \alpha^2(4\beta + 3\theta)/8 - \alpha^2/8 = \alpha^2[7 - 4R\lambda - 4\beta - 3\theta]/8 > \alpha^2(1 + \theta)/8 = b_{11}(t)$$

$$\text{при} \quad 0 < \lambda < \min\left\{\frac{2\beta(\beta + \theta) - 2\alpha}{R(\beta + \theta)(\beta + \theta)}; \frac{3 - 2\beta - 3\theta}{2R}\right\},$$

$$b_{24}(t) = 2(4\beta + 3\theta) + (\beta + \theta)/2 - b_{21}^a(t) < (17\beta + 13\theta)/2 = b_{25}(t),$$

$$b_{44}(t) \le \frac{\beta + \theta}{2} + 2 = \frac{\beta + \theta + 4}{2} = b_{45}(t).$$

С учетом этих оценок и (4.30а), (4.34а), из (4.35) получим

$$\begin{gathered} b_{11}(t)\|\mathbf{u}''\|^2 \le b_{25}(t)\|\mathbf{u}'\|^2 + b_{45}(t)\|\mathbf{u} - \mathbf{u}^*\|^2, \quad t \ge 0, \quad \mathbf{u}^* \in Q_*, \\ \|\mathbf{u}''\|^2 \le [b_{11}(t)]^{-1}\left(C_3(\xi, \mathbf{u}^*)[b_{25}(t)[b_{32}(t)]^{-1} + b_{45}(t)q(t)]\right). \end{gathered} \tag{4.36а}$$

С учетом (4.30), (4.34) из (4.36а) следует

$$\overline{\lim_{x\to\infty}} \|\mathbf{u}''(t)\|^2 \le C_4(\xi, \mathbf{u}^*), \quad \mathbf{u}^* \in U^*, \tag{4.36б}$$

где

$$C_4(\xi, \mathbf{u}^*) = C_3(\xi, \mathbf{u}^*)(b_{11}^0)^{-1}\left(b_{25}^0(b_{32}^0)^{-1} + b_{45}^0\alpha_0^{-1}\right),$$

$$b_{11}^0 = \alpha_0^2(1 + \theta_0)/8, \quad b_{25}^0 = (17\beta_0 + 13\theta_0)/2, \quad b_{45}^0 = (\beta_0 + \theta_0 + 4)/2.$$

Далее, в силу (4.30а), (4.30) траектория $\mathbf{u}(t)$ ограничена, а в силу (4.33) имеем

$$\underline{\lim_{t\to\infty}}\left(\|\mathbf{u}''(t)\|^2 + \|\mathbf{u}'(t)\|^2 + \|\mathbf{u}(t) - \mathbf{u}^*\|^2\right) = 0 \quad \forall \mathbf{u}^* \in U^*$$

и существует подпоследовательность $\{t_i\}$, что

$$\|\mathbf{u}''(t_i)\| \to 0, \quad \|\mathbf{u}'(t_i)\| \to 0, \quad \|\mathbf{u}(t_i) - \mathbf{u}^*\| \to 0, \quad i \to \infty, \quad \mathbf{u}^* \in U^*. \tag{4.37}$$

Если в (4.10) положим $t = t_i$, учтем (4.9) и для $t_i \ge t_1$ обозначим

$$C_5(t_i, \mathbf{u}^*) = b_3(t_i)\|\mathbf{u}'(t_i)\|^2 + 2\alpha(t_i)(\mathbf{u}'(t_i), \mathbf{u}(t_i) - \mathbf{u}^*) + b_{43}(t_i)\|\mathbf{u}(t_i) - \mathbf{u}^*\|^2,$$

то в пределе, с учетом (4.30), (4.34), (4.36), (4.37) получим,

$$C_5(t_i, \mathbf{u}^*) \to 0, \quad i \to \infty, \quad \mathbf{u}^* \in U^*.$$

Тогда с учетом (4.37) из (4.33) следует третье доказываемое соотношение из (4.3), а из (4.30), (4.34), (4.36), (4.37) следует четвертое соотношение из (4.3).

Теорема 1 доказана.

5. ОЦЕНКА СКОРОСТИ СХОДИМОСТИ НПОЭКМС (2.1)

Получим оценки скорости сходимости метода (2.1), (4.1) для выпукло-вогнутой функции.

Теорема 2. *Пусть выполнены все условия теоремы* 1, *включая* (4.1)–(4.3), *леммы* 1 *и* 2. *Тогда траектория* $\{\mathbf{x}(t);\mathbf{u}(t)\}$ НПОЭКМС (2.1), (4.1)–(4.3) *сходится к седловой точке* $(\mathbf{x}^*;\mathbf{u}^*)\in Q_*\times U^*$ *задачи* (1.1) $\forall t\geq 0$ *с экспоненциальной скоростью с оценками*:

$$\|\mathbf{x}(t)-\mathbf{x}^*\|\leq\{2C_{21}p_2(t)\}^{1/2}, \tag{5.1}$$

$$\|\mathbf{x}'(t)\|\leq 2\left(C_{21}(\alpha(t)\beta(t))^{-1}\right)^{1/2}, \tag{5.2}$$

$$\|\mathbf{x}''\|\leq[a_{11}(t)]^{-1/2}\left\{2C_{21}[2a_{25}(t)(\alpha\beta)^{-1}+a_{45}(t)p_2(t)]\right\}^{1/2}, \tag{5.3}$$

$$\|\mathbf{u}(t)-\mathbf{u}^*\|\leq[C_{31}q_2(t)]^{1/2}, \tag{5.4}$$

$$\begin{gathered}\|\mathbf{u}'(t)\|\leq(C_{31}/b_{32}(t))^{1/2},\\ \|\mathbf{u}''(t)\|\leq[b_{11}(t)]^{-1/2}\left\{C_{31}[b_{25}(t)(b_{32}(t))^{-1}+b_{45}(t)q_2(t)]\right\}^{1/2},\end{gathered} \tag{5.5}$$

где

$$\begin{gathered}C_{21}=a_{31}(0)\left\|\mathbf{x}^1\right\|^2+\alpha(0)(\mathbf{x}^1,\mathbf{x}^0-\mathbf{x}^*)+(a_{41}(0)-\frac{1}{2}\alpha'(0))\left\|\mathbf{x}^0-\mathbf{x}^*\right\|^2,\\ C_{31}=b_3(0)\left\|\mathbf{u}^1\right\|^2+2\alpha(0)(\mathbf{u}^1,\mathbf{u}^0-\mathbf{u}^*)+b_{43}(0)\left\|\mathbf{u}^0-\mathbf{u}^*\right\|^2,\end{gathered} \tag{5.6}$$

$$p_2(t)=\frac{1}{\alpha_0 e(t)}\int_0^t e(s)ds+\frac{e(0)\left\|\mathbf{x}^0-\mathbf{x}^*\right\|^2}{2C_{21}e(t)},\quad q_2(t)=\frac{1}{\alpha_0 g(t)}\int_0^t g(s)ds+\frac{g(\xi)\|\mathbf{u}(\xi)-\mathbf{u}^*\|^2}{C_{31}g(t)},$$

$$e(t)=\exp\left(\int_0^t a_{43}(s)\alpha^{-1}(s)ds\right),\quad g(t)=\exp\left(\int_0^t b_{43}(s)\alpha^{-1}(s)ds\right),\quad a_{11},a_{25},a_{31},$$

$a_{41},a_{43},a_{45},b_{11},b_{25},b_{31},b_{41},b_{42},b_{45}$ *из теоремы* 1, *а их значения при* $t=0$.

Доказательство. Заметим, что при выполнении всех условий теоремы 2, выкладки и результаты теоремы 1 о сходимости метода (2.1) справедливы (обозначения коэффициентов, совпадающих с полученными в теореме 1, сохраняем; новые коэффициенты C_{21}, C_{31} приведены в формулировке теоремы 2). Проведя аналоги выкладок из теоремы 1 от (4.4), но теперь при $\xi=0$, получим аналог (4.11)

$$\|\mathbf{x}(t)-\mathbf{x}^*\|^2\leq\frac{2C_{21}}{\alpha e(t)}\int_0^t e(s)ds+\frac{e(0)\left\|\mathbf{x}^0-\mathbf{x}^*\right\|^2}{e(t)}\leq 2C_{21}p_2(t),\quad t\geq 0. \tag{5.7}$$

Из (5.7) следует оценка (5.1).

Проведем выкладки, аналогичные проведенным в теореме 1 от (4.13) до (4.15б), только при $\xi=0$, тогда аналог (4.15б) имеет вид

$$\|\mathbf{x}'(t)\|^2\leq 4C_{21}(\alpha(t)\beta(t))^{-1},\quad t\geq 0. \tag{5.8}$$

Из (5.8) следует оценка (5.2).

Для получения (5.3) вычислим аналоги (4.16)–(4.18) для $\xi=0$, тогда (сохраняя из теоремы 1 и обозначения, за исключением $p_2(t)$ и $q_2(t)$), с учетом (5.7), (5.8) получим

$$\|\mathbf{x}''\|^2\leq[a_{11}(t)]^{-1}\left(2C_{21}[2a_{25}(t)(\alpha\beta)^{-1}+a_{45}(t)p_2(t)]\right),\quad t\geq 0, \tag{5.9}$$

где $p_2(t)$ из (5.7) (см. условия теоремы 2). Из (5.9) следует оценка (5.3).

Теперь проведем аналоги выкладок из теоремы 1 по $\mathbf{u}(t)$, но при $\xi = 0$. Вычислив аналоги для (4.26)–(4.29), затем получим аналог неравенства (4.30а)

$$\left\|\mathbf{u}(t) - \mathbf{u}^*\right\|^2 \le \frac{C_{31}}{\alpha g(t)} \int_0^t e(s)ds + \frac{g(0)\left\|\mathbf{u}^0 - \mathbf{u}^*\right\|^2}{g(t)} \le C_{31} q_2(t), \quad t \ge 0. \tag{5.10}$$

Из неравенства (5.10) следует оценка (5.4).

Далее получим оценку (5.5). Проведем аналоги выкладок из теоремы 1 после (4.30) до получения неравенства (4.34а), только при $\xi = 0$, с учетом неравенства (5.10), оценок коэффициентов, получим

$$\left\|\mathbf{u}'(t)\right\|^2 \le C_{31}/b_{32}(t), \quad t > 0. \tag{5.11}$$

Из (5.11) следует оценка (5.5).

Для вычисления оценки (5.6) продолжим аналоги выкладок из теоремы 1 после (4.34) до получения (4.36а), только при $\xi = 0$; с учетом неравенств (5.10), (5.11) получим

$$\left\|\mathbf{u}''(t)\right\|^2 \le [b_{11}(t)]^{-1}\left\{C_{31}[b_{25}(t)(b_{32}(t))^{-1} + b_{45}(t)q_2(t)]\right\}, \quad t > 0, \tag{5.12}$$

где $q_2(t)$ из (5.10). Из неравенства (5.12) следует оценка (5.6).

Теорема 2 доказана.

6. ВЫВОДЫ

Исследованный в данной работе НПОЭКМС (2.1) продолжает на непрерывные проекционные ЭГМ второго порядка *для решения седловых задач* идею использования операторов переменной метрики, впервые воплощенную в НПММ первого порядка в работе [4]. НПОЭКМС (2.1) обладает достоинствами, присущими и НПММ, и методам переменной метрики; он имеет лучшую точность и скорость сходимости в окрестности седловой точки.

СПИСОК ЛИТЕРАТУРЫ

1. *Эрроу К.Дж., Гурвиц Л., Удзава Х.* Исследования по линейному и нелинейному программированию. М.: Изд-во иностр. лит., 1962.
2. *Корпелевич Г.М.* Экстраполяционные градиентные методы и их связь с модифицированными функциями Лагранжа // Экономика и матем. методы. 1983. Т. 19. Вып. 4. С. 694–703.
3. *Антипин А.С.* Градиентный и экстраградиентный подходы в билинейном и равновесном программировании. М.: Изд-во ВЦ РАН, 2002.
4. *Антипин А.С., Васильев Ф.П.* О непрерывном методе минимизации в пространствах с переменной метрикой // Известия вузов. Математика. 1995. № 12(403). С. 3–9.
5. *Малинов В.Г.* О проекционном квазиньютоновском обобщенном двухшаговом методе минимизации и оптимизации траектории летательного аппарата // Ж. Средневолжского матем. общества. 2010. Т. 12. № 4. С. 37–48.
6. *Малинов В.Г.* Версия непрерывного проекционного метода минимизации второго порядка с переменной метрикой // Ж. Средневолжского матем. общества. 2014. Т. 16. № 1. С. 121–134.
7. *Малинов В.Г.* О версии обобщенного экстраградиентного квазиньютоновского метода решения седловых и других задач // IX Московская международная конференция по исследованию операций (ORM2018): Москва. 22–27 октября 2018 г.: Труды. В двух томах. Том II. М.: МАКС ПРЕСС, 2018. С. 124–126.
8. *Васильев Ф.П.* Численные методы решения экстремальных задач. М.: Наука, 1988.
9. *Карманов В.Г.* Математическое программирование. М.: Наука, 1975.
10. *Антипин А.С., Хамраева З.С.* Управляемые седловые дифференциальные градиентные методы 2-го порядка. М.: Вычислительный Центр РАН, 1996.
11. *Антипин А.С.* Управляемые дифференциальные градиентные методы второго порядка для решения равновесных задач // Дифференц. ур-ния. 1999. Т. 35. № 5. С. 590–599.

EDN: GXAXUM
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 790–808

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

УДК 519.622

МЕТОДЫ ESDIRK ТРЕТЬЕГО И ЧЕТВЕРТОГО ПОРЯДКОВ ДЛЯ ЖЕСТКИХ И ДИФФЕРЕНЦИАЛЬНО-АЛГЕБРАИЧЕСКИХ ЗАДАЧ

© 2022 г. Л. М. Скворцов

105005 Москва, 2-я Бауманская, 5, МГТУ им. Н.Э. Баумана, Россия

e-mail: lm_skvo@rambler.ru

Поступила в редакцию 16.11.2021 г.
Переработанный вариант 14.12.2021 г.
Принята к публикации 11.01.2022 г.

Рассматриваются жестко точные однократно диагонально-неявные методы Рунге–Кутты с явной первой стадией (ESDIRK), предназначенные для решения жестких обыкновенных дифференциальных уравнений (ОДУ) и дифференциально-алгебраических уравнений (ДАУ). Достоинство этих методов – простая реализация, но они имеют только второй стадийный порядок, что ограничивает возможность построения эффективных методов высоких порядков. Методы ESDIRK наиболее эффективны при расчетах с невысокой точностью, достаточной для решения большинства практических задач. Поэтому в статье рассматриваются методы 3-го и 4-го порядков, позволяющие получить решение с малыми вычислительными затратами при умеренных требованиях к точности. Предложены новые методы, удовлетворяющие некоторым дополнительным условиям, что позволяет эффективно решать не только жесткие ОДУ, но и ДАУ индексов 2 и 3. Уделено внимание реализации методов с автоматическим выбором размера шага, и приведены результаты численных экспериментов. Библ. 36. Фиг. 2. Табл. 12.

Ключевые слова: обыкновенные дифференциальные уравнения, жесткая задача Коши, дифференциально-алгебраические уравнения индексов 2 и 3, диагонально-неявные методы Рунге–Кутты, ESDIRK.

DOI: 10.31857/S004446692205012X

1. ВВЕДЕНИЕ

Диагонально-неявные методы Рунге–Кутты (DIRK, см. [1]–[18]) относятся к неявным методам, которые наиболее просто реализуются. Благодаря этому методы DIRK широко применяются в прикладных вычислениях. Они используются для решения жестких ОДУ, ДАУ индексов 2 и 3 (см. [3], [4], [10]–[12], [14]), уравнений в частных производных (см. [15], [19]–[21]), а также входят в состав явно-неявных аддитивных методов (см. [21], [22]). Методы DIRK реализованы в программных продуктах MATLAB и SimInTech (см. [23]).

Один шаг метода DIRK при решении задачи Коши

$$\mathbf{y}' = \mathbf{f}\left(t, \mathbf{y}\right), \quad \mathbf{y}\left(t_0\right) = \mathbf{y}_0$$

выполняется согласно формулам

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\sum_{i=1}^{s} b_i\mathbf{F}_i, \quad \mathbf{F}_i = \mathbf{f}\left(t_n + hc_i, \mathbf{Y}_i\right), \quad \mathbf{Y}_i = \mathbf{y}_n + h\sum_{j=1}^{i} a_{ij}\mathbf{F}_j,$$

где h – размер шага, s – число стадий, $\mathbf{Y}_i$ и $\mathbf{F}_i$ – стадийные значения и их производные. Таблица коэффициентов (таблица Бутчера) метода DIRK имеет вид

$$\begin{array}{c|cccc} c_1 & a_{11} & & & \\ c_2 & a_{21} & a_{22} & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} = \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^{\mathrm{T}} \end{array}$$

(нулевые элементы матрицы **A** обычно опускают).

Практическое применение нашли однократно диагонально-неявные методы Рунге–Кутты (SDIRK – Singly DIRK) и аналогичные методы с явной первой стадией (ESDIRK – Explicit first stage SDIRK). Методы SDIRK имеют $a_{ii} = \gamma, i = 1,\ldots,s$, а методы ESDIRK имеют $a_{11} = 0$, $a_{ii} = \gamma$, $i = 2,\ldots,s$. В [1] исследовалась сходимость методов SDIRK при решении жесткого уравнения Протеро–Робинсона (см. [24]). В результате был сделан вывод о преимуществе методов, имеющих $b_i = a_{si}, i = 1,\ldots,s$ (такие методы Рунге–Кутты называют *жестко точными*).

Большое значение при решении жестких ОДУ и ДАУ высших индексов имеет также *стадийный порядок* – наибольшее целое q, для которого выполняются равенства

$$\mathbf{A}\mathbf{c}^{k-1} = \mathbf{c}^k/k, \quad \mathbf{b}^{\mathrm{T}}\mathbf{c}^{k-1} = 1/k, \quad k = 1,\ldots,q$$

(здесь и далее предполагаем покомпонентное выполнение операций возведения вектора в степень и умножения векторов). Стадийный порядок методов SDIRK не может быть выше 1-го, а методов ESDIRK – выше 2-го.

Простейший тест для методов решения ОДУ – уравнение Далквиста $y' = \lambda y$. Один шаг решения этого уравнения методом Рунге–Кутты запишется в виде $y_{n+1} = R(h\lambda)\,y_n$, где $R(z)$ – функция устойчивости, вычисляемая по формуле

$$R(z) = 1 + z\mathbf{b}^{\mathrm{T}}(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}, \quad \mathbf{e} = [1,\ldots,1]^{\mathrm{T}}, \quad \mathbf{I} = \mathrm{diag}(\mathbf{e}).$$

Метод называется $A(\alpha)$-*устойчивым*, если $|R(z)| \le 1$ при $|\arg(-z)| \le \alpha$, и A-*устойчивым*, если $\alpha = 90°$. Если при этом $R(\infty) = 0$, то метод называется $L(\alpha)$-*устойчивым* либо L-*устойчивым* (при $\alpha = 90°$). Жестко точные методы SDIRK удовлетворяют условию $R(\infty) = 0$. Потребуем выполнения этого условия также и для методов ESDIRK. Требование L-устойчивости часто является завышенным, поэтому будем рассматривать также и $L(\alpha)$-устойчивые методы при значении α, близком к 90°.

Согласно [25], метод называется L-*затухающим порядка* $\mu > 0$, если его функция устойчивости удовлетворяет соотношению $|R(z)| = O(z^{-\mu})$ при $z \to \infty$. $L(\alpha)$-устойчивый метод с порядком L-затухания $\mu > 1$ будем обозначать как $L\mu(\alpha)$-устойчивый. На основании численных экспериментов мы убедились, что повышенный порядок L-затухания ($\mu > 1$) не дает преимущества при решении жестких задач. Действительно, любой $L(\alpha)$-устойчивый метод обеспечит порядок L-затухания μ, если принять μ шагов за один шаг. Однако $L2(\alpha)$-устойчивый метод позволяет реализовать эффективный контроль ошибки при решении ДАУ индекса 3. Поэтому наряду с $L(\alpha)$-устойчивыми методами рассмотрим также и $L2(\alpha)$-устойчивый метод.

Тестовое сравнение методов SDIRK и ESDIRK было выполнено в [7]–[10], [14], где показано преимущество методов ESDIRK при решении жестких ОДУ и ДАУ индексов 2 и 3. Это преимущество объясняется более высоким стадийным порядком и наличием дополнительных коэффициентов матрицы **A** при таком же числе неявных стадий. Поэтому мы рассматриваем жестко точные $L(\alpha)$-устойчивые методы ESDIRK 2-го стадийного порядка, и в дальнейшем под методами ESDIRK подразумеваем именно такие методы.

Простейший метод ESDIRK, обладающий перечисленными свойствами, имеет 2-й порядок и задается таблицей Бутчера

$$\begin{array}{c|ccc} 0 & 0 & & \\ 2\gamma & \gamma & \gamma & \\ 1 & \beta & \beta & \gamma \\ \hline & \beta & \beta & \gamma \end{array}, \quad \gamma = 1 - \frac{\sqrt{2}}{2}, \quad \beta = \frac{\sqrt{2}}{4}.$$

Этот метод можно интерпретировать как последовательное применение правила трапеций (TR) и формулы дифференцирования назад 2-го порядка (BDF2), поэтому он получил название TR-BDF2. Благодаря простым расчетным формулам и высокой эффективности, этот метод широко применяется в практических вычислениях, и ему посвящено множество работ, среди которых [17], [18]. Однако метод TR-BDF2 эффективен только при вычислениях с низкой точностью, а для получения более точного результата следует использовать методы более высоких порядков.

Четырехстадийный метод ESDIRK 3-го порядка можно получить, задав γ равным одному из корней многочлена $1 - 9\gamma + 18\gamma^2 - 6\gamma^3$. Задав $\gamma = 0.4358...$, получаем L-устойчивый метод, а задав

$\gamma = 0.1589...$, получаем более точный, но $L(75.6°)$-устойчивый метод. Такие методы рассматривались в [4]–[7]. Последующие исследования показали, что можно построить более эффективные методы, увеличив число стадий. Поэтому в настоящей статье рассматриваются пяти- и шестистадийные методы порядков 3 и 4.

2. ФУНКЦИЯ УСТОЙЧИВОСТИ

Построение метода ESDIRK обычно начинается с выбора значения диагонального элемента γ. Пусть $r = s - 1$ – число неявных стадий, а p – порядок метода. Примем $p = r - 1$ и $R(\infty) = 0$, тогда функция устойчивости метода ESDIRK однозначно определяется значением γ и имеет вид

$$R(z) = \frac{1}{1-\gamma z} + \sum_{i=1}^{r-1} D_i(\gamma)\frac{z^i}{(1-\gamma z)^{i+1}}, \quad D_i(\gamma) = \sum_{j=0}^{i}\frac{(-\gamma)^j}{(i-j)!}\binom{i}{j}. \tag{2.1}$$

Функция (2.1) аппроксимирует экспоненту с порядком $p = r - 1$. Разложив выражение $e^z - R(z)$ в ряд Тейлора, получаем

$$e^z - R(z) = \frac{C_r}{r!}z^r + \frac{C_{r+1}}{(r+1)!}z^{r+1} + O(z^{r+2}), \tag{2.2}$$

где коэффициенты C_r и C_{r+1} совпадают с одними из коэффициентов погрешности метода. Заметим, что при $\gamma = 0$, $r \le 5$ получаем функцию устойчивости явного p-стадийного метода порядка $p = r - 1$, тогда $C_r = C_{r+1} = 1$.

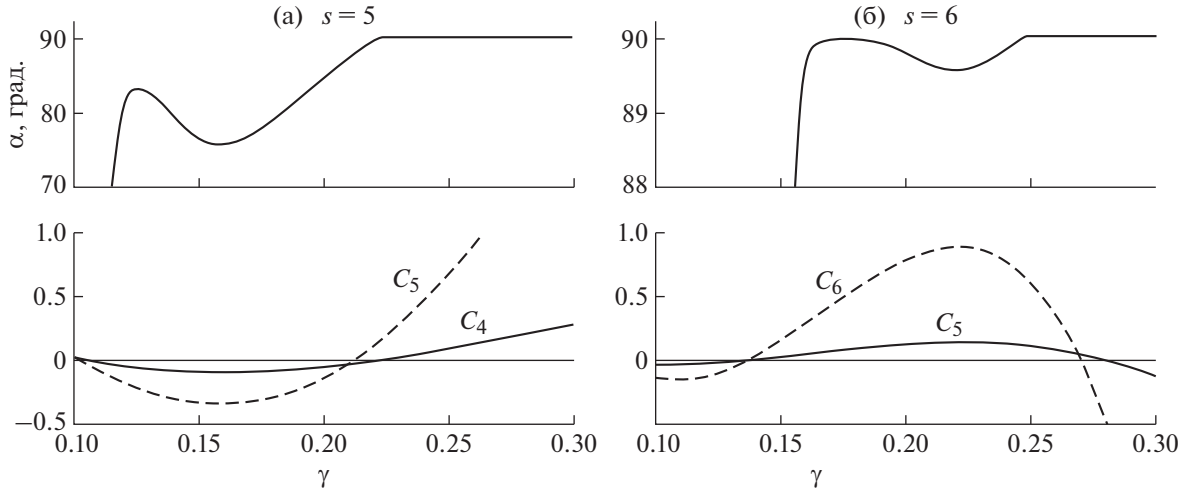
При выборе подходящих значений γ исходим из того, что метод должен иметь достаточно большой сектор устойчивости (примем $\alpha > 75°$). Это требование задает ограничение величины γ снизу. Кроме того, должна обеспечиваться достаточно высокая точность решения уравнения Далквиста (зададим ограничения $|C_r| < 1$, $|C_{r+1}| < 1$, т.е. точность должна быть заведомо выше, чем у явного метода порядка $r - 1$, полученного при $\gamma = 0$). Из условия $0 \le c_i \le 1$ получаем также $\gamma \le 0.5$. Эти требования ограничивают величину γ сверху. Исходя из перечисленных требований, получаем ограничения значения γ для пятистадийных методов в виде $0.117 < \gamma < 0.263$, а для шестистадийных методов – в виде $0.143 < \gamma < 0.289$.

Исследуем зависимости угла α и коэффициентов C_r, C_{r+1} от γ. Для пятистадийных методов ($r = 4$) с функцией устойчивости (2.1) получаем

$$C_4 = 24D_4(\gamma), \quad C_5 = 1 - 200\gamma^2 + 1200\gamma^3 - 1800\gamma^4 + 480\gamma^5.$$

На фиг. 1а приведены графики зависимостей α, C_4 и C_5 от γ. В общем случае при выборе соответствующих коэффициентов такие методы имеют 3-й порядок. А если задать γ равным одному из 4 значений, удовлетворяющих условию $C_4 = 0$, то метод может иметь 4-й порядок. Наилучшая точность обеспечивается при наименьшем из этих значении ($\gamma = 0.1064...$), но такой метод не является даже $L(0)$-устойчивым. Таким образом, выбор γ сводится к компромиссу между точностью и устойчивостью.

В первых пяти строках табл. 1 приведены характеристики функции устойчивости для некоторых значений γ, пригодных для построения пятистадийных $L(\alpha)$-устойчивых методов 3-го порядка. Значение $\gamma = 0.125$ примерно соответствует локальному максимуму зависимости $\alpha(\gamma)$. Близкое к этому значение $\gamma = 0.1288...$ (один из корней многочлена $1 - 12\gamma + 36\gamma^2 - 24\gamma^3$) обеспечивает 2-й порядок L-затухания. Метод будет L-устойчивым, если $0.2236... \le \gamma \le 0.5728...$ (см. [2]). Значение $\gamma = 0.225$ близко к левой границе этого интервала и использовалось в двух методах из [13]. А значение $\gamma = 0.5728...$ (правая граница интервала) является единственным, при котором метод L-устойчив и имеет 4-й порядок. В результате сравнения характеристик мы выбрали значение $\gamma = 0.2204...$, при котором метод имеет 4-й порядок (см. строку 1 в табл. 2). Метод с таким значением γ был предложен в [7] и является наиболее эффективным среди пятистадийных $L(\alpha)$-устойчивых методов. Он не является L-устойчивым, но ниже будет показано, что даже при решении жесткой задачи с чисто мнимым спектром матрицы Якоби он показывает приемлемые результаты. В табл. 1 (строки 3, 4, 6, 7) приведены также значения γ методов, построенных в разд. 4, 5 и удовлетворяющих дополнительным условиям порядка для ДАУ индексов 2 и 3.



Фиг. 1.

Рассмотрим теперь функцию $R(z)$ шестистадийных методов 4-го порядка. В этом случае коэффициенты погрешности в (2.2) находим по формулам

$$C_5 = 120D_5(\gamma), \quad C_6 = 1 - 450\gamma^2 + 4800\gamma^3 - 16\,200\gamma^4 + 17\,280\gamma^5 - 3600\gamma^6.$$

На фиг. 1б приведены графики зависимостей α, C_5 и C_6 от γ на интересующем нас интервале, а в строках 2–5 табл. 2 приведены характеристики функции устойчивости при некоторых значениях γ. В пределах этого интервала метод будет L-устойчивым при $\gamma \geq 0.2479...$ (см. [2]). Наиболее удобно близкое к граничному значение $\gamma = 0.25$, которое использовалось при построении методов 4-го порядка в [2], [7]–[10], [13], [14], [21]. В [2] рекомендовалось также значение $\gamma = 4/15 = 0.2666...$, при котором C_5 и C_6 малы. Отметим, что при $0.164 \leq \gamma \leq 0.191$ коэффициенты погрешности невелики, а $\alpha > 89.9°$. Значение $\gamma = 1/6$ – наиболее удобное из этого интервала. Методы с таким γ предлагались в [9], [10], [14]. Близкое к этому значение $\gamma = 0.1744...$ (один из корней многочлена $1 - 20\gamma + 120\gamma^2 - 240\gamma^3 + 120\gamma^4$) обеспечивает 2-й порядок L-затухания.

3. ТОЧНОСТЬ

Главным показателем точности методов численного решения ОДУ является порядок аппроксимации. Условия, обеспечивающие порядок p метода Рунге–Кутты, можно представить в виде

$$e(T_{ij}) = 0, \quad i = 1, \ldots, p, \quad j = 1, \ldots, \nu_i, \tag{3.1}$$

где $e(T_{ij})$ – коэффициенты погрешности метода; T_{ij} – корневые деревья порядка i, соответствующие этим коэффициентам; ν_i – число различных деревьев порядка i. Для $i \leq 6$ имеем $\nu_1 = \nu_2 = 1$,

Таблица 1. Характеристики функций устойчивости методов 3-го порядка

№	s	γ	Устойчивость	C_4	C_5
1	5	0.125	L(83.12°)	–0.0566	–0.206
2	5	0.1288	$L2$(82.90°)	–0.0691	–0.233
3	5	0.1815	L(79.35°)	–0.0800	–0.272
4	5	0.2164	L(88.81°)	–0.0107	0.076
5	5	0.225	L(90°)	0.0130	0.207
6	6	1/6	$L2$(88.91°)	–0.0185	–0.0185
7	6	0.2	L(90°)	0.0096	0.170

Таблица 2. Характеристики функций устойчивости методов 4-го порядка

№	s	γ	Устойчивость	C_5	C_6
1	5	0.2204	$L(89.55°)$	0.135	0.878
2	6	1/6	$L(89.95°)$	0.059	0.367
3	6	0.1744	$L2(89.97°)$	0.076	0.476
4	6	0.25	$L(90°)$	0.102	0.590
5	6	4/15	$L(90°)$	0.050	0.109

$\nu_3 = 2$, $\nu_4 = 4$, $\nu_5 = 9$, $\nu_6 = 20$. Условия порядка до 5-го включительно вместе с соответствующими деревьями приведены в [14], [26].

В общем случае все коэффициенты погрешности различны, но если стадийный порядок больше 1, то число различных коэффициентов сокращается. Для методов 2-го стадийного порядка справедливы равенства

$$e(T_{31}) = e(T_{32}); \quad e(T_{41}) = e(T_{42}), \quad e(T_{43}) = e(T_{44});$$
$$e(T_{51}) = e(T_{52}) = e(T_{55}), \quad e(T_{53}) = e(T_{54}), \quad e(T_{56}) = e(T_{57}), \quad e(T_{58}) = e(T_{59}).$$

Для методов ESDIRK из условия 1-го стадийного порядка получаем

$$a_{i1} = c_i - \sum_{j=2}^{i-1} a_{ij} - \gamma, \quad i = 2, \ldots, s. \tag{3.2}$$

Коэффициенты a_{i1} не входят во все остальные условия, поэтому их вычисляют в последнюю очередь. Из условия 2-го стадийного порядка имеем

$$c_2 = 2\gamma, \quad a_{i2} = \frac{1}{4\gamma}\left[c_i^2 - 2\left(\sum_{j=3}^{i-1} a_{ij}c_j + \gamma c_i\right)\right], \quad i = 3, \ldots, s. \tag{3.3}$$

Остальные коэффициенты находим исходя из условий порядка (3.1) для $2 < i \le p$, необходимого условия $L(\alpha)$-устойчивости $R(\infty) = 0$ и некоторых дополнительных условий (например, условий порядка для ДАУ индексов 2 и 3).

Основная трудность при построении методов высоких порядков (4 и выше) заключается в необходимости обеспечить выполнение большого числа алгебраических условий. Учет диагональной формы матрицы $\mathbf{A}$ позволяет упростить эти условия, в результате они становятся не сложнее, чем для явных методов. Упрощенные условия порядка для методов ESDIRK до 5-го порядка включительно приведены в [9], [14]. Пусть выполняются условия 2-го стадийного порядка (3.2), (3.3). Тогда для обеспечения 3-го порядка метода дополнительно должно выполняться условие

$$\sum_{i=2}^{s-1} b_i c_i^2 = \frac{1}{3} - \gamma, \tag{3.4}$$

а для обеспечения 4-го порядка должны выполняться также и условия

$$\sum_{i=2}^{s-1} b_i c_i^3 = \frac{1}{4} - \gamma, \quad \sum_{i=4}^{s-1} b_i \left[\sum_{j=3}^{i-1} a_{ij}\left(\sum_{k=2}^{j-1} a_{jk}c_k\right)\right] = \frac{1}{24} - \frac{1}{2}\gamma + \frac{3}{2}\gamma^2 - \gamma^3. \tag{3.5}$$

Кроме выполнения условий порядка потребуем выполнения необходимого условия $L(\alpha)$-устойчивости, которое запишется в виде

$$R(\infty) = 1 - \tilde{\mathbf{b}}^{\mathrm{T}}\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{c}} = 1 - \mathbf{e}_r{}^{\mathrm{T}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{c}} = 0, \tag{3.6}$$

где

$$\tilde{\mathbf{A}} = \begin{bmatrix} \gamma & 0 & \cdots & 0 & 0 \\ a_{32} & \gamma & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{s-1,2} & a_{s-1,3} & \cdots & \gamma & 0 \\ a_{s2} & a_{s3} & \cdots & a_{s,s-1} & \gamma \end{bmatrix}, \quad \tilde{\mathbf{b}} = \begin{bmatrix} a_{s2} \\ a_{s3} \\ \vdots \\ a_{s,s-1} \\ \gamma \end{bmatrix}, \quad \tilde{\mathbf{c}} = \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_{s-1} \\ c_s \end{bmatrix}, \quad \mathbf{e}_r = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

При $p = s - 1$ выполнение этого условия обеспечивается, если γ – корень многочлена $D_p(\gamma)$, а при $p = s - 2$ условие (3.6) приводится к виду

$$a_{s,s-1}a_{s-1,s-2}\cdots a_{32}c_2 = \gamma D_p(\gamma). \tag{3.7}$$

Точность решения ОДУ методом порядка p зависит, прежде всего, от размера шага и от коэффициентов погрешности $e(T_{p+1,i})$. Используя свободные коэффициенты метода как оптимизируемые параметры, можно повысить точность метода, минимизировав коэффициенты погрешности. Такой подход является обычным при построении методов Рунге–Кутты.

При решении жестких ОДУ реальный порядок метода может быть ниже классического порядка, что приводит к заметному снижению точности решения. Для исследования этого явления, которое получило известность как феномен снижения порядка, Протеро и Робинсон (см. [24]) исследовали уравнение

$$y' = \lambda(y - \varphi(t)) + \varphi'(t), \quad y(t_0) = \varphi(t_0), \tag{3.8}$$

с точным решением $y(t) = \varphi(t)$. Используя разложение $\varphi(t)$ в ряд Тейлора, получаем локальную ошибку метода стадийного порядка q в виде

$$\delta_1 = \varphi(t_0 + h) - y_1 = \sum_{i=q+1}^{\infty} e_i(h\lambda)\frac{d^i\varphi(t_0)}{dt^i}\frac{h^i}{i!},$$

где $e_i(z) = z\mathbf{b}^{\mathrm{T}}(\mathbf{I} - z\mathbf{A})^{-1}(\mathbf{c}^i - i\mathbf{A}\mathbf{c}^{i-1}) + (1 - i\mathbf{b}^{\mathrm{T}}\mathbf{c}^{i-1})$ – предложенные в [7] функции погрешности. Глобальная ошибка выражается формулой

$$\varphi(t_{n+1}) - y_{n+1} = R(h\lambda)(\varphi(t_n) - y_n) + \delta_{n+1},$$

где δ_{n+1} – локальная ошибка на $(n+1)$-м шаге. Для жестко точных методов с явной 1-й стадией функции погрешности можно представить в виде

$$e_i(z) = \mathbf{e}_r{}^{\mathrm{T}}\left(\mathbf{I} - z\tilde{\mathbf{A}}\right)^{-1}\left(\tilde{\mathbf{c}}^i - i\tilde{\mathbf{A}}\tilde{\mathbf{c}}^{i-1}\right).$$

В [8], [9], [11], [14] были рассмотрены также функции погрешности $e_{ij}(z)$, полученные при анализе ошибок решения простейших модельных уравнений, отличных от уравнения Протеро–Робинсона. Но при этом для всех j имеем $e_{q+1,j}(z) \equiv e_{q+1}(z)$. Таким образом, функция $e_{q+1}(z)$ задает главный член погрешности при решении жестких модельных уравнений, поэтому далее будем рассматривать только функцию $e_3(z)$.

Анализ ошибки численного решения уравнения (3.8) в зависимости от величины $z = h\lambda$ показал важность понятий жесткой точности и стадийного порядка для эффективного решения жестких задач. Жесткая точность обеспечивает малую ошибку при больших по модулю значениях z, а высокий стадийный порядок ограничивает ошибку при умеренных z. Снижение точности и порядка тем заметнее, чем больше разность между классическим порядком p и стадийным порядком q. Можно ограничить снижение точности, если минимизировать функцию $e_3(z)$. Методы ESDIRK с минимизированной функцией погрешности предлагались в [7]–[9], [14]. В [9], [14], [27] были рассмотрены также методы, имеющие $e_3(z) \equiv 0$, но они требуют выполнения двух дополнительных стадий.

Обсудим теперь точность решения ДАУ методами ESDIRK. Системы ДАУ индекса 1 имеют вид

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{0} = \mathbf{g}(\mathbf{y}, \mathbf{z}),$$

Таблица 3. Необходимые условия для порядков сходимости компонент ДАУ

Условие	ДАУ индекса 2		ДАУ индекса 3		
	$p_y = 4$	$p_z = 3$	$p_y = 3$	$p_z = 3$	$p_u = 2$
(3.4)	+	+	+	+	–
(3.5)	+	–	–	–	–
(3.9)	+	+	+	+	+
(3.10)	–	–	+	+	+
(3.11)	–	–	+	+	–
(3.12)	+	–	–	–	–

где матрица $\mathbf{g_z} = \partial \mathbf{g}(\mathbf{y}, \mathbf{z})/\partial \mathbf{z}$ обратима в окрестности решения. Жестко точные методы, к которым относятся ESDIRK, обеспечивают точное выполнение алгебраического соотношения, поэтому порядки сходимости дифференциальных и алгебраических компонент совпадают с порядком метода: $p_y = p_z = p$.

Систему ДАУ индекса 2 можно привести к виду

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{0} = \mathbf{g}(\mathbf{y}),$$

где матрица $\mathbf{g_y f_z}$ обратима в окрестности решения. Как следствие теоремы 5.2 из [28], порядки сходимости соответствующих компонент при решении таких задач методами ESDIRK (при $p \geq 3$, $q = 2$) следующие: $p_y = \min(p, q+1) = 3$, $p_z = q = 2$.

Систему ДАУ индекса 3 можно представить в виде

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{z}' = \mathbf{k}(\mathbf{y}, \mathbf{z}, \mathbf{u}), \quad \mathbf{0} = \mathbf{g}(\mathbf{y}),$$

где матрица $\mathbf{g_y f_z k_u}$ обратима в окрестности решения. В [29] получены порядки сходимости компонент $\mathbf{y}$, $\mathbf{z}$, $\mathbf{u}$ при решении таких задач методами с обратимой матрицей $\mathbf{A}$. Для методов с явной 1-й стадией аналогичных результатов мы не нашли, но численные эксперименты с методами ESDIRK давали оценки порядков $\tilde{p}_y = \tilde{p}_z = 2$, $\tilde{p}_u = 1$.

В [10], [11], [14] для исследования сходимости решения ДАУ индексов 2 и 3 использовались простые модельные уравнения, позволившие получить дополнительные условия, необходимые для повышения порядков сходимости различных компонент ДАУ. Для методов ESDIRK при $R(\infty) = 0$, $q = 2$ эти условия имеют вид

$$\mathbf{e}_r^{\mathrm{T}} \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{c}}^3 = 3; \tag{3.9}$$

$$\mathbf{e}_r^{\mathrm{T}} \tilde{\mathbf{A}}^{-2} \tilde{\mathbf{c}}^3 = 6; \tag{3.10}$$

$$\tilde{\mathbf{b}}^{\mathrm{T}} (\tilde{\mathbf{c}} \cdot (\tilde{\mathbf{A}}^{-2} \tilde{\mathbf{c}}^3)) = 2, \quad \tilde{\mathbf{b}}^{\mathrm{T}} (\tilde{\mathbf{A}}^{-2} \tilde{\mathbf{c}}^3)^2 = 12; \tag{3.11}$$

$$\tilde{\mathbf{b}}^{\mathrm{T}} (\tilde{\mathbf{c}} \cdot (\tilde{\mathbf{A}}^{-1} \tilde{\mathbf{c}}^3)) = 3/4. \tag{3.12}$$

В табл. 3 приведены необходимые условия сходимости с заданным порядком для компонент ДАУ, где необходимое условие для каждой компоненты отмечено знаком +. Эти условия являются необходимыми для рассмотренных в [10], [11], [14] модельных уравнений, а значит, и для уравнений более общего вида. На ряде тестовых задач мы убедились, что выполнение этих условий действительно обеспечивает указанные порядки, но у нас нет доказательства, что эти условия являются также и достаточными.

4. ПЯТИСТАДИЙНЫЕ МЕТОДЫ

Чтобы различать рассматриваемые методы, условимся обозначать их в виде ESDIRK$sp(\gamma)$, где s – число стадий, p – порядок, γ – диагональный элемент.

Пятистадийный метод 4-го порядка должен удовлетворять условиям (3.3)–(3.5), из которых получаем

$$c_2 = 2\gamma, \quad a_{32} = \frac{c_3^2 - 2\gamma c_3}{4\gamma}, \quad \begin{bmatrix} b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} a_{52} \\ a_{53} \\ a_{54} \end{bmatrix} = \begin{bmatrix} c_2 & c_3 & c_4 \\ c_2^2 & c_3^2 & c_4^2 \\ c_2^3 & c_3^3 & c_4^3 \end{bmatrix}^{-1} \begin{bmatrix} 1/2 - \gamma \\ 1/3 - \gamma \\ 1/4 - \gamma \end{bmatrix},$$
$$a_{43} = \frac{1 - 12\gamma + 36\gamma^2 - 24\gamma^3}{24 b_4 a_{32} c_2}, \quad a_{42} = \frac{c_4^2 - 2(a_{43}c_3 + \gamma c_4)}{4\gamma}, \tag{4.1}$$

после чего коэффициенты a_{i1} находим из (3.2). Такой метод будет иметь $R(\infty) = 0$, если γ – корень многочлена $D_4(\gamma)$. Мы выбрали значение

$$\gamma = 0.22042841025921, \tag{4.2}$$

которое считаем наиболее подходящим (тогда метод $L(89.55°)$-устойчив и имеет малые значения двух из девяти коэффициентов погрешности 5-го порядка: $e(T_{58}) = e(T_{59}) = 0.135$). Альтернативное значение $\gamma = 0.5728...$ обеспечивает L-устойчивость, но тогда $e(T_{58}) = e(T_{59}) = -3.27$ и $c_2 > 1$.

У нас остались два свободных параметра – c_3 и c_4. Их подбором можно минимизировать остальные коэффициенты погрешности, но это приводит к большим значениям коэффициентов метода, что нежелательно. Более заметный эффект получим, минимизировав функцию погрешности. При $p = r = s - 1$ функция $e_3(z)$ метода ESDIRK зависит только от γ и c_3 и имеет вид

$$e_3(z) = \frac{z^{p-2}}{(1-\gamma z)^p} 2\gamma D_{p-1}(\gamma)[c_3 - (c_3^* + \gamma) - \gamma z(c_3 - c_3^*)], \quad c_3^* = 4\gamma + \gamma^2 \frac{D_{p-2}(\gamma)}{D_{p-1}(\gamma)}.$$

В [9], [14] было показано, что неравенство $c_3^* \le c_3 \le c_3^* + \gamma$ задает множество всех (Парето-оптимальных) значений c_3, изменяя которые невозможно уменьшить функцию $|e_3(z)|$ сразу во всех точках левой полуплоскости. Отметим также, что значение $c_3 = c_3^*$ обеспечивает выполнение равенства (3.9).

При заданном значении γ (4.2) интервал оптимальных значений c_3 получаем в виде $0.701 \le c_3 \le 0.921$. Мы выбрали

$$c_3 = (2 + \sqrt{2})\gamma \approx 0.753, \quad c_4 = \frac{a + \gamma\sqrt{b}}{48\gamma^3 - 72\gamma^2 + 24\gamma - 2} \approx 0.601,$$
$$a = (6\gamma^2 - 6\gamma + 1)c_3 + 96\gamma^3 - 100\gamma^2 + 27\gamma - 2, \tag{4.3}$$
$$b = (19\,872\gamma^3 - 17\,808\gamma^2 + 4160\gamma - 264)c_3 - 26\,784\gamma^3 + 24\,128\gamma^2 - 5656\gamma + 361.$$

Эти значения обеспечивают попадание c_3 в оптимальный интервал, а также L-устойчивость 3-й и 4-й стадий, тогда

$$a_{i1} = a_{i2}, \quad i = 2, \ldots, 5. \tag{4.4}$$

Метод, задаваемый формулами (4.1)–(4.4), был предложен в [7], численные значения его коэффициентов приведены в [7], [9], [14]. Обозначим его через ESDIRK54(0.220).

Рассмотрим теперь построение пятистадийных методов, обладающих повышенной точностью при решении ДАУ индексов 2 и 3. Методы 4-го порядка не подходят для этого вследствие недостаточного числа свободных коэффициентов. Поэтому снизим порядок до 3-го и используем условия (3.2)–(3.4), (3.7), (3.9)–(3.11). В этом случае число алгебраических условий совпадает с числом коэффициентов метода. Мы нашли 12 методов, которые удовлетворяют этим условиям. Они разбиваются на 3 группы в зависимости от значений b_2 и b_3: 1) $b_2 = 0$, $b_3 \ne 0$ – 6 методов; 2) $b_2 = 0$, $b_3 = 0$ – 3 метода; 3) $b_2 \ne 0$, $b_3 \ne 0$ – 3 метода. Для всех методов c_3 задается формулой

$$c_3 = \frac{(1 - 6\gamma + 2\gamma^2)(3\gamma - 6\gamma^2)}{1 - 9\gamma + 18\gamma^2 - 6\gamma^3}. \tag{4.5}$$

Из всех 12 методов только два удовлетворяют нашим требованиям. Первый из них принадлежит 1-й группе, в которой γ – один из корней многочлена $1 - 24\gamma + 186\gamma^2 - 600\gamma^3 + 828\gamma^4 - 432\gamma^5 + 72\gamma^6$. Его коэффициенты находим по формулам (3.2), (3.3), (4.5),

$$\gamma = 0.18157222316139, \quad c_4 = \frac{1 - 20\gamma + 135\gamma^2 - 366\gamma^3 + 378\gamma^4 - 72\gamma^5}{1 - 15\gamma + 81\gamma^2 - 180\gamma^3 + 162\gamma^4 - 36\gamma^5}, \quad b_2 = 0,$$

$$b_3 = \frac{(3 - 6\gamma)c_4 - 2 + 6\gamma}{c_3(c_4 - c_3)}, \quad b_4 = \frac{(3 - 6\gamma)c_3 - 2 + 6\gamma}{c_4(c_3 - c_4)}, \quad a_{43} = \frac{\gamma(1 - 9\gamma + 18\gamma^2 - 6\gamma^3)}{6b_4a_{32}c_2}.$$

Обозначим этот метод через ESDIRK53(0.182).

Второй метод принадлежит 2-й группе, в которой γ – один из трех вещественных корней многочлена $2 - 36\gamma + 201\gamma^2 - 432\gamma^3 + 360\gamma^4 - 72\gamma^5$ (два других корня – комплексно-сопряженные). Его коэффициенты находим по формулам (3.2), (3.3), (4.5),

$$\gamma = 0.21646827973787, \quad c_4 = \frac{2(1 - 3\gamma)}{3(1 - 2\gamma)}, \quad b_2 = b_3 = 0,$$

$$b_4 = \frac{3(1 - 2\gamma)^2}{4(1 - 3\gamma)}, \quad a_{43} = \frac{\gamma(1 - 9\gamma + 18\gamma^2 - 6\gamma^3)}{6b_4a_{32}c_2}.$$

Обозначим этот метод через ESDIRK53(0.216).

5. ШЕСТИСТАДИЙНЫЕ МЕТОДЫ

При построении шестистадийных методов 4-го порядка часто используют значение $\gamma = 1/4$, которое обеспечивает L-устойчивость и малые коэффициенты погрешности. В [13] при таком значении γ свободные параметры метода выбирались из условий L-устойчивости внутренних стадий и минимизации коэффициентов погрешности. Построенный метод ESDIRK4(3)6L [2]SA (см. [13], табл. 16) “is recommended as a good default method for solving stiff problems at moderate error tolerances”. В наших обозначениях это метод ESDIRK64(1/4).

В [10], [14] рассматривались шестистадийные методы ESDIRK 4-го порядка, удовлетворяющие условиям (3.9)–(3.12). При $\gamma = 1/6$ такие методы образуют однопараметрическое семейство со свободным параметром c_4. Из этого семейства мы выбрали метод, имеющий $c_4 = 2/3$ и таблицу Бутчера

0	0					
1/3	1/6	1/6				
8/15	31/150	4/25	1/6			
2/3	23/88	8/99	125/792	1/6		
1/2	61/384	13/72	125/1152	−11/96	1/6	
1	1/6	0	0	0	2/3	1/6
b_i	1/6	0	0	0	2/3	1/6

Обозначим этот метод через ESDIRK64(1/6).

Рассмотрим также шестистадийные методы 3-го порядка, удовлетворяющие условиям (3.9)–(3.11). С уменьшением порядка появились дополнительные свободные параметры, которые позволили обеспечить удобную и эффективную реализацию методов с контролем ошибки. Приведем два таких метода.

Таблица 4. Характеристики методов ESDIRK

Метод $sp(\gamma)$	α	$\|e(T_{p+1})\|$	$\|e(T_{p+2})\|$	$\|e_3(z)\|_{\mathrm{R}}$	$\|e_3(z)\|_{\mathrm{C}}$
ESDIRK53(0.182)	79.35°	0.180	0.78	0.0027	0.0083
ESDIRK53(0.216)	88.81°	0.084	0.44	0.0058	0.0179
ESDIRK63(1/6)	88.91°	0.026	0.05	0.0040	0.0138
ESDIRK63(1/5)	90°	0.015	0.28	0.0013	0.0052
ESDIRK54(0.220)	89.55°	0.25	1.64	0.0020	0.0116
ESDIRK64(1/6)	89.95°	0.13	0.72	0.0016	0.0086
ESDIRK64(1/4)	90°	0.16	1.11	0.0114	0.0326

Первый метод является $L2(88.91°)$-устойчивым (см. строку 6 в табл. 1) и имеет таблицу коэффициентов

$$\begin{array}{c|cccccc} 0 & 0 & & & & & \\ 1/3 & 1/6 & 1/6 & & & & \\ 2/3 & 1/6 & 1/3 & 1/6 & & & \\ 1 & 1/3 & 0 & 1/2 & 1/6 & & \\ 1 & 7/16 & 0 & 3/16 & 5/24 & 1/6 & \\ 1 & 1/8 & 3/8 & 3/8 & 1/360 & -2/45 & 1/6 \\ \hline b_i & 1/8 & 3/8 & 3/8 & 1/360 & -2/45 & 1/6 \end{array}. \tag{5.1}$$

Обозначим его через ESDIRK63(1/6).

Второй метод является L-устойчивым (см. строку 7 в табл. 1). При его построении мы старались минимизировать функцию $e_3(z)$. Полученный метод ESDIRK63(1/5) имеет таблицу коэффициентов

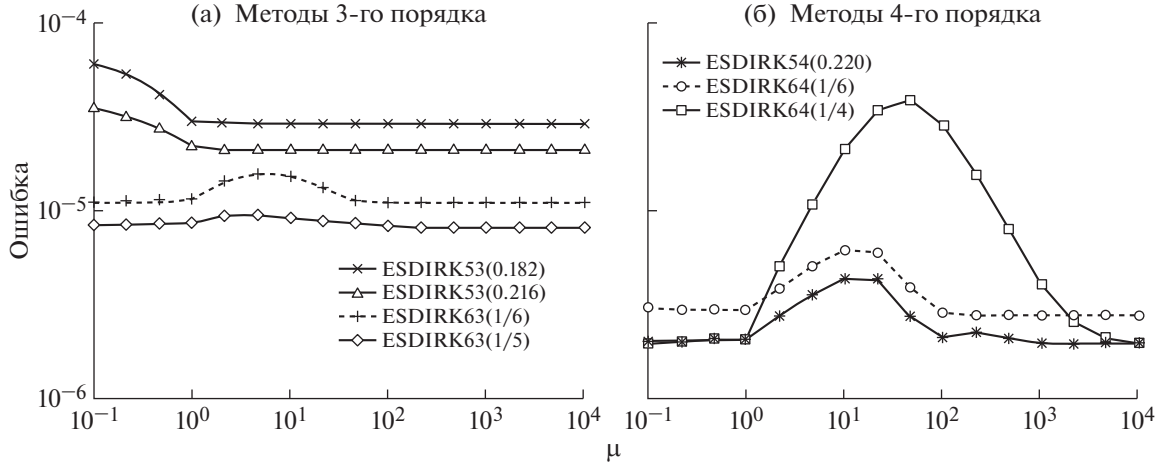
$$\begin{array}{c|cccccc} 0 & 0 & & & & & \\ 2/5 & 1/5 & 1/5 & & & & \\ 4/5 & 1/5 & 2/5 & 1/5 & & & \\ 0 & -877/8040 & -731/4020 & 731/8040 & 1/5 & & \\ 1 & 257423/2807040 & 59/1920 & 1381/3840 & 7437/23392 & 1/5 & \\ 1 & 5047/29240 & 8/15 & 29/120 & -4489/109650 & -8/75 & 1/5 \\ \hline b_i & 5047/29240 & 8/15 & 29/120 & -4489/109650 & -8/75 & 1/5 \end{array}.$$

Основные характеристики рассмотренных методов приведены в табл. 4, где

$$\|e(T_i)\| = \left(\sum_{j=1}^{\nu_i} e(T_{ij})^2\right)^{1/2}, \quad \|e_3(z)\|_{\mathrm{R}} = \max\left(|e_3(x)|,\ x \le 0\right),$$
$$\|e_3(z)\|_{\mathrm{C}} = \max\left(|e_3(z)|,\ \mathrm{Re}\, z \le 0\right) = \max\left(|e_3(iy)|,\ y \ge 0\right).$$

6. РЕЗУЛЬТАТЫ РАСЧЕТОВ С ПОСТОЯННЫМ РАЗМЕРОМ ШАГА

Чтобы обеспечить одинаковые условия для всех методов, задаем размер шага таким, чтобы на всем интервале было выполнено заданное число неявных стадий Nr. В этом случае будет выполнено $N = Nr/r$ шагов размером $h = T/N$. На каждой неявной стадии выполняем достаточно большое число итераций, заведомо обеспечивающее сходимость.



Фиг. 2.

Посмотрим сначала, как влияет жесткость задачи на точность решения. Для этого используем задачу

$$\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \begin{bmatrix} a & b \\ b & a \end{bmatrix} \begin{bmatrix} y_1 - \sin t \\ y_2 - \cos t \end{bmatrix} + \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix}, \quad \mathbf{y}(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$
$$a = -(\mu + 1)/2, \quad b = (\mu - 1)/2, \quad 0 \le t \le 2\pi,$$

решение которой $y_1(t) = \sin t$, $y_2(t) = \cos t$, а собственные числа матрицы Якоби равны -1 и $-\mu$. Задаем $Nr = 120$ и вычисляем максимальное значение евклидовой нормы ошибки на всем интервале интегрирования. Полученные результаты (зависимости ошибки от μ) приведены на фиг. 2. Большой "горб" кривой для ESDIRK64(1/4) объясняется поведением функции погрешности $e_3(z)$, которая у жестко точных методов достигает максимального значения при умеренных значениях z. У других методов такой горб проявляется значительно меньше или совсем отсутствует, что объясняется меньшими значениями функции погрешности и меньшей разностью $p - q$ у методов 3-го порядка.

Вторая задача – тест PLATE (ее описание дано в [2]). Она содержит 80 переменных и имеет комплексный спектр матрицы Якоби. Интегрирование выполняем на интервале [0, 7] с размерами шага $h_1 = 7/N$ и $h_2 = 0.1h_1$, где $N = 280/r$. Точность решения оцениваем величиной

$$scd = -\lg\left(\max_i \left|\frac{y_i - \tilde{y}_i}{y_i}\right|\right), \tag{6.1}$$

где y_i – точное, а $\tilde{y}_i$ – численное решение по i-й компоненте в конце интервала интегрирования. Величина *scd* (significant correct digits) показывает число правильных значащих цифр численного решения. Результаты приведены в табл. 5. Видно, что порядок метода практически не влияет на точность решения этой задачи, и лучшие результаты показывают методы, имеющие малые нормы функции погрешности (см. табл. 4). Такие результаты характерны для умеренно жестких задач, а также для задач с распределенным спектром матрицы Якоби, к которым относится PLATE.

Таблица 5. Значения *scd* при решении задачи PLATE

h	Метод ESDIRK						
	53(0.182)	53(0.216)	63(1/6)	63(1/5)	54(0.220)	64(1/6)	64(1/4)
h_1	3.68	3.51	3.43	3.91	3.77	3.67	2.78
$0.1h_1$	6.32	5.95	5.87	6.33	6.29	6.38	5.49

Таблица 6. Результаты решения задачи (6.2) при $h = \text{const}$

Метод	e_y	e_z	$\tilde{p}_y$	$\tilde{p}_z$
ESDIRK53(0.182)	1.55×10^{-5}	7.55×10^{-5}	3.05	3.04
ESDIRK53(0.216)	7.96×10^{-6}	7.93×10^{-5}	3.04	3.00
ESDIRK63(1/6)	1.26×10^{-5}	2.02×10^{-4}	3.06	2.99
ESDIRK63(1/5)	1.13×10^{-5}	4.92×10^{-4}	3.01	2.99
ESDIRK54(0.220)	4.61×10^{-6}	3.31×10^{-4}	3.08	2.02
ESDIRK64(1/6)	1.15×10^{-6}	1.20×10^{-4}	3.98	3.01
ESDIRK64(1/4)	1.65×10^{-5}	2.67×10^{-3}	3.00	2.00

Таблица 7. Результаты решения задачи (6.3) при $h = \text{const}$

Метод	e_y	e_z	e_u	$\tilde{p}_y$	$\tilde{p}_z$	$\tilde{p}_u$
ESDIRK53(0.182)	6.88×10^{-6}	5.95×10^{-6}	8.26×10^{-4}	3.01	3.01	2.00
ESDIRK53(0.216)	3.70×10^{-6}	2.26×10^{-6}	4.30×10^{-4}	3.00	3.00	2.00
ESDIRK63(1/6)	3.04×10^{-6}	2.18×10^{-6}	4.66×10^{-4}	3.03	3.04	2.00
ESDIRK63(1/5)	1.43×10^{-6}	4.35×10^{-6}	1.52×10^{-3}	3.03	3.00	2.00
ESDIRK54(0.220)	5.50×10^{-5}	5.56×10^{-5}	8.57×10^{-3}	2.00	2.01	1.00
ESDIRK64(1/6)	4.74×10^{-6}	3.31×10^{-6}	1.88×10^{-3}	2.98	3.00	2.00
ESDIRK64(1/4)	2.18×10^{-4}	2.60×10^{-4}	7.61×10^{-3}	2.00	2.01	1.06

Еще две задачи – ДАУ индексов 2 и 3. Задача индекса 2 задается уравнениями

$$\begin{gathered} y_1' = y_2 z, \quad y_2' = y_1 (z - 2\cos t), \quad 2y_1 y_2 = \sin(2\sin t), \\ y_1(0) = 0, \quad y_2(0) = z(0) = 1, \quad 0 \le t \le 2\pi \end{gathered} \tag{6.2}$$

и имеет решение $y_1(t) = \sin(\sin t)$, $y_2(t) = \cos(\sin t)$, $z(t) = \cos t$. Задаем $Nr = 200$ и вычисляем e_y и e_z – максимальные ошибки соответствующих компонент на всем интервале (для вектора $\mathbf{y} = (y_1, y_2)^{\mathrm{T}}$ используем евклидову норму ошибки). Вычисляем также оценки порядков сходимости $\tilde{p}_y$ и $\tilde{p}_z$ по соответствующим компонентам (для этого используем ошибки, полученные при размерах шага h и $h/2$). Результаты приведены в табл. 6.

Задача индекса 3 задается уравнениями

$$\begin{gathered} y_1' = z_1, \quad y_2' = z_2, \quad z_1' = -y_1 u - y_2 \sin t, \quad z_2' = -y_2 u + y_1 \sin t, \\ y_1^2 + y_2^2 = 1, \\ y_1(0) = z_2(0) = 0, \quad y_2(0) = z_1(0) = u(0) = 1, \quad 0 \le t \le 2\pi \end{gathered} \tag{6.3}$$

и имеет решение $y_1(t) = \sin(\sin t)$, $y_2(t) = \cos(\sin t)$, $z_1(t) = \cos(\sin t) \cdot \cos t$, $z_2(t) = -\sin(\sin t) \cdot \cos t$, $u(t) = \cos^2 t$. Задаем $Nr = 1000$, и аналогично предыдущей задаче вычисляем ошибки и оценки порядков сходимости. Результаты приведены в табл. 7.

Из приведенных результатов следует, что методы ESDIRK53(0.220) и ESDIRK64(1/4), которые не удовлетворяют условиям (3.9)–(3.11), уступают остальным методам, для которых эти условия выполняются. Аналогичные результаты были получены и при решении других задач индексов 2 и 3.

7. РЕАЛИЗАЦИЯ МЕТОДОВ С КОНТРОЛЕМ ОШИБКИ

Для реализации с переменным размером шага были выбраны методы 3-го порядка ESDIRK63(1/6) и ESDIRK63(1/5) (как более удобные для реализации) и все три метода 4-го порядка. Используем экономичную схему реализации с двухшаговым прогнозом (см. [14], [30], [31]). Рассмотрим эту схему применительно к системе ДАУ

$$\mathbf{y}' = \mathbf{f}(t,\mathbf{y},\mathbf{z}), \quad \mathbf{0} = \mathbf{g}(t,\mathbf{y},\mathbf{z}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \quad \mathbf{z}(t_0) = \mathbf{z}_0. \tag{7.1}$$

Формулы $(n+1)$-го шага решения этой системы методом ESDIRK запишутся в виде

$$\left.\begin{gathered} \mathbf{F}_1 = \mathbf{f}_n, \\ \mathbf{Y}_i = \mathbf{y}_n + h\sum_{j=1}^{i-1} a_{ij}\mathbf{F}_j + h\gamma\mathbf{F}_i\ , \quad \mathbf{F}_i = \mathbf{f}\left(t_n + c_i h,\ \mathbf{Y}_i, \mathbf{Z}_i\right), \\ \mathbf{g}\left(t_n + c_i h,\ \mathbf{Y}_i,\ \mathbf{Z}_i\right) = \mathbf{0} \end{gathered}\right\} \quad i = 2, \ldots, s, \tag{7.2}$$

$$\mathbf{y}_{n+1} = \mathbf{Y}_s\ , \quad \mathbf{z}_{n+1} = \mathbf{Z}_s, \quad \mathbf{f}_{n+1} = \mathbf{F}_s,$$

где выполнение неявных стадий сводится к решению нелинейных алгебраических уравнений.

Обозначим через $\mathbf{f}_\mathbf{y}$, $\mathbf{f}_\mathbf{z}$, $\mathbf{g}_\mathbf{y}$ и $\mathbf{g}_\mathbf{z}$ соответствующие матрицы частных производных, вычисленные в некоторой точке численного решения (предполагается, что эти матрицы не изменяются в течение нескольких шагов). Итерации метода Ньютона при реализации i-й стадии запишутся в виде

$$\begin{bmatrix} \Delta\mathbf{Y}_i^k \\ \Delta\mathbf{Z}_i^k \end{bmatrix} = \begin{bmatrix} \Delta\mathbf{Y}_i^{k-1} \\ \Delta\mathbf{Z}_i^{k-1} \end{bmatrix} + \begin{bmatrix} \mathbf{I} - h\gamma\mathbf{f}_\mathbf{y} & -h\gamma\mathbf{f}_\mathbf{z} \\ -\mathbf{g}_\mathbf{y} & -\mathbf{g}_\mathbf{z} \end{bmatrix}^{-1} \begin{bmatrix} h\sum_{j=1}^{i-1} a_{ij}\mathbf{F}_j + h\gamma\mathbf{F}_i^{k-1} - \Delta\mathbf{Y}_i^{k-1} \\ \mathbf{G}_i^{k-1} \end{bmatrix},$$

$$\mathbf{Y}_i^k = \mathbf{y}_n + \Delta\mathbf{Y}_i^k, \quad \mathbf{Z}_i^k = \mathbf{z}_n + \Delta\mathbf{Z}_i^k, \quad k = 1,\ldots,m;$$

$$\mathbf{F}_i^k = \mathbf{f}(t_n + c_i h,\ \mathbf{Y}_i^k, \mathbf{Z}_i^k), \quad \mathbf{G}_i^k = \mathbf{g}(t_n + c_i h,\ \mathbf{Y}_i^k, \mathbf{Z}_i^k), \quad k = 1,\ldots,m-1;$$

$$\mathbf{Y}_i = \mathbf{Y}_i^m, \quad \mathbf{Z}_i = \mathbf{Z}_i^m,$$

где m – число итераций.

Перед началом итераций нужно задать начальные значения $\Delta\mathbf{Y}_i^0 = \mathbf{Y}_i^0 - \mathbf{y}_n$, $\Delta\mathbf{Z}_i^0 = \mathbf{Z}_i^0 - \mathbf{z}_n$, $\mathbf{F}_i^0$, $\mathbf{G}_i^0$. Для уменьшения числа итераций применяем двухшаговый прогноз, задаваемый в виде

$$\mathbf{Y}_i^0 = \sum_{j=1}^{s-1} \alpha_{ij}\bar{\mathbf{Y}}_j + \sum_{j=1}^{i-1} \beta_{ij}\mathbf{Y}_j, \quad \mathbf{Z}_i^0 = \sum_{j=1}^{s-1} \alpha_{ij}\bar{\mathbf{Z}}_j + \sum_{j=1}^{i-1} \beta_{ij}\mathbf{Z}_j,$$

где $\bar{\mathbf{Y}}_j$, $\bar{\mathbf{Z}}_j$ – стадийные значения предыдущего шага (на 1-м шаге принимаем $\bar{\mathbf{Y}}_j = \mathbf{y}_0$, $\bar{\mathbf{Z}}_j = \mathbf{z}_0$). Значения $\mathbf{F}_i^0$ и $\mathbf{G}_i^0$ обычно вычисляют как правые части в (7.1) при $\mathbf{y} = \mathbf{Y}_i^0$, $\mathbf{z} = \mathbf{Z}_i^0$. В экономичной схеме вместо этого используем такой же прогноз, как для переменных, в результате получаем

$$\mathbf{F}_i^0 = \sum_{j=1}^{s-1} \alpha_{ij}\bar{\mathbf{F}}_j + \sum_{j=1}^{i-1} \beta_{ij}\mathbf{F}_j, \quad \mathbf{G}_i^0 = \mathbf{0}. \tag{7.3}$$

После выполнения итераций следует вычислить значение $\mathbf{F}_i$, которое будет использовано на последующих стадиях. Определяем его из формулы вычисления $\mathbf{Y}_i$ в (7.2), откуда

$$\mathbf{F}_i = \frac{1}{\gamma}\left(\frac{\Delta\mathbf{Y}_i}{h} - \sum_{j=1}^{i-1} a_{ij}\mathbf{F}_j\right). \tag{7.4}$$

Использование (7.3), (7.4) позволяет сэкономить одно вычисление правой части на каждой неявной стадии, а для жестких задач (как показали эксперименты) дает более точный результат по сравнению со стандартной схемой.

Остановимся на формулах прогноза. Для методов ESDIRK прогноз 2-го порядка формируем как значение интерполяционного многочлена, построенного по уже вычисленным трем стадийным значениям. Пусть для формирования прогноза используются два стадийных значения предыдущего шага: $\bar{\mathbf{Y}}_i$ и $\bar{\mathbf{Y}}_j$ (мы задаем $i = 1$ и j – наибольшее значение, при котором $0.5 \le c_j < 1$).

Принимаем также $w = h/\overline{h}$, где $\overline{h}$ – размер предыдущего шага. Формулы прогноза на стадиях 2, 3, 4 запишутся в виде

$$\mathbf{Y}_2^0 = \alpha_{2i}\overline{\mathbf{Y}}_i + \alpha_{2j}\overline{\mathbf{Y}}_j + \beta_{21}\mathbf{Y}_1,$$

$$\alpha_{2i} = \frac{(wc_2 - c_j + 1)wc_2}{(c_i - c_j)(c_i - 1)}, \quad \alpha_{2j} = \frac{(wc_2 - c_i + 1)wc_2}{(c_j - c_i)(c_j - 1)}, \quad \beta_{21} = 1 - \alpha_{2i} - \alpha_{2j};$$

$$\mathbf{Y}_3^0 = \alpha_{3j}\overline{\mathbf{Y}}_j + \beta_{31}\mathbf{Y}_1 + \beta_{32}\mathbf{Y}_2,$$

$$\beta_{31} = \frac{c_3 - c_2}{c_2}\left(\frac{wc_3}{c_j - 1} - 1\right), \quad \beta_{32} = \frac{c_3(wc_3 - c_j + 1)}{c_2(wc_2 - c_j + 1)}, \quad \alpha_{3j} = 1 - \beta_{31} - \beta_{32};$$

$$\mathbf{Y}_4^0 = \beta_{41}\mathbf{Y}_1 + \beta_{42}\mathbf{Y}_2 + \beta_{43}\mathbf{Y}_3,$$

$$\beta_{42} = \frac{c_4(c_4 - c_3)}{c_2(c_2 - c_3)}, \quad \beta_{43} = \frac{c_4(c_4 - c_2)}{c_3(c_3 - c_2)}, \quad \beta_{41} = 1 - \beta_{42} - \beta_{43}.$$

На 5-й стадии имеет смысл формировать прогноз 3-го порядка в виде

$$\mathbf{Y}_5^0 = \beta_{51}\mathbf{Y}_1 + \beta_{52}\mathbf{Y}_2 + \beta_{53}\mathbf{Y}_3 + \beta_{54}\mathbf{Y}_4, \quad \beta_{51} = 1 - \beta_{52} - \beta_{53} - \beta_{54},$$

где коэффициенты β_{52}, β_{53}, β_{54} находим из уравнений

$$\beta_{52}c_2 + \beta_{53}c_3 + \beta_{54}c_4 = c_5,$$
$$\beta_{52}c_2^2 + \beta_{53}c_3^2 + \beta_{54}c_4^2 = c_5^2,$$
$$\beta_{53}a_{32}c_2^2 + \beta_{54}(a_{42}c_2^2 + a_{43}c_3^2) = a_{52}c_2^2 + a_{53}c_3^2 + a_{54}c_4^2$$

(под порядком прогноза понимают порядок аппроксимации стадийного значения формулой прогноза).

На 6-й стадии к условиям порядка можно добавить условие L-сходимости прогноза. Для методов ESDIRK64(1/6) и ESDIRK64(1/4) такой прогноз 3-го порядка принимаем в виде

$$\mathbf{Y}_6^0 = \beta_{61}\mathbf{Y}_1 + \cdots + \beta_{65}\mathbf{Y}_5, \quad \beta_{61} = 1 - \beta_{62} - \cdots - \beta_{65},$$

где $\beta_{62}, \ldots, \beta_{65}$ находим из системы линейных уравнений

$$\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\mathbf{c}} = 1, \quad \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\mathbf{c}}^2 = 1, \quad \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\mathbf{A}}\hat{\mathbf{c}}^2 = 1/3, \quad \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\mathbf{A}}^{-1}\hat{\mathbf{c}} = 1,$$
$$\hat{\boldsymbol{\beta}} = \begin{bmatrix}\beta_{62}\\ \beta_{63}\\ \beta_{64}\\ \beta_{65}\end{bmatrix}, \quad \hat{\mathbf{c}} = \begin{bmatrix}c_2\\ c_3\\ c_4\\ c_5\end{bmatrix}, \quad \hat{\mathbf{A}} = \begin{bmatrix}\gamma & 0 & 0 & 0\\ a_{32} & \gamma & 0 & 0\\ a_{42} & a_{43} & \gamma & 0\\ a_{52} & a_{53} & a_{54} & \gamma\end{bmatrix}.$$

Для метода ESDIRK63(1/6) принимаем $\mathbf{Y}_6^0 = (\mathbf{Y}_4 + 2\mathbf{Y}_5)/3$, что обеспечивает L-сходимость прогноза, а для ESDIRK63(1/5) принимаем $\mathbf{Y}_6^0 = \mathbf{Y}_5$.

Формулу прогноза последней стадии используем также и для формирования нормированной оценки ошибки, которую принимаем в виде

$$err = \max_i\left(\frac{|\delta y_i|}{Rtol \times \max(|y_{n,i}|, |y_{n+1,i}|) + Atol}\right), \quad \delta\mathbf{y} = K(\mathbf{y}_{n+1} - \mathbf{Y}_s^0),$$

где $Rtol$ – допустимая относительная ошибка, $Atol$ – допустимая абсолютная ошибка. Такая оценка соответствует применению вложенной формулы, записанной в виде $\hat{\mathbf{y}}_{n+1} = \mathbf{y}_{n+1} - \delta\mathbf{y}$. Принимаем следующие значения коэффициента K: ESDIRK54(0.220) – $K = 0.5$, ESDIRK64(1/6) – $K = 0.125$, ESDIRK64(1/4) – $K = 0.0712123$ (это значение эквивалентно вложенной формуле в [13], табл. 16), ESDIRK63(1/6) и ESDIRK63(1/5) – $K = 0.25$. Шаг считаем успешным, если $err \le 1$. Размер следующего шага принимаем в виде $h_{\text{new}} = fac \times err^{-1/p}h$, где $fac = 0.7$ для методов 3-го порядка и $fac = 0.75$ для методов 4-го порядка.

Сходимость итерационных схем решения задачи Коши исследовалась в [32], [33]. При определенных условиях каждая итерация с неточной матрицей Якоби повышает порядок схемы на 1 до тех пор, пока не будет достигнут порядок метода. Численные эксперименты показали, что при использовании экономичной схемы и приведенных формул прогноза для решения жестких ОДУ достаточно выполнить две итерации, первая из которых не требует вычисления правой части. На последней стадии добавляем 3-ю итерацию, которая позволяет получить оценку сходимости итераций в виде $\theta = \delta_3/\delta_2$, где δ_i – норма приращения стадийных значений на i-й итерации. Перерасчет матрицы Якоби выполняем, если $\theta > 0.1$ и при этом δ_3 не очень мало – это позволяет исключить перерасчет при незначительном изменении матрицы.

При решении ДАУ высших индексов сходимость итераций более медленная, поэтому следует предусмотреть возможность выполнения большего числа итераций. В этом случае усложняется также контроль ошибки для переменных высших индексов. Исследуем точность этих компонент на примере системы

$$0 = y - \varphi(t), \quad y' = z, \quad z' = u,$$
$$y_0 = \varphi(t_0), \quad z_0 = \varphi'(t_0), \quad u_0 = \varphi''(t_0),$$

где переменная z имеет индекс 2, а переменная u – индекс 3. Глобальные ошибки решения этих уравнений запишутся в виде

$$\begin{aligned}
\varphi_{n+1} - y_{n+1} &= \alpha_0(\varphi_n - y_n) + \delta y_{n+1},\\
\varphi'_{n+1} - z_{n+1} &= \alpha_0(\varphi'_n - z_n) + h^{-1}\alpha_1(\varphi_n - y_n) + \delta z_{n+1},\\
\varphi''_{n+1} - u_{n+1} &= \alpha_0(\varphi''_n - u_n) + h^{-1}\alpha_1(\varphi'_n - z_n) + h^{-2}\alpha_2(\varphi_n - y_n) + \delta u_{n+1},
\end{aligned} \tag{7.5}$$

где $\alpha_0 = R(\infty)$, $\alpha_1 = \lim\limits_{z\to\infty} z(R(z) - \alpha_0)$, $\alpha_2 = \lim\limits_{z\to\infty} z[z(R(z) - \alpha_0) - \alpha_1]$, δy_{n+1}, δz_{n+1}, δu_{n+1} – локальные ошибки соответствующих компонент. Для жестко точных методов $y_n = \varphi_n$ и $\delta y_{n+1} = 0$.

Аналогичные выражения получаем для вложенной формулы (заменяем α_i, δy_{n+1}, δz_{n+1}, δu_{n+1} соответствующими значениями $\hat\alpha_i$, $\delta\hat y_{n+1}$, $\delta\hat z_{n+1}$, $\delta\hat u_{n+1}$ вложенного метода). Из приведенных выражений следует, что в оценке ошибки появляются составляющие, пропорциональные h^{-1}, а также составляющие, пропорциональные только глобальной ошибке. В результате размер шага может уменьшаться до нуля, что приводит к аварийной остановке численного решения. Чтобы этого не происходило, в [34] предлагалось специальным образом масштабировать либо игнорировать оценки ошибки для переменных высших индексов.

Заметим, что если $\hat\alpha_0 = \alpha_0$, $\hat\alpha_1 = \alpha_1$ и при этом $\alpha_0 = \alpha_1 = 0$, то глобальные ошибки будут равны локальным ошибкам. В этом случае можно применять обычный контроль ошибки для всех переменных, не опасаясь аварийной остановки. Вложенная пара такого типа была построена на основе метода ESDIRK63(1/6), в который добавлена еще одна (6-я) стадия специально для оценивания ошибки. Эта же стадия используется как прогноз $\mathbf{Y}_7^0 = \mathbf{Y}_6$ для заключительной стадии. Полученный метод имеет таблицу Бутчера

0	0						
1/3	1/6	1/6					
2/3	1/6	1/3	1/6				
1	1/3	0	1/2	1/6			
1	7/16	0	3/16	5/24	1/6		
1	7/48	17/48	17/48	1/80	−1/30	1/6	
1	1/8	3/8	3/8	1/360	−2/45	0	1/6
b_i	1/8	3/8	3/8	1/360	−2/45	0	1/6
$\hat b_i$	7/48	17/48	17/48	1/80	−1/30	1/6	0

.

Здесь основной шестистадийный метод (5.1) и вложенная формула образуют семистадийный метод, который обозначим через ESDIRK73(1/6).

Таблица 8. Результаты решения задачи VDPOL

Метод	*Tol*	*scd*	*mescd*	*Nf*	*NJ*
ESDIRK54(0.220)	10^{-3}	3.23	3.56	1256	67
	10^{-4}	4.05	4.38	1676	70
ESDIRK64(1/6)	10^{-3}	3.11	3.43	1213	55
	10^{-4}	3.89	4.21	1477	70
ESDIRK64(1/4)	10^{-3}	2.26	2.59	1123	41
	10^{-4}	3.68	3.89	1465	44
ESDIRK73(1/6)	10^{-3}	3.84	4.16	1531	39
	10^{-4}	3.92	4.24	2611	55
ESDIRK73(1/5)	10^{-3}	2.63	2.96	1669	36
	10^{-4}	4.45	4.78	2683	53
RADAU5	10^{-4}	4.96	5.28	2253	162

Аналогичным образом на основе метода ESDIRK63(1/5) построен метод ESDIRK73(1/5), 6-я стадия которого является вложенной формулой для оценивания ошибки и имеет коэффициенты

$$\hat{b}_i = (2065/11008, 1019/1920, 869/3840, -5293/103200, -7/75, 1/5)$$

(но в этом случае $\hat{\alpha}_1 = \alpha_1 \neq 0$). Численные эксперименты показали, что методы (вложенные пары) ESDIRK73(1/6) и ESDIRK73(1/5) более эффективны, чем соответствующие методы (пары) ESDIRK63(1/6) и ESDIRK63(1/5), поэтому далее приводим результаты методов ESDIRK73(1/6) и ESDIRK73(1/5).

8. РЕЗУЛЬТАТЫ РАСЧЕТОВ С ПЕРЕМЕННЫМ РАЗМЕРОМ ШАГА

При реализации всех методов используем одинаковое число итераций и одинаковые условия обновления матрицы Якоби. При решении ОДУ выполняем две итерации экономичной схемы на предварительных стадиях и три итерации на последней стадии. Для решения ДАУ индекса 3 такого числа итераций оказалось недостаточно, в этом случае выполняем три итерации на каждой стадии и вычисляем матрицу Якоби на каждом шаге. Вложенные формулы (6-е стадии) в методах ESDIRK73(1/6) и ESDIRK73(1/6) реализуем, делая одну итерацию (без вычисления правой части) при решении ОДУ и две итерации при решении ДАУ.

Для решения с переменным шагом были выбраны четыре жестких теста: VDPOL, HIRES, PLATE и BEAM, подробные описания которых приведены в [2], [35]. Интервалы интегрирования и размеры начального шага берем, как в [35], [36]. Для всех задач принимаем $Rtol = Atol = Tol$, где $Rtol$ – относительная допустимая ошибка, $Atol$ – абсолютная допустимая ошибка. Как и в [35], [36] точность решения оцениваем значениями *scd* (6.1) и

$$mescd = -\lg\left(\max_i\left(\frac{|y_i - \tilde{y}_i|}{Atol/Rtol + |y_i|}\right)\right),$$

где y_i – точное, а $\tilde{y}_i$ – численное решение по i-й компоненте в конце интервала интегрирования. При решении задачи BEAM значение *scd* вычисляем для первых 40 (из 80) переменных, поскольку “they refer to the physically important quantities” [35]. Каждую задачу решаем при двух значениях Tol: 10^{-3} и 10^{-4}, но если при $Tol = 10^{-3}$ точность решения низкая ($scd < 1$), то приводим результаты при $Tol = 10^{-4}$, 10^{-5}. Вычислительные затраты оцениванием числом вычислений правой части Nf и числом вычислений матрицы Якоби NJ. Результаты расчетов приведены в табл. 8–11. Приводим также некоторые результаты решателя RADAU5, взятые из [36]. Результаты решения этих задач другими известными методами приведены в [35], [36].

Приведенные результаты подтверждают высокую эффективность методов ESDIRK при решении жестких задач с умеренной точностью. Задача BEAM имеет чисто мнимый спектр матрицы Якоби. По этой причине многие известные методы требуют большого объема вычислений

Таблица 9. Результаты решения задачи HIRES

Метод	*Tol*	*scd*	*mescd*	*Nf*	*NJ*
ESDIRK54(0.220)	10^{-3}	1.95	4.16	136	12
	10^{-4}	3.07	5.42	176	12
ESDIRK64(1/6)	10^{-4}	1.62	3.83	175	10
	10^{-5}	2.28	4.49	235	12
ESDIRK64(1/4)	10^{-4}	0.92	3.13	163	9
	10^{-5}	2.49	5.03	229	11
ESDIRK73(1/6)	10^{-4}	1.91	4.11	295	10
	10^{-5}	1.96	4.17	415	9
ESDIRK73(1/5)	10^{-4}	1.79	4.00	217	15
	10^{-5}	3.14	5.35	421	18
RADAU5	10^{-5}	1.35	3.55	381	23

Таблица 10. Результаты решения задачи PLATE

Метод	*Tol*	*scd*	*mescd*	*Nf*	*NJ*
ESDIRK54(0.220)	10^{-3}	2.83	4.53	101	1
	10^{-4}	3.72	5.61	216	1
ESDIRK64(1/6)	10^{-3}	1.78	3.92	97	1
	10^{-4}	3.18	5.13	211	1
ESDIRK64(1/4)	10^{-4}	1.87	3.81	121	1
	10^{-5}	2.39	4.40	253	1
ESDIRK73(1/6)	10^{-3}	1.18	3.32	79	1
	10^{-4}	2.96	4.90	175	1
ESDIRK73(1/5)	10^{-4}	2.46	4.60	133	1
	10^{-5}	4.38	6.49	307	1
RADAU5	10^{-4}	1.62	3.77	74	4

Таблица 11. Результаты решения задачи BEAM

Метод	*Tol*	*scd*	*mescd*	*Nf*	*NJ*
ESDIRK54(0.220)	10^{-3}	2.25	2.54	296	5
	10^{-4}	3.53	3.10	626	5
ESDIRK64(1/6)	10^{-3}	1.69	2.51	151	4
	10^{-4}	2.96	2.50	321	4
ESDIRK64(1/4)	10^{-3}	1.45	2.27	116	5
	10^{-4}	2.04	2.42	211	6
ESDIRK73(1/6)	10^{-3}	1.86	2.35	259	4
	10^{-4}	3.00	2.99	841	3
ESDIRK73(1/5)	10^{-3}	1.84	2.59	193	7
	10^{-4}	2.65	2.58	469	4
RADAU5	10^{-4}	2.49	3.57	406	43

Таблица 12. Результаты решения задачи (6.3)

Метод	*Tol*	e_y	e_z	e_u	*Nf*	*NJ*
ESDIRK54(0.220)	10^{-5}	3.55×10^{-4}	3.67×10^{-4}	1.90×10^{-2}	801	99
	10^{-6}	7.55×10^{-5}	7.64×10^{-5}	8.79×10^{-3}	1673	208
ESDIRK64(1/6)	10^{-3}	2.76×10^{-4}	2.54×10^{-4}	2.02×10^{-2}	711	71
	10^{-4}	3.89×10^{-6}	8.48×10^{-6}	3.78×10^{-3}	2271	226
ESDIRK64(1/4)	10^{-4}	4.58×10^{-4}	5.95×10^{-4}	1.91×10^{-2}	1351	133
	10^{-5}	5.14×10^{-5}	6.15×10^{-5}	4.08×10^{-3}	4361	436
ESDIRK73(1/6)	10^{-3}	1.13×10^{-4}	1.10×10^{-4}	6.67×10^{-3}	749	68
	10^{-4}	4.70×10^{-6}	3.70×10^{-6}	1.34×10^{-3}	1970	179
ESDIRK73(1/5)	10^{-3}	5.25×10^{-4}	7.16×10^{-4}	2.94×10^{-2}	551	50
	10^{-4}	9.16×10^{-6}	2.11×10^{-5}	3.59×10^{-3}	1706	154

при решении этой задачи даже с невысокой точностью (см. [35], [36]). Рассмотренные методы ESDIRK (в том числе и $L(\alpha)$-устойчивые) успешно решают эту задачу с малыми вычислительными затратами. Этот факт показывает, что строгое соблюдение условия L-устойчивости не является необходимым при построении эффективных методов, достаточно, чтобы метод был $L(\alpha)$-устойчивым при α, близком к 90°.

В табл. 12 приведены результаты решения системы ДАУ индекса 3 (6.3). Задаем $Rtol = Tol$, $Atol = 10^{-4}Tol$, $h_0 = Tol$ и вычисляем ошибки по соответствующим компонентам. Из всех методов только ESDIRK73(1/6) обеспечил устойчивое управление размером шага при контроле ошибки по всем компонентам. Поэтому в методах ESDIRK64(1/6), ESDIRK64(1/4) и ESDIRK73(1/5) контроль ошибки выполняем по y- и z-компонентам, а в методе ESDIRK54(0.220) – только по y-компоненте (в отличие от других методов, вложенная формула в ESDIRK54(0.220) имеет $\hat{\alpha}_0 \neq 0$). Чтобы получить приемлемые результаты, пришлось уменьшить значение *Tol* для метода ESDIRK54(0.220) на 2 порядка и для ESDIRK64(1/4) на порядок. Три других метода оказались более эффективными, при этом для решения задач индекса 3 удобнее всего использовать метод ESDIRK73(1/6), позволяющий осуществлять контроль ошибки по всем переменным.

СПИСОК ЛИТЕРАТУРЫ

1. *Alexander R.* Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s // SIAM J. Numer. Anal. 1977. V. 14. № 6. P. 1006–1021.
2. *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. М.: Мир, 1999.
3. *Cameron F., Palmroth M., Piche R.* Quasi stage order conditions for SDIRK methods // Appl. Numer. Math. 2002. V. 42. № 1–3. P. 61–75.
4. *Williams R., Burrage K., Cameron I., Kerr M.* A four-stage index 2 diagonally implicit Runge–Kutta method // Appl. Numer. Math. 2002. V. 40. № 3. P. 415–432.
5. *Alexander R.* Design and implementation of DIRK integrators for stiff systems // Appl. Numer. Math. 2003. V. 46. № 1. P. 1–17.
6. *Kværnø A.* Singly diagonally implicit Runge–Kutta methods with an explicit first stage // BIT. 2004. V. 44. № 3. P. 489–502.
7. *Скворцов Л.М.* Диагонально неявные FSAL-методы Рунге–Кутты для жестких и дифференциально-алгебраических систем // Матем. моделирование. 2002. Т. 14. № 2. С. 3–17.
8. *Скворцов Л.М.* Точность методов Рунге–Кутты при решении жестких задач // Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 9. С. 1374–1384.
9. *Скворцов Л.М.* Диагонально неявные методы Рунге–Кутты для жестких задач // Ж. вычисл. матем. и матем. физ. 2006. Т. 46. № 12. С. 2209–2222.
10. *Скворцов Л.М.* Диагонально-неявные методы Рунге–Кутты для дифференциально-алгебраических уравнений индексов 2 и 3 // Ж. вычисл. матем. и матем. физ. 2010. Т. 50. № 6. С. 1047–1059.
11. *Скворцов Л.М.* Модельные уравнения для исследования точности методов Рунге–Кутты // Матем. моделирование. 2010. Т. 22. № 5. С. 146–160.

12. *Rang J.* An analysis of the Prothero–Robinson example for constructing new adaptive ESDIRK methods of order 3 and 4 // Appl. Numer. Math. 2015. V. 94. P. 75–87.

13. *Kennedy C.A., Carpenter M.H.* Diagonally implicit Runge–Kutta methods for ordinary differential equations // A Rev. NASA report NASA/TM-2016-219173. 2016.

14. *Скворцов Л.М.* Численное решение обыкновенных дифференциальных и дифференциально-алгебраических уравнений. М: ДМК Пресс, 2018.

15. *Boom P.D., Zingg D.W.* Optimization of high-order diagonally-implicit Runge–Kutta methods // J. Comput. Phys. 2018. V. 371. P. 168–191.

16. *Kennedy C.A., Carpenter M.H.* Diagonally implicit Runge–Kutta methods for stiff ODEs // Appl. Numer. Math. 2019. V. 146. P. 221–244.

17. *Hosea M.E., Shampine L.F.* Analysis and implementation of TR-BDF2 // Appl. Numer. Math. 1996. V. 20. № 1–2. P. 21–37.

18. *Bonaventura L., Marmol M.G.* The TR-BDF method for second order problems in structural mechanics // Comput. Math. Appl. 2021. V. 92. P. 13–26.

19. *Брагин М.Д., Рогов Б.В.* Метод итерируемой приближенной факторизации операторов высокоточной бикомпактной схемы для систем многомерных неоднородных квазилинейных уравнений гиперболического типа // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 3. С. 313–325.

20. *Рогов Б.В., Чикиткин А.В.* О сходимости и точности метода итерируемой приближенной факторизации операторов многомерных высокоточных бикомпактных схем // Матем. моделирование. 2019. Т. 31. № 12. С. 119–144.

21. *Kennedy C.A., Carpenter M.H.* Additive Runge–Kutta schemes for convection–diffusion–reaction equations // Appl. Numer. Math. 2003. V. 44. P. 139–181.

22. *Kennedy C.A., Carpenter M.H.* Higher-order additive Runge–Kutta schemes for ordinary differential equations // Appl. Numer. Math. 2019. V. 136. P. 183–205.

23. *Карташов Б.А., Шабаев Е.А., Козлов О.С., Щекатуров А.М.* Среда динамического моделирования технических систем SimInTech. М: ДМК Пресс, 2017.

24. *Prothero A., Robinson A.* On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations // Math. Comput. 1974. V. 28. № 125. P. 145–162.

25. *Кочетков К.А., Ширков П.Д. L*-затухающие ROW-методы третьего порядка точности // Ж. вычисл. матем. и матем. физ. 1997. Т. 37. № 6. С. 699–710.

26. *Хайрер Э., Нёрсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990.

27. *Скворцов Л.М.* Как избежать снижения точности и порядка методов Рунге–Кутты при решении жестких задач // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 7. С. 1126–1141.

28. *Jay L.* Convergence of a class of Runge–Kutta methods for differential-algebraic systems of index 2 // BIT. 1993. V. 33. № 1. P. 137–150.

29. *Jay L.* Convergence of Runge–Kutta methods for differential-algebraic systems of index 3 // Appl. Numer. Math. 1995. V. 17. № 2. P. 97–118.

30. *Скворцов Л.М.* Экономичная схема реализации неявных методов Рунге–Кутты // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 11. С. 2008–2018.

31. *Скворцов Л.М., Козлов О.С.* Эффективная реализация диагонально-неявных методов Рунге–Кутты // Матем. моделирование. 2014. Т. 26. № 1. С. 96–108.

32. *Куликов Г.Ю.* Теоремы сходимости для итерационных методов Рунге–Кутты с постоянным шагом интегрирования // Ж. вычисл. матем. и матем. физ. 1996. Т. 36. № 8. С. 73–89.

33. *Jackson K.R., Kværnø A., Nørsett S.P.* An analysis of the order of Runge–Kutta methods that use an iterative scheme to compute their internal stage values // BIT. 1996. V. 36. № 4. P. 713–765.

34. *Hairer E., Lubich Ch., Roche M.* The numerical solution of differential-algebraic systems by Runge–Kutta methods. Berlin: Springer-Verlag, 1989.

35. *Mazzia F., Magherini C.* Test set for initial value problem solvers. Release 2.4. 2008. URL: http://pitagora.dm.uniba.it/~testset/report/testset.pdf.

36. The codes BiM and BiMD home page. 2014. URL: http://web.math.unifi.it/users/brugnano/BiM/.

EDN: TPPQBC
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 809–822

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 519.63

РЕШЕНИЕ ВНЕШНЕЙ КРАЕВОЙ ЗАДАЧИ ДЛЯ УРАВНЕНИЯ ГЕЛЬМГОЛЬЦА ДЕКОМПОЗИЦИЕЙ ОБЛАСТИ С ПЕРЕСЕЧЕНИЕМ[1)]

© 2022 г. А. В. Петухов[1], А. О. Савченко[1,*]

[1]630090 Новосибирск, пр-т Акад. Лаврентьева, 6, Институт вычислительной математики и математической геофизики СО РАН, Россия

**e-mail: savch@ommfao1.sscc.ru*

Поступила в редакцию 05.03.2020 г.
Переработанный вариант 20.07.2021 г.
Принята к публикации 14.01.2022 г.

Предложен и исследован метод решения внешней трехмерной краевой задачи для уравнения Гельмгольца, основанный на декомпозиции внешней области с пересечением. Предлагаемый подход основан на применении альтернирующего метода Шварца с последовательным решением внутренней и внешней краевой задачи в подобластях с пересечением, на смежных границах которых ставятся итерируемые интерфейсные условия. Найдены достаточные условия сходимости метода в случае отрицательного коэффициента в уравнении Гельмгольца. Проведено исследование сходимости частного случая проблемы, позволяющее сделать вывод о применимости предложенного подхода для решения задачи с произвольным волновым числом. Предложенный метод апробирован численным решением задач, с применением метода конечных объемов для решения внутренних краевых задач и формулы Грина для решения внешних краевых задач. Скорость сходимости итераций и достигаемая точность вычислений иллюстрируются на серии вычислительных экспериментов. Проведен анализ выбора параметров декомпозиции, обеспечивающих сходимость метода. Библ. 14. Фиг. 1. Табл. 2.

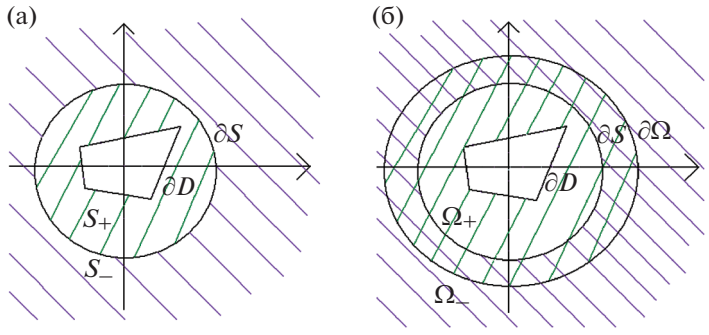
Ключевые слова: уравнение Гельмгольца, внешняя краевая задача, декомпозиция области, формула Грина.

DOI: 10.31857/S0044466922050118

ВВЕДЕНИЕ

Многие приложения в физике связаны с решением трехмерного уравнения Гельмгольца в неограниченной области. Применение стандартных технологий решения этой проблемы таких, как методы конечных элементов или конечных объемов, сопряжены со сложностями, связанными с неограниченностью расчетной области и необходимостью удовлетворять условию излучения на бесконечности. Существует большое количество численных методов решения внешних краевых задач в неограниченной области [1]. Одним из наиболее известных является представление искомого решения в виде комбинации потенциалов простого и двойного слоя, в которых плотности потенциалов находятся решением интегральных уравнений Фредгольма на границе области [2]–[4]. Другой подход для решения внешних краевых задач состоит в комбинированном применении метода конечных элементов и интегрального метода путем декомпозиции области на составляющие, допускающие применение в них этих методов. В 80-х годах прошлого века K. Feng и D. Yu предложили разбить всю область решения внешней краевой задачи на две подобласти: без пересечения или с пересечением (см. [5]). При декомпозиции без пересечения (см. фиг. 1а) вся область вне поверхности тела ∂D разбивается на ограниченную область S_+ между сферой ∂S поверхностью ∂D, и на область S_-, ограниченную только снизу сферой ∂S. При декомпозиции с пересечением (фиг. 1б) ограниченная область Ω_+ находится между поверхностью $\partial\Omega$ и поверхностью ∂D, а неограниченная область Ω_-, так же как и в декомпозиции без пересечения, ограничена только снизу сферой ∂S.

[1)]Работа выполнена при финансовой подержке в рамках государственного задания ИВМиМГ СО РАН (проект 0251-2021-0001).



Фиг. 1. Декомпозиция области без пересечения (а) и с пересечением (б). Области S_+ и Ω_+ соответствуют фону ///; области S_- и Ω_- соответствуют фону \\\ .

Для решения исходной задачи был предложен альтернирующий метод Шварца, в котором внутренняя задача решалась методом конечных элементов, а условия согласования ставились на внешней поверхности. В [5] и других работах [6]–[8] для численного решения задачи было отдано предпочтение декомпозиции области без пересечения (фиг. 1а). В предложенном методе нет необходимости искать решение внешней задачи в области вне сферической поверхности. Достаточно найти только значения нормальной производной на этой поверхности, если на ней заданы значения функции. Производные находятся вычислением значений оператора Пуанкаре–Стеклова, которые могут быть определены в виде ряда [6].

Необходимо отметить, что при решении внешней краевой задачи для искусственной поверхности ∂S необходимо вычислять интегралы с сингулярными ядрами, что вызывает необходимость использовать для корректного численного вычисления таких интегралов квадратуры специального вида. В настоящей работе для решения задачи предлагается использовать декомпозицию области с пересечением (фиг. 1б). В этом случае при решении внешней краевой задачи все вычисляемые интегралы не имеют сингулярностей, что значительно упрощает использование квадратур для их численного вычисления.

Предлагаемый в настоящей работе подход к решению внешней краевой задачи Дирихле для уравнения Гельмгольца состоит в нахождении приближенных значений искомой функции и ее нормальной производной на поверхности вспомогательной сферы ∂S, заключающей в себя исходную внутреннюю границу ∂D, поскольку тогда можно найти значения функции в другой точке области, используя формулу Грина. При этом внешняя граница $\partial\Omega$ ограниченной подобласти Ω_+ не обязательно является сферой и выбирается из условия экономии вычислительных ресурсов и получения наиболее точного решения на сфере ∂S. Для решения задачи применяется альтернирующий метод Шварца, в котором последовательно пересчитываются граничные условия на сфере ∂S и внешней вспомогательной поверхности $\partial\Omega$. На каждой итерации метода решаются внутренняя и внешняя краевые задачи. Решение внутренней краевой задачи производится в области, ограниченной внешней вспомогательной поверхностью $\partial\Omega$ и исходной границей ∂D (область Ω_+ на фиг. 1б), и находятся приближенные значения искомой функции и ее нормальной производной на поверхности сферы ∂S. Для решения внешней задачи в области Ω_-, ограниченной снизу сферой ∂S, в отличие от рассмотренных выше методов, предлагается использовать формулу Грина. Такой подход имеет следующие преимущества. Во-первых, формула Грина является более универсальной, и позволяет находить решения уравнения Гельмгольца не только при положительном коэффициенте в уравнении, но и при произвольном, в том числе комплексном коэффициенте. Во-вторых, эта формула не представлена в виде ряда, как в рассмотренных выше методах, и поэтому нет необходимости аппроксимировать этот ряд частичной суммой, внося тем самым дополнительную погрешность в численный метод. Для применения формулы Грина необходимо знать значения функции и ее нормальной производной на сфере ∂S, а применение метода конечных объемов при решении задачи в области Ω_+ позволяет получить и те, и другие значения на сфере ∂S. Отметим, что в предлагаемом итерационном алгоритме искомыми значениями при решении внутренней краевой задачи будут значения функции и ее нормальной

производной только на сфере ∂S, а при решении внешней краевой задачи – только значения функции на поверхности $\partial\Omega$. В статье получены достаточные условия сходимости предложенного итерационного метода и оценка для производной решения уравнения Гельмгольца при отрицательном коэффициенте в этом уравнении.

Проведено исследование сходимости обобщения предложенного метода, состоящего во введении релаксационного процесса при определении итерируемых приближенных значений функции и ее нормальной производной на поверхности вспомогательной сферы. Исследована сходимость метода для частного случая рассмотренной проблемы, что позволило сделать выводы об ограничении применимости предложенного подхода для решения общей проблемы с произвольным волновым числом. Предложенный метод иллюстрируется результатами численных экспериментов для решения внешней краевой задачи с отрицательным коэффициентом. Проведен анализ и получены рекомендации для выбора надлежащих параметров метода, обеспечивающих его сходимость.

1. ПОСТАНОВКА ЗАДАЧИ И МЕТОД РЕШЕНИЯ

Рассмотрим открытую область D в пространстве $\mathbb{R}^3$, ограниченную поверхностью ∂D. Внешняя краевая задача для уравнения Гельмгольца состоит в нахождении функции $u\in C^1(\mathbb{R}^3\setminus D)\cap C^2(\mathbb{R}^3\setminus\bar{D})$, удовлетворяющей уравнению

$$\Delta u(\mathbf{r}) + k^2 u(\mathbf{r}) = 0, \quad \mathbf{r}\in\mathbb{R}^3\setminus\bar{D}, \tag{1}$$

краевому условию

$$u(\mathbf{r}) = f(\mathbf{r}), \quad \mathbf{r}\in\partial D, \tag{2}$$

и условию излучения на бесконечности

$$\lim_{r\to\infty} r\left(\frac{\partial u}{\partial r} - iku\right) = 0. \tag{3}$$

Задача (1)–(3) предполагается решенной, если удастся найти значения искомой функции u_S и ее нормальной производной u_n на сфере ∂S, являющейся границей шара $\bar{S} = S\cup\partial S$, $\bar{D}\subset S$. Действительно, в этом случае можно найти значения функции в произвольной точке пространства $\mathbf{r}\in\mathbb{R}^3\setminus\bar{D}$ по формуле Грина:

$$u(\mathbf{r}) = \frac{1}{4\pi}\int_{\partial S}\left[u_n\frac{e^{ikR}}{R} - u_S\frac{\partial}{\partial n}\left(\frac{e^{ikR}}{R}\right)\right]d s, \tag{4}$$

где r_0 – радиус сферы ∂S, $\mathbf{r}_0\in\partial S$, $R = |\mathbf{r}-\mathbf{r}_0| = \sqrt{r^2 - 2r_0 r\cos\gamma + r_0^2}$, γ – угол между векторами $\mathbf{r}$ и $\mathbf{r}_0$, $\cos\gamma = \cos\theta\cos\theta_0 + \sin\theta\sin\theta_0\cos(\varphi-\varphi_0)$, $\dfrac{\partial}{\partial n}$ – производная по нормали к поверхности в точке $\mathbf{r}_0$.

Формула (4) является основой для предлагаемого итерационного метода решения задачи (1)–(3), который заключается в последовательном решении вспомогательных краевых задач в ограниченной и неограниченной области.

Введем дополнительную ограниченную область Ω_0 с границей $\partial\Omega_0$ такую, что $\bar{S}\subset\Omega_0$. Будем решать методом конечных объемов внутреннюю краевую задачу в области $\Omega_0\setminus\bar{D}$ при заданных граничных условиях на ∂D и пересчитываемых на каждой итерации краевых условиях на $\partial\Omega_0$. На первой итерации метода задаем, например, нулевые граничные условия на поверхности $\partial\Omega_0$ и решаем внутреннюю краевую задачу в области $\Omega_0\setminus\bar{D}$. При этом нас будут интересовать полученные значения для функции и ее нормальной производной только на поверхности ∂S. По этим значениям определим новые граничные условия на внешней границе расчетной области $\partial\Omega_0$ по формуле (4), и проведем аналогично вторую и последующие итерации. Отметим, что на разных шагах этого процесса расчетные области для внутренних краевых задач можно выбирать дина-

мическим образом, т.е. на j-й итерации вместо Ω_0 определить область Ω_{j+1} с границей $\partial\Omega_{j+1}$, $j = 0,1,\ldots,J$. Количество итераций J определяется по условию сходимости метода с заданной точностью $\varepsilon \ll 1$. Выбор области Ω_j, заключающей в себя шар $\overline{S}$, на второй и всех последующих итерациях производится из соображений экономии вычислительных ресурсов и возможности получения наиболее точного решения на поверхности шара. При этом условие включения $S \subset \Omega_j$ остается в силе. На всех последующих итерациях решается внутренняя краевая задача в области $\Omega_j \backslash \overline{D}$ и по формуле (4) восстанавливаются новые граничные условия Дирихле на поверхности $\partial\Omega_j$.

2. СХОДИМОСТЬ ИТЕРАЦИОННОГО МЕТОДА

Исследуем сходимость предложенного метода для решения уравнения

$$\Delta u(\mathbf{r}) - k^2 u(\mathbf{r}) = 0 \tag{5}$$

с граничными условиями (2), (3), где k – вещественное число.

2.1. Итерационные формулы для искомого приближенного решения и его погрешности

Запишем формулу (4) в виде

$$u(\mathbf{r}) = \int\limits_{\partial S}\left[G(\mathbf{r},\mathbf{r}_0)\frac{\partial u}{\partial n}(\mathbf{r}_0) - u(\mathbf{r}_0)\frac{\partial}{\partial n}(G(\mathbf{r},\mathbf{r}_0))\right] d\,s, \quad \mathbf{r}_0 \in \partial S,$$

где $G(\mathbf{r},\mathbf{r}_0) = \dfrac{1}{4\pi}\dfrac{e^{-kR}}{R}$. Тогда итерационный процесс определяется следующим образом:

$$\begin{gathered}\Delta u^{j+1}(\mathbf{r}) - k^2 u^{j+1}(\mathbf{r}) = 0, \quad \mathbf{r} \in \Omega_j \backslash \overline{D},\\ u^{j+1}(\mathbf{r}) = \Phi^j(\mathbf{r}), \quad \mathbf{r} \in \partial\Omega_j,\\ u^{j+1}(\mathbf{r}) = f(\mathbf{r}), \quad \mathbf{r} \in \partial D, \quad j = 0,1,\ldots,\end{gathered} \tag{6}$$

где

$$\Phi^0(\mathbf{r}) = 0, \quad \Phi^j(\mathbf{r}) = \int\limits_{\partial S}\left[G(\mathbf{r},\mathbf{r}_0)\frac{\partial u^j}{\partial n}(\mathbf{r}_0) - u^j(\mathbf{r}_0)\frac{\partial}{\partial n}(G(\mathbf{r},\mathbf{r}_0))\right] ds, \tag{7}$$

$$\mathbf{r} \in \partial\Omega_j, \quad \mathbf{r}_0 \in \partial S, \quad k = 1,2,\ldots.$$

Определим погрешность метода как

$$\omega^j(\mathbf{r}) = u^j(\mathbf{r}) - u(\mathbf{r}), \quad \mathbf{r} \in \overline{\Omega}_{j-1} \backslash \overline{D}, \tag{8}$$

$$\omega_n^j(\mathbf{r}_0) = \frac{\partial u^j}{\partial n}(\mathbf{r}_0) - \frac{\partial u}{\partial n}(\mathbf{r}_0), \quad \mathbf{r}_0 \in \partial S.$$

Тогда получим

$$\omega^{j+1}(\mathbf{r}) = \int\limits_{\partial S}\left[G(\mathbf{r},\mathbf{r}_0)\omega_n^j(\mathbf{r}_0) - \omega^j(\mathbf{r}_0)\frac{\partial}{\partial n}(G(\mathbf{r},\mathbf{r}_0))\right] d\,s \equiv \varphi^j(\mathbf{r}), \quad \mathbf{r} \in \partial\Omega_j, \tag{9}$$

и погрешность будет удовлетворять уравнениям

$$\begin{gathered}\Delta\omega^{j+1}(\mathbf{r}) - k^2\omega^{j+1}(\mathbf{r}) = 0, \quad \mathbf{r} \in \Omega_j \backslash \overline{D},\\ \omega^{j+1}(\mathbf{r}) = \varphi^j(\mathbf{r}), \quad \mathbf{r} \in \partial\Omega_j,\\ \omega^{j+1}(\mathbf{r}) = 0, \quad \mathbf{r} \in \partial D.\end{gathered} \tag{10}$$

2.2. Оценка производной по нормали

Для определения условия сходимости метода необходимо найти оценку для производной по нормали от решения уравнения Гельмгольца на поверхности ∂S. Обозначим через ρ_0 расстояние между сферой и границей расчетной области, $\rho_0 = \min\limits_{\mathbf{r},\,\mathbf{r}_0} |\mathbf{r} - \mathbf{r}_0|$, $\mathbf{r} \in \partial\,\Omega_j \cup \partial\, D$, $\mathbf{r}_0 \in \partial\, S$; а через u_l производную от функции u по направлению l, $u_l = \dfrac{\partial u}{\partial l}$. Если функция u удовлетворяет уравнению Гельмгольца, то и функция u_l будет удовлетворять такому же уравнению, и для любых δ, $\delta \le \rho_0$ справедлива теорема о среднем для решения уравнения Гельмгольца (см. [9])

$$\int\limits_{\partial S_\delta} u_l \, ds_\delta = u_l(\mathbf{r}_0)\frac{4\pi\delta}{k}\sin(k\delta), \tag{11}$$

где $\partial\, S_\delta$ – сфера радиуса δ с центром в точке $\mathbf{r}_0$. Проинтегрируем уравнение (11) от 0 до ρ по переменной δ, $0 \le \delta \le \rho \le \rho_0$. Тогда получим

$$I_\rho = \int\limits_{S_\rho} u_l \, dv_\rho = 4\pi u_l(\mathbf{r}_0)\,\alpha(k,\rho), \tag{12}$$

где S_ρ – шар радиуса ρ, и

$$\alpha(k,\rho) = \frac{\sin(k\rho)}{k^3} - \frac{\rho\cos(k\rho)}{k^2}. \tag{13}$$

Интегрируя равенство (12) по частям, получим

$$I_\rho = \int\limits_{\partial S_\rho} u\frac{\partial l}{\partial n_S}\, d\, s_\rho,$$

где n_S – нормаль в текущей точке на сфере $\partial\, S_\rho$.

Произведем оценку интеграла I_ρ, что позволит не только получить искомую оценку производной в точке $\mathbf{r}_0$, но и уточнить оценку для производной гармонической функции, полученной в [9]. Выберем ось Z в декартовой системе координат, совпадающей с направлением l. Тогда получим

$$I_\rho = \rho^2 \int\limits_0^{2\pi}\int\limits_0^{\pi} u(\theta,\varphi)\sin\theta\cos\theta d\theta\, d\varphi.$$

Для оценки $|I_\rho|$ разобьем промежуток интегрирования по переменной θ на интервалы $[0,\, \pi/2]$ и $[\pi/2,\, \pi]$, на которых $\cos\theta$ – знакопостоянная функция. Тогда имеем

$$|I_\rho| \le 2\pi\rho^2 \max_{\mathbf{r}\in\partial S_\rho}|u(\mathbf{r})|\left\{\int\limits_0^{\pi/2}\sin\theta\cos\theta d\theta - \int\limits_{\pi/2}^{\pi}\sin\theta\cos\theta d\theta\right\} = 2\pi\rho^2 \max_{\mathbf{r}\in\partial S_\rho}|u(\mathbf{r})|.$$

Отсюда, с учетом (12), следует оценка

$$|u_l(\mathbf{r}_0)| \le \frac{\rho^2}{2\alpha(k,\rho)}\max_{\mathbf{r}\in\partial S_\rho}|u(\mathbf{r})|, \tag{14}$$

где $\alpha(k,\rho)$ определено формулой (13).

Заметим, что правая часть неравенства (14) является функцией, зависящей от произвольного радиуса шара ρ, $\rho \le \rho_0$, и поэтому естественно выбрать этот радиус таким образом, чтобы правая часть в (14) принимала наименьшее значение. При относительно больших значениях k это можно сделать достаточно просто. Введем обозначение: $x = k\rho$. Тогда, с учетом (13), получим

$$\frac{\rho^2}{\alpha(k,\rho)} = \frac{k\ x^2}{\sin x - x\cos x}. \tag{15}$$

Функция в правой части формулы (15) имеет минимум при $x_* \approx 2.08$ и принимает в точке x_* значение $\rho_*^2/\alpha(k,\rho_*) \approx 2.3k$, где $\rho_* = x_*/k$. Тогда неравенство (14) примет вид

$$|u_l(\mathbf{r}_0)| \le 1.15k \max_{\mathbf{r}\in\partial S_{\rho_*}} |u(\mathbf{r})|. \tag{16}$$

Если $\rho_* > \rho_0$, то в силу убывания функции в правой части формулы (15) на промежутке $(0, x_*]$ выбираем $\rho = \rho_0$, и неравенство (14) примет вид

$$|u_l(\mathbf{r}_0)| \le \frac{\rho_0^2}{2\alpha(k,\rho_0)} \max_{\mathbf{r}\in\partial S_{\rho_0}} |u(\mathbf{r})| = \frac{k\,x_0^2}{2(\sin x_0 - x_0\cos x_0)} \max_{\mathbf{r}\in\partial S_{\rho_0}} |u(\mathbf{r})|, \tag{17}$$

где $x_0 = k\rho_0$.

Нетрудно показать, что $\lim\limits_{k\to 0} \dfrac{\rho_0^2}{\alpha(k,\rho_0)} = \dfrac{3}{\rho_0}$. Отсюда следует, что при $k = 0$ оценка производной для уравнения Лапласа в (17) в 2 раза лучше оценки $|u_x(\mathbf{r}_0)| \le \dfrac{3M}{\rho_0}$, приведенной в [9], где M – максимальное по модулю значение функции $u(\mathbf{r})$ в заданной области.

2.3. Сходимость итерационного метода

Найдем явный вид производной по нормали от фундаментального решения, а также интегралы от фундаментального решения и от его нормальной производной по поверхности сферы.

$$\frac{\partial G}{\partial n} = -\frac{\partial G}{\partial R}\frac{\partial R}{\partial n} = e^{-kR}\frac{1+kR}{R^2}\cos(\widehat{R,n}) = e^{-kR}\frac{(1+kR)}{R^3}\frac{(r_0^2 + R^2 - r^2)}{2r_0},$$

$$I_0(\mathbf{r}) = \int\limits_{\partial S} G(\mathbf{r},\mathbf{r}_0)\,ds = \frac{r_0}{2k\,r}e^{-k\,r}[e^{kr_0} - e^{-kr_0}], \tag{18}$$

$$I_1(\mathbf{r}) = \int\limits_{\partial S} \frac{\partial}{\partial n}(G(\mathbf{r},\mathbf{r}_0))\,ds = \frac{1}{2k\,r}e^{-k\,r}[(1-k\,r_0)e^{k\,r_0} - (1+k\,r_0)e^{-k\,r_0}], \tag{19}$$

где $r = |\mathbf{r}|$. Нетрудно заметить, что $I_0(\mathbf{r}) > 0$, $I_1(\mathbf{r}) < 0$, и обе функции убывают по модулю с возрастанием аргумента r.

Оценим норму погрешности метода на $j+1$-й итерации через норму погрешности на j-й итерации. Ввиду того, что для уравнения Гельмгольца вида (5) применим принцип максимума, из (9) следует неравенство

$$\max_{\mathbf{r}\in\bar\Omega_j\setminus D} |\omega^{j+1}(\mathbf{r})| \le \max_{\mathbf{r}\in\partial S} |\omega_n^j(\mathbf{r})| \max_{\mathbf{r}\in\partial\Omega_j} I_0(\mathbf{r}) + \max_{\mathbf{r}\in\partial S} |\omega^j(\mathbf{r})| \max_{\mathbf{r}\in\partial\Omega_j} |I_1(\mathbf{r})|,$$

где интегралы $I_0(\mathbf{r})$ и $I_1(\mathbf{r})$ определены формулами (18), (19). Отсюда, принимая во внимание формулы (16) и (17), получаем окончательную оценку

$$\max_{\mathbf{r}\in\bar\Omega_j\setminus D} |\omega^{j+1}(\mathbf{r})| \le \mathrm{M}(k,d,r_0,x_0) \max_{\mathbf{r}\in\bar\Omega_{j-1}\setminus D} |\omega^j(\mathbf{r})|,$$

где

$$\mathrm{M}(k,d,r_0,x_0) = \frac{e^{-k\,d}}{2d}\left[e^{k\,r_0}\left(r_0 + r_0\beta - \frac{1}{k}\right) + e^{-k\,r_0}\left(r_0 - r_0\beta + \frac{1}{k}\right)\right], \tag{20}$$

$$d = \min_{\mathbf{r}\in\partial\Omega_j} |\mathbf{r}|, \quad x_0 = k\rho_0,$$

$$\beta = \beta(k,x_0) = \begin{cases} \dfrac{kx_0^2}{2(\sin x_0 - x_0\cos x_0)}, & x_0 < 2.08, \\ 1.15k, & x_0 \ge 2.08. \end{cases}$$

Таким образом, достаточным условием сходимости итерационного процесса для решения уравнения (5) будет

$$\mathrm{M}(k,d,r_0,x_0)<1. \tag{21}$$

2.4. Сходимость модифицированного метода

Для ускорения сходимости метода декомпозиции в [11], [12] была предложена релаксация при определении итерируемого приближения на поверхности раздела. В этих работах исследовалась декомпозиция области без пересечения для решения внутренних краевых задач. В [6] метод релаксации в сочетании с декомпозицией без пересечения был применен для решения внешней краевой задачи для уравнения Гельмгольца. Исследуем применение релаксации для ускорения сходимости решения уравнения (5) с условиями (2) и (3).

Пусть значения функции λ^{j+1} и ее нормальной производной $\dfrac{\partial\lambda^{j+1}}{\partial n}$ на сфере ∂S на $j+1$-й итерации метода определяются не только решением внутренней задачи в области $\Omega_j\backslash\overline{D}$, но и их значениями на j-й итерации:

$$\lambda^{j+1}(\mathbf{r}_0)=(1-\eta_j)u^{j+1}(\mathbf{r})\big|_{\partial S}+\eta_j\lambda^j(\mathbf{r}_0),\quad \mathbf{r}_0\in\partial S, \tag{22}$$

где η_j – релаксационный параметр на j-й итерации. Аналогичную формулу определим и для нормальной производной функции λ^{j+1} на сфере ∂S. Значения λ^0 и $\dfrac{\partial\lambda^0}{\partial n}$ соответствуют решению на сфере ∂S внутренней краевой задачи в области $\Omega_0\backslash\overline{D}$ с нулевыми граничными условиями на $\partial\Omega_0$.

Заметим, что итерационный процесс (6) является частным случаем метода декомпозиции с релаксацией при $\eta_j=0$.

Формула Грина (7) примет в этом случае следующий вид:

$$\Phi^j(\mathbf{r})=\int\limits_{\partial S}\left[G(\mathbf{r},\mathbf{r}_0)\frac{\partial\lambda^j}{\partial n}(\mathbf{r}_0)-\lambda^j(\mathbf{r}_0)\frac{\partial}{\partial n}(G(\mathbf{r},\mathbf{r}_0))\right]ds.$$

Определим погрешность метода

$$\omega_\lambda^j(\mathbf{r}_0)=\lambda^j(\mathbf{r}_0)-u(\mathbf{r}_0), \tag{23}$$

$$\omega_{\lambda n}^j(\mathbf{r}_0)=\frac{\partial\lambda^j}{\partial n}(\mathbf{r}_0)-\frac{\partial u}{\partial n}(\mathbf{r}_0),\quad \mathbf{r}_0\in\partial S.$$

Тогда из (22) и (23) следует, что

$$\omega_\lambda^{j+1}(\mathbf{r}_0)=(1-\eta_j)\,\omega^{j+1}(\mathbf{r}_0)+\eta_j\,\omega_\lambda^j(\mathbf{r}_0). \tag{24}$$

Аналогичное равенство справедливо и для погрешности $\omega_{\lambda n}^j(\mathbf{r}_0)$. Поскольку $\omega^{j+1}(\mathbf{r}_0)$ и $\omega_\lambda^j(\mathbf{r}_0)$ – ограниченные функции при $\mathbf{r}_0\in\partial S$, то существует такое число M_0, зависящее от входных параметров задачи, что выполняется неравенство

$$\max_{\mathbf{r}_0\in\partial S}\left|\omega^{j+1}(\mathbf{r}_0)\right|\le\mathrm{M}_0\max_{\mathbf{r}_0\in\partial S}\left|\omega_\lambda^j(\mathbf{r}_0)\right|. \tag{25}$$

Тогда из (24) и (25) следует, что

$$\max_{\mathbf{r}_0\in\partial S}\left|\omega_\lambda^{j+1}(\mathbf{r}_0)\right|\le\mathrm{M}_1\max_{\mathbf{r}_0\in\partial S}\left|\omega_\lambda^j(\mathbf{r}_0)\right|,$$

где $\mathrm{M}_1=|\eta_j|+|1-\eta_j|\,\mathrm{M}_0$. Нетрудно получить, что для выполнения условия $\mathrm{M}_1<1$ необходимо, чтобы $\eta_j<1$ и $\mathrm{M}_0<1$. Для таких областей изменения значений параметров η_j и M_0 минимум числа M_1 будет достигаться при $\eta_j=0$, что соответствует изначально рассмотренному в п. 2.1 ме-

тоду решения уравнения вида (5) без релаксации. Таким образом, можно сделать вывод, что модификация (22) для решения уравнения (5) нецелесообразна.

3. АНАЛИЗ СХОДИМОСТИ ЧАСТНОГО РЕШЕНИЯ С ПРОИЗВОЛЬНЫМ ВОЛНОВЫМ ЧИСЛОМ

В предыдущем разделе исследована сходимость метода для решения частного случая задачи (1)–(3) с отрицательным коэффициентом в уравнении Гельмгольца. В этом разделе будет исследована сходимость частного решения с произвольным волновым числом, что позволит сделать вывод о нецелесообразности использования предложенного метода декомпозиции с пересечением для решения задач с произвольными волновыми числами.

Проведем анализ сходимости альтернирующего метода Шварца для решения задачи (1)–(3) для частного случая, когда поверхность ∂D является сферой радиуса r_D, функция f в формуле (2) – константа, а решением задачи (1)–(3) является функция $u(r) = \dfrac{e^{ikr}}{r}$.

Пусть на сфере ∂S задано некоторое начальное значение искомой функции, являющейся константой: $u^0(r_0) = V_R^0 + iV_I^0 = \text{const}$. Выберем в качестве области Ω_0 шар, ограниченный сферой радиуса r_Ω, $r_D < r_0 < r_\Omega$. Исследование сходимости сведется к следующим шагам.

Шаг 1. По заданному начальному значению искомой функции на сфере ∂S решим внешнюю задачу и найдем значение функции на сфере $\partial\Omega_0$.

Шаг 2. Решим внутреннюю задачу в области $\Omega_0\backslash\overline{D}$ с учетом полученного значения на $\partial\Omega_0$ и заданного на ∂D, и найдем новое значение для искомой функции на сфере ∂S.

Шаг 3. Сравним новые и начальные значения на сфере ∂S с точными значениями и определим условия для убывания погрешности решения.

Детализируем предложенные шаги в данной постановке задачи.

Шаг 1. Решением внешней задачи для уравнения Гельмгольца вне сферы ∂S является функция $(C_R + iC_I)\dfrac{e^{ikr}}{r}$, где

$$C_R = r_0(V_R^0\cos(kr_0) + V_I^0\sin(kr_0)), \quad C_I = r_0(-V_R^0\sin(kr_0) + V_I^0\cos(kr_0)),$$

поэтому

$$u^1(r_\Omega) = \frac{r_0}{r_\Omega}\{V_R^0\cos\phi - V_I^0\sin\phi + i[V_R^0\sin\phi + V_I^0\cos\phi]\}, \tag{26}$$

где $\phi = k(r_\Omega - r_0)$.

Шаг 2. Решением внутренней задачи для уравнения Гельмгольца в области $\Omega_0\backslash\overline{D}$ с граничными условиями $u^1(r_\Omega) = U_R + iU_I$ и $u^1(r_D) = u_R + iu_I$ является функция

$$u^1(r) = [a\cos(kr) + b\sin(kr) + i\{c\cos(kr) + d\sin(kr)\}]\frac{1}{r\sin(k(r_\Omega - r_D))}, \tag{27}$$

где

$$a = r_D u_R\sin(kr_\Omega) - r_\Omega U_R\sin(kr_D), \quad b = r_\Omega U_R\cos(kr_D) - r_D u_R\cos(kr_\Omega),$$
$$c = r_D u_I\sin(kr_\Omega) - r_\Omega U_I\sin(kr_D), \quad d = r_\Omega U_I\cos(kr_D) - r_D u_I\cos(kr_\Omega).$$

Если значения U_R и U_I определены формулой (26), а значения функции на сфере ∂D равны

$$u_R = \frac{\cos(kr_D)}{r_D}, \quad u_I = \frac{\sin(kr_D)}{r_D},$$

то функция $u^1(r)$ из формулы (27) примет на сфере ∂S значение

$$u^1(r_0) = \frac{V_R^1 + iV_I^1}{\sin(k(r_\Omega - r_D))},$$

где

$$V_R^1 = a_R V_R^0 + b_R V_I^0 + c_R, \quad V_I^1 = a_I V_R^0 + b_I V_I^0 + c_I, \tag{28}$$
$$a_R = \sin(k(r_0 - r_D))\cos\phi, \quad b_R = -\sin(k(r_0 - r_D))\sin\phi,$$
$$a_I = \sin(k(r_0 - r_D))\sin\phi, \quad b_I = \sin(k(r_0 - r_D))\cos\phi,$$
$$c_R = \sin\phi\cos(k r_D)/r_0, \quad c_I = \sin\phi\sin(k r_D)/r_0.$$

Шаг 3. Перейдем к уравнениям для погрешностей решения. Введем обозначения:

$$\varepsilon_R^i = V_R^i - \cos(k r_0)/r_0, \quad \varepsilon_I^i = V_I^i - \sin(k r_0)/r_0, \quad i = 0,1. \tag{29}$$

Тогда $\varepsilon_R^1 = a_R\varepsilon_R^0 + b_R\varepsilon_I^0$, $\varepsilon_I^1 = a_I\varepsilon_R^0 + b_I\varepsilon_I^0$, или

$$\varepsilon_R^1 = \chi(\varepsilon_R^0\cos\phi - \varepsilon_I^0\sin\phi),$$
$$\varepsilon_I^1 = \chi(\varepsilon_R^0\sin\phi + \varepsilon_I^0\cos\phi), \tag{30}$$

где

$$\chi = \frac{\sin(k(r_0 - r_D))}{\sin(k(r_\Omega - r_D))}. \tag{31}$$

Формула (30) задает преобразование вращения, помноженное на постоянный множитель χ, который и определяет сходимость метода для рассмотренного частного случая.

Условие сходимости $|\chi| < 1$ для рассмотренной простейшей задачи имеет непростую структуру ввиду наличия бесконечного количества интервалов расходимости. Для более общего случая условия сходимости, скорее всего, примут еще более сложный вид. По этой причине авторы не рекомендуют решать уравнение Гельмгольца с произвольным волновым числом, отличное от вида (5), предложенным методом.

Рассмотрим условие сходимости модельной задачи с модификацией предложенного метода вида (22). Пусть $\eta_j = \eta = \text{const}$. Формулы (28) для данной модификации метода будут:

$$V_R^1 = \eta V_R^0 + (1-\eta)(a_R V_R^0 + b_R V_I^0 + c_R),$$
$$V_I^1 = \eta V_I^0 + (1-\eta)(a_I V_R^0 + b_I V_I^0 + c_I).$$

Перейдя к погрешностям решения ε_R^i и ε_I^i в формулах (29), получим

$$\varepsilon_R^1 = \eta\varepsilon_R^0 + (1-\eta)\chi[\varepsilon_R^0\cos\phi - \varepsilon_I^0\sin\phi],$$
$$\varepsilon_I^1 = \eta\varepsilon_I^0 + (1-\eta)\chi[\varepsilon_R^0\sin\phi + \varepsilon_I^0\cos\phi].$$

Выберем параметр η так, чтобы модуль погрешности $|\varepsilon^1| = \sqrt{(\varepsilon_R^1)^2 + (\varepsilon_I^1)^2}$ был минимальным. Нетрудно показать, что оптимальное значение параметра η_{opt} не зависит от $|\varepsilon^0|$ и равно

$$\eta_{\text{opt}} = \frac{\chi^2 - \chi\cos\phi}{1 + \chi^2 - 2\chi\cos\phi}. \tag{32}$$

Преобразовав формулу (32) с учетом явного вида параметров χ и ϕ, получим

$$\eta_{\text{opt}} = -\frac{\sin(k(r_0 - r_D))\cos(k(r_\Omega - r_D))}{\sin(k(r_\Omega - r_0))}.$$

Рассмотрим убывание погрешности метода. Условие $\rho = |\varepsilon^1|/|\varepsilon^0| < 1$ равносильно

$$\eta^2 + (1-\eta)^2\chi^2 + 2\eta(1-\eta)\chi\cos\phi < 1. \tag{33}$$

Выбрав в качестве параметра η в формуле (33) значение η_{opt} из (32), получим

$$\rho_{\text{opt}} = \frac{\chi\left|\sin\phi\right|}{\sqrt{1-2\chi\cos\phi+\chi^2}}.$$

Воспользовавшись явным видом параметра χ в формуле (31), получим окончательно

$$\rho_{\text{opt}} = \left|\sin\left(k\left(r_0 - r_D\right)\right)\right|.$$

Таким образом, преимущество более общего подхода (22) для решения задачи состоит в том, что в нем существует оптимальное значение параметра релаксации $\eta = \eta\left(k, r_D, r_0, r_\Omega\right)$ в (22), обеспечивающее сходимость метода при любых значениях входных параметров k, r_D, r_0, r_Ω, за исключением случаев, когда $k\left(r_0 - r_D\right) = \pi n/2$, где n – любое целое нечетное число.

Полученные значения интервалов и скорости сходимости метода позволяют предположить, что в случае произвольной исходной поверхности решение внешней задачи предложенными методами имеет аналогичный характер или удовлетворяет еще более сложным условиям. А именно: сходимость метода (6), (7) для решения задачи (1)–(3) имеет место только на интервалах, где выполняются определенные соотношения между входными параметрами, характеризующими поверхности ∂D и $\partial\Omega$, волновым числом k, и радиусом вспомогательной сферы r_0. Для модификации метода с релаксацией (22) сходимость будет иметь место при некоторых значениях параметров η_j этого метода. Выбор соотношений для исходных параметров в первом варианте метода и оптимального значения параметра для второго варианта, необходимых для их сходимости в общем случае, является пока открытой проблемой. Таким образом, рекомендация о нецелесообразности использования также и модифицированного метода с релаксацией для решения задачи (1)–(3) с произвольным волновым числом остается в силе.

4. АЛГОРИТМ НАХОЖДЕНИЯ ПРИБЛИЖЕННОГО РЕШЕНИЯ

Приведем принципиальную структуру программы для нахождения приближенного решения задачи (5), (2), (3).

Шаг 1. Задание начальных данных:

– задание трехмерной неравномерной структурированной тетраэдральной сетки, приближенно описывающей трехмерную замкнутую область $\Omega_j \backslash D$;

– задание равномерного разбиения сферы ∂S;

– задание критерия останова для итерационного решения внутренней краевой задачи ∂_i;

– задание критерия останова для внешнего итерационного процесса ∂_0;

Шаг 2. По заданной сетке проводится аппроксимация внутренней краевой задачи (6) хорошо известным методом барицентрических конечных объемов [12], [13], которая приводит к системе линейных алгебраических уравнений относительно приближенных значений искомой функции u_h в открытой области $\Omega_j \backslash \bar{D} \backslash \partial\Omega_j$, значений $u_{\partial\Omega_j}$ в узлах сетки на поверхности $\partial\Omega_j$, и заданных значений $u_{\partial D} = \left(f\left(\mathbf{r}\right)\right)_h$ в узлах сетки на поверхности ∂D:

$$\mathrm{A}u_h = \mathrm{B}_{\partial\Omega_j} u_{\partial\Omega_j} + \mathrm{B}_{\partial D} u_{\partial D}, \tag{34}$$

где A – матрица баланса для внутренних узлов сетки, а $\mathrm{B}_{\partial D}$ и $\mathrm{B}_{\partial\Omega_j}$ – матрицы учета краевых условий типа Дирихле от значений в узлах, лежащих на границах ∂D и $\partial\Omega_j$. Данное представление СЛАУ обусловлено тем, что позволяет отказаться от аппроксимации внутренней краевой задачи на каждом шаге внешнего итерационного процесса, ограничившись лишь пересчетом вклада значений функций в узлах, лежащих на границе $\partial\Omega_j$ в правой части СЛАУ (34).

Шаг 3. Внешний итерационный процесс:

– методом сопряженных градиентов в подпространствах Крылова [14] решается СЛАУ для внутренней краевой задачи

$$\mathrm{A}\,u_h^j = F_h^j, \quad j = 0,1,\ldots,$$

где $F_h^j = \mathrm{B}_{\partial\Omega_j} u_{\partial\Omega_j}^{j-1} + \mathrm{B}_{\partial D} u_{\partial D}$, $u_{\partial\Omega_0}^{-1} \equiv 0$. По полученному приближенному решению u_h^j проводится интерполяция приближенных значений функции $u_{S,h}^j$ и ее нормальной производной $u_{n,h}^j$ на сфере ∂S в точках, необходимых для решения внешней краевой задачи;

– по полученным значениям функции $u_{S,h}^j$ и ее нормальной производной $u_{n,h}^j$ численно находятся приближенные значения $u_{\partial\Omega_{j+1}}^j$ на границе $\partial\Omega_{j+1}$, для чего интеграл (4), записанный в сферических координатах в виде двойного интеграла, аппроксимируется квадратурной формулой Симпсона по каждой из переменных. По этим значениям на границе $\partial\Omega_{j+1}$ рассчитывается правая часть F_h^{j+1} для СЛАУ внутренней краевой задачи;

– выполняется проверка на критерий останова внешнего итерационного процесса:

если выполняется:

$$\left\| F_h^{j+1} - F_h^j \right\|_\infty \le \partial_0, \tag{35}$$

то внешний итерационный процесс завершается, иначе повторяем шаг 3.

5. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ И ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Будем искать численное решение внешней краевой задачи (5) с граничными условиями (2), (3), где k – вещественное число, имеющее следующее точное решение

$$u(x,y,z) = e^{-k r_1}/r_1 + e^{-k r_2}/r_2, \tag{36}$$

или

$$u(x,y,z) = z e^{-k r} (k + 1/r)/r^2, \tag{37}$$

где

$$r = \sqrt{x^2+y^2+z^2}, \quad r_1 = \sqrt{(x-x_0)^2+y^2+z^2}, \quad r_2 = \sqrt{(x+x_0)^2+y^2+z^2}, \quad x_0 = 0.1.$$

Зададим на гранях куба с ребрами, равными 0.4, граничные условия в соответствии с точным решением (36) или (37). Будем полагать, что куб, а также и все последующие в рассмотрении области, ограниченные выбранными поверхностями, имеют центр симметрии в начале координат. Выберем границу внешней расчетной области $\partial\Omega_0$, которую оставим неизменной на каждой итерации, в виде поверхности куба с ребрами, равными 2, на поверхности которого зададим нулевые граничные условия. В области $\Omega_0 \backslash \overline{D}$, ограниченной поверхностями этих кубов, решается внутренняя задача Дирихле и находятся значения функции и ее нормальной производной на сфере ∂S, лежащей в области $\Omega_0 \backslash \overline{D}$. По этим значениям определяются новые граничные условия на поверхности $\partial\Omega_0$ по формуле (4). Далее снова решается задача Дирихле в расчетной области $\Omega_0 \backslash \overline{D}$, и определяются новые приближенные значения функции и ее нормальной производной на сфере. Поверхность сферы, на которой ищется приближенное решение, является неизменной на всех итерациях в каждом численном эксперименте, однако может варьироваться в разных экспериментах.

Построение конечно-объемных аппроксимаций внутренней задачи Дирихле проводилось на последовательности трех равномерных сгущающихся сеток. Приведем число узлов и конечных элементов для каждой из сеток:

редкая сетка L = 29666, Ltet = 160704;

средняя сетка L = 225650, Ltet = 1285632;

густая сетка L = 1759794, Ltet = 10285056.

Обозначим через ε_v, ε_d и δ_v, δ_d среднеквадратичные и максимальные погрешности отклонения от точного решения и от точной нормальной производной на сфере ∂S, вычисленные по формулам

$$\varepsilon_v = \sqrt{\sum_{i,j} \left(u(\theta_j, \varphi_i) - \tilde{u}(\theta_j, \varphi_i) \right)^2 \Big/ \sum_{i,j} u^2(\theta_j, \varphi_i)},$$

Таблица 1. Зависимость погрешностей решения и нормальной производной от вида сетки

Сетка/Погрешность	ε_v	δ_v	ε_d	δ_d
Редкая	0.0149	0.0230	0.1290	0.2300
Средняя	0.0039	0.0065	0.0549	0.1004
Густая	0.0009	0.0013	0.0310	0.0557

$$\delta_v = \max_{i,j} \left| u(\theta_j, \varphi_i) - \tilde{u}(\theta_j, \varphi_i) \right| \Big/ \max_{i,j} \left| u(\theta_j, \varphi_i) \right|,$$

$$\varepsilon_d = \sqrt{\sum_{i,j} \left(\frac{\partial u}{\partial n}(\theta_j, \varphi_i) - \frac{\partial \tilde{u}}{\partial n}(\theta_j, \varphi_i) \right)^2 \Big/ \sum_{i,j} \left(\frac{\partial u}{\partial n} \right)^2 (\theta_j, \varphi_i)},$$

$$\delta_d = \max_{i,j} \left| \frac{\partial u}{\partial n}(\theta_j, \varphi_i) - \frac{\partial \tilde{u}}{\partial n}(\theta_j, \varphi_i) \right| \Big/ \max_{i,j} \left| \frac{\partial u}{\partial n}(\theta_j, \varphi_i) \right|,$$

где

$$u(\theta_j, \varphi_i), \quad \frac{\partial u}{\partial n}(\theta_j, \varphi_i) \quad \text{и} \quad \tilde{u}(\theta_j, \varphi_i), \quad \frac{\partial \tilde{u}}{\partial n}(\theta_j, \varphi_i)$$

являются точными и приближенными значениями искомой функции и ее нормальной производной в узлах θ_j, φ_i сетки в сферических координатах на сфере.

В табл. 1 представлены погрешности вычисления функции и ее нормальной производной на сфере радиуса $r_0 = 0.5$ при $k^2 = 1$, $N_\theta = N_\varphi = 17$. Число внешних итераций N_{it} для достижения критерия останова (35) итерационного процесса при $\partial_0 = 10^{-7}$ было $N_{it} = 3$. Точное решение определялось формулой (37). Вычисления проводились на редкой, средней и подробной сетках, в которых длина ребра элементарного тетраэдра в методе конечных объемов уменьшалась приблизительно в 2 раза при переходе к более подробной сетке. Заметим, что увеличение количества точек на сфере не приводило к существенному уменьшению вычисляемых погрешностей, однако требовало дополнительных временных затрат. Это обстоятельство иллюстрирует то, что погрешность решения задачи обусловлена, прежде всего, погрешностью решения ее внутренней части.

Результаты, приведенные в таблице, демонстрируют квадратичную сходимость полученного приближенного решения и линейную сходимость для его производной при сгущении сетки и отсутствие зависимости числа внешних итераций от шага сетки.

В следующем численном эксперименте определялись погрешности метода при разных значениях радиуса вспомогательной сферы r_0 и разных коэффициентах k^2 в уравнении Гельмгольца. Вычисления проводились на подробной сетке, точное решение было задано формулой (36). Максимальные погрешности приближенного значения функции δ_v, ее нормальной производной на сфере δ_d и количество внешних итераций N_{it}, полученные при достижении условия (35), где $\partial_0 = 10^{-7}$ для разных значений радиуса вспомогательной сферы, приведены в табл. 2.

Для анализа полученных значений воспользуемся формулой (20) для коэффициента уменьшения погрешности $\mathrm{M}(k, d, r_0, x_0)$ при переходе к следующей итерации метода. При заданных значениях параметров исходной задачи этот коэффициент будет меньше единицы при $0.487 \le r_0 \le 0.696$ для $k^2 = 1$, и при $0.362 < r_0 < 0.846$ для $k^2 = 10$. Полученные значения для погрешностей позволяют сделать вывод, что сходимость метода имеет место при больших значениях радиуса вспомогательной сферы, вплоть до $r_0 < 0.95$. Это обстоятельство не противоречит условию сходимости метода (21), поскольку оно является достаточным.

Для обеспечения надежности численных вычислений необходимо выбрать параметры метода, характеризующие расположение внешней поверхности и вспомогательной сферы таким образом, чтобы условие сходимости метода (21) было выполнено. Если при некотором выборе внешней границы расчетной области $\partial\Omega_j$ условие $\mathrm{M} < 1$ не выполняется для любых радиусов

Таблица 2. Зависимость погрешностей решения задачи и числа внешних итераций от радиуса сферы и коэффициента в уравнении Гельмгольца

r_0	$k^2 = 1$			$k^2 = 10$		
	N_{it}	δ_v	δ_d	N_{it}	δ_v	δ_d
0.4	4	0.0016	0.0506	4	0.0017	0.0520
0.5	5	0.0007	0.0502	4	0.0010	0.0506
0.6	5	0.0012	0.0431	4	0.0013	0.0489
0.7	5	0.0015	0.0325	4	0.0014	0.0366
0.8	5	0.0118	0.0511	4	0.0092	0.0473
0.9	7	0.0609	0.1019	7	0.0526	0.0831
0.95	14	0.1754	0.9836	12	0.1562	0.6748

вспомогательных сфер, расположенных между поверхностями ∂D и $\partial\Omega_j$, то необходимо переместить внешнюю поверхность $\partial\Omega_j$ дальше от исходной поверхности ∂D. Действительно, множитель $e^{-kd}/2d$ в формуле (20) быстро убывает с увеличением параметра d, где d – расстояние от начала координат до поверхности $\partial\Omega_j$. Однако рост этого параметра влечет увеличение расчетной области для решения внутренней краевой задачи, что является нежелательным ввиду использования большего объема вычислительных ресурсов для решения задачи. Другим параметром декомпозиции является радиус вспомогательной сферы r_0. Этот параметр связан со вторым множителем в формуле (20). При расположении сферы вблизи поверхности ∂D или $\partial\Omega_j$ параметры ρ_0 и x_0 уменьшаются, что приводит к увеличению коэффициента β, значит, и всего коэффициента M. По этой причине целесообразно размещать вспомогательную сферу ∂S где-нибудь посередине между поверхностями ∂D и $\partial\Omega_j$ для уменьшения коэффициента β, который монотонно убывает при $0 < x < 2.08$.

Итак, подытожим сказанное о выборе параметров метода. Для уменьшения времени счета и экономии вычислительных ресурсов выбираем внешнюю поверхность $\partial\Omega_j$, расположенную относительно близко к исходной поверхности ∂D. После этого по формуле (20) производим вычисления коэффициента M при разных значениях радиуса вспомогательной сферы r_0. Если условие (21) при выбранном расположении вспомогательной поверхности $\partial\Omega_j$ не выполнено для всех сфер, расположенных между поверхностями ∂D и $\partial\Omega_j$, то отодвигаем поверхность $\partial\Omega_j$ дальше от поверхности ∂D и снова производим вычисления коэффициента M для нового набора значений радиусов сфер r_0 между этими областями. Такую процедуру необходимо проводить до выполнения условия (21), после чего можно приступать к численным расчетам по решению задачи (5) с условиями (2) и (3).

СПИСОК ЛИТЕРАТУРЫ

1. *Givoli D.* Numerical methods for problems in infinite domains Amsterdam, Netherlands: Elsevior, 1992.
2. *Engleder S., Steinbach O.* Stabilized boundary element methods for exterior Helmholtz problems // Numer. Math. 2008. V. 110. Issue 2. P. 145–160.
3. *Kleefeld A., Lin Tsu-Chu.* Boundary element collocation method for solving the exterior Neumann problem for Helmholt's equation in three dimensions // Electronic Transactions on Numerical Analysis. 2012. V. 39. P. 113–143.
4. *Aziz A., Dorr M., Kellogg R.* A new approximation method for the Helmholtz equation in an exterior domain // SIAM J.Numer.Anal. 1982. V. 19. № 5.
5. *Yu D.* Natural Boundary Integral Method and Its Applications. Netherlands: Springer, 2002.
6. *Chen Q., Liu B., Du Q.* A D-N Alternating Algorithm for Solving 3D Exterior Helmholtz Problems // Math. Problems in Engng, 2014. https://doi.org/10.1155/2014/418426
7. *Du Q., Yu D.* Schwarz alternating method based on natural boundary reduction for time-dependent problems on unbounded domains // Commun. Numer. Meth. Engng. 2004. V. 20. P. 363–378.

8. *Jia Z., Wu J., Yu D.* A coupled natural boundary element and finite element method for solving a 3-dimensional exterior Helmholtz problem // Mathematica Numerica Sinica, 2001. V. 23. № 3. P. 357–368.
9. *Курант Р., Гилберт Д.* Методы математической физики. Т. 2. М.: Мир, 1964.
10. *Marini L.D., Quarteroni A.* A relaxation procedure for domain decomposition methods using finite elements // Numer. Math. 1989. 55. P. 575–598.
11. *Quarteroni A., Valli A.* Domain decomposition methods for partial differential equations. Oxford: Clarendon Press, 1999.
12. *Petukhov A.V.* The Barycentric Finite Volume Method for 3D Helmholtz Complex Equation // Optoelectronics, Instrumentation and Data Processing. 2007. V. 43. № 2. P. 182–191.
13. *Ильин В.П.* Методы конечных разностей и конечных объемов для эллиптических уравнений Новосибирск: Изд-во Ин-та математики, 2000.
14. *Ильин В.П.* Методы и технологии конечных элементов Новосибирск: Изд-во ИВМиМГ СО РАН, 2007.

EDN: COLUDD
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 823–837

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.54

О РЕШЕНИИ ОДНОЙ ЗАДАЧИ О КОНФОРМНОМ ОТОБРАЖЕНИИ ПРИ ПОМОЩИ ФУНКЦИЙ ВЕЙЕРШТРАССА[1)]

© 2022 г. М. Смирнов[1,2,*]

[1]119333 Москва, ул. Губкина, 8, Ин-т вычисл. математики РАН, Россия

[2]119991 Москва, Ленинские горы, МГУ, Россия

**e-mail: matsmir98@gmail.com*

Поступила в редакцию 15.09.2021 г.
Переработанный вариант 25.11.2021 г.
Принята к публикации 14.01.2022 г.

Рассматривается задача о конформном отображении сечения канала, заполненного пористым материалом, под плотиной прямоугольного сечения на верхнюю полуплоскость. Подобные задачи возникают при расчете течения жидкости в гидротехнических сооружениях. В качестве метода решения используется представление эллиптического интеграла Кристоффеля–Шварца через функции Вейерштрасса. Для расчета используется ряд Тейлора для сигма-функции, коэффициенты которого определяются рекуррентно. Получена простая формула для конформного отображения, зависящая от четырех параметров и использующая сигма-функцию. Для конкретной области проведен численный эксперимент. Рассмотрено вырождение области, состоящее в стремлении к нулю толщины плотины, и показано, что полученная формула имеет предел, осуществляющий решение предельной задачи. Приведено уточненное доказательство рекуррентной формулы Вейерштрасса для коэффициентов ряда Тейлора сигма-функции. Библ. 17. Фиг. 5.

Ключевые слова: конформные отображения, интеграл Кристоффеля–Шварца, эллиптические функции, сигма-функция Вейерштрасса, вырождение функций Вейерштрасса.

DOI: 10.31857/S0044466922050131

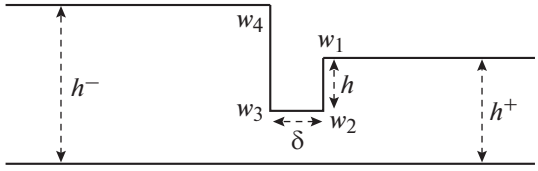
1. ВВЕДЕНИЕ

В данной статье рассматривается одна область $\Omega \subset \mathbb{C}$, граница которой является ломаной с углами, кратными $\pi/2$. Данная область моделирует сечение канала под плотиной. Расчет течения жидкости по такому каналу сводится к задаче о конформном отображении Ω на верхнюю полуплоскость. Решение подобных задач дается интегралом Кристоффеля–Шварца (см., например, [1] или [2]), который в рассматриваемом случае живет на эллиптической римановой поверхности. В работе найдена простая формула, выражающая интеграл через сигма-функцию Вейерштрасса (см., например, [3] или [4]). Благодаря такому подходу отпадает необходимость использовать численное интегрирование, а параметры отображения находятся из достаточно простой системы нелинейных уравнений, что существенно упрощает численное решение.

Аналогичные задачи уже рассматривались в [5]–[9], где для эффективного представления интеграла Кристоффеля–Шварца использовались тэта-функции (см., например, [10]). В статье [9] также исследовалось применение функций Лауричеллы к этим задачам, а кроме того, было проведено практическое сравнение различных подходов к решению.

Основным преимуществом использования функций Вейерштрасса перед тэта-функциями является наличие у них предельных значений при вырождениях поверхности. В работе анализируется поведение построенного конформного отображения при условии, что толщина тела плотины стремится к нулю. Оказывается, что конформные отображения имеют предел, причем являющийся решением возникающей в пределе задачи. Таким образом, показано, что построенное решение устойчиво при рассмотренном вырождении.

[1)]Работа выполнена в части доказательства рекуррентной формулы Вейерштрасса для коэффициентов ряда Тейлора сигма-функции и оценки их роста при финансовой поддержке отделения ИВМ РАН Московского центра фундаментальной и прикладной математики (соглашение номер 075-15-2019-1624). Остальная часть исследования выполнена при финансовой поддержке РНФ (проект номер 21-11-00325).



Фиг. 1. Область Ω.

Описанное свойство сигма-функции Вейерштрасса теряет свой смысл, если использовать для вычислений стандартный метод, выражающий сигма-функцию через тэта-функцию (поскольку она не выдерживает вырождения). Таким образом, возникает необходимость использовать независимый метод вычисления сигма-функций. В данной работе используется выражение для коэффициентов ее разложения в ряд Тейлора, полученное Вейерштрассом (см. [11]). Поскольку изложенное там доказательство, по всей видимости, не полно (в один момент используется голоморфность сигма-функции по трем переменным в окрестности нуля, что не очевидно), мы приводим более подробное доказательство в приложении. Вышеописанной формулы, однако, недостаточно для окончательного численного решения, так как ряды Тейлора не подходят для вычислений при больших аргументах (а именно такая необходимость возникает при вырождении). Таким образом, остается нерешенной еще задача о построении эффективного вычислительного метода для сигма-функции, не зависимого от тэта-функций. При наличии такого метода появится возможность строить устойчивые при различных вырождениях формулы и использовать их в вычислениях. Результаты настоящей работы иллюстрируют необходимость построения подобных методов.

Задачи, в которых возникают гиперэллиптические римановы поверхности высокого рода, рассмотренные, например, в [7], [8], можно также решать при помощи теории сигма-функций, развитой Клейном и Бейкером в [12] и [13] соответственно (более подробное изложение см. в [14]). Есть надежда, что удастся доказать устойчивость формул, выражающих решение вышеупомянутых задач через сигма-функции высокого рода. Таким образом, построение рекуррентных формул типа Вейерштрасса (которые известны для рода 1 и 2; см. [14]) и методов вычисления для сигма-функций могут оказаться крайне полезными для прикладных задач.

2. ФОРМУЛИРОВКА ЗАДАЧИ И ЕЕ ПРОИСХОЖДЕНИЕ

Рассмотрим область Ω в комплексной плоскости, изображенную на фиг. 1. Снизу она ограничена прямой, а сверху – ломаной с четырьмя вершинами: w_1, w_2, w_3, w_4 (удобно считать, что у этой области также еще две вершины находятся в $\pm\infty$). Будем считать, что прямая, ограничивающая снизу эту область, параллельна вещественной оси, а вершина w_4 расположена в нуле. Тогда данная область определяется четырьмя вещественными параметрами h^-, h^+, h, δ, где h – длина отрезка $[w_1, w_2]$, δ – длина отрезка $[w_2, w_3]$, а h^- и h^+ – расстояние от прямой, ограничивающей область снизу до w_4 и w_1 соответственно. Параметры положительны и удовлетворяют неравенствам $h^- - h^+ + h > 0$, что соответствует положительности длины отрезка $[w_3, w_4]$, и $h < h^+$. Область определяется этими параметрами однозначно.

Области, подобные Ω, возникают в задачах, связанных с расчетом течения жидкости через пористый материал под плотиной. Поскольку течение неразрывно и подчиняется закону Дарси, давление p является в Ω гармонической функцией. Считая, что отрезки $[w_1, w_2]$, $[w_2, w_3]$, $[w_3, w_4]$, а также дно канала непроницаемы для жидкости, получаем естественные граничные условия: нормальная производная $\partial p/\partial n$ обращается в нуль на непроницаемых кусках границы, в то время как на оставшихся сегментах (т.е. на полупрямых, выходящих из w_1 и w_4) функция p локально постоянна.

Рассмотрим в области Ω такую функцию q, что $f = p + iq$ голоморфна (такая функция существует, так как Ω односвязно). Условие обращения в нуль нормальной производной функции p, как нетрудно видеть, равносильно постоянству функции q на соответствующем сегменте границы. Отсюда следует, что, если в качестве f взять функцию, конформно отображающую Ω на прямоугольник так, чтобы в его вершины перешли точки w_1, w_4 и пара бесконечно удаленных вер-

шин области Ω, то $p = \mathrm{Re} f$ будет решением исходной краевой задачи. Функция $q = \mathrm{Im} f$ называется функцией тока. Ее линии уровня являются линиями тока жидкости под плотиной. Ясно, что для решения обозначенной задачи достаточно решить задачу о конформном отображении Ω на верхнюю полуплоскость $\mathbb{C}_+$.

В дальнейшем задача о конформном отображении будет решена явно при помощи аппарата эллиптических функций Вейерштрасса. Ниже показан расчет линий тока в области Ω, полученный с помощью построенного в настоящей работе метода.

3. РЕШЕНИЕ ЗАДАЧИ О КОНФОРМНОМ ОТОБРАЖЕНИИ

3.1. Общий вид решения и определение параметров

Поскольку Ω односвязна, существует конформное отображение $W : \mathbb{C}_+ \to \Omega$, где $\mathbb{C}_+ = \{z \in \mathbb{C} : \mathrm{Im} z > 0\}$ – верхняя полуплоскость (см., например, [15] или [1]). Используя, если нужно, подходящий автоморфизм области $\mathbb{C}_+$, можно добиться того, чтобы в точку w_4 при отображении W (точнее при его продолжении на границу) переходила точка ∞. Тогда, по теореме Кристоффеля–Шварца (см. [2]), найдутся такие $x^- < x^+ < x_1 < x_2 < x_3 \in \mathbb{R}$ и $C \in \mathbb{C}$, что

$$dW = \phi = C \frac{\sqrt{(x - x_2)(x - x_3)}}{(x - x^-)(x - x^+)\sqrt{x - x_1}} dx. \tag{3.1}$$

Замечание 1. Здесь x_i – прообраз вершины w_i при отображении W, а x^- и x^+ – точки на границе верхней полуплоскости, в которых W уходит на бесконечность (прообразы бесконечно удаленных вершин).

Дифференциальную форму ϕ можно рассматривать на гиперэллиптической римановой поверхности V рода 1, заданной уравнением $y^2 = F(x) = 4(x - x_1)(x - x_2)(x - x_3)$. Используя, если нужно, сдвиг верхней полуплоскости мы можем, не ограничивая общности, считать, что $x_1 + x_2 + x_3 = 0$. Отсюда $F(x) = 4x^3 - g_2 x - g_3$ для некоторых вещественных g_2, g_3 (определяемых числами x_1, x_2, x_3). На этой поверхности ϕ можно переписать в виде

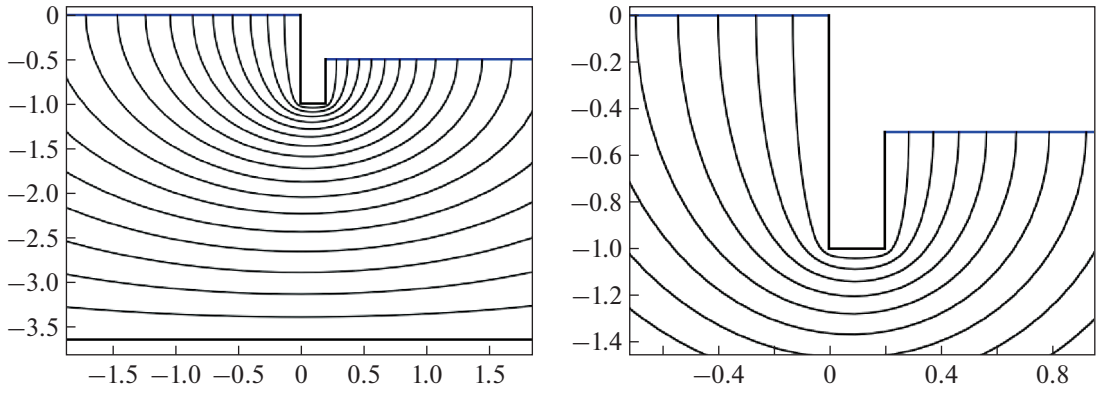
$$\phi = 2C \frac{(x - x_2)(x - x_3)}{y(x - x^-)(x - x^+)} dx. \tag{3.2}$$

Зафиксируем ветвь функции $\sqrt{F(x)}$ в области, полученной из $\mathbb{C}$ выбрасыванием отрезка $[x_1, x_2]$ и полупрямой $[x_3, \infty]$. Будем считать, что эта ветвь положительна при стремлении к полупрямой (x_3, ∞) из верхней полуплоскости. Вспоминая, что dx/y – голоморфная (всюду отличная от нуля) форма на V, получаем, что ϕ имеет два нуля кратности 2 в точках $(x_2, 0)$ и $(x_3, 0)$, а также четыре простых полюса в точках $(x^-, \pm\sqrt{F(x^-)})$ и $(x^+, \pm\sqrt{F(x^+)})$. Заметим теперь, что вычеты этой формы в данных полюсах равны $\pm h^-/\pi$ и $\mp h^+/\pi$ соответственно.

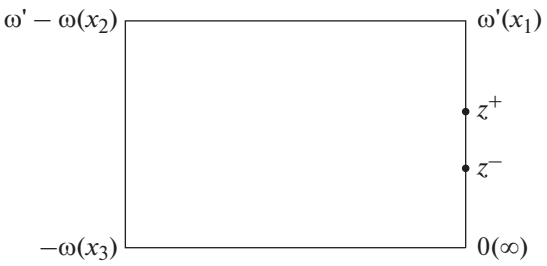
Теперь воспользуемся отображением Абеля (см., например, [16]), которое отождествляет V с $\mathrm{Jac}(V)$ (как обычно, начальной точкой положим бесконечно удаленную точку, а в качестве базиса голоморфных форм – dx/y). Введем полупериоды

$$\omega = \int_{x_1}^{x_2} \frac{dx}{y}, \quad \omega' = -\int_{x_2}^{x_3} \frac{dx}{y},$$

а также величины $\eta = \zeta(\omega)$ и $\eta' = \zeta(\omega')$, где ζ – дзета-функция Вейерштрасса (см. [3]). Нетрудно видеть, что $\omega, \eta \in \mathbb{R}$ и $\omega', \eta' \in i\mathbb{R}$. Совокупность точек $(x, \sqrt{F(x)})$, где $x \in \mathbb{C}_+$ при этом отображении перейдет в прямоугольник с вершинами $0, \omega', \omega' - \omega, -\omega$. Обозначим образы точек $(x^-, \sqrt{F(x^-)})$ и $(x^+, \sqrt{F(x^+)})$ через z^- и z^+ соответственно (см. фиг. 3, где в скобках указаны прообразы соответствующих точек). Образы точек $(x^-, -\sqrt{F(x^-)})$ и $(x^+, -\sqrt{F(x^+)})$ равны в этом случае $-z^-$ и $-z^+$. Рассмотрим дифференциальную форму ψ на торе, в которую ϕ переходит при отождествлении V с



Фиг. 2. Линии тока в области Ω.



Фиг. 3. Образ верхней полуплоскости при отображении Абеля.

$\mathrm{Jac}(V)$. Эта форма имеет 4 простых полюса в точках $\pm z^-$ и $\pm z^+$, причем имеет в этих точках вычеты $\pm h^-/\pi$ и $\mp h^+/\pi$ соответственно.

Воспользуемся теперь описанным в [3] методом представления эллиптических функций через функции Вейерштрасса. Для этого рассмотрим мероморфную функцию

$$g(z) = \frac{h^-}{\pi}(\zeta(z - z^-) - \zeta(z + z^-)) - \frac{h^+}{\pi}(\zeta(z - z^+) - \zeta(z + z^+)). \tag{3.3}$$

Из соотношений квазипериодичности функции ζ (см. [3]), легко вывести, что g – эллиптическая функция. Форма $g(z)dz$ имеет точно такие же (простые) полюса, что и ψ, причем в этих полюсах имеет те же вычеты. Значит, $\psi - g(z)dz$ – голоморфная форма на торе. Поскольку на торе пространство голоморфных 1-форм одномерно, получаем, что $\psi - g(z)dz = Ddz$, где D – некоторая константа (причем $D \in i\mathbb{R}$).

Теперь обратимся к отображению W. Ясно, что

$$W(x) = -\int_x^\infty \phi.$$

Таким образом, положим

$$Q(z) = \int_0^z \psi. \tag{3.4}$$

Очевидно, что $W(x)$ совпадает с $Q(z)$, где z – образ точки $(x, \sqrt{F(x)})$ при отображении Абеля. Значит, Q осуществляет конформное отображение прямоугольника с вершинами $0, \omega', \omega' - \omega, -\omega$

на Ω, причем ω' переходит в w_1, $\omega' - \omega$ переходит в w_2, а $-\omega$ – в w_3 (а 0 – в w_4). Теперь мы можем записать систему уравнений, следующих из полученных ранее соотношений:

$$g(-\omega) + D = 0, \quad g(\omega' - \omega) + D = 0, \tag{3.5}$$

$$Q(\omega' - \omega) - Q(\omega') = -ih, \quad Q(-\omega) - Q(\omega' - \omega) = -\delta. \tag{3.6}$$

Замечание 2. Первая пара уравнений получается из того, что ϕ имеет нули в точках $(x_2, 0)$ и $(x_3, 0)$, а вторая пара получается из равенств $w_3 - w_2 = -\delta$, $w_2 - w_1 = -ih$.

Остается найти разумную формулу для Q. Вспомним, что ζ – логарифмическая производная функции σ. Отсюда легко получаем, что Q имеет вид

$$Q(z) = Dz + \frac{h^-}{\pi}\ln\left(\frac{\sigma(z - z^-)}{\sigma(z + z^-)}\right) - \frac{h^+}{\pi}\ln\left(\frac{\sigma(z - z^+)}{\sigma(z + z^+)}\right) - i(h^- - h^+), \tag{3.7}$$

где взята ветвь функции ln, определенная на плоскости с разрезом по мнимой отрицательной полуоси, принимающая в 1 значение 0. Подставляя в (3.5) выражение для g из (3.3), а также формулу (3.7) в (3.6), применяя при этом свойства квазипериодичности σ-функции (см., например, [3]), получаем следующую систему уравнений:

$$\begin{gathered}
-D\omega - \frac{2h^+}{\pi}\eta z^+ + \frac{2h^-}{\pi}\eta z^- = -ih, \\
-D\omega' - \frac{2h^+}{\pi}\eta' z^+ + \frac{2h^-}{\pi}\eta' z^- = -\delta, \\
D + \frac{h^-}{\pi}(\zeta(\omega - z^-) - \zeta(\omega + z^-)) - \frac{h^+}{\pi}(\zeta(\omega - z^+) - \zeta(\omega + z^+)) = 0, \\
D + \frac{h^-}{\pi}(\zeta(\omega' + \omega - z^-) - \zeta(\omega' + \omega + z^-)) - \frac{h^+}{\pi}(\zeta(\omega' + \omega - z^+) - \zeta(\omega' + \omega + z^+)) = 0.
\end{gathered} \tag{3.8}$$

В записанной системе уравнений на текущий момент имеется пять независимых переменных (величины $\omega, \omega', \eta, \eta'$ определяются через g_2 и g_3): g_2, g_3, D, z^+, z^- (первые два вещественны, а остальные – чисто мнимые) и 4 уравнения (3.8) (среди этих уравнений первое, третье и четвертое – чисто мнимые, а второе – вещественное). Поэтому для определения параметров естественно рассмотреть какое-нибудь однопараметрическое семейство кривых, среди которых заведомо найдется подходящая, т.е. рассмотреть функции $g_2 = g_2(\gamma)$ и $g_3 = g_3(\gamma)$ и использовать систему (3.8) для определения параметров γ, D, z^+, z^-.

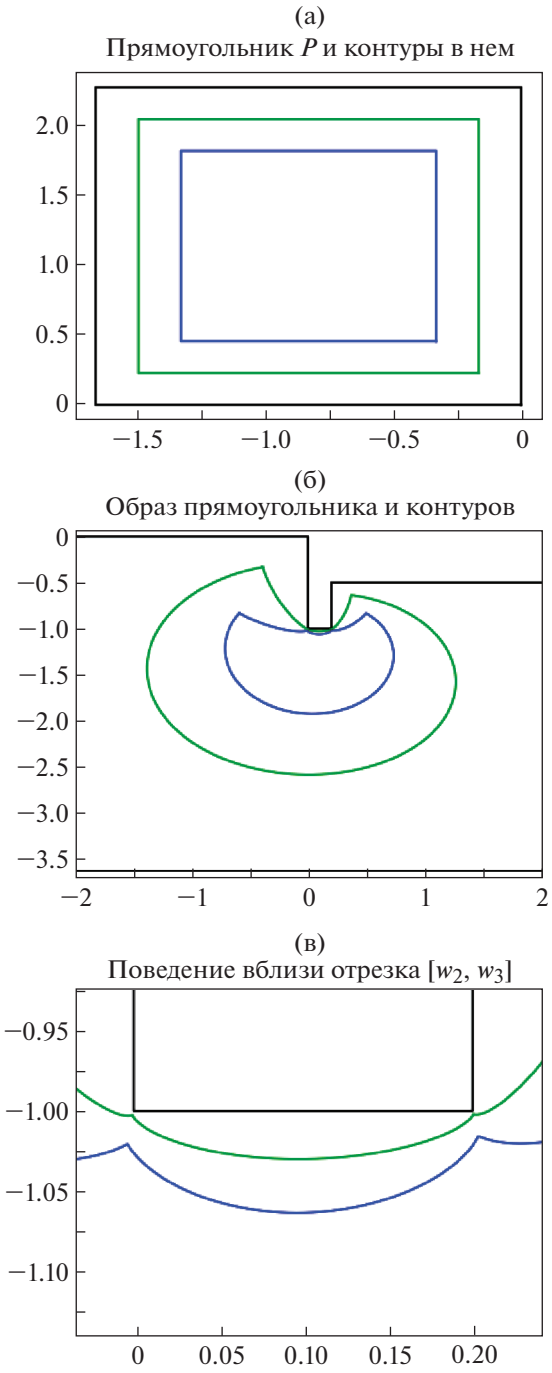
В дальнейшем мы будем использовать семейство кривых, заданных корнями полинома F: $x_1 = \gamma - 1/2$, $x_2 = -2\gamma$, $x_3 = \gamma + 1/2$, $\gamma \in (-1/6, 1/6)$ (более подробно изучать это семейство мы будем при анализе поведения решения при вырождении $\delta \to 0$). Это семейство соответствует нормировке $x_3 - x_1 = 1$ с учетом уже имеющегося тождества $x_1 + x_2 + x_3 = 0$.

3.2. О численной реализации

Для численной реализации было принято решение использовать явное вычисление сигма-функции Вейерштрасса через ее параметры g_2, g_3 при помощи ее ряда Тейлора (см. [11] или ниже теорему П.1). Ясно, что для эффективного решения системы (3.8) требуется вычислять все входящие в нее величины и их производные по параметрам. В конечном счете все сводится к вычислению величин ω, ω' и их производных по g_2, g_3, а также функции ζ и ее производных по z, g_2, g_3. Поскольку

$$\zeta = \frac{1}{\sigma}\frac{\partial\sigma}{\partial z},$$

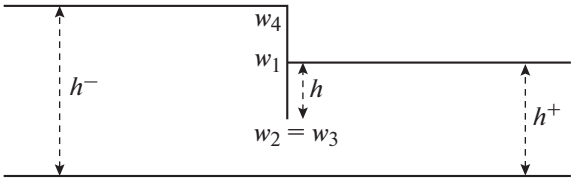
задача о вычислении ζ-функции и ее производных решается тривиально. Чтобы вычислить ω, надо заметить, что σ-функция обращается в ноль только в точках решетки $\{2m\omega + 2n\omega' : n, m \in \mathbb{Z}\}$, причем все эти нули просты. Эффективный способ локализовать простой нуль z_0 голоморфной функции f – найти интеграл от функции $zf'(z)/2\pi i f(z)$ по контуру, охватывающему точку z_0. Чтобы найти подходящий контур, можно применить аналог бинарного поиска, пользуясь тем, что



Фиг. 4. Конформное отображение Q.

$\omega \geq \pi/2$. Используя сформулированный метод либо напрямую, либо для приближенного вычисления нуля и последующего применения методов решения уравнений, легко построить эффективный и точный алгоритм вычисления ω (и ω'). Для вычисления их производных можно продифференцировать интеграл от $z\sigma'(z)/\sigma(z)$ по g_2 или g_3, а затем вычислить его явно, найдя вычет полученной функции в нуле функции σ. Таким образом, решение системы уравнений (3.8) можно полностью свести к вычислению сигма-функции Вейерштрасса и ее производных по z, g_2 и g_3.

Продемонстрируем решение конкретной задачи при помощи данного метода. Положим, $h^+ = \pi$, $h^- = \pi + 0.5$, $h = 0.5$, $\delta = 0.2$. В качестве однопараметрического семейства кривых, среди



Фиг. 5. Область $\widetilde{\Omega}$.

которых мы будем искать решение, положим: $x_1 = \gamma - 1/2$, $x_2 = -2\gamma$, $x_3 = \gamma + 1/2$, $\gamma \in (-1/6, 1/6)$. Решение системы (3.8) дает вектор

$$(\gamma, D, z^+, z^-) = (0.1051616134, 0.0203152915i, 1.3043479103i, 0.7195735824i).$$

При данном γ имеем $\omega = 1.6518996331$, $\omega' = 2.2939120295i$. На фиг. 4 далее демонстрируется образ прямоугольника P с вершинами $0, \omega', \omega' - \omega, -\omega$ при отображении Q.

4. УСТОЙЧИВОСТЬ РЕШЕНИЯ ПРИ ВЫРОЖДЕНИИ ОБЛАСТИ

Здесь мы рассмотрим задачу о конформном отображении верхней полуплоскости на область $\widetilde{\Omega}$, которая из Ω получается вырождением $\delta \to 0$ (см. фиг. 5) и проанализируем поведение решения при этом вырождении, считая, что никаких других вырождений не происходит (т.е. считая, что величины $h^-, h^+, h, h^- + h - h^+, h^+ - h$ имеют конечные положительные пределы). Область $\widetilde{\Omega}$ задается теперь тремя параметрами h, h^+ и h^-. Конформное отображение верхней полуплоскости на область $\widetilde{\Omega}$ можно искать способом, аналогичным уже разобранному (через теорему Кристоффеля–Шварца). В этом случае, поскольку соответствующая поверхность будет иметь род 0, решение выразится в элементарных функциях. Другой же способ (рассматриваемый здесь) состоит в том, чтобы формулу (3.7) приспособить для данного случая, пользуясь тем, что σ-функция определена в том числе при тех значениях g_2 и g_3, при которых полином $F(x) = 4x^3 - g_2 x - g_3$ имеет кратные корни. Естественно предположить, что искомое отображение получается предельным переходом, при котором пара корней полинома, переходящих в w_2 и w_3, склеивается в одну точку. Вместе с этим также будет установлена устойчивость построенного решения при $\delta \to 0$.

4.1. Склеивание пары корней полинома

Снова рассмотрим семейство кривых, зависящих от параметра $\gamma \in (-1/6, 1/6)$ следующим образом: $F_\gamma(x) = 4(x - x_1(\gamma))(x - x_2(\gamma))(x - x_3(\gamma))$, где $x_1(\gamma) = \gamma - 1/2$, $x_2(\gamma) = -2\gamma$, $x_3(\gamma) = \gamma + 1/2$. При $\gamma \to -1/6$ будем иметь склейку корней x_2 и x_3. Предельные значения g_2 и g_3 равны соответственно $4/3$ и $-8/27$. Для каждого γ можно определить $\omega(\gamma)$, $\omega'(\gamma)$, $\eta(\gamma)$, $\eta'(\gamma)$. В дальнейшем мы будем опускать зависимость величин от параметра γ.

Лемма 4.1. *При* $\gamma \to -1/6$ *имеем*

$$\omega, \eta \to \infty, \quad \omega' \to \frac{i\pi}{2}, \quad \eta' \to -\frac{i\pi}{6}, \quad \frac{\eta}{\omega} \to -\frac{1}{3}. \tag{4.1}$$

Кроме того,

$$\sigma\left(z, \frac{4}{3}, -\frac{8}{27}\right) = e^{-\frac{z^2}{6}} \sinh(z), \quad \zeta\left(z, \frac{4}{3}, -\frac{8}{27}\right) = \coth(z) - \frac{z}{3}, \quad \wp\left(z, \frac{4}{3}, -\frac{8}{27}\right) = \frac{1}{\sinh^2(z)} + \frac{1}{3}. \tag{4.2}$$

Наконец, существует такой $\varepsilon > 0$, *что при* $\gamma + 1/6 < \varepsilon$ *имеет место оценка*

$$-c_1 \ln(\gamma + 1/6) \le \omega(\gamma) \le -c_2 \ln(\gamma + 1/6), \tag{4.3}$$

где $0 < c_1 < c_2$.

Доказательство. Предельные переходы (4.1) легко получаются из интегральных формул:

$$\omega(\gamma) = \frac{1}{2}\int\limits_{\gamma-\frac{1}{2}}^{-2\gamma} \frac{dx}{\sqrt{(x-\gamma-1/2)(x-\gamma+1/2)(x+2\gamma)}},$$

$$\omega'(\gamma) = \frac{1}{2}\int\limits_{-2\gamma}^{\gamma+\frac{1}{2}} \frac{dx}{\sqrt{-(x-\gamma-1/2)(x-\gamma+1/2)(x+2\gamma)}},$$

$$\eta(\gamma) = -\frac{1}{2}\int\limits_{\gamma-\frac{1}{2}}^{-2\gamma} \frac{xdx}{\sqrt{(x-\gamma-1/2)(x-\gamma+1/2)(x+2\gamma)}},$$

$$\eta'(\gamma) = -\frac{1}{2}\int\limits_{-2\gamma}^{\gamma+\frac{1}{2}} \frac{xdx}{\sqrt{-(x-\gamma-1/2)(x-\gamma+1/2)(x+2\gamma)}}.$$

Для вывода формул (4.2) можно рассмотреть предельный переход при $\gamma \to -1/6$, воспользовавшись формулой, представляющей σ в виде бесконечного произведения (см. [3]). Имеем

$$\sigma\left(z, \frac{4}{3}, -\frac{8}{27}\right) = z\prod_{n\neq 0}\left(1-\frac{z}{in\pi}\right)e^{\frac{z}{in\pi}-\frac{z^2}{2n^2\pi^2}}.$$

Пользуясь классическими тождествами

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}, \quad \prod_{n=1}^{\infty}\left(1-\frac{x^2}{n^2\pi^2}\right) = \frac{\sin(x)}{x},$$

получаем первое тождество из (4.2). Остальные элементарно получаются из равенств $\zeta(z) = \sigma'(z)/\sigma(z)$, $\wp(z) = -\zeta'(z)$.

Теперь оценим рост $\omega(\gamma)$. Для этого запишем равенство

$$\omega(\gamma) = \frac{1}{2}\int\limits_{0}^{1/2-3\gamma} \frac{dt}{\sqrt{t(1-t)(1/2-3\gamma-t)}}.$$

Заметим, что при малых γ, интеграл по отрезку $[0,1/2]$ имеет ограниченное поведение, а потому

$$\omega(\gamma) \sim \frac{1}{2}\int\limits_{1/2}^{1/2-3\gamma} \frac{dt}{\sqrt{t(1-t)(1/2-3\gamma-t)}}.$$

Оценка интеграла в правой части далее не составляет труда, поскольку множитель $1/\sqrt{t}$ ограничен снизу и сверху положительными константами, а после его исключения остается вычисляемый интеграл. Лемма доказана.

Из того, что $\omega' \to i\pi/2$, а $\omega \to \infty$, естественно предположить, что формула (3.7) может давать конформное отображение полуполосы

$$S = \{z \in \mathbb{C} : \mathrm{Re}z < 0, \mathrm{Im}z \in (0,\pi/2)\}$$

на область $\widetilde{\Omega}$. Положим

$$\widetilde{Q}(z) = Dz + \frac{h^-}{\pi}\ln\left(\frac{\sigma(z-z^-)}{\sigma(z+z^-)}\right) - \frac{h^+}{\pi}\ln\left(\frac{\sigma(z-z^+)}{\sigma(z+z^+)}\right) - i(h^- - h^+), \tag{4.4}$$

где σ взята при $g_2 = 4/3$, $g_3 = -8/27$, а $D \in \mathbb{C}$, $z^-, z^+ \in (0, i\pi/2)$ – параметры. Подставляя выражение (4.2) для σ в (4.4), получаем, что

$$\widetilde{Q}(z) = z\left(D + \frac{2h^-z^-}{3\pi} - \frac{2h^+z^+}{3\pi}\right) + \frac{h^-}{\pi}\ln\frac{\sinh(z-z^-)}{\sinh(z+z^-)} - \frac{h^+}{\pi}\ln\frac{\sinh(z-z^+)}{\sinh(z+z^+)} - i(h^- - h^+). \tag{4.5}$$

Легко проверяется, что при наличии у $\widetilde{Q}$ ненулевого линейного слагаемого, эта функция не имеет предела при $\mathrm{Re} z \to -\infty$. Если же

$$D + \frac{2h^- z^-}{3\pi} - \frac{2h^+ z^+}{3\pi} = 0,$$

то она имеет предел, равный $2(h^- z^- - h^+ z^+)/\pi - i(h^- - h^+)$. Таким образом, для того, чтобы $\widetilde{Q}$ конформно отображала S на $\widetilde{\Omega}$, необходимо выполнение условий

$$\begin{gathered} D + \frac{2h^- z^-}{3\pi} - \frac{2h^+ z^+}{3\pi} = 0, \\ h^- z^- - h^+ z^+ = -\frac{ih\pi}{2}. \end{gathered} \tag{4.6}$$

Очевидно, что условий (4.6) достаточно для этого при дополнительном предположении о том, что производная функции $\widetilde{Q}$ не обращается в ноль на множестве S и на его границе. Достаточно громоздкими (но элементарными) выкладками проверяется, что при выполнении (4.6) для этого необходимо и достаточно выполнение равенства

$$h^- \sinh(2z^-) = h^+ \sinh(2z^+). \tag{4.7}$$

Уравнения (4.6) и (4.7), таким образом, определяют параметры D, z^-, z^+, при которых $\widetilde{Q}$ осуществляет искомое конформное отображение.

Покажем, что построенная формула редуцируется к интегралу Кристоффеля–Шварца в верхней полуплоскости после замены переменной $x = \wp(z) = 1/\sinh^2(z) + 1/3$. Переходя в формуле (4.4) к новой переменной и учитывая, что линейное слагаемое равно нулю, получаем для конформного отображения $\mathbb{C}_+$ на $\widetilde{\Omega}$ выражение

$$\widetilde{W}(x) = \frac{h^-}{\pi} \ln\left(\frac{\sqrt{x^- + 2/3} - \sqrt{x + 2/3}}{\sqrt{x^- + 2/3} + \sqrt{x + 2/3}}\right) + \frac{h^+}{\pi} \ln\left(\frac{\sqrt{x^+ + 2/3} - \sqrt{x + 2/3}}{\sqrt{x^+ + 2/3} + \sqrt{x + 2/3}}\right). \tag{4.8}$$

Полученные ранее уравнения на параметры z^- и z^+ переписываются в виде

$$\begin{gathered} \frac{h^-\sqrt{x^- + 1}}{x^-} = \frac{h^+\sqrt{x^+ + 1}}{x^+}, \\ h^- \ln\left(\frac{1 + \sqrt{x^- + 4/3}}{\sqrt{x^- + 1/3}}\right) + h^+ \ln\left(\frac{1 + \sqrt{x^+ + 4/3}}{\sqrt{x^+ + 1/3}}\right) = -\frac{ih\pi}{2}. \end{gathered} \tag{4.9}$$

Дифференцируя $\widetilde{W}$ и используя первое из уравнений (4.9), приходим к равенству

$$d\widetilde{W} = \frac{h^-\sqrt{x^- + 1} - h^+\sqrt{x^+ + 1}}{\pi} \frac{(x - 1/3)}{\sqrt{x + 2/3}(x - x^-)(x - x^+)}.$$

Таким образом, получен явный вид константы в интеграле Кристоффеля–Шварца. Неизвестными остаются параметры x^- и x^+, которые можно определить из системы (4.9).

4.2. Предельный переход

Рассмотрим последовательность областей, заданных параметрами $(h_n^-, h_n^+, h_n, \delta_n)$. Предположим, что эти параметры имеют предел $(h_{\lim}^-, h_{\lim}^+, h_{\lim}, 0)$. При этом также будем считать, что $h_{\lim}^+ - h_{\lim} > 0$ и $h_{\lim}^- - h_{\lim}^+ + h_{\lim} > 0$. Мы докажем, что параметры $(D_n, \gamma_n, z_n^-, z_n^+)$, полученные решением системы (8) для соответствующих параметров области, имеют предел $(D_{\lim}, -1/6, z_{\lim}^-, z_{\lim}^+)$, причем параметры $(D_{\lim}, z_{\lim}^-, z_{\lim}^+)$ удовлетворяют уравнениям (4.6) и (4.7). Таким образом, с учетом того, что сигма-функция Вейерштрасса является целой, получаем, что построенное решение задачи о конформном отображении устойчиво.

Последующее доказательство достаточно длинное и техническое, а потому мы позволим себе опускать большую часть выкладок.

Нам для оценок также потребуются параметры $x_n^-, x_n^+, x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, C_n$ отображения W_n.

Лемма 4.2. *Предположим, что* $\gamma_n \to -1/6$*, причем* $\delta_n\omega(\gamma_n) \to 0$. *Тогда описанный выше предельный переход имеет место.*

Доказательство. Заметим, что из того, что $\gamma_n \to -1/6$ следует, что последовательности z_n^- и z_n^+ ограничены. Из первого уравнения системы (3.8) следует, что последовательность D_n тоже ограничена. Переходя к подпоследовательностям, можно считать, что данные последовательности сходятся (если удастся доказать, что пределы удовлетворяют уравнениям (4.6) и (4.7), то в силу единственности получим, что все подпоследовательности сходятся к одному и тому же пределу, что влечет сходимость исходной последовательности). Первое уравнение из системы (4.6) получается предельным переходом из второго уравнения системы (3.8) с учетом результатов леммы 4.1. Умножая первое уравнение из системы (3.8) на ω', а второе на ω, вычитая и переходя к пределу, получаем второе уравнение из (4.6) (слагаемое $\delta\omega$ по условию стремится к нулю).

Теперь выведем уравнение (4.7) для предела подпоследовательности. Напомним некоторые обозначения теории эллиптических функций (см. [3]):

$$\zeta_2(z) = \zeta(z + \omega) - \eta,$$
$$\zeta_3(z) = \zeta(z + \omega + \omega') + \eta + \eta'.$$

Эти функции связаны с функциями σ_2, σ_3:

$$\zeta_k = \frac{1}{\sigma_k}\frac{d\sigma_k}{dz} = \frac{d\ln\sigma_k}{dz}.$$

Наконец, $\sigma_k = \sigma\sqrt{\wp - x_k}$. Последние два уравнения системы (3.8) можно записать в виде

$$D + \frac{h^-}{\pi}(\zeta_k(-z^-) - \zeta(z^-)) - \frac{h^+}{\pi}(\zeta_k(-z^+) - \zeta(z^+)) = 0, \quad k = 2,3. \tag{4.10}$$

Запишем

$$\zeta_2(z) - \zeta_3(z) = \frac{\sigma'(z)\sqrt{\wp - x_2} + \wp'(z)\sigma(z)(\wp - x_2)^{-1/2}}{\sigma(z)\sqrt{\wp - x_2}} - \frac{\sigma'(z)\sqrt{\wp - x_3} + \wp'(z)\sigma(z)(\wp - x_3)^{-1/2}}{\sigma(z)\sqrt{\wp - x_3}},$$

откуда следует

$$\zeta_2(z) - \zeta_3(z) = \frac{\wp'(z)(x_2 - x_3)}{(\wp - x_2)(\wp - x_3)}. \tag{4.11}$$

Уравнение (4.7) выводится предельным переходом (при помощи леммы 4.1) из уравнений (4.10) (из которых надо исключить константу D) и подстановкой вместо $\zeta_2 - \zeta_3$ формулы (4.11) Лемма доказана.

Лемма 4.3. *Имеют место неравенства* $|C_n| \geq a_1$, $|C_n| \leq a_2\sqrt{x_3^{(n)} - x_n^-}$ *для некоторых положительных констант* a_1, a_2. *Кроме того, последовательность* x_n^+ *ограничена снизу.*

Доказательство. Оценки на $C^{(n)}$ достаточно элементарно получаются из равенства

$$|C_n|\int_{x_3^{(n)}}^{+\infty} \frac{\sqrt{(x - x_2^{(n)})(x - x_3^{(n)})}dx}{\sqrt{x - x_1^{(n)}}(x - x_n^-)(x - x_n^+)} = h_n^- - h_n^+ + h_n.$$

Чтобы доказать ограниченность последовательности x_n^+ снизу, надо рассмотреть равенство

$$|C_n|\int_{x_1^{(n)}}^{x_2^{(n)}} \frac{\sqrt{(x_2^{(n)} - x)(x_3^{(n)} - x)}dx}{\sqrt{x - x_1^{(n)}}(x - x_n^-)(x - x_n^+)} = h_n.$$

Лемма 4.4. *Предположим, что последовательность x_n^- ограничена снизу. Тогда найдутся такие постоянные $0 < b_1 < b_2$, что $b_1\sqrt{\delta_n} \le |x_2^{(n)} - x_3^{(n)}| \le b_2\sqrt{\delta_n}$.*

Доказательство. Это следует из элементарных оценок на интеграл в равенстве

$$|C_n| \int_{x_2^{(n)}}^{x_3^{(n)}} \frac{\sqrt{(x - x_2^{(n)})(x_3^{(n)} - x)}dx}{\sqrt{x - x_1^{(n)}}(x - x_n^-)(x - x_n^+)} = \delta_n.$$

Из доказанных лемм следует, что достаточно доказать ограниченность последовательности x_n^- снизу (необходимо также принять во внимание асимптотику (4.3)). Предположим, что это не так. Переходя к подпоследовательности, можем считать, что $x_n^- \to -\infty$ и что последовательности $x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, x_n^+$ сходятся.

Лемма 4.5. *В сделанных выше предположениях имеем $x_3^{(n)} - x_2^{(n)} \to 0$, $x_1^{(n)} - x_n^+ \to 0$.*

Доказательство. Из равенства

$$|C_n| \frac{\sqrt{(x_2^{(n)} - x_n^+)(x_3^{(n)} - x_n^+)}}{\sqrt{x_1^{(n)} - x_n^+}(x_n^+ - x_n^-)} = \frac{h_n^+}{\pi}$$

легко выводится сходимость $x_1^{(n)} - x_n^+ \to 0$.

Для доказательства сходимости $x_3^{(n)} - x_2^{(n)} \to 0$ перейдем к параметрам $(D_n, \gamma_n, z_n^-, z_n^+)$. Предположим, что $x_2^{(n)} - x_1^{(n)} \to 0$. Тогда $\omega'(\gamma_n), \eta'(\gamma_n) \to \infty$, а ω и η имеют конечные пределы. Кроме того, из тождества Лежандра (см. [3] или [4]) следует, что

$$\lim_{n\to\infty} \frac{\omega'(\gamma_n)}{\eta'(\gamma_n)} = \lim_{n\to\infty} \frac{\omega(\gamma_n)}{\eta(\gamma_n)}.$$

Записывая первую пару уравнений из (3.8), получаем

$$\lim_{n\to\infty} \left(\frac{2h_n^+}{\pi} z_n^+ \left(\frac{\omega'}{\eta'} - \frac{\omega}{\eta} \right) - i\frac{h_n}{\omega} \right) = 0.$$

Отсюда получаем, что

$$\lim_{n\to\infty} \left(\frac{h_n^+ z_n^+}{\omega'} - h_n \right) = 0,$$

а значит, в пределе $h_{\lim} \ge h_{\lim}^+$. Это противоречит исходным предположениям.

Теперь предположим, что $x_2^{(n)} - x_1^{(n)} \not\to 0$ и $x_3^{(n)} - x_2^{(n)} \not\to 0$. Тогда оба периода ω и ω' имеют конечные пределы. При этом $z_n^- \to 0$, $z_n^+ \to \omega'(\gamma_{\lim})$. Очевидно, что последовательность D_n также имеет предел и, записывая в пределе второе уравнение из (3.8), имеем

$$-D_{\lim}\omega' - \frac{2h_{\lim}^+ \omega' \eta'}{\pi} = 0.$$

Подставляя в первое уравнение, получаем

$$\frac{2h_{\lim}^+}{\pi}\omega\eta' - \frac{2h_{\lim}^+}{\pi}\omega'\eta = -ih_{\lim},$$

что дает равенство $h_{\lim}^+ = h_{\lim}$. Лемма доказана.

Теперь у нас достаточно информации, чтобы вывести противоречие из того, что $x_n^- \to -\infty$. Для этого мы будем отслеживать асимптотики некоторых последовательностей (далее эквивалентность последовательностей используется для обозначения того, что их отношение стремится к 1).

Из равенства

$$|C_n|\frac{\sqrt{(x_2^{(n)}-x_n^-)(x_3^{(n)}-x_n^-)}}{\sqrt{x_1^{(n)}-x_n^-}(x_n^+-x_n^-)}=\frac{h_n^-}{\pi}$$

следует, что

$$|C_n|\sim\frac{h_n^-}{\pi}\sqrt{|x_n^-|}. \tag{4.12}$$

С другой стороны, имеем

$$|C_n|\frac{\sqrt{(x_2^{(n)}-x_n^+)(x_3^{(n)}-x_n^+)}}{\sqrt{x_1^{(n)}-x_n^+}(x_n^+-x_n^-)}=\frac{h_n^+}{\pi},$$

откуда следует, что

$$\sqrt{x_1^{(n)}-x_n^+}\sim\frac{h_n^-}{h_n^+}\frac{1}{\sqrt{|x_n^-|}}. \tag{4.13}$$

Теперь обратимся к равенству

$$|C_n|\int\limits_{x_1^{(n)}}^{x_2^{(n)}}\frac{\sqrt{(x_2^{(n)}-x)(x_3^{(n)}-x)}dx}{\sqrt{x-x_1^{(n)}}(x-x_n^-)(x-x_n^+)}=h_n.$$

Пользуясь (4.12), легко показать, что последовательность в левой части ведет себя эквивалентно последовательности

$$\frac{h_n^-}{\pi\sqrt{|x_n^-|}}\int\limits_{x_1^{(n)}}^{x_2^{(n)}}\frac{\sqrt{(x_2^{(n)}-x)(x_3^{(n)}-x)}dx}{\sqrt{x-x_1^{(n)}}(x-x_n^+)}.$$

Далее, при помощи замены переменной получаем

$$\int\limits_{x_1^{(n)}}^{x_2^{(n)}}\frac{\sqrt{(x_2^{(n)}-x)(x_3^{(n)}-x)}dx}{\sqrt{x-x_1^{(n)}}(x-x_n^+)}=\int\limits_0^{x_2^{(n)}-x_1^{(n)}}\frac{\sqrt{(x_2^{(n)}-x_1^{(n)}-x)(1-x)}dx}{\sqrt{x}(x+x_1^{(n)}-x_n^+)}.$$

Оказывается, что асимптотика последнего интеграла не зависит от характера сходимости $|x_2^{(n)}-x_1^{(n)}|\to 1$. Именно, для любых последовательностей $\alpha_n\to 1$ и $a_n\to 0$ имеет место эквивалентность

$$\int\limits_0^{\alpha_n}\frac{\sqrt{(\alpha_n-x)(1-x)}dx}{\sqrt{x}(x+a_n)}\sim\int\limits_0^1\frac{(1-x)dx}{\sqrt{x}(x+a_n)}\sim\frac{\pi}{\sqrt{a_n}}.$$

Наконец, с учетом (4.13), получаем

$$h_n=|C_n|\int\limits_{x_1^{(n)}}^{x_2^{(n)}}\frac{\sqrt{(x_2^{(n)}-x)(x_3^{(n)}-x)}dx}{\sqrt{x-x_1^{(n)}}(x-x_n^-)(x-x_n^+)}\sim\frac{h_n^-}{\pi\sqrt{|x_n^-|}}\frac{\pi}{\sqrt{x_1^{(n)}-x_n^+}}\sim h_n^+.$$

Получилось противоречие с тем, что $h_{\lim}<h_{\lim}^+$.

5. ЗАКЛЮЧЕНИЕ

Для конформного отображения многоугольной области Ω получено простое выражение через сигма-функцию Вейерштрасса. Для конкретного примера вычислены параметры отображения и проведены численные расчеты. Исследовано поведение при вырождении области и показано, что построенная формула устойчива и имеет предел, осуществляющий решение предельной задачи.

Дальнейшее направление исследований может быть связано как с построением и анализом решений для аналогичных задач, например, связанных с поверхностями рода 2, так и с развитием теории сигма-функций: построением рекуррентных формул для произвольного рода и развитием методов вычисления, не основанных на теории тэта-функций.

Автор выражает благодарность А. Богатыреву и О. Григорьеву за постановку задачи и полезные обсуждения, а также К. Малкову за помощь в компьютерной реализации вычислений. Также автор благодарит центр дополнительного профессионального образования “Университет Сириус” за приглашение на образовательный модуль “Вычислительные технологии, многомерный анализ данных и моделирование” в 2021 г., в ходе которого были получены некоторые из результатов данной статьи.

ПРИЛОЖЕНИЕ

О КОЭФФИЦИЕНТАХ РЯДА ТЕЙЛОРА СИГМА-ФУНКЦИИ ВЕЙЕРШТРАССА

Здесь мы приведем доказательство того, что сигма-функция является целой функцией трех переменных, и выведем рекуррентную формулу для коэффициентов ее разложения в ряд Тейлора, впервые полученную Вейерштрассом в [11]. Доказательство, приведенное там, имеет пробел, связанный с голоморфностью сигма-функции в окрестности нуля. Возможно, данный факт можно доказать при помощи независимого рассуждения, но поскольку Вейерштрасс не приводит никаких ссылок (и вообще не затрагивает этот вопрос), было решено привести здесь более полное доказательство.

Из соотношения однородности

$$\sigma\left(\frac{z}{\lambda},\lambda^4 g_2,\lambda^6 g_3\right)=\lambda\sigma(z,g_2,g_3)$$

легко вывести следующее дифференциальное уравнение на σ-функцию:

$$z\frac{\partial\sigma}{\partial z}-4g_2\frac{\partial\sigma}{\partial g_2}-6g_3\frac{\partial\sigma}{\partial g_3}-\sigma=0. \tag{П.1}$$

Далее, используя определение σ-функции и стандартное дифференциальное уравнение для $\wp$-функции, можно вывести уравнение (доказательство см. в [17])

$$\frac{\partial^2\sigma}{\partial z^2}-12g_3\frac{\partial\sigma}{\partial g_2}-\frac{2}{3}g_2^2\frac{\partial\sigma}{\partial g_3}+\frac{1}{12}g_2z^2\sigma=0. \tag{П.2}$$

Предположим, что f – целая функция трех переменных (z,g_2,g_3), удовлетворяющая уравнениям (П.1) и (П.2). Выведем соотношения между коэффициентами f_{mnk} разложения f в ряд Тейлора:

$$f=\sum_{m,n,k=0}^{\infty}f_{mnk}g_2^m g_3^n z^k.$$

Уравнение (П.1) легко приводит к уравнениям $f_{mnk}=0$, если $k\neq 4m+6n+1$. Значит, f можно записать в виде

$$f=\sum_{m,n=0}^{\infty}a_{mn}g_2^m g_3^n z^{4m+6n+1}.$$

Теперь, подставляя выражение для f в уравнение (П.2), получаем соотношение

$$a_{mn}=\frac{12(m+1)a_{m+1,n-1}+\frac{2}{3}(n+1)a_{m-2,n+1}-\frac{1}{12}a_{m-1,n}}{(4m+6n+1)(4m+6n)}, \tag{П.3}$$

в котором для удобства символ a_{mn} определен нулем для случая, когда m или n отрицательно. Легко видеть, что (П.3) определяет двойную последовательность a_{mn} однозначно, если задано a_{00}. Для доказательства введем на парах неотрицательных целых чисел (m,n) отношение порядка: $(m,n)\le(m',n')$, если $m+n<m'+n'$ или если $m+n=m'+n'$ и $n<n'$. Ясно, что этим задано отно-

шение полного порядка на $\mathbb{Z}_+ \times \mathbb{Z}_+$, причем в формуле (П.3) индексы членов последовательности a_{mn}, находящихся в правой части, все строго меньше пары (m, n). Тем самым доказано, что соотношение (П.3) рекуррентно определяет a_{mn}, если задать a_{00}.

Если бы сигма-функция была целой функцией трех переменных, или, хотя бы, была голоморфной в окрестности нуля, то рекуррентное соотношение (П.3) для ее коэффициентов ряда Тейлора было бы доказано. Проблема состоит в том, что областью определения σ-функции служит множество $\{(z, g_2, g_3) \in \mathbb{C}^3 : g_2^3 - 27g_3^2 \neq 0\}$. Последующие рассуждения доказывают целость функции σ и рекуррентное соотношение (П.3).

Замечание 3. Как известно (см., например, [3] или [4]), условие $g_2^3 - 27g_3^2 \neq 0$ равносильно простоте корней полинома $4x^3 - g_2x - g_3$.

Лемма П.1. *Пусть последовательность a_{nm} удовлетворяет рекуррентному соотношению* (П.3). *Тогда для любого $q > (28 + \sqrt{811})/36 \approx 1.569$ существует такое $C > 0$, что*

$$|a_{mn}| \leq C \frac{q^{2m+3n}}{(2m+3n)!}. \tag{П.4}$$

Доказательство. Подставляя в (П.3) оценку и преобразовывая, легко показать, что для существования константы достаточно выполнения неравенства

$$6\frac{m+1}{4m+6n+1}\frac{q^{2m+3n-1}}{(2m+3n)!} + \frac{n+1}{6(4m+6n+1)}\frac{q^{2m+3n-1}}{(2m+3n)!} + \frac{q^{2m+3n-2}}{48(2m+3n)!} \leq \frac{q^{2m+3n}}{(2m+3n)!}$$

начиная с некоторого индекса (m, n) (в смысле введенного ранее отношения порядка). Для этого, в свою очередь, достаточно выполнения неравенства

$$\frac{1}{48} + q\left(\frac{3}{2} + \frac{1}{18}\right) < q^2.$$

Решая квадратное уравнение, получаем требуемое утверждение. Лемма доказана.

Из леммы П.1 следует, что можно определить целую функцию

$$h(z, g_2, g_3) = \sum_{m,n=0}^{\infty} a_{mn} g_2^m g_3^n z^{4m+6n+1},$$

где коэффициенты a_{mn} определены рекуррентным соотношением (П.3) и начальным условием $a_{00} = 1$. Мы докажем, что $h \equiv \sigma$ для таких значений (g_2, g_3), что $g_2^3 - 27g_3^2 \neq 0$.

Лемма П.2. *Пусть f – голоморфная функция переменных (z, g_2, g_3), заданная на множестве $\mathbb{C} \times U$, где $U \subset \mathbb{C}^2$ открыто, и удовлетворяющая там уравнению* (П.2). *Предположим также, что f нечетна по z. Тогда f представима в виде ряда*

$$f(z, g_2, g_3) = \sum_{n=0}^{\infty} c_n(g_2, g_3) z^{2n+1}, \tag{П.5}$$

причем на U выполнено рекуррентное соотношение

$$(2n+3)(2n+2)c_{n+1} - 12g_3\frac{\partial c_n}{\partial g_2} - \frac{2}{3}g_2^2\frac{\partial c_n}{\partial g_3} + \frac{1}{12}g_2 c_{n-1}, \tag{П.6}$$

где $n \geq 0$ (при $n = 0$ положим $c_{n-1} = 0$).

Доказательство. Действительно, представимость f в виде ряда следует из того, что f – целая функция по z. Ее коэффициенты $c_n(g_2, g_3)$ имеют вид

$$c_n(g_2, g_3) = \frac{1}{(2n+1)!}\frac{\partial^{2n+1} f}{\partial z^{2n+1}}\bigg|_{z=0}.$$

Легко видеть, что ряд (П.5) можно почленно дифференцировать, а потому его можно подставить в уравнение (П.2). Собирая коэффициент при z^{2n+1}, получаем равенство (П.6). Лемма доказана.

Рекуррентное соотношение (П.6) можно использовать для доказательства того, что σ и h совпадают на области определения σ-функции. Действительно, если первый член в разложении этих функций по z совпадает, то эти функции также совпадают (заметим, что они обе нечетны по z). Действительно, $\partial\sigma/\partial z|_{z=0} \equiv \partial h/\partial z|_{z=0} \equiv 1$. Значит, h – аналитическое продолжение σ-функции до целой функции переменных (z, g_2, g_3). Этим доказана следующая теорема.

Теорема П.1 (Вейерштрасс). *Имеет место представление σ-функции в виде ряда*

$$\sigma(z, g_2, g_3) = \sum_{m,n=0}^{\infty} a_{mn} g_2^m g_3^n z^{4m+6n+1}, \qquad (\text{П.7})$$

где коэффициенты a_{mn} определены рекуррентным соотношением (П.3) *и начальным условием $a_{00} = 1$.*

СПИСОК ЛИТЕРАТУРЫ

1. *Лаврентьев М.А., Шабат Б.В.* Методы теории функций комплексного переменного. М.: Наука, 1987.
2. *Driscoll T.A., Trefethen L.N.* Schwartz-Cristoffel mapping. Cambridge: Cambridge University Press, 2002.
3. *Ахиезер Н.И.* Элементы теории эллиптических функций. М.: Наука, 1970.
4. *Chandrasekharan K.* Elliptic functions. Berlin: Springer-Verlag, 1985.
5. *Bogatyrev A.B., Hassner M., Yarmolich D.* An exact analytical-expression for the read sensor signal in magnetic data storage channels // Contemp. Math. 2010. V. 523. P. 155–160.
6. *Bogatyrev A.B.* The conformal mapping of rectangular heptagons // Mat. Sb. 2012. V. 203. N. 12. P. 35–56.
7. *Bogatyrev A.B., Grigor'ev O.A.* Conformal mapping of rectangular heptagons // Comput. Methods Funct. Theory. 2018. V. 18. N. 2. P. 221–238.
8. *Богатырев А.Б., Григорьев О.А.* Фильтрация под ступенчатой плотиной и римановы тета-функции // Труды матем. института им. В.А. Стеклова. 2020. V. 311. P. 14–26.
9. *Bezrodnykh S.I., Bogatyrev A.B., Goreinov S.A., Grigor'ev O.A., Hakula H., Vuorinen M.* On capacity computation for symmetric polygonal condensers // J. Comput. Appl. Math. 2019. V. 361. P. 271–282.
10. *Farkas H.M., Kra I.* Riemann surfaces. New-York: Springer-Verlag, 1992.
11. *Weierstrass K.* Zur Theorie der elliptischen Funktionen // Sitzungsberichte der Akademie der Wissenshaften zu Berlin. 1882. V. 1. P. 443–451.
12. *Klein F.* Uber hyperelliptische Sigmafunktionen // Math. Ann. 1886. V. 27. P. 431–464.
13. *Baker H.F.* On the hyperelliptic sigma-functions // Amer. J. Math. 1898. V. 20. P. 301–384.
14. *Buchstaber V.M., Leykin D.V., Enolskii V.Z.* Kleinian functions, hyperelliptic Jacobians and applications // Reviews in Mathematics and Math. Physics. 1997. V. 10. N. 2. P. 3–120.
15. *Картан А.* Элементарная теория функций одного и нескольких комплексных переменных. М.: Изд-во иностр. лит., 1963.
16. *Форстер О.* Римановы поверхности. М.: Мир, 1980.
17. *Halphen G.H.* Traite des fonctions elliptiques et de leurs applications. T. 1. Paris: Gauthier-Villars, 1886.

EDN: OUAAJH
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 838–853

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 519.64

ИССЛЕДОВАНИЕ ПРИБЛИЖЕННОГО РЕШЕНИЯ ОДНОГО КЛАССА СИСТЕМ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

© 2022 г. Э. Г. Халилов

AZ 1010 Баку, пр-т Азадлыг, 20, Азербайджанский Государственный Университет Нефти и Промышленности, Азербайджан

e-mail: elnurkhalil@mail.ru

Поступила в редакцию 20.08.2021 г.
Переработанный вариант 20.08.2021 г.
Принята к публикации 17.11.2021 г.

Дано обоснование метода коллокации для системы интегральных уравнений граничной задачи сопряжения для уравнения Гельмгольца в двухмерном пространстве. Построены квадратурные формулы для потенциалов простого и двойного слоев и нормальной производной потенциала простого слоя. В определенно выбранных точках система интегральных уравнений заменяется системой алгебраических уравнений, при этом устанавливаются существование и единственность решения системы алгебраических уравнений. Доказывается сходимость решения системы алгебраических уравнений к точному решению системы интегральных уравнений и указывается скорость сходимости метода. Кроме того, построена последовательность, сходящаяся к точному решению граничной задачи сопряжения. Библ. 16.

Ключевые слова: граничная задача сопряжения, уравнение Гельмгольца, система интегральных уравнений, потенциалы простого и двойного слоев, функция Ханкеля, квадратурные формулы, метод коллокации.

DOI: 10.31857/S0044466922050064

1. ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

Пусть $D \subset R^2$ – ограниченная область с дважды непрерывно дифференцируемой границей L. Следует указать, что (см. [1, с. 112]) математическая формулировка задачи дифракции акустических волн на теле D с различными акустическими характеристиками в $R^2\backslash\overline{D}$ и D приводит к задаче сопряжения, которая заключается в следующем: найти две функции $u \in C^{(2)}\left(R^2\backslash\overline{D}\right) \cap C\left(R^2\backslash D\right)$ и $u_0 \in C^{(2)}(D) \cap C(\overline{D})$, обладающие нормальной производной в смысле равномерной сходимости и удовлетворяющие уравнениям Гельмгольца $\Delta u + k^2 u = 0$ в $R^2\backslash\overline{D}$ и $\Delta u_0 + k^2 u_0 = 0$ в D, условию излучения Зоммерфельда

$$\left(\frac{x}{|x|}, \operatorname{grad} u(x)\right) - iku(x) = o\left(\frac{1}{|x|^{1/2}}\right), \quad x \to \infty,$$

равномерно по всем направлениям $x/|x|$ и условиям сопряжения

$$\mu u - \mu_0 u_0 = f \text{ на } L,$$

$$\frac{\partial u}{\partial \nu} - \frac{\partial u_0}{\partial \nu} = g \text{ на } L,$$

где Δ – оператор Лапласа, k и k_0 – волновые числа, причем $\operatorname{Im} k \geq 0$ и $\operatorname{Im} k_0 \geq 0$, $\nu(y)$ – внешняя единичная нормаль в точке $y \in L$, f и g – заданные непрерывные функции на L, а μ и μ_0 – заданные комплексные числа, причем $\mu + \mu_0 \neq 0$. Отметим, что с физической точки зрения надлежащий выбор постоянных μ и μ_0 гарантирует непрерывность давления и нормальной скорости акустических волн при переходе через границу L.

Пусть $\Phi_k(x,y)$ – фундаментальное решение уравнения Гельмгольца, т.е.

$$\Phi_k(x,y)=\begin{cases}\dfrac{1}{2\pi}\ln\dfrac{1}{|x-y|} & \text{при} \quad k=0,\\ \dfrac{i}{4}H_0^{(1)}(k|x-y|) & \text{при} \quad k\neq 0,\end{cases}$$

где $H_0^{(1)}$ – функция Ханкеля I рода нулевого порядка, определяемая формулой $H_0^{(1)}(z)=J_0(z)+iN_0(z)$,

$$J_0(z)=\sum_{m=0}^{\infty}\frac{(-1)^m}{(m!)^2}\left(\frac{z}{2}\right)^{2m}$$

есть функция Бесселя нулевого порядка,

$$N_0(z)=\frac{2}{\pi}\left(\ln\frac{z}{2}+C\right)J_0(z)+\sum_{m=1}^{\infty}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}}{(m!)^2}\left(\frac{z}{2}\right)^{2m}$$

есть функция Неймана нулевого порядка, а $C=0.57721\ldots$ – постоянная Эйлера. Кресс и Роч (см. [2]) доказали, что комбинация потенциалов простого и двойного слоев

$$u(x)=\int_L\left\{\frac{\partial\Phi_k(x,y)}{\partial\nu(y)}\psi(y)+\mu\Phi_k(x,y)\varphi(y)\right\}dL_y,\quad x\in R^2\backslash\overline{D},$$

$$u_0(x)=\int_L\left\{\frac{\partial\Phi_{k_0}(x,y)}{\partial\nu(y)}\psi(y)+\mu_0\Phi_{k_0}(x,y)\varphi(y)\right\}dL_y,\quad x\in D,$$

с непрерывными плотностями ψ и φ, является решением задачи сопряжения, если ψ и φ являются решениями системы интегральных уравнений

$$\begin{aligned}&\psi+\left(\frac{\mu}{\mu+\mu_0}K-\frac{\mu_0}{\mu+\mu_0}K_0\right)\psi+\left(\frac{\mu^2}{\mu+\mu_0}S-\frac{\mu_0^2}{\mu+\mu_0}S_0\right)\varphi=\frac{2f}{\mu+\mu_0},\\ &\varphi-\frac{1}{\mu+\mu_0}(T-T_0)\psi-\left(\frac{\mu}{\mu+\mu_0}\tilde{K}-\frac{\mu_0}{\mu+\mu_0}\tilde{K}_0\right)\varphi=-\frac{2g}{\mu+\mu_0},\end{aligned}\tag{1.1}$$

где

$$(S\varphi)(x)=2\int_L\Phi_k(x,y)\varphi(y)\,dL_y,\quad x\in L,\tag{1.2}$$

$$(K\psi)(x)=2\int_L\frac{\partial\Phi_k(x,y)}{\partial\nu(y)}\psi(y)\,dL_y,\quad x\in L,\tag{1.3}$$

$$(\tilde{K}\varphi)(x)=2\int_L\frac{\partial\Phi_k(x,y)}{\partial\nu(x)}\varphi(y)\,dL_y,\quad x\in L,\tag{1.4}$$

$$((T-T_0)\psi)(x)=2\int_L\frac{\partial}{\partial\nu(x)}\left(\frac{\partial\left(\Phi_k(x,y)-\Phi_{k_0}(x,y)\right)}{\partial\nu(y)}\right)\psi(y)\,dL_y,\quad x\in L,\tag{1.5}$$

и

$$\Phi_{k_0}(x,y)=\Phi_k(x,y)\big|_{k=k_0},\quad S_0=S\big|_{k=k_0},\quad K_0=K\big|_{k=k_0},\quad \tilde{K}_0=\tilde{K}\big|_{k=k_0}.$$

Отметим, что ряд работ посвящены исследованию приближенных решений интегральных уравнений различных краевых задач для уравнения Гельмгольца (см. [3]–[7]), а в работе же [8] дано обоснование метода коллокации для системы интегральных уравнений задачи сопряжения для уравнения Гельмгольца в трехмерном пространстве. Однако до сих пор не исследованы приближенные решения задачи сопряжения для уравнения Гельмгольца в двухмерном пространстве

методом интегральных уравнений. Как известно, в трехмерном пространстве фундаментальное решение уравнения Гельмгольца имеет вид

$$\Phi_k(x,y)=\frac{\exp(ik|x-y|)}{4\pi|x-y|},\quad x,y\in R^3,\quad x\neq y,$$

и поэтому интегральные операторы, участвующие в системе (1.1) строго отличаются от интегральных операторов, участвующих в системе интегральных уравнений для задачи сопряжения для уравнения Гельмгольца в трехмерном пространстве. Кроме того, в [9] построена квадратурная формула для логарифмических потенциалов простого и двойного слоев, а в [10] построена квадратурная формула для потенциалов простого и двойного слоев. Однако в [10] для построения квадратурных формул использована асимптотическая формула для функций Ханкеля I рода нулевого порядка, которая не дает возможность определить скорость сходимости этих квадратурных формул. Поэтому более практичным способом построения квадратурных формул для потенциалов простого и двойного слоев, а также исследование приближенного решения задачи сопряжения для уравнения Гельмгольца в двухмерном пространстве методом системы интегральных уравнений (1.1) имеет важные значения, чему и посвящена настоящая заметка.

2. ПОСТРОЕНИЕ КВАДРАТУРНЫХ ФОРМУЛЫ ДЛЯ ИНТЕГРАЛОВ (1.2)–(1.5)

Предположим, что замкнутая и дважды непрерывно дифференцируемая кривая $L\subset R^2$ задана параметрическим уравнением $x(t)=(x_1(t),x_2(t))$, $t\in[a,b]$. Разобьем промежуток $[a,b]$ на $n>2M_0(b-a)/d$ равных частей: $t_p=a+\dfrac{(b-a)p}{n}$, $p=\overline{0,n}$, где $M_0=\max\limits_{t\in[a,b]}\sqrt{\left(x_1'(t)\right)^2+\left(x_2'(t)\right)^2}<+\infty$ (см. [11, с. 560]) и d – стандартный радиус (см. [12, с. 400]). В качестве опорных точек возьмем $x(\tau_p)$, $p=\overline{1,n}$, где $\tau_p=a+\dfrac{(b-a)(2p-1)}{2n}$. Тогда кривая L разбивается на элементарные части: $L=\bigcup_{p=1}^{n}L_p$, где $L_p=\{x(t):t_{p-1}\le t\le t_p\}$.

Известно, что (см. [9])

(1) $\forall p\in\{1,2,\ldots,n\}$: $r_p(n)\sim R_p(n)$, где $r_p(n)=\min\{|x(\tau_p)-x(t_{p-1})|,|x(t_p)-x(\tau_p)|\}$, $R_p(n)=\max\{|x(\tau_p)-x(t_{p-1})|,|x(t_p)-x(\tau_p)|\}$, а запись $a(n)\sim b(n)$ означает, что

$$C_1\le\frac{a(n)}{b(n)}\le C_2,$$

где C_1 и C_2 – положительные постоянные, не зависящие от n;

(2) $\forall p\in\{1,2,\ldots,n\}$: $R_p(n)\le d/2$;

(3) $\forall p,j\in\{1,2,\ldots,n\}$: $r_j(n)\sim r_p(n)$;

(4) $r(n)\sim R(n)\sim\dfrac{1}{n}$, где $R(n)=\max\limits_{p=1,n}R_p(n)$, $r(n)=\min\limits_{p=1,n}r_p(n)$.

В дальнейшем такое разбиение будем называть разбиением кривой L на “регулярные” элементарные части.

Поступая точно также, как и в доказательстве леммы 2.1 работы [13], можно показать справедливость следующей леммы.

Лемма 1. *Существуют такие постоянные $C_0'>0$ и $C_1'>0$, не зависящие от n, для которых при $\forall p,j\in\{1,2,\ldots,n\}$, $j\neq p$, и $\forall y\in L_j$ справедливы следующие неравенства*:

$$C_0'|y-x(\tau_p)|\le|x(\tau_j)-x(\tau_p)|\le C_1'|y-x(\tau_p)|.$$

Через $C(L)$ обозначим пространство всех непрерывных функций на L с нормой $\|\varphi\|_\infty=\max\limits_{x\in L}|\varphi(x)|$, и для функции $\varphi\in C(L)$ вводим модуль непрерывности вида

$$\omega(\varphi,\delta)=\max_{\substack{|x-y|\le\delta\\ x,y\in L}}|\varphi(x)-\varphi(y)|,\qquad\delta>0.$$

Сначала построим квадратурную формулу для интеграла (1.2). Пусть

$$\Phi_k^n(x,y) = \frac{i}{4} H_{0,n}^{(1)}(k|x-y|), \quad x,y \in L, \quad x \neq y,$$

где

$$H_{0,n}^{(1)}(z) = J_{0,n}(z) + iN_{0,n}(z),$$

$$J_{0,n}(z) = \sum_{m=0}^{n} \frac{(-1)^m}{(m!)^2} \left(\frac{z}{2}\right)^{2m},$$

и

$$N_{0,n}(z) = \frac{2}{\pi}\left(\ln\frac{z}{2} + C\right) J_{0,n}(z) + \sum_{m=1}^{n} \left(\sum_{l=1}^{m} \frac{1}{l}\right) \frac{(-1)^{m+1}}{(m!)^2} \left(\frac{z}{2}\right)^{2m}.$$

Теорема 1. *Пусть L — замкнутая и дважды непрерывно дифференцируемая кривая в R^2 и $\varphi \in C(L)$. Тогда выражение*

$$(S_n\varphi)(x(\tau_p)) = \frac{2(b-a)}{n} \sum_{\substack{j=1 \\ j\neq p}}^{n} \Phi_k^n(x(\tau_p), x(\tau_j)) \sqrt{\left(x_1'(\tau_j)\right)^2 + \left(x_2'(\tau_j)\right)^2}\, \varphi(x(\tau_j))$$

в опорных точках $x(\tau_p)$, $p = \overline{1,n}$, является квадратурной формулой для интеграла (1.2)*, причем справедлива следующая оценка*:

$$\max_{p=\overline{1,n}} |(S\varphi)(x(\tau_p)) - (S_n\varphi)(x(\tau_p))| \leq M\left(\omega(\varphi, 1/n) + \|\varphi\|_\infty \frac{\ln n}{n}\right).$$

(Здесь и далее через M будем обозначать положительные постоянные, разные в различных неравенствах.)

Доказательство. Несложно заметить, что

$$(S\varphi)(x(\tau_p)) - (S_n\varphi)(x(\tau_p)) = 2\int_{L_p} \Phi_k(x(\tau_p), y)\varphi(y) dL_y +$$

$$+ 2\sum_{\substack{j=1 \\ j\neq p}}^{n} \int_{L_j} \left(\Phi_k(x(\tau_p), y) - \Phi_k^n(x(\tau_p), x(\tau_j))\right)\varphi(y)\, dL_y +$$

$$+ 2\sum_{\substack{j=1 \\ j\neq p}}^{n} \int_{L_j} \Phi_k^n(x(\tau_p), x(\tau_j))(\varphi(y) - \varphi(x(\tau_j)))\, dL_y +$$

$$+ 2\sum_{\substack{j=1 \\ j\neq p}}^{n} \int_{t_{j-1}}^{t_j} \Phi_k^n(x(\tau_p), x(\tau_j)) \left(\sqrt{\left(x_1'(t)\right)^2 + \left(x_2'(t)\right)^2} - \sqrt{\left(x_1'(\tau_j)\right)^2 + \left(x_2'(\tau_j)\right)^2}\right) \varphi(x(\tau_j))\, dt.$$

Слагаемые в последнем равенстве обозначим через $h_1^n(x(\tau_p))$, $h_2^n(x(\tau_p))$, $h_3^n(x(\tau_p))$ и $h_4^n(x(\tau_p))$ соответственно.

Очевидно, что

$$|J_0(k|x-y|)| \leq \sum_{m=0}^{\infty} \frac{(|k|\operatorname{diam} L)^{2m}}{4^m (m!)^2} \leq M \quad \forall x,y \in L, \tag{2.1}$$

и

$$\left|\sum_{m=1}^{\infty} \left(\sum_{l=1}^{m} \frac{1}{l}\right) \frac{(-1)^{m+1}}{(m!)^2} \left(\frac{k|x-y|}{2}\right)^{2m}\right| \leq \sum_{m=1}^{\infty} \left(\sum_{l=1}^{m} \frac{1}{l}\right) \frac{(|k|\operatorname{diam} L)^{2m}}{4^m (m!)^2} \leq M \quad \forall x,y \in L, \tag{2.2}$$

следовательно,

$$|\Phi_k(x,y)| \le M\,|\ln|x-y|| \quad \forall x,y\in L, \quad x\ne y. \tag{2.3}$$

Тогда, применяя формулу вычисления криволинейного интеграла, находим

$$\left|h_1^n(x(\tau_p))\right| \le 2\|\varphi\|_\infty \int\limits_{L_p} |\Phi_k(x(\tau_p),y)|dL_y \le M\|\varphi\|_\infty \int\limits_0^{R(n)} |\ln\tau|\,d\tau \le M\|\varphi\|_\infty R(n)|\ln R(n)|.$$

Пусть $y\in L_j$ и $j\ne p$. Учитывая лемму 1, имеем

$$\left||x(\tau_p)-y|^q - |x(\tau_p)-x(\tau_j)|^q\right| \le Mq|x(\tau_j)-y||x(\tau_p)-y|^{q-1} \le MqR(n)(\operatorname{diam}L)^{q-1} \tag{2.4}$$

и

$$\begin{aligned}|\ln(k|x(\tau_p)-y|) - \ln(k|x(\tau_p)-x(\tau_j)|)| &= \left|\ln\left(1+\frac{|x(\tau_p)-x(\tau_j)|-|x(\tau_p)-y|}{|x(\tau_p)-y|}\right)\right| \le \\ &\le \left|\ln\left(1+\frac{|x(\tau_j)-y|}{|x(\tau_p)-y|}\right)\right| \le M\frac{R(n)}{|x(\tau_p)-y|},\end{aligned} \tag{2.5}$$

где $q\in \mathrm{N}$. Тогда, принимая во внимание неравенства (2.1), (2.2), (2.4) и (2.5), получаем, что

$$\begin{aligned}&|\Phi_k(x(\tau_p),y)-\Phi_k(x(\tau_p),x(\tau_j))| \le \\ &\le \frac{1}{4}\left|\sum_{m=0}^{\infty}\frac{(-1)^m}{(m!)^2}\left(\left(\frac{k|x(\tau_p)-y|}{2}\right)^{2m}-\left(\frac{k|x(\tau_p)-x(\tau_j)|}{2}\right)^{2m}\right)\right| + \\ &+\frac{1}{2\pi}\left|\left(\ln\frac{k|x(\tau_p)-x(\tau_j)|}{2}+C\right)\sum_{m=0}^{\infty}\frac{(-1)^m}{(m!)^2}\left(\left(\frac{k|x(\tau_p)-y|}{2}\right)^{2m}-\left(\frac{k|x(\tau_p)-x(\tau_j)|}{2}\right)^{2m}\right)\right| + \\ &+\frac{1}{2\pi}\left|\left(\ln(k|x(\tau_p)-y|)-\ln(k|x(\tau_p)-x(\tau_j)|)\right)\sum_{m=0}^{\infty}\frac{(-1)^m}{(m!)^2}\left(\frac{k|x(\tau_p)-y|}{2}\right)^{2m}\right| + \\ &+\frac{1}{4}\left|\sum_{m=1}^{\infty}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}}{(m!)^2}\left(\left(\frac{k|x(\tau_p)-y|}{2}\right)^{2m}-\left(\frac{k|x(\tau_p)-x(\tau_j)|}{2}\right)^{2m}\right)\right| \le \frac{MR(n)}{|x(\tau_p)-y|}.\end{aligned}$$

Кроме того, учитывая неравенства

$$|J_0(k|x-y|)-J_{0,n}(k|x-y|)| \le \sum_{m=n+1}^{\infty}\frac{|k|^{2m}|x-y|^{2m}}{4^m(m!)^2} \le \frac{M}{(n+1)!} \quad \forall x,y\in L, \tag{2.6}$$

и

$$|N_0(k|x-y|)-N_{0,n}(k|x-y|)| \le \frac{M|\ln|x-y||}{(n+1)!} \quad \forall x,y\in L, \tag{2.7}$$

имеем

$$\left|\Phi_k(x(\tau_p),x(\tau_j))-\Phi_k^n(x(\tau_p),x(\tau_j))\right| \le \frac{M|\ln|x(\tau_p)-x(\tau_j)||}{(n+1)!} \le \frac{M}{(n+1)!|x(\tau_p)-y|}.$$

В результате находим, что

$$\begin{aligned}&\left|\Phi_k(x(\tau_p),y)-\Phi_k^n(x(\tau_p),x(\tau_j))\right| \le |\Phi_k(x(\tau_p),y)-\Phi_k(x(\tau_p),x(\tau_j))| + \\ &+\left|\Phi_k(x(\tau_p),x(\tau_j))-\Phi_k^n(x(\tau_p),x(\tau_j))\right| \le \frac{M}{|x(\tau_p)-y|}\left(R(n)+\frac{1}{(n+1)!}\right).\end{aligned}$$

Следовательно,

$$\left|h_2^n(x(\tau_p))\right| \le M\|\varphi\|_\infty\left(R(n)+\frac{1}{(n+1)!}\right)\int\limits_{r(n)}^{\operatorname{diam}L}\frac{d\tau}{\tau} \le M\|\varphi\|_\infty\left(R(n)+\frac{1}{(n+1)!}\right)\left|\ln R(n)\right|.$$

Из неравенства (2.3), получаем, что

$$\int\limits_L \left|\Phi_k(x,y)\right| dL_y$$

сходится как несобственный и

$$\int\limits_L \left|\Phi_k(x,y)\right| dL_y \le M \quad \forall x \in L.$$

Тогда из неравенства (2.6) и (2.7) получим

$$\int\limits_L \left|\Phi_k^n(x,y)\right| dL_y \le \int\limits_L \left|\Phi_k(x,y)\right| dL_y + \int\limits_L \left|\Phi_k(x,y)-\Phi_k^n(x,y)\right| dL_y \le M \quad \forall x \in L, \quad \forall n \in \mathrm{N}.$$

В итоге, принимая во внимание лемму 1, получаем, что

$$\left|h_3^n(x(\tau_p))\right| \le M\omega(\varphi,R(n))\int\limits_L \left|\Phi_k^n(x(\tau_p),y)\right| dL_y \le M\omega(\varphi,R(n)).$$

Очевидно, что

$$\left|\sqrt{\left(x_1'(t)\right)^2+\left(x_2'(t)\right)^2}-\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}\right| \le MR(n) \quad \forall t \in [t_{j-1},t_j]. \tag{2.8}$$

Пусть $y \in L_j$ и $j \ne p$. Учитывая леммы 1 и неравенства (2.1) и (2.2), имеем

$$\left|J_{0,n}(k|x(\tau_p)-x(\tau_j)|)\right| \le \sum_{m=0}^{n}\frac{(|k|\operatorname{diam}L)^{2m}}{4^m(m!)^2} \le M \quad \forall n \in \mathrm{N},$$

и

$$\left|N_{0,n}(k|x(\tau_p)-x(\tau_j)|)\right| \le M\left|\ln|x(\tau_p)-y|\right| \quad \forall n \in \mathrm{N},$$

следовательно,

$$\left|\Phi_k^n(x(\tau_p),x(\tau_j))\right| \le M\left|\ln|x(\tau_p)-y|\right| \quad \forall n \in \mathrm{N}.$$

Отсюда получаем, что

$$\left|h_4^n(x(\tau_p))\right| \le M\|\varphi\|_\infty R(n)\sum_{\substack{j=1\\ j\ne p}}^{n}\int\limits_{t_{j-1}}^{t_j}\left|\Phi_k^n(x(\tau_p),x(\tau_j))\right| dt \le$$

$$\le M\|\varphi\|_\infty R(n)\sum_{\substack{j=1\\ j\ne p}}^{n}\int\limits_{L_j}\left|\Phi_k^n(x(\tau_p),x(\tau_j))\right| dL_y \le M\|\varphi\|_\infty R(n)\int\limits_L\left|\ln|x(\tau_p)-y|\right| dL_y \le M\|\varphi\|_\infty R(n).$$

В результате, суммируя полученные оценки для выражений $h_1^n(x(\tau_p))$, $h_2^n(x(\tau_p))$, $h_3^n(x(\tau_p))$ и $h_4^n(x(\tau_p))$, и, принимая во внимание соотношение $R(n) \sim \frac{1}{n}$, доказываем справедливость теоремы 1.

Теперь построим квадратурную формулу для интеграла (1.3). Нетрудно показать, что

$$\frac{\partial\Phi_k^n(x,y)}{\partial\nu(y)} = \frac{i}{4}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(y)}+i\frac{\partial N_{0,n}(k|x-y|)}{\partial\nu(y)}\right),$$

где

$$\frac{\partial J_{0,n}(k|x-y|)}{\partial \nu(y)} = (y-x,\nu(y))\sum_{m=1}^{n}\frac{(-1)^m k^{2m}|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}$$

и

$$\frac{\partial N_{0,n}(k|x-y|)}{\partial \nu(y)} = \frac{2}{\pi}\left(\ln\frac{k|x-y|}{2}+C\right)\frac{\partial J_{0,n}(k|x-y|)}{\partial \nu(y)} + \frac{2(y-x,\nu(y))}{\pi|x-y|^2}J_{0,n}(k|x-y|) +$$
$$+ (y-x,\nu(y))\sum_{m=1}^{n}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}k^{2m}|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}.$$

Справедлива следующая

Теорема 2. *Пусть L — замкнутая и дважды непрерывно дифференцируемая кривая в R^2 и $\psi \in C(L)$. Тогда выражение*

$$(K_n\psi)(x(\tau_p)) = \frac{2(b-a)}{n}\sum_{\substack{j=1\\ j\neq p}}^{n}\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}\,\psi(x(\tau_j))$$

в опорных точках $x(\tau_p)$, $p=\overline{1,n}$, является квадратурной формулой для интеграла (1.3)*, причем справедлива следующая оценка*:

$$\max_{p=1,n}\left|(K\psi)(x(\tau_p))-(K_n\psi)(x(\tau_p))\right| \le M\left(\omega(\psi,1/n)+\|\psi\|_\infty\frac{\ln n}{n}\right).$$

Доказательство. Нетрудно увидеть, что

$$(K\psi)(x(\tau_p))-(K_n\psi)(x(\tau_p)) = 2\int_{L_p}\frac{\partial\Phi_k(x(\tau_p),y)}{\partial\nu(y)}\psi(y)dL_y +$$
$$+ 2\sum_{\substack{j=1\\ j\neq p}}^{n}\int_{L_j}\left(\frac{\partial\Phi_k(x(\tau_p),y)}{\partial\nu(y)}-\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right)\psi(y)\,dL_y +$$
$$+ 2\sum_{\substack{j=1\\ j\neq p}}^{n}\int_{L_j}\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\left(\psi(y)-\psi(x(\tau_j))\right)dL_y +$$
$$+ 2\sum_{\substack{j=1\\ j\neq p}}^{n}\int_{t_{j-1}}^{t_j}\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\left(\sqrt{\left(x_1'(t)\right)^2+\left(x_2'(t)\right)^2}-\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}\right)\psi(x(\tau_j))\,dt.$$

Слагаемые в последнем равенстве обозначим через $\delta_1^n(x(\tau_p))$, $\delta_2^n(x(\tau_p))$, $\delta_3^n(x(\tau_p))$ и $\delta_4^n(x(\tau_p))$ соответственно.

Легко вычислить, что

$$\frac{\partial\Phi_k(x,y)}{\partial\nu(y)} = \frac{i}{4}\left(\frac{\partial J_0(k|x-y|)}{\partial\nu(y)}+i\frac{\partial N_0(k|x-y|)}{\partial\nu(y)}\right),$$

здесь

$$\frac{\partial J_0(k|x-y|)}{\partial\nu(y)} = (y-x,\nu(y))\sum_{m=1}^{\infty}\frac{(-1)^m k^{2m}|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}$$

и

$$\frac{\partial N_0(k|x-y|)}{\partial \nu(y)} = \frac{2}{\pi}\left(\ln\frac{k|x-y|}{2} + C\right)\frac{\partial J_0(k|x-y|)}{\partial \nu(y)} + \frac{2(y-x,\nu(y))}{\pi|x-y|^2}J_0(k|x-y|) +$$
$$+ (y-x,\nu(y))\sum_{m=1}^{\infty}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}k^{2m}|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}.$$

Так как (см. [12, с. 403])

$$|(y-x,\nu(y))| \le M|x-y|^2, \tag{2.9}$$

то

$$\left|\frac{\partial J_0(k|x-y|)}{\partial \nu(y)}\right| \le M|x-y|^2 \tag{2.10}$$

и

$$\left|\frac{\partial N_0(k|x-y|)}{\partial \nu(y)}\right| \le M\left(|x-y|^2|\ln|x-y|| + |x-y|^2 + 1\right), \tag{2.11}$$

а значит,

$$\left|\frac{\partial \Phi_k(x,y)}{\partial \nu(y)}\right| \le M \quad \forall\, x, y \in L, \quad x \ne y. \tag{2.12}$$

Тогда, учитывая формулу вычисления криволинейного интеграла, получаем

$$\left|\delta_1^n(x(\tau_p))\right| \le M\|\psi\|_\infty \int_0^{R(n)} d\tau \le M\|\psi\|_\infty R(n).$$

Пусть $y \in L_j$ и $j \ne p$. Из леммы 1 и неравенства (2.9) очевидно, что

$$\begin{gathered}|(y-x(\tau_p),\nu(y)) - (x(\tau_j)-x(\tau_p),\nu(x(\tau_j)))| = |(y-x(\tau_j),\nu(y))| + \\ + |(x(\tau_j)-x(\tau_p),\nu(y)-\nu(x(\tau_j)))| \le M|y-x(\tau_p)|R(n).\end{gathered} \tag{2.13}$$

Тогда, учитывая неравенства (2.4), получаем, что

$$\begin{gathered}\left|\frac{\partial J_0(k|x(\tau_p)-y|)}{\partial \nu(y)} - \frac{\partial J_0(k|x(\tau_p)-x(\tau_j)|)}{\partial \nu(x(\tau_j))}\right| \le \\ \le |(y-x(\tau_p),\nu(y)) - (x(\tau_j)-x(\tau_p),\nu(x(\tau_j)))| \times \\ \times \sum_{m=1}^{\infty}\frac{|k|^{2m}|x(\tau_p)-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!} + |(x(\tau_j)-x(\tau_p),\nu(x(\tau_j)))| \times \\ \times \sum_{m=1}^{\infty}\frac{|k|^{2m}\left||x(\tau_p)-x(\tau_j)|^{2m-2} - |x(\tau_p)-y|^{2m-2}\right|}{2^{2m-1}(m-1)!\,m!} \le M|y-x(\tau_p)|R(n).\end{gathered} \tag{2.14}$$

Кроме того, из леммы 1 и неравенств (2.9) и (2.13) имеем

$$\left|\frac{(y-x(\tau_p),\nu(y))}{|x(\tau_p)-y|^2} - \frac{(x(\tau_j)-x(\tau_p),\nu(x(\tau_j)))}{|x(\tau_p)-x(\tau_j)|^2}\right| \le \frac{MR(n)}{|x(\tau_p)-y|}.$$

Тогда, принимая во внимание неравенства (2.1), (2.5), (2.10), (2.11), (2.13) и (2.14), нетрудно показать, что

$$\left|\frac{\partial N_0(k|x(\tau_p)-y|)}{\partial \nu(y)} - \frac{\partial N_0(k|x(\tau_p)-x(\tau_j)|)}{\partial \nu(x(\tau_j))}\right| \le \frac{MR(n)}{|x(\tau_p)-y|}.$$

В результате находим

$$\left|\frac{\partial\Phi_k(x(\tau_p),y)}{\partial\nu(y)}-\frac{\partial\Phi_k(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|\le\frac{MR(n)}{|x(\tau_p)-y|}.$$

Также, учитывая неравенство

$$\left|\frac{\partial\Phi_k(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}-\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|\le\frac{M\left|\ln|x(\tau_p)-y|\right|}{n!},\tag{2.15}$$

получаем, что

$$\left|\frac{\partial\Phi_k(x(\tau_p),y)}{\partial\nu(y)}-\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|\le M\left(\frac{R(n)}{|x(\tau_p)-y|}+\frac{\left|\ln|x(\tau_p)-y|\right|}{n!}\right).$$

В итоге

$$\left|\delta_2^n(x(\tau_p))\right|\le M\|\psi\|_\infty\left(R(n)\int\limits_{r(n)}^{\operatorname{diam}L}\frac{d\tau}{\tau}+\frac{1}{n!}\int\limits_{r(n)}^{\operatorname{diam}L}|\ln\tau|\,d\tau\right)\le M\|\psi\|_\infty\left(R(n)|\ln R(n)|+\frac{1}{n!}\right).$$

Пусть $y\in L_j$ и $j\ne p$. Так как из леммы 1 и неравенства (2.12) и (2.15) очевидно, что

$$\begin{aligned}\left|\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|&\le\left|\frac{\partial\Phi_k(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|+\left|\frac{\partial\Phi_k(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}-\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|\le\\&\le\frac{M\left|\ln|x(\tau_p)-y|\right|}{n!}\quad\forall n\in\mathrm{N},\end{aligned}\tag{2.16}$$

тогда

$$\left|\delta_3^n(x(\tau_p))\right|\le 2\omega(\psi,R(n))\int_L\left|\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(y)}\right|dL_y\le M\omega(\psi,R(n)).$$

Кроме того, учитывая леммы 1 и неравенства (2.8) и (2.16), получаем

$$\begin{aligned}\left|\delta_4^n(x(\tau_p))\right|&\le M\|\psi\|_\infty R(n)\sum_{\substack{j=1\\j\ne p}}^{n}\int\limits_{t_{j-1}}^{t_j}\left|\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|dt\le\\&\le M\|\psi\|_\infty R(n)\int_L\left|\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right|dL_y\le M\|\psi\|_\infty R(n).\end{aligned}$$

В результате, суммируя полученные оценки для выражений $\delta_1^n(x(\tau_p))$, $\delta_2^n(x(\tau_p))$, $\delta_3^n(x(\tau_p))$ и $\delta_4^n(x(\tau_p))$, и учитывая соотношение $R(n)\sim\frac{1}{n}$, получаем доказательство теоремы 2.

Очевидно, что

$$\frac{\partial\Phi_k^n(x,y)}{\partial\nu(x)}=\frac{i}{4}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(x)}+i\frac{\partial N_{0,n}(k|x-y|)}{\partial\nu(x)}\right),$$

где

$$\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(x)}=(x-y,\nu(x))\sum_{m=1}^{n}\frac{(-1)^m k^{2m}|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}$$

и

$$\frac{\partial N_{0,n}(k|x-y|)}{\partial \nu(x)} = \frac{2}{\pi}\left(\ln\frac{k|x-y|}{2} + C\right)\frac{\partial J_{0,n}(k|x-y|)}{\partial \nu(x)} + \frac{2(x-y,\nu(x))}{\pi|x-y|^2}J_{0,n}(k|x-y|) +$$
$$+ (x-y,\nu(x))\sum_{m=1}^{n}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}k^{2m}|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}.$$

Тогда, поступая точно также, как и в доказательстве теоремы 2, можно доказать следующую теорему.

Теорема 3. *Пусть L – замкнутая и дважды непрерывно дифференцируемая кривая в R^2 и $\varphi \in C(L)$. Тогда выражение*

$$(\tilde{K}_n\varphi)(x(\tau_p)) = \frac{2(b-a)}{n}\sum_{\substack{j=1\\ j\neq p}}^{n}\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_p))}\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}\,\varphi(x(\tau_j))$$

в опорных точках $x(\tau_p)$, $p = \overline{1,n}$, является квадратурной формулой для интеграла (1.4)*, причем справедлива следующая оценка*:

$$\max_{p=1,n}\left|(\tilde{K}\varphi)(x(\tau_p)) - (\tilde{K}_n\varphi)(x(\tau_p))\right| \le M\left(\omega(\varphi,1/n) + \|\varphi\|_\infty\frac{\ln n}{n}\right).$$

Кроме того, можно убедиться, что

$$\frac{\partial}{\partial\nu(x)}\left(\frac{\partial\Phi_k^n(x,y)}{\partial\nu(y)} - \frac{\partial\Phi_{k_0}^n(x,y)}{\partial\nu(y)}\right) = \frac{i}{4}\frac{\partial}{\partial\nu(x)}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(y)} - \frac{\partial J_{0,n}(k_0|x-y|)}{\partial\nu(y)}\right) -$$
$$- \frac{1}{4}\frac{\partial}{\partial\nu(x)}\left(\frac{\partial N_{0,n}(k|x-y|)}{\partial\nu(y)} - \frac{\partial N_{0,n}(k_0|x-y|)}{\partial\nu(y)}\right),$$

где

$$\frac{\partial}{\partial\nu(x)}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(y)} - \frac{\partial J_{0,n}(k_0|x-y|)}{\partial\nu(y)}\right) = (\nu(x),\nu(y))\sum_{m=1}^{n}\frac{(-1)^{m+1}\left(k^{2m}-k_0^{2m}\right)|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!} +$$
$$+ (y-x,\nu(y))(x-y,\nu(x))\sum_{m=2}^{n}\frac{(-1)^m\left(k^{2m}-k_0^{2m}\right)|x-y|^{2m-4}}{2^{2m-2}(m-2)!\,m!}$$

и

$$\frac{\partial}{\partial\nu(x)}\left(\frac{\partial N_{0,n}(k|x-y|)}{\partial\nu(y)} - \frac{\partial N_{0,n}(k_0|x-y|)}{\partial\nu(y)}\right) =$$
$$= \frac{2}{\pi}(\ln k - \ln k_0)(y-x,\nu(y))(x-y,\nu(x))\sum_{m=2}^{n}\frac{(-1)^m k^{2m}|x-y|^{2m-4}}{2^{2m-2}(m-2)!\,m!} +$$
$$+ \frac{2(x-y,\nu(x))}{\pi|x-y|^2}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(y)} - \frac{\partial J_{0,n}(k_0|x-y|)}{\partial\nu(y)}\right) +$$
$$+ \frac{2}{\pi}\left(\ln\frac{k_0|x-y|}{2} + C\right)\frac{\partial}{\partial\nu(x)}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(y)} - \frac{\partial J_{0,n}(k_0|x-y|)}{\partial\nu(y)}\right) -$$
$$- \frac{2(\nu(x),\nu(y))|x-y|^2 + 4(y-x,\nu(y))(x-y,\nu(x))}{\pi|x-y|^4}\sum_{m=1}^{n}\frac{(-1)^m\left(k^{2m}-k_0^{2m}\right)|x-y|^{2m}}{2^{2m}(m!)^2} +$$
$$+ \frac{2(y-x,\nu(y))}{\pi|x-y|^2}\left(\frac{\partial J_{0,n}(k|x-y|)}{\partial\nu(x)} - \frac{\partial J_{0,n}(k_0|x-y|)}{\partial\nu(x)}\right) -$$

$$-(\nu(x),\nu(y))\sum_{m=1}^{n}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}\left(k^{2m}-k_0^{2m}\right)|x-y|^{2m-2}}{2^{2m-1}(m-1)!\,m!}+$$

$$+(y-x,\nu(y))(x-y,\nu(x))\sum_{m=2}^{n}\left(\sum_{l=1}^{m}\frac{1}{l}\right)\frac{(-1)^{m+1}\left(k^{2m}-k_0^{2m}\right)|x-y|^{2m-4}}{2^{2m-2}(m-2)!\,m!}.$$

Тогда также справедлива следующая

Теорема 4. *Пусть L – замкнутая и дважды непрерывно дифференцируемая кривая в R^2 и $\psi\in C(L)$. Тогда выражение*

$$\left((T-T_0)_n\,\psi\right)(x(\tau_p))=\frac{2(b-a)}{n}\times$$

$$\times\sum_{\substack{j=1\\ j\neq p}}^{n}\frac{\partial}{\partial\nu(x(\tau_p))}\left(\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}-\frac{\partial\Phi_{k_0}^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right)\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}\,\psi(x(\tau_j))$$

в опорных точках $x(\tau_p)$, $p=\overline{1,n}$, является квадратурной формулой для интеграла (1.5)*, причем справедлива следующая оценка*:

$$\max_{p=1,n}\left|\left((T-T_0)\psi\right)(x(\tau_p))-\left((T-T_0)_n\,\psi\right)(x(\tau_p))\right|\le M\left(\omega(\psi,1/n)+\|\psi\|_\infty\frac{\ln n}{n}\right).$$

3. ОБОСНОВАНИЕ МЕТОДА КОЛЛОКАЦИИ ДЛЯ СИСТЕМЫ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ (1.1)

Пусть C^{2n} – пространство $2n$-мерных векторов $z^{2n}=\left(z_1^{2n},z_2^{2n},\ldots,z_{2n}^{2n}\right)^{\mathrm{T}}$, $z_l^{2n}\in C$, $l=\overline{1,2n}$, с нормой $\left\|z^{2n}\right\|=\max\limits_{l=1,2n}\left|z_l^{2n}\right|$, где запись "$a^{\mathrm{T}}$" означает транспонировку вектора a. Рассмотрим $2n$-мерную матрицу $A^{2n}=\left(a_{pj}\right)_{p,j=1}^{2n}$ с элементами

$$a_{pj}=0\quad\text{при}\quad p=\overline{1,n},\quad j=\overline{1,n}\quad\text{и}\quad p=j;$$

$$a_{pj}=\frac{2(b-a)\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}}{(\mu+\mu_0)n}\left(\mu\frac{\partial\Phi_k^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}-\mu_0\frac{\partial\Phi_{k_0}^n(x(\tau_p),x(\tau_j))}{\partial\nu(x(\tau_j))}\right)$$

$$\text{при}\quad p=\overline{1,n},\quad j=\overline{1,n}\quad\text{и}\quad p\neq j;$$

$$a_{pj}=0\quad\text{при}\quad p=\overline{1,n},\quad j=\overline{n+1,\,2n}\quad\text{и}\quad p=j-n;$$

$$a_{pj}=\frac{2(b-a)\sqrt{\left(x_1'(\tau_{j-n})\right)^2+\left(x_2'(\tau_{j-n})\right)^2}}{(\mu+\mu_0)n}\left(\mu^2\Phi_k^n(x(\tau_p),x(\tau_{j-n}))-\mu_0^2\Phi_{k_0}^n(x(\tau_p),x(\tau_{j-n}))\right)$$

$$\text{при}\quad p=\overline{1,n},\quad j=\overline{n+1,2n}\quad\text{и}\quad p\neq j-n;$$

$$a_{pj}=0\quad\text{при}\quad p=\overline{n+1,\,2n},\quad j=\overline{1,n}\quad\text{и}\quad p=j+n;$$

$$a_{pj}=\frac{2(b-a)\sqrt{\left(x_1'(\tau_j)\right)^2+\left(x_2'(\tau_j)\right)^2}}{(\mu+\mu_0)n}\times$$

$$\times\frac{\partial}{\partial\nu(x(\tau_{p-n}))}\left(\frac{\partial\left(\Phi_{k_0}^n(x(\tau_{p-n}),x(\tau_j))-\Phi_k^n(x(\tau_{p-n}),x(\tau_j))\right)}{\partial\nu(x(\tau_j))}\right)$$

при $p = \overline{n+1, 2n}$, $j = \overline{1, n}$ и $p \neq j + n$;

$a_{pj} = 0$ при $p = \overline{n+1, 2n}$, $j = \overline{n+1, 2n}$ и $p = j$;

$$a_{pj} = \frac{2(b-a)\sqrt{\left(x_1'(\tau_{j-n})\right)^2 + \left(x_2'(\tau_{j-n})\right)^2}}{(\mu + \mu_0)n}\left(\mu_0 \frac{\partial \Phi_{k_0}^n(x(\tau_{p-n}), x(\tau_{j-n}))}{\partial \nu(x(\tau_{p-n}))} - \mu \frac{\partial \Phi_k^n(x(\tau_{p-n}), x(\tau_{j-n}))}{\partial \nu(x(\tau_{p-n}))}\right)$$

при $p = \overline{n+1, 2n}$, $j = \overline{n+1, 2n}$ и $p \neq j$.

Если через z_p^{2n}, $p = \overline{1, n}$, обозначим приближенные значения $\psi(x(\tau_p))$, а через z_{p+n}^{2n}, $p = \overline{1, n}$, приближенные значения $\varphi(x(\tau_p))$, то, используя построенные квадратурные формулы для интегралов (1.2)–(1.5), система интегральных уравнений (1.1) заменяется системой алгебраических уравнений относительно $z^{2n} \in C^{2n}$, которую запишем в виде

$$\begin{aligned} z_p^{2n} + \sum_{j=1}^{2n} a_{pj} z_j^{2n} &= \frac{2f(x(\tau_p))}{\mu + \mu_0}, \quad p = \overline{1, n}, \\ z_p^{2n} + \sum_{j=1}^{2n} a_{pj} z_j^{2n} &= -\frac{2g(x(\tau_{p-n}))}{\mu + \mu_0}, \quad p = \overline{n+1, 2n} \end{aligned} \tag{3.1}$$

Теперь сформулируем основной результат данной работы.

Теорема 5. *Пусть функции f и g непрерывны на кривой L. Тогда уравнения* (1.1) *и* (3.1) *имеют единственные решения $(\psi_*, \varphi_*) \in C(L) \times C(L)$ и $w^{2n} \in C^{2n}$ $(n \geq n_0)$ соответственно, причем справедливы следующие оценки*:

$$\max_{p=1,n} \left| w_p^{2n} - \psi_*(x(\tau_p)) \right| \leq M\left(\omega(f, 1/n) + \omega(g, 1/n) + \frac{\ln n}{n}\right),$$

$$\max_{p=1,n} \left| w_{p+n}^{2n} - \varphi_*(x(\tau_p)) \right| \leq M\left(\omega(f, 1/n) + \omega(g, 1/n) + \frac{\ln n}{n}\right).$$

Доказательство. Для обоснования метода коллокации будем пользоваться теоремой Г.М. Вайникко о сходимости для линейных операторных уравнений (см. [14]). Для этого сначала запишем уравнения (1.1) и (3.1) в операторном виде.

Отметим, что $C(L) \times C(L)$ является банаховым пространством с нормой $\|\rho\|_1 = \max\{\|\psi\|_\infty, \|\varphi\|_\infty\}$. Рассмотрим матричный оператор 2-го порядка

$$A = \begin{pmatrix} \dfrac{\mu}{\mu + \mu_0} K - \dfrac{\mu_0}{\mu + \mu_0} K_0 & \dfrac{\mu^2}{\mu + \mu_0} S - \dfrac{\mu_0^2}{\mu + \mu_0} S_0 \\ \dfrac{1}{\mu + \mu_0}(T_0 - T) & \dfrac{\mu_0}{\mu + \mu_0} \tilde{K}_0 - \dfrac{\mu}{\mu + \mu_0} \tilde{K} \end{pmatrix},$$

определенный в пространстве $C(L) \times C(L)$. Тогда систему интегральных уравнений (1.1) можно переписать в виде

$$(I + A)\rho = \chi, \tag{3.2}$$

а систему алгебраических уравнений (3.1) в виде

$$\left(I^{2n} + A^{2n}\right) z^{2n} = \chi^{2n}, \tag{3.3}$$

где I – единичный оператор на $C(L) \times C(L)$,

$$\rho = \begin{pmatrix} \psi \\ \varphi \end{pmatrix}, \quad \chi = \frac{2}{\mu + \mu_0} \begin{pmatrix} f \\ -g \end{pmatrix},$$

I^{2n} – единичная матрица $2n$-го порядка, $\chi^{2n} = p^{2n}\chi$, а $p^{2n}: C(L)\times C(L) \to C^{2n}$ – линейный ограниченный оператор, определяемый формулой

$$p^{2n}\rho = p^{2n}\begin{pmatrix}\psi\\ \varphi\end{pmatrix} = (\psi(x(\tau_1)), \psi(x(\tau_2)), \ldots, \psi(x(\tau_n)), \varphi(x(\tau_1)), \varphi(x(\tau_2)), \ldots, \phi(x(\tau_n)))^{\mathrm{T}}.$$

Теперь проверим выполнение условий теоремы 4.2 из работы [14], при этом обозначения и необходимые определения и предложения возьмем из [14]. В работе [2] доказано, что система интегральных уравнений (1.1) однозначно разрешима в пространстве $C(L)\times C(L)$, т.е. $\mathrm{Ker}(I + A) = \{0\}$. Кроме того, операторы $I^{2n} + A^{2n}$ фредгольмовы с нулевым индексом. Принимая во внимание способ разбиения кривой L на "регулярные" элементарные части, получаем, что для любого $\rho \in C(L)\times C(L)$ справедливо следующее равенство:

$$\lim_{n\to\infty}\left\|p^{2n}\rho\right\| = \lim_{n\to\infty}\max\left\{\max_{l=1,n}|\psi(x(\tau_l))|, \max_{l=1,n}|\varphi(x(\tau_l))|\right\} = \max\left\{\max_{x\in L}|\psi(x)|, \max_{x\in L}|\varphi(x)|\right\} = \|\rho\|_1.$$

Следовательно, система операторов $P = \left\{p^{2n}\right\}$ является связывающей для пространств $C(L)\times C(L)$ и C^{2n}. Тогда $\chi^{2n} \overset{P}{\to} \chi$ и принимая во внимание теоремы 1– 4, получаем, что по определению 2.1 из работы [14] $I^{2n} + A^{2n} \overset{PP}{\to} I + A$. Так как по определению 3.2 из [14] $I^{2n} \to I$ устойчиво, то по предложению 3.5 и по определению 3.3 из [14] осталось проверить условие компактности, которое ввиду предложения 1.1 из [14] равносильно условию: $\forall\left\{z^{2n}\right\}$, $z^{2n} \in C^{2n}$, $\left\|z^{2n}\right\| \le M$, существует относительно компактная последовательность $\left\{A_{2n}z^{2n}\right\} \subset C(L)\times C(L)$ такая, что

$$\left\|A^{2n}z^{2n} - p^{2n}\left(A_{2n}z^{2n}\right)\right\| \to 0 \quad \text{при} \quad n \to \infty.$$

В качестве $\left\{A_{2n}z^{2n}\right\}$ выберем последовательность

$$\left(A_{2n}z^{2n}\right)(x) = \begin{pmatrix}\sum\limits_{j=1}^{2n} a_j^{(1)}(x)\, z_j^{2n}\\ \sum\limits_{j=1}^{2n} a_j^{(2)}(x)\, z_j^{2n}\end{pmatrix},$$

где

$$a_j^{(1)}(x) = \frac{2}{\mu+\mu_0}\left(\mu\int_{L_j}\frac{\partial\Phi_k^n(x,y)}{\partial\nu(y)}dL_y - \mu_0\int_{L_j}\frac{\partial\Phi_{k_0}^n(x,y)}{\partial\nu(y)}dL_y\right) \quad \text{при} \quad j = \overline{1,n},$$

$$a_j^{(1)}(x) = \frac{2}{\mu+\mu_0}\left(\mu^2\int_{L_{j-n}}\Phi_k^n(x,y)\,dL_y - \mu_0^2\int_{L_{j-n}}\Phi_{k_0}^n(x,y)\,dL_y\right) \quad \text{при} \quad j = \overline{n+1,2n},$$

$$a_j^{(2)}(x) = \frac{2}{\mu+\mu_0}\int_{L_j}\frac{\partial}{\partial\nu(x)}\left(\frac{\partial\left(\Phi_{k_0}^n(x,y) - \Phi_k^n(x,y)\right)}{\partial\nu(y)}\right)dL_y \quad \text{при} \quad j = \overline{1,n},$$

$$a_j^{(2)}(x) = \frac{2}{\mu+\mu_0}\left(\mu_0\int_{L_{j-n}}\frac{\partial\Phi_{k_0}^n(x,y)}{\partial\nu(x)}dL_y - \mu\int_{L_{j-n}}\frac{\partial\Phi_k^n(x,y)}{\partial\nu(x)}dL_y\right) \quad \text{при} \quad j = \overline{n+1,\,2n}.$$

Из неравенства (2.1), (2.2) и (2.9) очевидно, что для любых точек $x, y \in L$, $x \ne y$, и для любого натурального числа n, справедливы следующие оценки:

$$\left|\Phi_k^n(x,y)\right| \le M\left|\ln|x-y|\right|, \quad \left|\frac{\partial\Phi_k^n(x,y)}{\partial\nu(y)}\right| \le M, \quad \left|\frac{\partial\Phi_k^n(x,y)}{\partial\nu(x)}\right| \le M$$

и

$$\left|\frac{\partial}{\partial\nu(x)}\left(\frac{\partial\left(\Phi_{k_0}^n(x,y)-\Phi_k^n(x,y)\right)}{\partial\nu(y)}\right)\right|\le M.$$

Отсюда получаем, что

$$\left|\sum_{j=1}^{2n}a_j^{(1)}(x)z_j^{2n}\right|\le\frac{2\left\|z^{2n}\right\|}{|\mu+\mu_0|}\int_L\left(|\mu|\left|\frac{\partial\Phi_k^n(x,y)}{\partial\nu(y)}\right|+\right.$$
$$\left.+|\mu_0|\left|\frac{\partial\Phi_{k_0}^n(x,y)}{\partial\nu(y)}\right|+|\mu|^2\left|\Phi_k^n(x,y)\right|+|\mu_0|^2\left|\Phi_{k_0}^n(x,y)\right|\right)dL_y\le M\left\|z^{2n}\right\|\quad\forall\ x\in L,$$

и

$$\left|\sum_{j=1}^{2n}a_j^{(2)}(x)z_j^{2n}\right|\le\frac{2\left\|z^{2n}\right\|}{|\mu+\mu_0|}\int_L\left(|\mu|\left|\frac{\partial\Phi_k^n(x,y)}{\partial\nu(x)}\right|+|\mu_0|\left|\frac{\partial\Phi_{k_0}^n(x,y)}{\partial\nu(x)}\right|+\right.$$
$$\left.+\left|\frac{\partial}{\partial\nu(x)}\left(\frac{\partial\left(\Phi_{k_0}^n(x,y)-\Phi_k^n(x,y)\right)}{\partial\nu(y)}\right)\right|\right)dL_y\le M\left\|z^{2n}\right\|\quad\forall\ x\in L.$$

Следовательно,

$$\left|\left(A_{2n}z^{2n}\right)(x)\right|\le M\left\|z^{2n}\right\|\quad\forall\ x\in L.$$

Тогда, принимая во внимание условие $\left\|z^N\right\|\le M$, получаем равномерную ограниченность последовательности $\left\{A_{2n}z^{2n}\right\}$.

Теперь возьмем любые точки $x',x''\in L$ такие, что $|x'-x''|=\delta<d/2$. Тогда, поступая точно также, как и в работе [15], можно показать, что

$$\left|\sum_{j=1}^{2n}a_j^{(1)}(x')z_j^{2n}-\sum_{j=1}^{2n}a_j^{(1)}(x'')z_j^{2n}\right|\le M\left\|z^{2n}\right\|\delta|\ln\delta|\quad\forall\ x',x''\in L,$$

и

$$\left|\sum_{j=1}^{2n}a_j^{(2)}(x')z_j^{2n}-\sum_{j=1}^{2n}a_j^{(2)}(x'')z_j^{2n}\right|\le M\left\|z^{2n}\right\|\delta|\ln\delta|\quad\forall\ x',x''\in L.$$

Следовательно,

$$\left|\left(A_{2n}z^{2n}\right)(x')-\left(A_{2n}z^{2n}\right)(x'')\right|\le M\left\|z^{2n}\right\||x'-x''||\ln|'x-x''\|\quad\forall\ x',x''\in L,$$

а значит, $\left\{A_{2n}z^{2n}\right\}\subset C(L)\times C(L)$. Отсюда непосредственно вытекает равностепенная непрерывность последовательности $\left\{A_{2n}z^{2n}\right\}$. Тогда из теоремы Арцеля следует относительная компактность последовательности $\left\{A_{2n}z^{2n}\right\}$. Кроме того, поступая точно также, как и в доказательствах теоремы 1 и 2, получим

$$\left\|A^{2n}z^{2n}-p^{2n}\left(A_{2n}z^{2n}\right)\right\|\to0\quad\text{при}\quad n\to\infty.$$

Тогда, применяя теорему 4.2 из работы [14], находим, что уравнения (3.2) и (3.3) имеют единственные решения $\rho_*=\begin{pmatrix}\psi_*\\ \varphi_*\end{pmatrix}\in C(L)\times C(L)$ и $w^{2n}\in C^{2N}$ $(n\ge n_0)$ соответственно, причем

$$c_1\delta_n\le\left\|w^{2n}-p^{2n}\rho_*\right\|\le c_2\delta_n,$$

где

$$c_1 = 1\Big/\sup_{n\ge n_0}\left\|I^{2n} + A^{2n}\right\| > 0, \quad c_2 = \sup_{n\ge n_0}\left\|\left(I^{2n} + A^{2n}\right)^{-1}\right\| < +\infty,$$

$$\delta_n = \left\|\left(I^{2n} + A^{2n}\right)\left(p^{2n}\rho_*\right) - \chi^{2n}\right\|.$$

Принимая во внимание равенство

$$\chi^{2n} = p^{2n}\chi = p^{2n}\rho_* + p^{2n}\left(A\rho_*\right)$$

и оценки погрешности построенных квадратурных формул для интегралов (1.2)–(1.5), имеем

$$\delta_n = \left\|A^{2n}\left(p^{2n}\rho_*\right) - p^{2n}\left(A\rho_*\right)\right\| \le M\left(\left\|\rho_*\right\|_1\frac{\ln n}{n} + \omega\left(\rho_*, 1/n\right)\right),$$

где

$$\omega\left(\rho_*,\delta\right) = \max_{\substack{|x-y|\le\delta\\ x,y\in L}}\sqrt{\left(\psi_*(x)-\psi_*(y)\right)^2 + \left(\varphi_*(x)-\varphi_*(y)\right)^2}, \quad \delta > 0.$$

Так как из неравенства (2.1), (2.2) и (2.9) ясно, что для любых точек $x, y \in L$, $x \ne y$,

$$\left|\Phi_k(x,y)\right| \le M\left|\ln|x-y|\right|, \quad \left|\frac{\partial\Phi_k(x,y)}{\partial\nu(y)}\right| \le M, \quad \left|\frac{\partial\Phi_k(x,y)}{\partial\nu(x)}\right| \le M$$

и

$$\left|\frac{\partial}{\partial\nu(x)}\left(\frac{\partial\left(\Phi_{k_0}(x,y)-\Phi_k(x,y)\right)}{\partial\nu(y)}\right)\right| \le M,$$

то, поступая точно также, как и в работе [16], можно показать, что

$$\left|\left(S\rho_*\right)(x') - \left(S\rho_*\right)(x'')\right| \le M\left\|\rho_*\right\|\left|x'-x''\right|\left|\ln|x'-x''|\right| \quad \forall\ x', x'' \in L,$$

$$\left|\left(K\rho_*\right)(x') - \left(K\rho_*\right)(x'')\right| \le M\left\|\rho_*\right\|\left|x'-x''\right|\left|\ln|x'-x''|\right| \quad \forall\ x', x'' \in L,$$

$$\left|\left(\tilde{K}\rho_*\right)(x') - \left(\tilde{K}\rho_*\right)(x'')\right| \le M\left\|\rho_*\right\|\left|x'-x''\right|\left|\ln|x'-x''|\right| \quad \forall\ x', x'' \in L,$$

и

$$\left|\left(\left(T-T_0\right)\rho_*\right)(x') - \left(\left(T-T_0\right)\rho_*\right)(x'')\right| \le M\left\|\rho_*\right\|\left|x'-x''\right|\left|\ln|x'-x''|\right| \quad \forall\ x', x'' \in L.$$

Следовательно,

$$\left|\left(A\rho_*\right)(x') - \left(A\rho_*\right)(x'')\right| \le M\left\|\rho_*\right\|\left|x'-x''\right|\left|\ln|x'-x''|\right| \quad \forall\ x', x'' \in L,$$

т.е.

$$\omega\left(A\rho_*, 1/n\right) \le M\left\|\rho_*\right\|_1\frac{\ln n}{n}.$$

Тогда, принимая во внимание неравенство

$$\omega\left(\rho_*,1/n\right) = \omega\left(\chi - A\rho_*, 1/n\right) \le \omega(\chi,1/n) + \omega\left(A\rho_*,1/n\right) \le \omega(f,1/n) + \omega(g,1/n) + M\left\|\rho_*\right\|_1\frac{\ln n}{n}$$

и

$$\left\|\rho_*\right\|_1 \le \left\|\left(I+A\right)^{-1}\right\|\left\|\chi\right\|_1,$$

получаем, что

$$\delta_n \le M\left(\omega(f,1/n) + \omega(g,1/n) + \frac{\ln n}{n}\right).$$

Теорема доказана.

Следствие 1. *Пусть* $x_* \in D$, $x^* \in R^2/\overline{D}$ *и* $w^{2n} = \left(w_1^{2n}, w_2^{2n}, \ldots, w_{2n}^{2n}\right)^{\mathrm{T}}$ *является решением системы алгебраических уравнений* (3.1). *Тогда последовательность*

$$u^n(x^*) = \frac{b-a}{n}\sum_{j=1}^{n}\left(\frac{\partial\Phi_k^n(x^*, x(\tau_j))}{\partial\nu(x(\tau_j))}w_j^{2n} + \mu\Phi_k^n(x^*, x(\tau_j))w_{n+j}^{2n}\right)\sqrt{\left(x_1'(\tau_j)\right)^2 + \left(x_2'(\tau_j)\right)^2}$$

сходится к $u(x^*)$, *а последовательность*

$$u_0^n(x_*) = \frac{b-a}{n}\sum_{j=1}^{n}\left(\frac{\partial\Phi_{k_0}^n(x_*, x(\tau_j))}{\partial\nu(x(\tau_j))}w_j^{2n} + \mu_0\Phi_{k_0}^n(x_*, x(\tau_j))w_{n+j}^{2n}\right)\sqrt{\left(x_1'(\tau_j)\right)^2 + \left(x_2'(\tau_j)\right)^2}$$

сходится к $u_0(x_*)$, *причем*

$$\left|u^n(x^*) - u(x^*)\right| \le M\left(\frac{\ln n}{n} + \omega(f, 1/n) + \omega(g, 1/n)\right),$$

$$\left|u_0^n(x_*) - u_0(x_*)\right| \le M\left(\frac{\ln n}{n} + \omega(f, 1/n) + \omega(g, 1/n)\right).$$

СПИСОК ЛИТЕРАТУРЫ

1. *Колтон Д., Кресс Р.* Методы интегральных уравнений в теории рассеяния. М.: Мир, 1987. 311 с.
2. *Kress R., Roach G.F.* Transmission problems Helmholtz equation // J. Math. Phys. 1978. V. 19. P. 1433–1437.
3. *Каширин А.А., Смагин С.И., Талтыкина М.Ю.* Применение мозаично-скелетонного метода при численном решении трехмерных задач Дирихле для уравнения Гельмгольца в интегральной форме // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 4. С. 625–638.
4. *Халилов Э.Г.* Обоснование метода коллокации для интегрального уравнения смешанной краевой задачи для уравнения Гельмгольца // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 7. С. 1340–1348.
5. *Harris P.J., Chen K.* On efficient preconditioners for iterative solution of a Galerkin boundary element equation for the three–dimensional exterior Helmholtz problem // J. Comp. Appl. Math. 2003. V. 156. P. 303–318.
6. *Khalilov E.H., Aliev A.R.* Justification of a quadrature method for an integral equation to the external Neumann problem for the Helmholtz equation // Math. Meth. Appl. Sci. 2018. V. 41. № 16. P. 6921–6933.
7. *Turc C., Boubendir Y., Riahi M.K.* Well–conditioned boundary integral equation formulations and Nyström discretizations for the solution of Helmholtz problems with impedance boundary conditions in two-dimensional Lipschitz domains // J. Integral Eq. Appl. 2017. V. 29. № 3. P. 441–472.
8. *Халилов Э.Г.* Обоснование метода коллокации для одного класса систем интегральных уравнений // Украинский матем. ж. 2017. Т. 69. № 6. С. 823–835.
9. *Khalilov E.H., Bakhshaliyeva M.N.* Quadrature formulas for simple and double layer logarithmic potentials // Proceed. of IMM of NAS of Azerbaijan. 2019. V. 45. № 1. P. 155–162.
10. *Kress R.* Boundary integral equations in time-harmonic acoustic scattering // Math. Comp. Modeling. 1991. V. 15. № 3–5. P. 229–243.
11. *Мусхелешвили Н.И.* Сингулярные интегральные уравнения. М.: Физматлит, 1962. 599 с.
12. *Владимиров В.С.* Уравнения математической физики. М.: Наука, 1976. 527 с.
13. *Халилов Э.Г.* Обоснование метода коллокации для одного класса поверхностных интегральных уравнений // Матем. заметки. 2020. Т. 107. № 4. С. 604–622.
14. *Вайникко Г.М.* Регулярная сходимость операторов и приближенное решение уравнений // Итоги науки и техники. Матем. анализ. 1979. Т. 16. С. 5–53.
15. *Бахшалыева М.Н., Халилов Э.Г.* Обоснование метода коллокации для интегрального уравнения внешней краевой задачи Дирихле для уравнения Лапласа // Ж. вычисл. матем. и матем. физ. 2021. Т. 61. № 6. С. 936–950.
16. *Халилов Э.Г., Бахшалыева М.Н.* Исследование приближенного решения интегрального уравнения, соответствующего смешанной краевой задаче для уравнения Лапласа // Уфимский матем. журн. 2021. Т. 13. № 1. С. 86–98.

EDN: LGWCWE
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 854–860

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.632.4

РЕШЕНИЕ ДВУМЕРНОЙ ОБРАТНОЙ ЗАДАЧИ КВАЗИСТАТИЧЕСКОЙ ЭЛАСТОГРАФИИ С ПОМОЩЬЮ МЕТОДА МАЛОГО ПАРАМЕТРА[1)]

© 2022 г. А. С. Леонов[1,*], Н. Н. Нефедов[2,], А. Н. Шаров[2,***], А. Г. Ягола[2,****]**

[1] 115409 Москва, Каширское ш., 31, НИЯУ "МИФИ", Россия

[2] 119992 Москва, Ленинские горы, МГУ, физический факультет, кафедра математики, Россия

**e-mail: asleonov@mephi.ru*

***e-mail: nefedov@phys.msu.ru*

****e-mail: scharov.aleksandr@physics.msu.ru*

*****e-mail: yagola@physics.msu.ru*

Поступила в редакцию 31.10.2021 г.
Переработанный вариант 31.10.2021 г.
Принята к публикации 14.01.2022 г.

Рассматривается прямая и обратная задачи двумерной квазистатической эластографии в рамках модели деформации исследуемой биологической ткани как упругого тела, подвергаемого поверхностному сжатию. Возникающие смещения ткани в приближении плоского линейного деформированного состояния упругого тела описываются краевой задачей для уравнений в частных производных с коэффициентами, которые определяются модулем Юнга и постоянным коэффициентом Пуассона ткани. Эта краевая задача содержит малый параметр, что позволяет решить ее с помощью теории регулярных возмущений уравнений в частных производных. Исследуется процедура такого решения и при некоторых предположениях выписываются простые формулы для решения как прямой, так и обратной задач двумерной квазистатической эластографии. Представлены результаты численных экспериментов по решению прямой и обратной модельных задач по предлагаемым формулам. Полученные результаты достаточно хорошо отражают модельные решения. Расчеты по формулам требуют долей микросекунд на персональном компьютере средней производительности для достаточно мелких сеток, так что предлагаемый подход с использованием малого параметра может быть применен при онкологической диагностике в реальном времени. Библ. 16. Фиг. 2.

Ключевые слова: двумерная квазистатическая эластография, обратные задачи, метод малого параметра, регуляризация.

DOI: 10.31857/S0044466922050076

1. ВВЕДЕНИЕ

В последние десятилетия бурно развивается отрасль онкологической диагностики, называемая *эластографией* (см. [1]–[5]). Она объединяет ряд специфических методов, основанных на различиях в механических характеристиках здоровой и опухолевой ткани определенных типов. Все эти методы структурно состоят из следующих этапов: воздействие на исследуемую часть тела поверхностными силами; измерение (или вычисление) возникающих деформаций исследуемой биологической ткани; определение характеристик упругости ткани путем решения обратной задачи. Деформации (смещения) тканей определяются по данным ультразвуковых исследований или с помощью магнитно-резонансного метода. Принципиальным моментом в такой схеме является решение обратной математической задачи в рамках некоторой модели биологической ткани.

Разработано множество таких моделей различного уровня сложности. Так как зачастую указанные обратные задачи решаются с использованием различной вычислительной техники (персональные компьютеры, вычислительные кластеры, суперкомпьютеры и т.д.), то наиболее перспективными для практической диагностики являются те математические модели, в которых об-

[1)]Работа выполнена при финансовой поддержке РНФ (проект 18-11-00042).

ратные задачи могут быть решены в реальном времени или близко к нему. Поэтому основные требования, предъявляемые к моделям, следующие: адекватное отражение процессов в биологических тканях при наиболее возможной простоте используемого математического аппарата; возможность постановки для таких моделей достаточно простой обратной задачи определения механических характеристик тканей.

В теории эластографии большое распространение получила модель, основанная на уравнениях квазистатической теории упругости (см. [5]–[7]). В ней участок исследуемой ткани, характеризуемый распределениями механических модулей, представляется как линейно-упругое изотропное тело, подвергаемое малым поверхностным сжатиям. Решение обратной задачи для такой модели позволяет найти по известным смещениям ткани распределение этих модулей в рассматриваемой области и, тем самым, найти характерные онкологические включения с повышенным значением модулей. Квазистатическая модель адекватно описывает исследуемые ткани, но обратная задача поиска механических модулей для нее оказывается достаточно сложной. Многочисленные численные эксперименты показали, что решить такую обратную задачу в реальном времени невозможно (см. [6]–[10]).

Поэтому весьма актуальным для эластографии представляется модификация упомянутой модели квазистатической теории упругости с целью ее возможного упрощения. Оказывается, это можно сделать с учетом того, что соответствующие дифференциальные уравнения содержат малый параметр. Целями данной работы являются разработка такой упрощенной модели в двумерном случае, постановка для нее обратной задачи и решение этой обратной задачи в реальном времени.

2. ПРЯМАЯ И ОБРАТНАЯ ЗАДАЧИ ДВУМЕРНОЙ ЭЛАСТОГРАФИИ В ПРИБЛИЖЕНИИ ПЛОСКОГО ДЕФОРМИРОВАННОГО СОСТОЯНИЯ

Будем считать, что исследуемый участок двумерного сечения биологической ткани моделируется областью $\Omega = (-\infty,\infty)\times(0,h) \subset R^2_{xy}$, а сама ткань характеризуется распределением модуля Юнга $E = E(x,y)$ и постоянным коэффициентом Пуассона $\nu = 0.495$. Этот слой подвергается на нижней поверхности воздействию направленной вертикально вверх силы, а верхняя поверхность слоя неподвижна. Такие предположения характерны для постановок задач двумерной эластографии (см., например, [11]). В модели двумерного плоского линейного деформированного состояния упругого тела Ω (см., например, [5]–[8]) связь горизонтальных и вертикальных смещений ткани $u(x,y), w(x,y)$ с распределением модуля Юнга и коэффициентом Пуассона описывается системой уравнений в частных производных вида

$$\begin{gathered}(1-\nu)\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x}\right)+\frac{1}{2}(1-2\nu)\frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial y}\right)+\nu\frac{\partial}{\partial x}\left(E\frac{\partial w}{\partial y}\right)+\frac{1}{2}(1-2\nu)\frac{\partial}{\partial y}\left(E\frac{\partial w}{\partial x}\right)=0,\\ \frac{1}{2}(1-2\nu)\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial y}\right)+\nu\frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial x}\right)+\frac{1}{2}(1-2\nu)\frac{\partial}{\partial x}\left(E\frac{\partial w}{\partial x}\right)+(1-\nu)\frac{\partial}{\partial y}\left(E\frac{\partial w}{\partial y}\right)=0,\\ (x,y)\in\Omega,\end{gathered} \tag{1}$$

с граничными условиями на $\Gamma_1 = \{(x,y): x\in(-\infty,\infty), y=0\}$ (нижняя граница полосы Ω):

$$\begin{gathered}\frac{1}{2}(1-2\nu)E(x,y)\frac{\partial u}{\partial y}+\frac{1}{2}(1-2\nu)E(x,y)\frac{\partial w}{\partial x}=0,\\ \nu E(x,y)\frac{\partial u}{\partial x}+(1-\nu)E(x,y)\frac{\partial w}{\partial y}=(1+\nu)(1-2\nu)f_y(x),\end{gathered} \tag{2}$$

и на $\Gamma_2 = \{(x,y): x\in(-\infty,\infty), y=h\}$: $u(x,y)=w(x,y)=0$ (верхняя граница полосы Ω). Здесь f_y есть вертикальная компонента давления на поверхность Γ_1. Определение смещений $u(x,y), w(x,y)$ по известным коэффициентам $E(x,y), \nu$ составляет *прямую задачу* двумерной эластографии. При условиях $f_y(x)\in C^1(\Gamma_1), E(x,y)\in C^1(\bar{\Omega})$ и $0<E_1\le E(x,y)\le E_2$ она однозначно разрешима на классе функций $u(x,y), w(x,y)\in W_2^1(\Omega)$ (это следует, например, из [12, гл. 2, 11], где получены более общие результаты). Здесь $E_{1,2}$ – известные константы. Соответствующие *обратные задачи* эластографии можно ставить по-разному, но все они заключаются в нахождении рас-

пределения модуля Юнга $E(x, y)$ по известным смещениям или функционалам от них (см. [5]–[11]). Ниже это будет уточнено.

Величину $\varepsilon = \frac{1}{2} - \nu \approx 0.005$ ($2\varepsilon \approx 0.01$) можно считать малым параметром. Тогда прямую задачу (1), (2) можно переписать в форме:

$$\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x} + E\frac{\partial w}{\partial y}\right) = -2\varepsilon\left[\frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial y} + E\frac{\partial w}{\partial x}\right) + \frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x} - E\frac{\partial w}{\partial y}\right)\right],$$
$$\frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial x} + E\frac{\partial w}{\partial y}\right) = -2\varepsilon\left[\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial y} + E\frac{\partial w}{\partial x}\right) - \frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial x} - E\frac{\partial w}{\partial y}\right)\right], \quad (x, y) \in \Omega,$$

с граничными условиями

$$\left(E(x, y)\frac{\partial u}{\partial x} + E(x, y)\frac{\partial w}{\partial y}\right)_{y=0} = 4\varepsilon(1+\nu)f_y + 2\varepsilon\left(E(x, y)\frac{\partial u}{\partial x} - E(x, y)\frac{\partial w}{\partial y}\right)_{y=0}, \quad (x, y) \in \Gamma_1,$$
$$\varepsilon\left(\frac{\partial u}{\partial y} + \frac{\partial w}{\partial x}\right)_{y=0} = 0, \quad (x, y) \in \Gamma_1; \quad u(x, y) = w(x, y) = 0, \quad (x, y) \in \Gamma_2.$$

В дальнейшем функцию давления, фигурирующую в первом граничном условии, будем обозначать как $F(x) = 4\varepsilon(1+\nu)f_y(x)$.

Выше была отмечена одна из важнейших проблем в практических приложениях прямой и обратной задач эластографии – *получить решение быстро и достаточно точно*. Тогда соответствующий алгоритм можно использовать в реальной диагностике. Численное решение прямой задачи (1), (2), основанное на методе конечных элементов, с этой точки зрения сравнительно эффективно и требует несколько секунд на ПК средней производительности при достаточно подробных сетках переменных x, y. Однако все известные нам методы решения двумерных обратных задач эластографии требуют значительно большего времени (от десятков минут до часов в зависимости от размеров сеток). Ниже предлагается метод приближенного решения прямой задачи как системы уравнений в частных производных с малым параметром. Полученные простые аналитические формулы для решений можно использовать для “быстрого” приближенного решения обратной задачи.

3. ИССЛЕДОВАНИЕ ПРЯМОЙ ЗАДАЧИ: НЕДООПРЕДЕЛЕННОСТЬ ЗАДАЧ ДЛЯ ПРИБЛИЖЕНИЙ МАЛОГО ПАРАМЕТРА, СОВМЕСТНОСТИ ЭТИХ ЗАДАЧ

Пусть $U = (u(x, y), w(x, y))$, где $u(x, y), w(x, y) \in W_2^1(\Omega)$. Введем дифференциальные операторы, действующие из $W_2^1(\Omega) \times W_2^1(\Omega)$ в $L_2(\Omega)$, предполагая, что коэффициент $E(x, y)$ это гарантирует:

$$L_0(U) = \left(\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x} + E\frac{\partial w}{\partial y}\right), \quad \frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial x} + E\frac{\partial w}{\partial y}\right)\right),$$
$$L_1(U) = \left(\frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial y} + E\frac{\partial w}{\partial x}\right) + \frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x} - E\frac{\partial w}{\partial y}\right), \quad \frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial y} + E\frac{\partial w}{\partial x}\right) - \frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial x} - E\frac{\partial w}{\partial y}\right)\right).$$

Определим также операторы граничных условий:

$$l_0(U) = \left(\left(E(x, y)\frac{\partial u}{\partial x} + E(x, y)\frac{\partial w}{\partial y}\right)_{y=0}, 0\right),$$
$$l_1(U) = \left(\left(E(x, y)\frac{\partial u}{\partial x} - E(x, y)\frac{\partial w}{\partial y}\right)_{y=0}, \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial w}{\partial x}\right)_{y=0}\right).$$

Тогда прямую задачу можно записать в форме:

$$\begin{aligned} &L_0(U) = -2\varepsilon L_1(U), \quad (x, y) \in \Omega, \\ &l_0(U) = 2\varepsilon l_1(U) + (F(x), 0), \quad x \in (-\infty, \infty), \\ &U|_{y=h} = 0. \end{aligned} \tag{3}$$

Решая (3) формально по методу малого параметра для регулярных возмущений (см., например, [13], [14]), получаем $U(x,y) = \sum_{n=0}^{\infty}(2\varepsilon)^n U_n(x,y)$. Здесь векторные функции $U_n(x,y)$, $n = 0,1,2\ldots$, суть решения задач

$$\begin{aligned} &L_0(U_0) = 0, \quad (x,y) \in \Omega, && L_0(U_{n+1}) = -L_1(U_n), \quad (x,y) \in \Omega, \\ &l_0(U_0) = (F,0), \quad x \in (-\infty,\infty), && l_0(U_{n+1}) = l_1(U_n), \quad x \in (-\infty,\infty), \quad n = 0,1,\ldots, \\ &U_0|_{y=h} = 0, && U_{n+1}|_{y=h} = 0, \end{aligned}$$

и предполагается, что все эти задачи имеют единственные решения в классе $W = W_2^1(\Omega)\times W_2^1(\Omega)$ с нормой $\|U\|_W^2 = \|u\|_{W_2^1}^2 + \|w\|_{W_2^1}^2$. Обозначим через $P(U)$ линейный оператор решения задач второго типа из этой группы. Тогда

$$\begin{gathered} U_1 = P(U_0), \quad U_2 = P(U_1) = P^2(U_0), \quad \ldots, \quad U_n = P^n(U_0), \quad \ldots \Rightarrow \\ U = U_0 + \sum_{n=1}^{\infty}(2\varepsilon)^n P^n(U_0) = \sum_{n=0}^{\infty}(2\varepsilon)^n P^n(U_0). \end{gathered} \tag{4}$$

Ряд Неймана (4) сходится, если формально $2\varepsilon\|P(U_0)\|_W < 1$. В этом случае для приближенного решения задачи (3) надо решить задачу для U_0 и затем проделать несколько итераций по вычислению величин $U_n = P(U_{n-1}),$, $n = 1,2,\ldots$.

Задача для U_0 имеет следующий вид:

$$\begin{aligned} &\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x} + E\frac{\partial w}{\partial y}\right) = 0, \quad \frac{\partial}{\partial y}\left(E\frac{\partial u}{\partial x} + E\frac{\partial w}{\partial y}\right) = 0, \quad (x,y) \in \Omega, \\ &\left(E(x,y)\frac{\partial u}{\partial x} + E(x,y)\frac{\partial w}{\partial y}\right)_{y=0} = F(x), \quad (x,y) \in \Gamma_1; \\ &u(x,y) = w(x,y) = 0, \quad (x,y) \in \Gamma_2. \end{aligned} \tag{5}$$

Из первых двух уравнений получим

$$E(x,y)\frac{\partial u}{\partial x} + E(x,y)\frac{\partial w}{\partial y} = C = \text{const}, \quad (x,y) \in \Omega, \tag{6}$$

а из граничного условия на Γ_1 тогда следует

$$\left(E(x,y)\frac{\partial u}{\partial x} + E(x,y)\frac{\partial w}{\partial y}\right)_{y=0} = C = F(x) \quad \forall x.$$

Это значит, что задача для U_0 разрешима только для постоянных давлений: $F(x) = F_0$. Даже при таком предположении задача оказывается недоопределенной, так как из (6) имеем

$$\begin{aligned} &E(x,y)\frac{\partial u}{\partial x} + E(x,y)\frac{\partial w}{\partial y} = F_0, \quad (x,y) \in \Omega, \\ &u(x,y) = w(x,y) = 0, \quad (x,y) \in \Gamma_2, \end{aligned}$$

и найти однозначно $u(x,y), w(x,y)$ нельзя. Отсюда следует, что для предлагаемой схемы решения краевой задачи (1), (2) в задачу для U_0 нужно вводить дополнительные предположения. Аналогичные проблемы возникают при решении задач для U_n, $n > 0$.

4. ДОПОЛНИТЕЛЬНЫЕ ПРЕДПОЛОЖЕНИЯ. АНАЛИТИЧЕСКОЕ РЕШЕНИЕ ЗАДАЧ ДЛЯ ПРИБЛИЖЕНИЙ МАЛОГО ПАРАМЕТРА. РЕШЕНИЕ ОБРАТНОЙ ЗАДАЧИ

Будем дополнительно предполагать, что горизонтальные смещения $u(x,y)$ “малы”: формально $u(x,y) = 0$, $(x,y) \in \Omega$. Тогда из (5) получим для нулевого приближения w_0:

$$\begin{aligned} &\frac{\partial}{\partial x}\left(E\frac{\partial w_0}{\partial y}\right) = 0, \quad \frac{\partial}{\partial y}\left(E\frac{\partial w_0}{\partial y}\right) = 0, \quad (x,y) \in \Omega, \\ &\left(E(x,y)\frac{\partial w_0}{\partial y}\right)_{y=0} = F(x), \quad (x,y) \in \Gamma_1; \quad w_0(x,y) = 0, \quad (x,y) \in \Gamma_2. \end{aligned}$$

Отсюда, как и при получении равенства (6), выводим, что $F(x) = F_0$ и

$$E(x,y)\frac{\partial w_0}{\partial y} = F_0, \quad (x,y) \in \Omega; \quad w_0(x,h) = 0. \tag{7}$$

Поэтому нулевое приближение вертикального смещения есть

$$w_0(x,y) = F_0 \int_h^y \frac{d\eta}{E(x,\eta)}. \tag{8}$$

Формулы (8) и (7) можно использовать при приближенном решении прямой и обратной задач эластографии в нулевом приближении.

Задача для следующего приближения $w_1(x,y)$ принимает вид

$$\begin{aligned} &L_0(U_1) = -L_1(U_0), \quad U_0 = (0,w_0), \quad U_1 = (0,w_1),\\ &l_0(U_1) = l_1(U_0), \quad x \in \Gamma_1,\\ &U_1|_{y=h} = 0 \end{aligned}$$

или

$$\begin{aligned} &\frac{\partial}{\partial x}\left(E(x,y)\frac{\partial w_1}{\partial y}\right) = -\frac{\partial}{\partial y}\left(E(x,y)\frac{\partial w_0}{\partial x}\right), \quad \frac{\partial}{\partial y}\left(E(x,y)\frac{\partial w_1}{\partial y}\right) = -\frac{\partial}{\partial x}\left(E(x,y)\frac{\partial w_0}{\partial x}\right), \quad (x,y) \in \Omega,\\ &E(x,y)\frac{\partial w_1}{\partial y}\bigg|_{y=0} = -E(x,y)\frac{\partial w_0}{\partial y}\,|_{y=0} = -F_0, \quad 0 = \frac{1}{2}\frac{\partial w_0}{\partial x}\bigg|_{y=0}, \quad w_1|_{y=h} = 0. \end{aligned} \tag{9}$$

Фигурирующие здесь уравнения в частных производных совместны не всегда. Например, для достаточно гладкого коэффициента $E(x,y)$, точнее при $E(x,y) \in C^3(\Omega)$, необходимо выполнение условия

$$\frac{\partial^2}{\partial x^2}\left(E(x,y)\frac{\partial w_0}{\partial x}\right) = \frac{\partial^2}{\partial y^2}\left(E(x,y)\frac{\partial w_0}{\partial x}\right), \quad (x,y) \in \Omega. \tag{10}$$

Равенство (10) есть ограничение на класс допустимых функций $E(x,y)$, которое достаточно трудно интерпретируется на практике. Сделаем более ясное дополнительное предположение, гарантирующее выполнение условия (10) – считаем, что модуль Юнга слабо зависит от x в следующем смысле:

$$\frac{\partial w_0}{\partial x} = \frac{\partial}{\partial x}\left(F_0 \int_h^y \frac{d\eta}{E(x,\eta)}\right) \approx 0, \quad (x,y) \in \Omega. \tag{11}$$

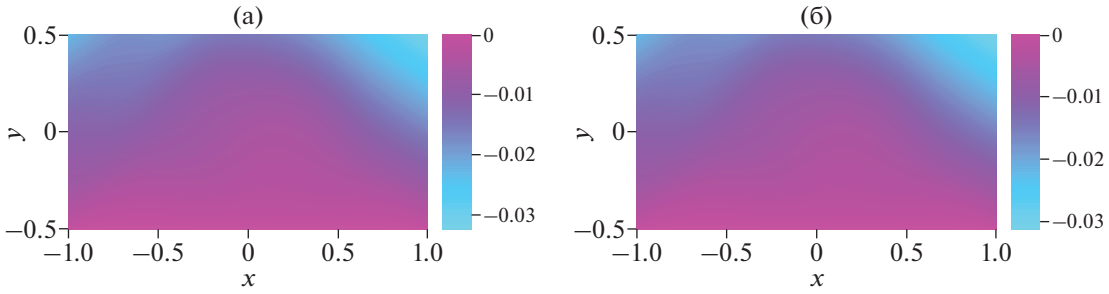
Тогда задача (9) переходит в задачу

$$\begin{aligned} &\frac{\partial}{\partial x}\left(E(x,y)\frac{\partial w_1}{\partial y}\right) = 0, \quad \frac{\partial}{\partial y}\left(E(x,y)\frac{\partial w_1}{\partial y}\right) = 0, \quad (x,y) \in \Omega,\\ &E(x,y)\frac{\partial w_1}{\partial y}\bigg|_{y=0} = -F_0, \quad w_1|_{y=h} = 0, \end{aligned}$$

схожую с задачей для w_0. Поэтому $w_1 = -w_0 = -F_0\int_h^y \frac{d\eta}{E(x,\eta)}$. Аналогичные рассуждения можно провести для всякого $w_n, n > 1 : w_n = -w_{n-1}$. При этом используется одно и то же предположение (11) о модуле Юнга. Соответственно, ряд Неймана (4) будет иметь вид

$$w = w_0 + 2\varepsilon w_1 + (2\varepsilon)^2 w_2 + \ldots + (2\varepsilon)^n w_n + \ldots = w_0 \sum_{n=0}^{\infty} (-1)^n (2\varepsilon)^n = \frac{w_0}{1+2\varepsilon} = \frac{F_0}{(1+2\varepsilon)} \int_h^y \frac{d\eta}{E(x,\eta)}. \tag{12}$$

Таким образом, при сделанных предположении малости смещений $u(x,y)$ и предположении (11) нулевое приближение рассматриваемой прямой задачи с точностью до нормирующего



Фиг. 1. (а) – "Точное" решение прямой задачи (1), (2) с помощью метода конечных элементов. (б) – Приближенное решение $w_0(x, y)$ прямой задачи по формуле (12).

множителя $\dfrac{1}{(1+2\varepsilon)} \approx 0.99$ совпадает с ее точным решением. Это позволяет использовать уравнения (8), (12) для приближенного нахождения функции $E(x, y)$ при решении обратной задачи. Например, это можно сделать в форме

$$E(x,y) = \frac{F_0}{(1+2\varepsilon)}\left(\frac{\partial w}{\partial y}\right)^{-1}, \quad (x,y) \in \Omega, \tag{13}$$

если вычислять частную производную с помощью регуляризованных методов, устойчивых к возмущениям данных $w(x, y)$ обратной задачи.

Отметим, что при эластографической диагностике часто важно знать не абсолютные значения модуля Юнга, а отношение $E(x, y)/E_0$, где E_0 – характерная величина модуля Юнга здоровой ткани. Существенное превышение единицы для такого отношения есть признак, на который следует обратить внимание при диагностике.

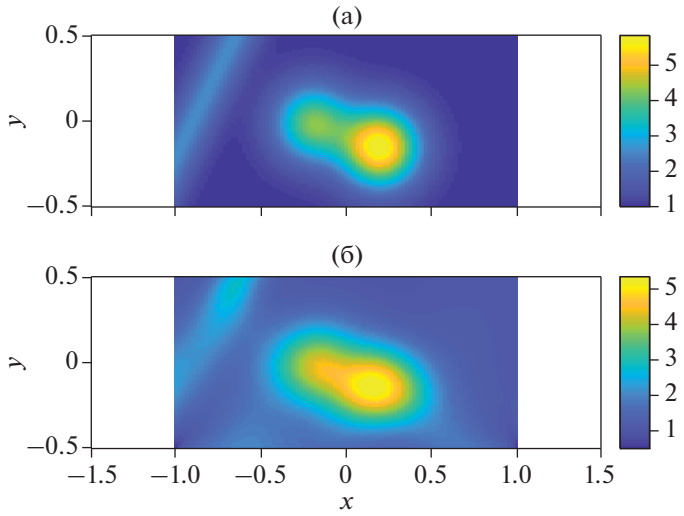
5. ЧИСЛЕННОЕ РЕШЕНИЕ ПРЯМОЙ И ОБРАТНОЙ ЗАДАЧ

Рассмотрим модельный пример, иллюстрирующий приведенную теорию, считая, что все полученные выше формулы обезразмерены. Пусть точное модельное распределение модуля Юнга представлено суммой двух круговых гауссианов и гауссиана, локализованного вдоль прямой:

$$E(x,y) = 32 + 100\exp\left(-\frac{(x+0.2)^2 + (y+0.01)^2}{0.05}\right) + 152\exp\left(-\frac{(x-0.2)^2 + (y+0.15)^2}{0.053}\right) + \\ + 50\exp\left(-\frac{(y-2(x+0.9))^2}{0.085}\right). \tag{14}$$

Для распределения (14) решим прямую задачу нахождения смещений $w(x, y)$ по заданному коэффициенту $E(x, y)$ в "перевернутой и сдвинутой" области конечного размера $\Omega = [-1,1]\times[-0.5,0.5]$. Полагаем, что к верхней границе этой области приложено давление $F_0 = 1$, а левый и правый края области свободны. Решение соответствующей задачи (1), (2), полученное с помощью метода конечных элементов с высокой точностью ("точное решение"), представлено на фиг. 1а. Для сравнения найдем приближенное решение прямой задачи по формуле (12), которая получена для бесконечной полосы при указанных в разд. 4 дополнительных предположениях (фиг. 1б). Видно, что оно, по крайней мере качественно, отражает основные особенности точного решения $w(x, y)$ прямой задачи.

Теперь решим обратную задачу: по данным $w(x, y)$, которые изображены на фиг. 1а, т.е. по решению точной прямой задачи (1), (2), восстановим модуль Юнга. Для этого применим приближенную формулу (13), используя в ней регуляризованную процедуру численного дифференцирования функции $w(x, y)$ (см., например, [15], [16]). Результат в сравнении с точным распределением модуля (14), нормированным на фоновое значение модуля Юнга $E_0 = \min\{E(x, y) : (x, y) \in \Omega\}$, показан на фиг. 2. Полученное приближенное решение обратной задачи с качественной точки зрения достаточно хорошо отражает структуру неоднородности мо-



Фиг. 2. Сравнение точного (а) и восстановленного (б) распределений модуля Юнга при приближенном решении обратной задачи по методу малого параметра. Модули нормированы на фоновое значение $E_{\min}$.

дуля Юнга. С количественной точки зрения оно с удовлетворительной точностью представляет величины отношения $E(x, y)/E_0$.

Расчет по формуле (13) на персональном компьютере средней производительности занимает 0.5 мс для сеток размера 100 × 100. Скорость расчета и достаточная точность получаемого приближенного решения позволяют надеяться, что этот подход можно использовать в реальной онкологической диагностике. Можно также использовать данное приближенное решение в качестве начального приближения при решении обратной задачи двумерной квазистатической эластографии для исходной системы (1), (2) по методам из работ [8]–[10].

СПИСОК ЛИТЕРАТУРЫ

1. *Gao L., Parker K., Lerner R. et al.* Imaging of the elastic properties of tissue – a review // Ultrasound Med. Biol. 1996. V. 22. P. 959–977.
2. *Ophir J., Alam S., Garra B. et al.* Elastography: ultrasonic estimation and imaging of the elastic properties of tissues // Proc. Inst. Mech. Eng. Part H: J. Eng. Med. 1999. V. 213. P. 203–233.
3. *Greenleaf J.F., Fatemi M., Insana M.* Selected methods for imaging elastic properties of biological tissues // Annu. Rev. Biomed. Eng. 2003. V. 5. P. 57–78.
4. *Parker K.J., Taylor L.S., Gracewski S. et al.* A unified view of imaging the elastic properties of tissue // J. Acoust. Soc. Am. 2005. V. 117. P. 2705–2712.
5. *Doyley M.* Model-based elastography: a survey of approaches to the inverse elasticity problem // Phys Med Biol. 2012. V. 57. P. R35–R73.
6. *Oberai A.A., Gokhale N.H., Feijoo G.R.* Solution of inverse problems in elasticity imaging using the adjoint method // Inverse Probl. 2003. V. 19. P. 297–313.
7. *Richards M., Barbone P., Oberai A.* Quantitative three-dimensional elasticity imaging from quasi-static deformation: a phantom study // Phys. Med. Biol. 2009. V. 54. P. 757–779.
8. *Leonov A.S., Sharov A.N., Yagola A.G.* A posteriori error estimates for numerical solutions to inverse problems of elastography // Inverse Probl. Sci. Eng. 2017. V. 25. P. 114–128.
9. *Leonov A.S., Sharov A.N., Yagola A.G.* Solution of the inverse elastography problen for parametric classes of inclusions with a posteriori error estimate // J. Inverse Ill-Posed Probl. 2017. V. 26. P. 1–7.
10. *Leonov A.S., Sharov A.N., Yagola A.G.* Solution of the three-dimensional inverse elastography problem for parametric classes of inclusions // Inverse Probl. Sci. Eng. 2021. V. 29. N. 8. P. 1055–1069.
11. *Rychagov M., Khaled W., Reichling S. et al.* Numerical modeling and experimental investigation of biomedical elastographic problem by using plane strain state model // Fortsch. Der Akustik. 2003. V. 29. P. 586–589.
12. *Ладыженская О.А.* Краевые задачи математической физики. М.: Наука, 1973.
13. *Тихонов А.Н., Васильева А.Б., Свешников А.Г.* Дифференц. ур-ния. М.: Наука, 1980.
14. *Треногин В.А.* Функциональный анализ. М.: Наука, 1980.
15. *Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г.* Численные методы решения некорректных задач. М.: Наука, 1990.
16. *Леонов А.С.* Решение некорректно поставленных обратных задач. Очерк теории, практические алгоритмы и демонстрации в МАТЛАБ. М.: Либроком, 2009.

EDN: DFDCGS
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 861–871

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.634

АНАЛИТИЧЕСКИЕ РЕШЕНИЯ МОДЕЛЬНЫХ КИНЕТИЧЕСКИХ УРАВНЕНИЙ ПЕРЕНОСА ИЗЛУЧЕНИЯ И УРАВНЕНИЯ ЭНЕРГИИ

© 2022 г. Н. Я. Моисеев[1,*], В. М. Шмаков[1,]**

[1] 456770 Снежинск, Челябинская обл., а/я 245, ул. Васильева, 13, ФГУП "РФЯЦ-ВНИИТФ им. акад. Е.И. Забабахина", Россия

**e-mail: nik.moiseev.43@mail.ru*

***e-mail: v.m.shmakov@vniitf.ru*

Поступила в редакцию 11.03.2021 г.
Переработанный вариант 12.10.2021 г.
Принята к публикации 14.01.2022 г.

Рассмотрена модельная система нестационарных кинетических уравнений переноса теплового излучения и уравнения энергии в многогрупповом изотропном приближении в средах с постоянными коэффициентами поглощения и кусочно-постоянной функцией Планка в группах. Для модельной системы уравнений получены аналитические решения для плоской геометрии и для шара. Аналитические решения получены путем перехода от решения сложной системы уравнений к решению более простой, которая имеет известное аналитическое решение. Обратный переход дает решение сложных уравнений. Приведены решения тестовых задач для плоского слоя и для шара. Библ. 6. Фиг. 10. Табл. 2.

Ключевые слова: кинетическое уравнение переноса теплового излучения и уравнения энергии, аналитические решения.

DOI: 10.31857/S004446692205009X

1. ВВЕДЕНИЕ

Кинетическое нестационарное уравнение переноса теплового излучения является интегро-дифференциальным уравнением. В работах [1]–[3] рассматриваются подходы к решению этого уравнения в различных предположениях относительно коэффициентов, что позволяет получать упрощенные уравнения переноса. Так, если коэффициенты в уравнении переноса положить равными нулю, то уравнение записывается в простейшей форме и моделирует перенос излучения в вакууме. Для уравнения переноса излучения в вакууме А.В. Вронским были получены аналитические решения, которые используются для отладки численных методов. Однако, если решаются уравнения переноса теплового излучения и уравнения энергии в средах с поглощением и переизлучением, то этих решений для отладки численных методов недостаточно.

Здесь рассмотрен подход к аналитическому решению модельных нестационарных кинетических уравнений переноса теплового излучения и уравнения энергии в многогрупповом изотропном приближении в средах с поглощением и переизлучением для плоской и сферически-симметричной геометрии. В основе подхода лежат аналитические решения уравнения переноса излучения в вакууме. Аналитические решения модельных уравнений выписываются в квадратурах, часть из которых выражается через элементарные функции, а другая содержит интегралы экспоненциального типа. Интегралы экспоненциального типа вычисляются по многоточечным квадратурам Гаусса [4] повышенной точности. Аналитические решения модельных задач подтверждены численными расчетами по методике [5].

2. ТОЧНЫЕ РЕШЕНИЯ МОДЕЛЬНЫХ КИНЕТИЧЕСКИХ НЕСТАЦИОНАРНЫХ УРАВНЕНИЙ ПЕРЕНОСА ИЗЛУЧЕНИЯ И УРАВНЕНИЯ ЭНЕРГИИ

2.1. Постановка задачи

Предположим, что коэффициенты рассеяния и функция источника равны нулю, плотность вещества $\rho = 1$. Систему кинетических нестационарных уравнений переноса теплового излучения и уравнения энергии в многогрупповом изотропном приближении запишем в виде [2]

$$\begin{gathered}\frac{1}{c}\frac{\partial I_g}{\partial t} + \mathbf{\Omega\nabla} I_g + \alpha_g I_g = 0.5\alpha_g B_g,\\ \frac{dE}{dt} = \sum_{g=1}^{G}\alpha_g\left(U_g - B_g\right),\\ U_g = \int_{-1}^{1} I_g d\mu,\end{gathered} \tag{2.1}$$

$$\mathbf{\Omega\nabla} = \begin{cases}\mu\dfrac{\partial}{\partial r} + \dfrac{1-\mu^2}{r}\dfrac{\partial}{\partial\mu} - \text{для сферически-симметричной геометрии,}\\ \mu\dfrac{\partial}{\partial x} - \text{для плоской геометрии.}\end{cases}$$

Здесь x, r и t – независимые переменные по пространству и времени соответственно, μ – косинус угла между направлением движения частиц и осью x или r, $-1 \le \mu \le 1$, g – индекс энергетической группы, I_g, E – неизвестные функции, $E(t,x)$ – удельная внутренняя энергия вещества, $T(t,x)$ – температура вещества, c – скорость света, $I_g(t,x,\mu)$ – спектральная интенсивность энергии излучения в группе g, $U_g(t,x)$ – спектральная плотность энергии излучения, умноженная на скорость света, $\alpha_g(x,T)$ – коэффициент поглощения,

$$B_g(T) = \frac{8\pi}{c^2\overline{h}^3}\int_{\varepsilon_g}^{\varepsilon_{g+1}}\frac{\varepsilon^3}{\exp(\varepsilon/T)-1}d\varepsilon$$

есть интенсивность равновесного излучения (функция Планка), умноженная на скорость света, $\overline{h}$ – постоянная Планка. Разностная сетка по энергии фотонов включает G независимых групп с энергиями $\varepsilon_1,\ldots,\varepsilon_g,\ldots,\varepsilon_G$. Разностная сетка по переменной μ с центрами $\mu_m = 0.5(\mu_{m+1/2} + \mu_{m-1/2})$ и шагами $\Delta\mu = \mu_{m+1/2} - \mu_{m-1/2}$, $m = 1,\ldots,M$, включает M направлений движения частиц (фотонов).

Система уравнений (2.1) замыкается уравнением состояния вещества в форме $E = E(T)$. Начальные условия: $T(0,x) = T_0(x)$, $I_g(0,x,\mu) = I_{g,0}(x,\mu)$. Требуется найти решение задачи Коши для $t > 0$ в областях $D = \{x_L \le x \le x_R, -1 \le \mu \le 1\}$, $D = \{0 \le r \le r_0, -1 \le \mu \le 1\}$ в случае плоской или сферически-симметричной геометрии соответственно.

2.2. Точные решения модельного уравнения переноса частиц в средах с поглощением

Для упрощения исследования системы уравнений (2.1) коэффициенты поглощения положим равными некоторым постоянным $\alpha_g(x,T) = \text{const}$. Функцию Планка аппроксимируем кусочно-постоянной функцией $B_g(T) = \text{const}$ в каждой группе. Уравнение состояния вещества возьмем в форме $E = 0.81T$. Систему кинетических уравнений переноса излучения и уравнения энергии с постоянными коэффициентами поглощения и кусочно-постоянной функцией Планка будем называть *модельной* системой уравнений. Для модельной системы уравнений рассмотрим решение задач Коши в плоской и сферически-симметричной геометриях. Для плоской геометрии начальные данные задаются в слое $|x| \le x_0$, для сферической – в шаре с радиусом r_0. В первом случае задачу будем называть задачей *о плоском слое*, во втором – задачей *о шаре*. Пусть в начальный мо-

мент времени на всей оси x задана температура вещества $T(0,x) = T_0(x)$ и задано однородно распределение частиц (фотонов)

$$I_g(0,x,\mu) = \begin{cases} I_{g,0}(x,\mu), & |x| \le x_0, \quad I_{g,0} \ge 0.5B_g(T_0), \\ 0.5B_g(T_0), & |x| > x_0. \end{cases}$$

Требуется определить интенсивность излучения, плотность энергии излучения и температуру вещества для $t > 0$.

Аналитические решения модельной системы уравнений (2.1) для решения поставленной задачи получим следующим образом. Выполнив в системе уравнений (2.1) преобразование

$$I_g = J_g \exp(-ct\alpha_g) + 0.5B_g, \qquad \widehat{U_g} = \int_{-1}^{1} J_g d\mu \tag{2.2}$$

получим эквивалентную систему уравнений

$$\frac{1}{c}\frac{\partial J_g}{\partial t} + \boldsymbol{\Omega}\nabla J_g = 0, \qquad \frac{dE}{dt} = \sum_{g=1}^{G} \alpha_g \widehat{U_g} \exp(-ct\alpha_g) \tag{2.3}$$

с начальными данными для уравнения переноса в (2.3)

$$J_g(0,x,\mu) = \begin{cases} I_{g,0} - 0.5B_g, & |x| \le x_0, \\ 0, & |x| > x_0. \end{cases}$$

Уравнение переноса в (2.3) можно интерпретировать как уравнение переноса излучения в вакууме, точные решения которого для интенсивности и плотности энергии излучения записываются в виде

$$J_g(t,x,\mu) = J_{g,0}\xi(t,x), \quad \widehat{U_g}(t,x) = U_{g,0}\xi(t,x) \tag{2.4}$$

соответственно. Подставив (2.4) в выражения (2.2), получим для интенсивности и плотности энергии излучения решения модельного уравнения переноса в (2.1), которые запишем в виде

$$I_g(t,x,\mu) = \gamma\xi(t,x) I_{g,0} + 0.5(1 - \gamma\xi(t,x)) B_g,$$
$$U_g(t,x) = \gamma\xi(t,x) U_{g,0} + (1 - \gamma\xi(t,x)) B_g, \tag{2.5}$$
$$\gamma = \exp(-ct\alpha_g)$$

соответственно. Отметим, что решения уравнения переноса излучения в вакууме находятся в плоскости (t,x) в областях, которые отделены друг от друга характеристиками

$$dx/dt = \pm c,$$

выходящими из точек x_0 и $-x_0$. Функция $\xi(t,x)$ вычисляется в задаче о плоском слое в шести областях следующим образом. Если $ct \le x_0$, то

$$\xi(t,x) = \begin{cases} 0, & -\infty < x < -x_0 - ct, \quad x_0 + ct < x < \infty, \quad \text{области } I \text{ и } II, \\ \dfrac{x_0 + ct + x}{2ct}, & -x_0 - ct \le x < -x_0 + ct, \quad \text{область } IV, \\ 1, & -x_0 + ct \le x \le x_0 - ct, \quad \text{область } III, \\ \dfrac{x_0 + ct - x}{2ct}, & x_0 - ct < x \le x_0 + ct, \quad \text{область } V. \end{cases}$$

Если $ct > x_0$, то

$$\xi(t,x) = \begin{cases} 0, & -\infty < x < -x_0 - ct, \quad x_0 + ct < x < \infty, \quad \text{области } I \text{ и } II, \\ \dfrac{x_0 + ct + x}{2ct}, & -x_0 - ct \le x < x_0 - ct, \quad \text{область } IV, \\ \dfrac{x_0}{ct}, & x_0 - ct \le x \le -x_0 + ct, \quad \text{область } VI, \\ \dfrac{x_0 + ct - x}{2ct}, & -x_0 + ct < x \le x_0 + ct, \quad \text{область } V. \end{cases}$$

Функция $\xi(t,r)$ вычисляется в задаче о шаре радиуса r_0 в четырех областях следующим образом. Если $ct \le r_0$, то

$$\xi(t,x) = \begin{cases} 1, & 0 \le r < r_0 - ct, \quad \text{область } I, \\ \dfrac{r_0^2 - (r-ct)^2}{4rct}, & r_0 - ct \le r < r_0 + ct, \quad \text{область } III, \\ 0, & r_0 + ct \le r < \infty, \quad \text{область } II. \end{cases}$$

Если $ct > r_0$, то

$$\xi(t,x) = \begin{cases} 0, & 0 \le r < -r_0 + ct, \quad \text{область } IV, \\ \dfrac{r_0^2 - (r-ct)^2}{4rct}, & -r_0 + ct \le r < r_0 + ct, \quad \text{область } III, \\ 0, & r_0 + ct \le r < \infty, \quad \text{область } II. \end{cases}$$

2.3. Аналитические решения модельного уравнения энергии в средах с поглощением

Аналитические решения модельного уравнения энергии получим следующим образом. Подставив выражение (2.5) для плотности энергии излучения в уравнение энергии в (2.1), получим обыкновенное дифференциальное уравнение в виде

$$\frac{dE}{dt} = \sum_{g=1}^{G} \alpha_g (U_{g,0} - B_g) \xi(t,x) \exp(-ct\alpha_g). \tag{2.6}$$

Проинтегрировав уравнение (2.6) на интервале $[0,t]$, получим для вычисления удельной внутренней энергии вещества выражение

$$E(t,x) = E(0,x) + \sum_{g=1}^{G} \int_0^t \alpha_g (U_{g,0} - B_g) \xi(\tau,x) \exp(-c\tau\alpha_g)\, d\tau. \tag{2.7}$$

Если $\xi(t,x) = 0$, то удельная внутренняя энергия вещества остается постоянной $E(t,x) = E(0,x)$ в задаче о плоском слое в областях I и II, в задаче о шаре – в областях II и IV. Если $\xi(t,x) = 1$, то интеграл в (2.7) берется в квадратурах точно. Взяв интеграл, получим для вычисления удельной внутренней энергии вещества в задачах о плоском слое в области III и о шаре в области I выражение

$$E(t,x) = E(0,x) + \frac{1}{c} \sum_{g=1}^{G} (1-\gamma)(U_{g,0} - B_g), \quad \gamma = \exp(-ct\alpha_g). \tag{2.8}$$

Если $\xi \ne 0$, то удельная внутренняя энергия вещества вычисляется в задаче о плоском слое в областях I, V и VI из выражений

$$E(t,x) = E(0,x) + 0.5 \sum_{g=1}^{G} \alpha_g (U_{g,0} - B_g) \int_0^t (x_0 + c\tau \pm x) \frac{\exp(-c\tau\alpha_g)}{c\tau}\, d\tau, \tag{2.9}$$

$$E(t,x) = E(0,x) + \sum_{g=1}^{G} \alpha_g (U_{g,0} - B_g) \int_0^t x_0 \frac{\exp(-c\tau\alpha_g)}{c\tau} d\tau \tag{2.10}$$

соответственно. Здесь плюс берется в области *IV*, минус – в области *V*. Удельная внутренняя энергия вещества вычисляется в задаче о шаре в области *III* из выражения

$$E(t,x) = E(0,x) + \sum_{g=1}^{G} \alpha_g (U_{g,0} - B_g) \int_0^t \frac{r_0^2 - (r - c\tau)^2}{4r} \frac{\exp(-c\tau\alpha_g)}{c\tau} d\tau. \tag{2.11}$$

Интегралы в выражениях (2.9)–(2.11) являются интегралами экспоненциального типа, не интегрируются в элементарных функциях, и имеют устранимую особенность в нуле. Пусть удельная внутренняя энергия вещества в задаче о плоском слое вычисляется в точке $A(t,x)$, которая находится в области *VI*, где $t > t_0$. Если $|x| \le x_0$ или $|x| > x_0$, то интеграл в уравнении (2.9) можно представить в виде суммы трех интегралов в областях *III*, *V*, *VI* или *II*, *V*, *VI* на интервалах по времени $[0,t_1]$, $[t_1,t_2]$, $[t_2,t]$ соответственно. Здесь $t_1 = (x_0 - x)/c$, $t_2 = (x_0 + x)/c$ – это времена прихода возмущений по характеристикам из точек x_0 и $-x_0$ в точку x соответственно. При таком подходе к вычислению интегралов особенность в нуле устраняется за счет того, что функция $\xi(t,x) = 1$ в области *III* и $\xi(t,x) = 0$ в областях *I*, *II*. Особенности в нуле при вычислении интегралов в уравнениях (2.10) и (2.11) устраняются аналогичным образом. Экспоненциальные интегралы вычисляются по какой-либо квадратурной формуле с заданной точностью. Если интегралы вычислять по квадратурным формулам правых прямоугольников, которые обозначим как (С1), методом средних (С2) или по двухточечной формуле Гаусса (С3), то погрешности вычисления будут порядка $O(\tau)$, $O(\tau^2)$, $O(\tau^3)$ соответственно.

3. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ РЕШЕНИЙ МОДЕЛЬНЫХ ЗАДАЧ

Аналитические решения модельных уравнений (2.1) проверялись путем решения простейших задач по переносу излучения в задачах о плоском слое и шаре. По энергии фотонов используется 15-групповое приближение из работы [6] с умноженными на 10 границами интервалов: [0, 3, 6, 8, 12, 15, 18, 24, 27, 30, 40, 50, 70, 90, 110, 150]. Средние значения энергии в группах относятся к центрам интервалов и равны полусумме значений энергий на границах интервалов.

В расчетах использовалась модельная функция Планка: [0.029, 0.202, 0.391, 0.609, 0.813, 0.927, 1.0, 0.977, 0.926, 0.762, 0.489, 0.2057, 0.051, 0.010, 0.001]. Модельная функция Планка построена путем усреднения обычной функции Планка в диапазоне температур $[0, T_{\max}]$ для $T_{\max} = 10$ кэВ и нормированием по максимальному значению. Для счета задач нормированная функция Планка умножается на положительный коэффициент. В дальнейшем такие функции будем различать по этим коэффициентам. Например, запись $B_g = 10$ определяет функцию Планка, которая получена из нормированной функции, умноженной на 10.

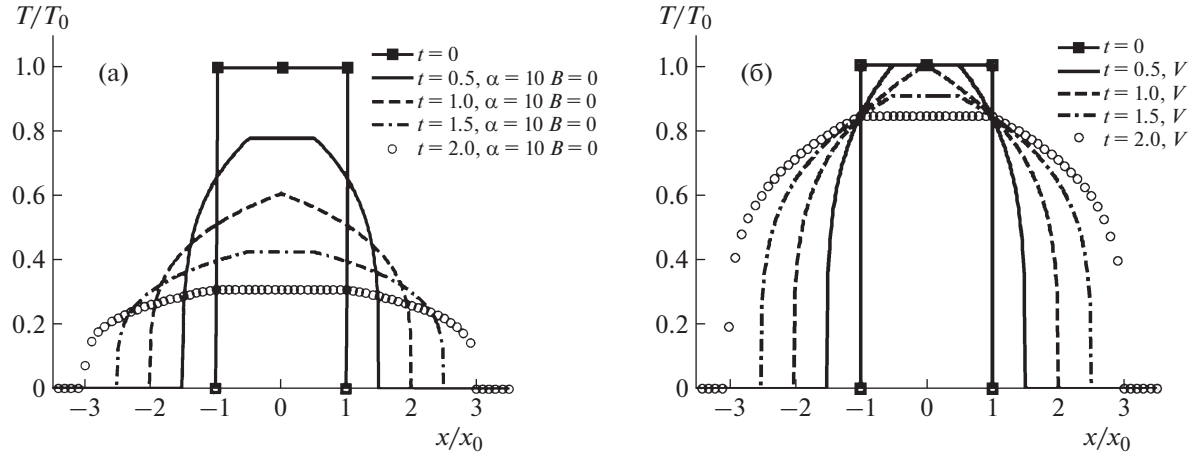
Решения находились на равномерной разностной сетке с шагом интегрирования по пространству равным $h = 0.002$ в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$. Удельная внутренняя энергия вещества вычислялась с контролем точности по квадратурным формулам С1, С2, С3. Контроль точности осуществлялся по результатам двух расчетов с шагами интегрирования по времени τ и 0.5τ из анализа условия

$$|I(\tau) - I(0.5\tau)| < \varepsilon.$$

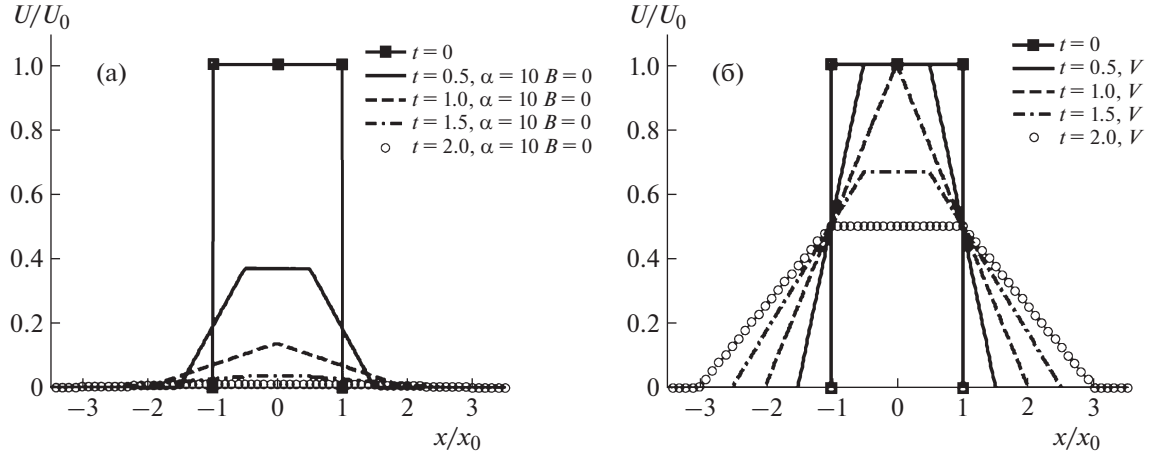
Здесь ε – константа сходимости. Если условие выполняется, то интеграл полагается равным $I(0.5\tau)$. Если условие не выполняется, то шаг интегрирования по времени уменьшается в два раза и вычисление интеграла повторяется [4]. Начальный шаг интегрирования $t_0 = x_0/c = 6.6(6)\text{e} - 5$.

3.1. Перенос излучения в оптически прозрачных средах с поглощением. Задача о плоском слое

Задача 1. Коэффициенты поглощения равны $\alpha_g = 0$ и $\alpha_g = 10$ при переносе излучения в вакууме и в среде с поглощением соответственно. Модельные функции Планка в группах равны



Фиг. 1. Зависимости от x/x_0 температуры излучения: (а) – в среде с поглощением, (б) – в вакууме.



Фиг. 2. Зависимости от x/x_0 плотности энергии излучения: (а) – в среде с поглощением, (б) – в вакууме.

$B_g = 0$. Начальные данные: если $|x| \le x_0 = 0.2$ см, то $I_0 = 500$ и $U_0 = 1000$, иначе $I_0 = 0$, $U_0 = 0$, температура вещества $T_0 = 0.001$ кэВ. Найти распределения температуры вещества и плотности энергии излучения на интервале $[-0.8, 0.8]$ в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

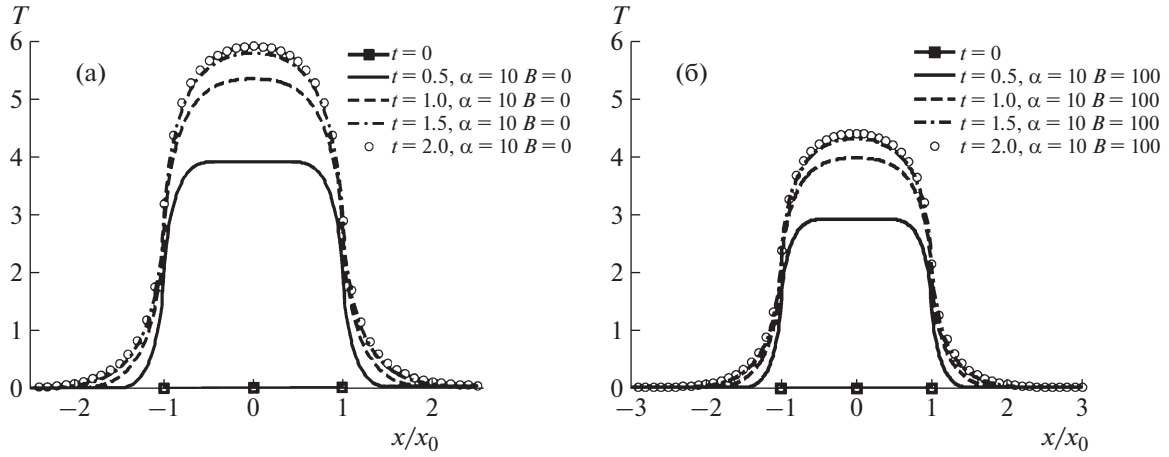
На фиг. 1–10 введены обозначения: V – обозначение вакуума, t в легендах на фигурах – отношение t/t_0.

На фиг. 1 представлены зависимости от x/x_0 температуры излучения в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$ в среде с поглощением и в вакууме.

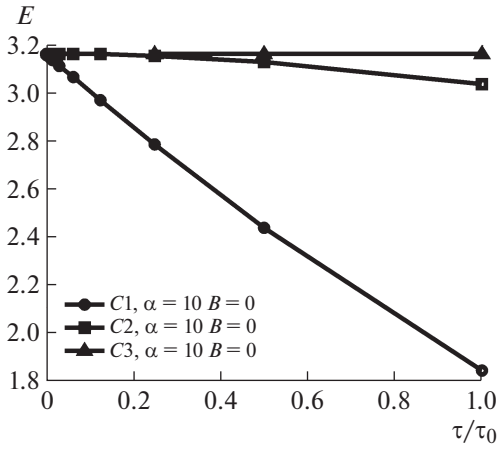
На фиг. 2 представлены зависимости от x/x_0 плотности энергии излучения в среде с поглощением и в вакууме в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

Из анализа графиков на фиг. 1 и 2 следует, что положения фронтов тепловых волн излучения в вакууме и в среде с поглощением совпадают между собой. Температуры и плотности энергии излучения существенно различаются.

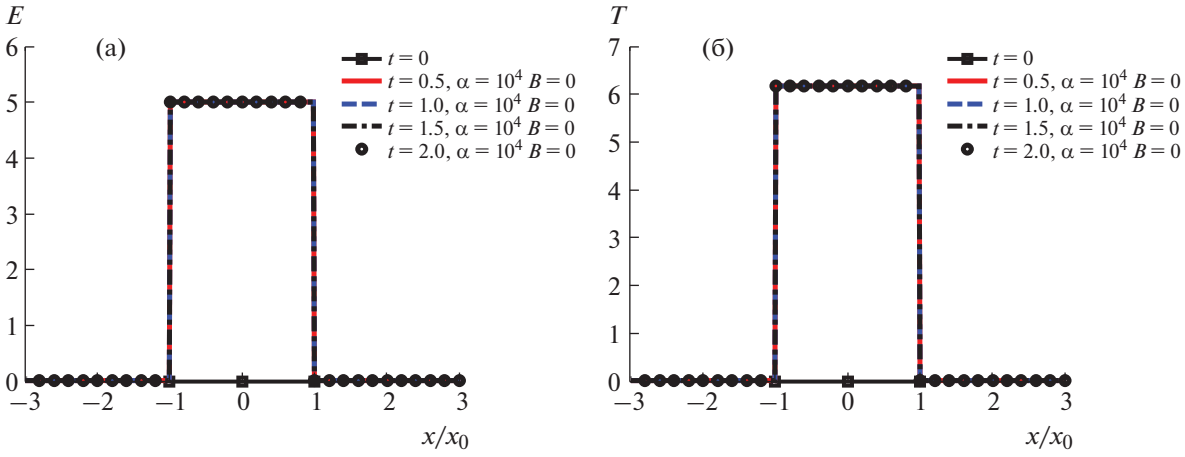
На фиг. 3 представлены зависимости от x/x_0 температуры вещества в среде с коэффициентом поглощения в группах $\alpha_g = 10$ и функциями Планка $B_g = 0$, $B_g = 100$ в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.



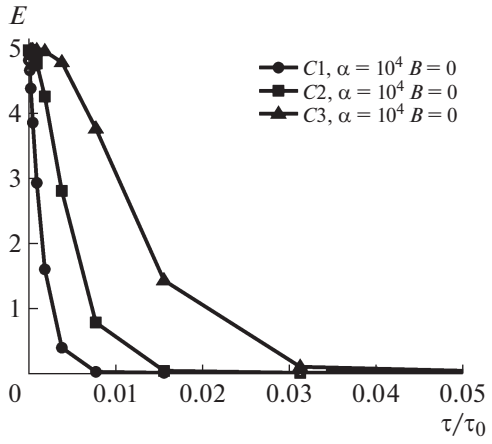
Фиг. 3. Зависимости от x/x_0 температуры вещества в среде с коэффициентом поглощением $\alpha_g = 10$: (а) – $B_g = 0$, (б) – $B_g = 100$.



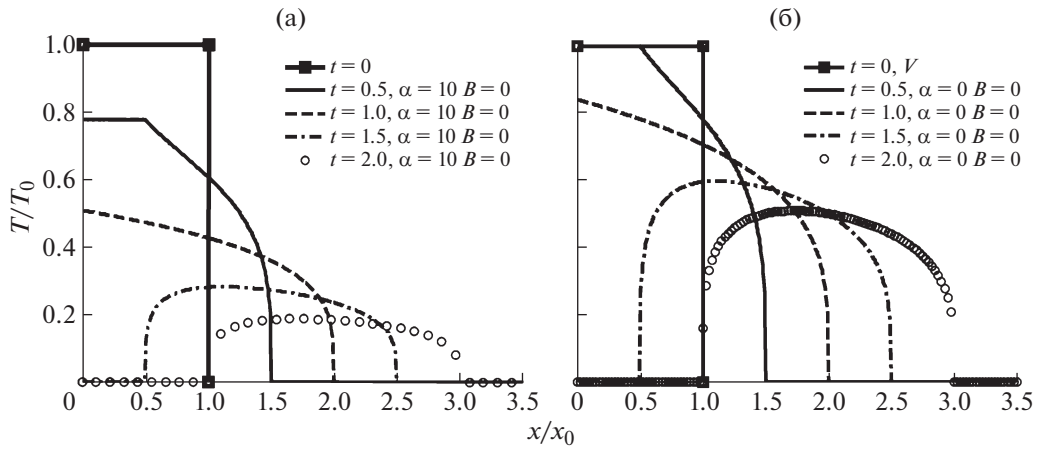
Фиг. 4. Зависимости от τ/τ_0 удельной внутренней энергии вещества: кружки – C1, квадраты – C2, треугольники – C3.



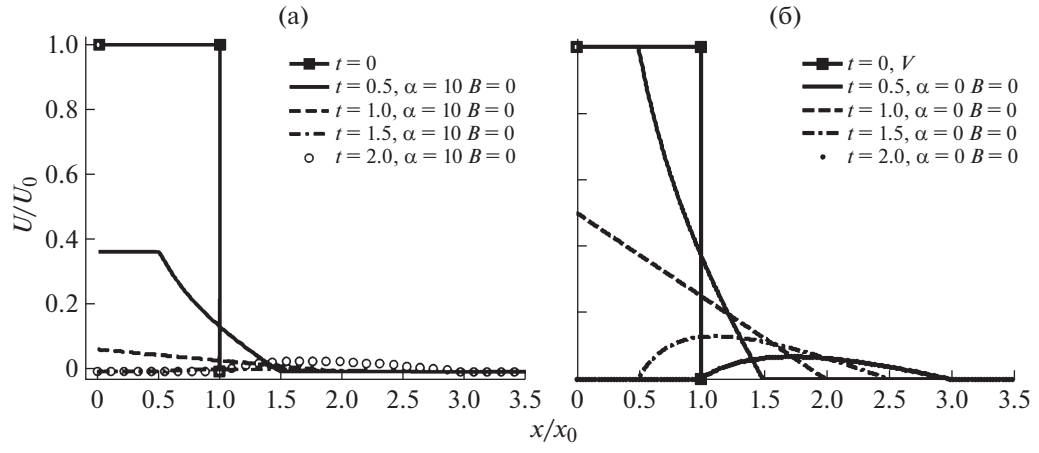
Фиг. 5. Зависимости от x/x_0: (а) – удельной внутренней энергии вещества, (б) – температуры вещества.



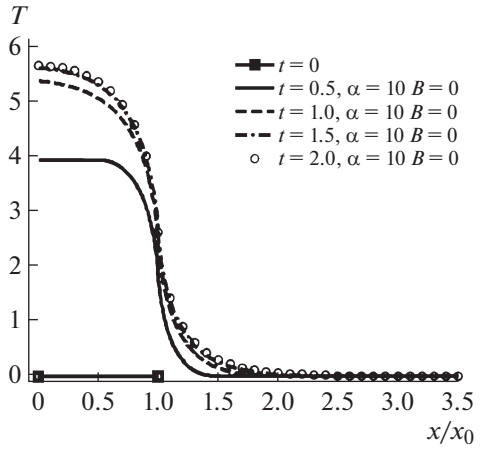
Фиг. 6. Зависимости от τ/τ_0 удельной внутренней энергии вещества.



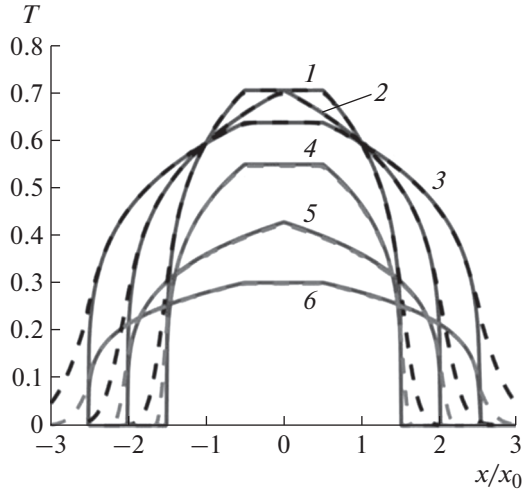
Фиг. 7. Профили температур излучения: (а) – в среде с поглощением, (б) – в вакууме.



Фиг. 8. Профили плотности энергии излучения: (а) – в среде с поглощения, (б) – в вакууме.



Фиг. 9. Профили температур вещества в среде с коэффициентом поглощения $\alpha_g = 10$ и функцией Планка $B_g = 0$.



Фиг. 10. Зависимости от x/x_0 температуры излучения: *1–3* – в вакууме, *4–6* – в среде с поглощением. Сплошные линии – точные решения, штриховые – численные решения.

Точные максимальные значения температуры, рассчитанные в задачах с функциями Планка $B_g = 0$, $B_g = 100$ в момент времени $t/t_0 = 0.5$, равны $T\left(B_g = 0\right) = 3.90297875820097$ и $T\left(B_g = 100\right) = 2.89551187902263$ соответственно.

На фиг. 4 представлены зависимости от τ/τ_0 удельной внутренней энергии вещества в расчетах на сходимость в момент времени $t/t_0 = 0.5$ при вычислении интегралов по квадратурным формулам C1, C2, C3.

Таблица 1

Метод	E $\varepsilon = 0.1$	τ_1	Число шагов	E $\varepsilon = 0.0001$	τ_2	Число шагов
C1	3.11147	1.4167 e-6	32	3.16050	2.034e-9	16384
C2	3.12792	1.66(2.6) e-5	2	3.16057	5.208e-7	64
C3	3.15989	3.33(2.3) e-5	1	3.16059	8.333e-6	4

Таблица 2

Метод	E $\varepsilon = 0.1$	τ_1	Число шагов	E $\varepsilon = 0.0001$	τ_2	Число шагов
C1	4.92409	1.0 e-9	32768	4.92409	5.0e-10	65536
C2	4.95067	1.63e-8	2048	4.99980	5.0e-10	32768
C3	4.9850	6.5e-8	512	4.99999	8.0e-9	4096

В табл. 1 приведены удельные внутренние энергии вещества, шаги интегрирования и число шагов по времени при вычислении интегралов в точке $x = 0.011$ с константами сходимости $\varepsilon = 0.1$, $\varepsilon = 0.0001$ и функцией Планка $B_g = 0$ в момент времени $t/t_0 = 0.5$.

3.2. Перенос излучения в оптически плотных средах с поглощением. Задача о плоском слое

Задача 2. Коэффициенты поглощения в группах равны $\alpha_g = 10000$, функция Планка – $B_g = 0$, удельная внутренняя энергия вещества – $E_0 = 0.00081$, температура – $T_0 = 0.001$ кэВ. Плотность энергии излучения: если $|x| \le x_0 = 0.2$ см, то $I_0 = 500$ и $U_0 = 1000$, иначе $I_0 = 0, U_0 = 0$, удельная внутренняя энергия излучения – $E_{0,U} = 5$. Требуется определить температуру вещества и излучения в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

Технология счета задачи была такой же, как и в задаче 1. Однако для достижения точного максимального значения температуры вещества потребовался меньший шаг интегрирования по времени, чем в задаче 1.

На фиг. 5 представлены зависимости от x/x_0 удельной внутренней энергии и температуры вещества в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

Результаты расчетов показали, что удельная внутренняя энергия излучения полностью перешла (преобразовалась) в удельную внутреннюю энергию вещества. Удельная внутренняя энергия вещества, рассчитанная по точным формулам (2.6) без учета E_0, равна $E_0 = 5$, температура вещества – T = 6.17383950617284.

На фиг. 6 представлены зависимости от τ/τ_0 удельной внутренней энергии вещества в расчетах на сходимость в момент времени $t/t_0 = 0.5$ при вычислении интегралов по квадратурным формулам C1, C2, C3.

В табл. 2 приведены удельные внутренние энергии вещества, шаги интегрирования и число шагов по времени при вычислении интегралов в точке $x = 0.011$ с константами сходимости $\varepsilon = 0.1$, $\varepsilon = 0.0001$ и функцией Планка $B_g = 0$ в момент времени $t/t_0 = 0.5$.

Вычисление интегралов по квадратурным формулам C1, C2, C3 можно интерпретировать как численное решение уравнения энергии по разностным схемам Эйлера первого порядка, предиктор-корректор второго порядка и по схеме повышенного третьего порядка точности соответственно. Из анализа графиков на фиг. 4, 6 и из табл. 1, 2 следует, что для достижения заданной точности с константами сходимости $\varepsilon = 0.1$ и $\varepsilon = 0.0001$ по схеме C3 число шагов в 2, 16 и 4, 8 раз меньше, чем по схеме C2, и в 32, 4096 и в 64, 16 раз меньше, чем по схеме C1 в задачах 1 и 2 по переносу излучения в прозрачных и плотных средах соответственно. По схеме C2 число шагов в 16, 256 и в 16, 2 раза меньше, чем по схеме C1. Поэтому разностные схемы Эйлера первого порядка точности для решения уравнения энергии не эффективны и малопригодны для серийного счета задач. Схемы C2, C3 существенно эффективнее, чем схемы первого порядка точности: шаг интегрирования по времени на один, два порядка больше, чем в схеме первого порядка. Обе схемы позволяют получить результаты с заданной точностью за приемлемое время счета.

3.3. Перенос излучения в оптически прозрачных средах с поглощением. Задача о шаре

Задача 3. Начальные данные такие же, как в задаче 1 и заданы в шаре $r \le r_0 = 0.2$ см. Найти распределения температуры вещества и плотности энергии излучения в шаре радиуса $r = 0.6$ см в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$. Шаг интегрирования по пространству $h = 0.001$. Начальный шаг интегрирования $t_0 = 6.6(6)\text{e} - 5$.

На фиг. 7 представлены профили температур излучения в среде с коэффициентом поглощения $\alpha_g = 10$ и функцией Планка $B_g = 0$ и в вакууме в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

На фиг. 8 представлены профили плотности энергии излучения в среде с коэффициентом поглощения $\alpha_g = 10$ и функцией Планка $B_g = 0$ и в вакууме в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

На фиг. 9 представлены профили температур вещества при переносе излучения в среде с коэффициентом поглощения $\alpha_g = 10$ и функцией Планка $B_g = 0$ в моменты времени $t/t_0 = 0.5, 1, 1.5, 2$.

4. РЕЗУЛЬТАТЫ РАСЧЕТОВ МОДЕЛЬНОЙ ЗАДАЧИ МОДИФИЦИРОВАННЫМ МЕТОДОМ РАСЩЕПЛЕНИЯ

В разделе для сравнения с точными решениями приведены численные решения модельной задачи 1, которые получены модифицированным методом расщепления, [5]. Численные решения уравнения переноса излучения получены на фиксированной разностной сетке с шагом по пространству $h = 0.002$ и с коэффициентом запаса устойчивости $k = 0.8$.

На фиг. 10 представлены профили температуры излучения в вакууме и в среде с поглощением в моменты времени $t/t_0 = 0.5, 1, 1.5$.

Из анализа графиков на фиг. 10 следует, что численные решения уравнения переноса излучения и уравнения энергии, которые получены модифицированным методом расщепления, согласуются с точными решениями. Видно, что размазывание профиля в окрестности оси при переносе излучения в вакууме больше, чем в среде с поглощением.

5. ЗАКЛЮЧЕНИЕ

Получены аналитические решения модельных нестационарных кинетических уравнений переноса излучения и энергии в изотропном многогрупповом приближении в средах с постоянными коэффициентами поглощения и кусочно-постоянной функцией Планка в группах в задачах о плоском слое и о шаре.

Аналитические решения модельных задач подтверждены численными решениями, которые получены модифицированным методом расщепления.

Показано, что разностные схемы Эйлера первого порядка точности для решения уравнения энергии не эффективны и малопригодны.

СПИСОК ЛИТЕРАТУРЫ

1. *Бай Ши-и*. Динамика излучающего газа. М.: Мир, 1968.
2. *Четверушкин Б.Н.* Математическое моделирование задач динамики излучающего газа. М.: Наука, 1985.
3. *Сушкевич Т.А.* Математические модели переноса излучения. М.: Бином, 2006.
4. *Каханер Д., Моулер К., Неш С.* Численные методы и программное обеспечение. М.: Мир, 2001.
5. *Моисеев Н.Я.* Модифицированный метод расщепления по физическим процессам для решения уравнений радиационной газовой динамики // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 2. С. 303–315.
6. *Гольдин В.Я.* Квазидиффузионный метод решения кинетического уравнения // Ж. вычисл. матем. и матем. физ. 1964. Т. 4. № 6. С. 1070–1087.

EDN: YKGSRE
ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2022, том 62, № 5, с. 872–888

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.633.6

О МОДЕЛИРОВАНИИ ЦИЛИНДРИЧЕСКОЙ МЕДЛЕННОЙ НЕОБЫКНОВЕННОЙ ВОЛНЫ В ХОЛОДНОЙ МАГНИТОАКТИВНОЙ ПЛАЗМЕ[1)]

© 2022 г. А. А. Фролов[1,*], Е. В. Чижонков[2,]**

[1] 119991 Москва, Ленинский пр-т, 53, Физический институт им. П.Н. Лебедева РАН, Россия

[2] 119899 Москва, Ленинские горы, МГУ им. М.В. Ломоносова, Россия

**e-mail: frolovaa@lebedev.ru*

***e-mail: chizhonk@mech.math.msu.su*

Поступила в редакцию 12.09.2021 г.
Переработанный вариант 10.11.2021 г.
Принята к публикации 14.01.2022 г.

Исследовано влияние внешнего магнитного поля на нерелятивистские цилиндрические плазменные колебания. Для инициализации медленной необыкновенной волны в магнитоактивной плазме предложен способ построения недостающих начальных условий на основе решения линейной задачи в виде рядов Фурье–Бесселя. С целью численного моделирования нелинейной волны построена схема метода конечных разностей второго порядка точности типа Мак-Кормака. Показано, что при учете внешнего магнитного поля ленгмюровские колебания трансформируются в медленную необыкновенную волну. При этом скорость волны увеличивается с ростом внешнего постоянного поля, что способствует выносу энергии из первоначальной области локализации колебаний. По этой причине известный эффект внеосевого опрокидывания наблюдается с запаздыванием по времени, а начиная с некоторого критического значения внешнего поля, перестает реализовываться совсем, т.е. формируется глобальное по времени гладкое решение. Библ. 22. Фиг. 12.

Ключевые слова: магнитоактивная плазма, плазменные колебания, медленная необыкновенная волна, ряды Фурье–Бесселя, численное моделирование, метод конечных разностей, эффект опрокидывания.

DOI: 10.31857/S0044466922050040

ВВЕДЕНИЕ

Полностью ионизованная плазма является сильно нелинейной средой, в которой даже относительно небольшие начальные коллективные смещения частиц могут приводить к колебаниям и волнам большой амплитуды [1]. Численному моделированию колебаний в холодной плазме, а также кильватерных волн, возбуждаемых коротким мощным лазерным импульсом, посвящена монография [2].

Сложность и разнообразие постановок задач резко возрастают, если рассматривать динамику магнитоактивной плазмы, т.е. плазмы, помещенной во внешнее магнитное поле. Даже в случае малых возмущений магнитоактивной плазмы, т.е. при рассмотрении только линейных волн, в классических учебниках и монографиях (см., например, [3], [4]) вводится их специальная классификация. В частности, в высокочастотной области спектра существуют три волны: обыкновенная, а также быстрая и медленная необыкновенная. Следут отметить, что внешнее магнитное поле порождает (индуцирует) магнитное поле в самой плазме, что требует соответствующего задания начальных условий для рассматриваемых дифференциальных уравнений. Эти условия в общем случае не могут быть произвольными. Например, в работе [5] показано, что при инициализации медленных необыкновенных волн (МНВ) тривиальные условия допустимы только в случае слабых внешних полей; в противном случае динамика волны может быть сильно искажена “неестественными” начальными условиями. По существу, такая ситуация приводит к постановке

[1)]Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках реализации программы Московского центра фундаментальной и прикладной математики (соглашение № 075-15-2019-1621).

новой вспомогательной задачи о разумной постановке недостающих (по сравнению с инициализацией колебаний без внешнего магнитного поля) начальных условий. В настоящей работе такая задача сначала формулируется, а затем решается в терминах рядов Фурье–Бесселя [6].

Следует также обратить внимание, что даже в случае зависимости решения от одной пространственной переменной, т.е. при наличии аксиальной симметрии, дифференциальные уравнения модели холодной плазмы достаточно сложны и громоздки. По этой причине возможности аналитического и асимптотического подходов здесь сильно ограничены, а в качестве основного инструмента исследования выступает численное моделирование. В свою очередь, это требует применения надежных и эффективных алгоритмов расчета. В настоящей работе с этой целью построена модификация классической схемы Мак-Кормака второго порядка точности, анализ и тестирование которой в плоском случае приведены в [7]. Напомним, что для колебаний и волн в случае аксиальной симметрии в плазме без магнитного поля характерным является эффект внеосевого опрокидывания [8], наблюдаемый по истечении нескольких периодов, что накладывает дополнительные требования на качество численного алгоритма [2].

Наконец, отметим актуальность тематики, связанной с МНВ. Их изучению в последнее время уделяется повышенное внимание (см., например, [9], [10]). Это напрямую связано с интенсивными экспериментальными и теоретическими исследованиями взаимодействия мощных лазерных импульсов с плазмой, которое имеет множество практических приложений от ядерной физики до медицины.

Статья имеет следующую структуру. Сначала приведена постановка задачи в эйлеровых переменных, включая начальные и асимптотические граничные условия, необходимые для описания временной эволюции плазменных колебаний и волн, порождаемых мощным коротким лазерным импульсом. Этих условий недостаточно для однозначного определения решения, поэтому в следующем разделе рассматривается вспомогательная задача. На основании ее решения с помощью рядов Фурье–Бесселя можно описать линейную МНВ, дискретный аналог которых хорошо приспособлен для практических вычислений недостающих начальных условий. В следующем разделе приводится описание численного метода решения нелинейной системы уравнений в частных производных типа Мак-Кормака, имеющего на гладких решениях второй порядок точности относительно параметров дискретизации. Наконец, самый большой раздел посвящен численным экспериментам. В нем приводятся примеры построенных начальных данных, которые зависят от напряженности внешнего магнитного поля и инициализируют МНВ. Далее в терминах функции, описывающей энергию линейной волны, иллюстрируется ее динамика во времени во внешнем магнитном поле различной напряженности. Особое внимание уделяется расчетам нелинейной нерелятивистской МНВ и ее сравнению с линейным случаем. Представляет отдельный интерес влияние внешнего магнитного поля на эффект внеосевого опрокидывания волны. Показано, что при значении магнитного поля, превышающего критическое, решение носит глобально гладкий по времени и пространству характер. В заключение систематизируются результаты проведенных исследований.

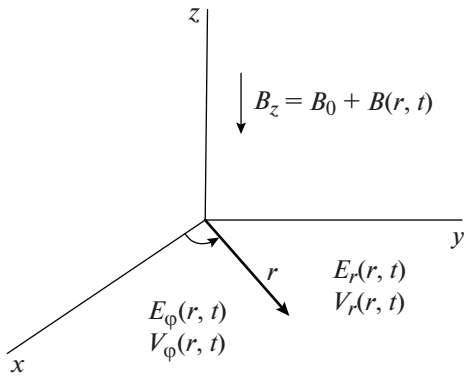
1. ОСНОВНЫЕ УРАВНЕНИЯ

Будем считать плазму холодной нерелятивистской электронной жидкостью, пренебрегая рекомбинационными эффектами и движением ионов. Тогда в векторной форме система описывающих ее гидродинамических уравнений совместно с уравнениями Максвелла будет иметь вид

$$\begin{gathered}\frac{\partial n}{\partial t}+\operatorname{div}(n\mathbf{V})=0,\quad \frac{\partial \mathbf{V}}{\partial t}+(\mathbf{V}\cdot\nabla)\mathbf{V}=\frac{e}{m}\Big(\mathbf{E}+\frac{1}{c}[\mathbf{V}\cdot\mathbf{B}]\Big),\\ \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t}=-\frac{4\pi}{c}en\mathbf{V}+\operatorname{rot}\mathbf{B},\quad \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}=-\operatorname{rot}\mathbf{E},\quad \operatorname{div}\mathbf{B}=0,\end{gathered}\tag{1}$$

где e, m – заряд и масса электрона (здесь заряд электрона имеет отрицательный знак: $e<0$), c – скорость света; n, $\mathbf{V}$ – плотность и скорость электронов; $\mathbf{E}$, $\mathbf{B}$ – векторы электрического и магнитного полей.

Система уравнений (1) является одной из простейших моделей плазмы, которую часто называют уравнениями гидродинамики “холодной” плазмы; она хорошо известна и достаточно подробно описана в учебниках и монографиях (см., например, [3], [4], [11], [12]).



Фиг. 1. Искомые функции и их зависимость от координат.

Получим из базовых уравнений рассматриваемой модели плазмы (1) систему, решения которой обладают аксиальной (цилиндрической) симметрией. Будем обозначать независимые переменные в цилиндрической системе координат обычным образом – (r, φ, z) и примем допущение, что плазма находится во внешнем магнитном поле $\mathbf{B}_0$, которое направлено вдоль оси z и не зависит от времени и пространства.

С целью проведения численного моделирования одномерных цилиндрических плазменных колебаний и волн во внешнем поле $\mathbf{B}_0 \equiv \text{const}$ базовые уравнения (1) можно существенно упростить. Будем считать, что

– решение определяется только r- и φ-компонентами вектор-функций $\mathbf{V}$, $\mathbf{E}$, а также z-компонентой вектора магнитного поля $\mathbf{B}$; при этом $B_z = B_0 + B_p$, где B_p – индуцированное магнитное поле в плазме;

– зависимость во всех указанных функциях и в электронной плотности n от переменных φ и z отсутствует, т.е. соответствующие частные производные равны нулю: $\partial/\partial\varphi = \partial/\partial z = 0$. Напомним, что $B_0 \equiv \text{const}$.

Качественная структура искомого решения представлена на фиг. 1, где $B(r,t) \equiv B_p(r,t)$. В рассматриваемом случае из системы (1) следуют нетривиальные уравнения:

$$\begin{gathered}\frac{\partial n}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}(rnV_r) = 0, \quad \frac{\partial V_r}{\partial t} + v_r\frac{\partial V_r}{\partial r} - \frac{V_\varphi^2}{r} = \frac{e}{m}\left[E_r + \frac{1}{c}V_\varphi(B_0 + B_p)\right], \\ \frac{\partial V_\varphi}{\partial t} + V_r\frac{\partial V_\varphi}{\partial r} + \frac{V_rV_\varphi}{r} = \frac{e}{m}\left[E_\varphi - \frac{1}{c}V_r(B_0 + B_p)\right], \\ \frac{1}{c}\frac{\partial B_p}{\partial t} + \frac{1}{r}\frac{\partial(rE_\varphi)}{\partial r} = 0, \quad \frac{\partial E_r}{\partial t} = -4\pi enV_r, \quad \frac{1}{c}\frac{\partial E_\varphi}{\partial t} = -\frac{4\pi e}{c}nV_\varphi - \frac{\partial B_p}{\partial r}.\end{gathered} \tag{2}$$

Введем безразмерные величины

$$\rho = k_pr, \quad \theta = \omega_pt, \quad \hat{N} = n/n_0, \quad \hat{V}_r = \frac{V_r}{c}, \quad \hat{V}_\varphi = \frac{V_\varphi}{c},$$
$$\hat{E}_r = -\frac{eE_r}{mc\omega_p}, \quad \hat{E}_\varphi = -\frac{eE_\varphi}{mc\omega_p}, \quad \hat{B} = -\frac{eB}{mc\omega_p}, \quad \hat{B} = \hat{B}_0 + \hat{B}_p,$$

где $\omega_p = \left(4\pi e^2n_0/m\right)^{1/2}$ – плазменная частота, n_0 – значение невозмущенной электронной плотности, $k_p = \omega_p/c$.

В новых переменных система (2) примет вид

$$\begin{gathered}\frac{\partial \hat{N}}{\partial \theta}+\frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \hat{N}\hat{V}_r\right)=0,\\ \frac{\partial \hat{V}_r}{\partial \theta}+\hat{V}_r\frac{\partial \hat{V}_r}{\partial \rho}-\frac{\hat{V}_\varphi^2}{\rho}=-\hat{E}_r-\hat{V}_\varphi\left(\hat{B}_0+\hat{B}_p\right),\\ \frac{\partial \hat{V}_\varphi}{\partial \theta}+\hat{V}_r\frac{\partial \hat{V}_\varphi}{\partial \rho}+\frac{\hat{V}_r\hat{V}_\varphi}{\rho}=-\hat{E}_\varphi+\hat{V}_r\left(\hat{B}_0+\hat{B}_p\right),\\ \frac{\partial \hat{B}_p}{\partial \theta}+\frac{1}{\rho}\frac{\partial\left(\rho \hat{E}_\varphi\right)}{\partial \rho}=0,\quad \frac{\partial \hat{E}_r}{\partial \theta}=\hat{N}\hat{V}_r,\quad \frac{\partial \hat{E}_\varphi}{\partial \theta}=\hat{N}\hat{V}_\varphi-\frac{\partial \hat{B}_p}{\partial \rho}.\end{gathered}\tag{3}$$

Из полученных уравнений

$$\frac{\partial \hat{N}}{\partial \theta}+\frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \hat{N}\hat{V}_r\right)=0,\quad \frac{\partial \hat{E}_r}{\partial \theta}=\hat{N}\hat{V}_r$$

следует

$$\frac{\partial}{\partial \theta}\left[\hat{N}+\frac{1}{\rho}\frac{\partial\left(\rho \hat{E}_r\right)}{\partial \rho}\right]=0.$$

Это соотношение справедливо как при отсутствии плазменных колебаний ($\hat{N}\equiv 1$, $\hat{E}_r\equiv 0$), так и при их наличии. Поэтому в традиционном предположении однородной плотности фонового заряда неподвижных ионов отсюда следует более простое выражение для электронной плотности $\hat{N}(\rho,\theta)$:

$$\hat{N}(\rho,\theta)=1-\frac{1}{\rho}\frac{\partial\left(\rho \hat{E}_r\right)}{\partial \rho}.\tag{4}$$

Воспользовавшись им, приходим к уравнениям, описывающим цилиндрические одномерные нерелятивистские плазменные колебания и волны:

$$\begin{gathered}\frac{\partial V_r}{\partial \theta}+V_r\frac{\partial V_r}{\partial \rho}-\frac{V_\varphi^2}{\rho}=-E_r-V_\varphi\left(B_0+B\right),\quad \frac{\partial V_\varphi}{\partial \theta}+V_r\frac{\partial V_\varphi}{\partial \rho}+\frac{V_rV_\varphi}{\rho}=-E_\varphi+V_r\left(B_0+B\right),\\ \frac{\partial B}{\partial \theta}+\frac{1}{\rho}\frac{\partial\left(\rho E_\varphi\right)}{\rho}=0,\quad \frac{\partial E_r}{\partial \theta}+\frac{V_r}{\rho}\frac{\partial\left(\rho E_r\right)}{\partial \rho}=V_r,\quad \frac{\partial E_\varphi}{\partial \theta}+\frac{V_\varphi}{\rho}\frac{\partial\left(\rho E_r\right)}{\partial \rho}+\frac{\partial B}{\partial \rho}=V_\varphi.\end{gathered}\tag{5}$$

Здесь для удобства у всех безразмерных искомых функций убран символ “крышка”, а также – нижний индекс “p” у индуцированного магнитного поля в плазме. Система (5) является основной для численного моделирования в настоящей работе. Отметим, что в дальнейшем будет использоваться полезная формула электронной плотности (4), в которой также убран символ “крышка”.

Приведем следствие уравнений (4), (5) – закон сохранения энергии (см. [13]):

$$\frac{\partial}{\partial \theta}\left\{\frac{E_r^2+E_\varphi^2+B_z^2}{2}+N\frac{V_r^2+V_\varphi^2}{2}\right\}+\frac{1}{\rho}\frac{\partial}{\partial \rho}\left\{\rho\left[E_\varphi B_z+NV_r\frac{V_r^2+V_\varphi^2}{2}\right]\right\}=0.\tag{6}$$

Здесь в качестве B_z можно понимать не только индуцированное поле $B(\rho,\theta)$, но и полное поле $B_0+B(\rho,\theta)$, так как способ получения уравнения (6) допускает оба варианта без изменения формы записи.

2. ЛИНЕЙНАЯ МЕДЛЕННАЯ НЕОБЫКНОВЕННАЯ ВОЛНА

2.1. Начальные и граничные условия

Рассмотрим возбуждение колебаний в постоянном внешнем магнитном поле $B_0 \equiv \text{const}$ в окрестности прямой $\rho = 0$. Положим, что скорость электронов V_r в начальный момент времени ($\theta = 0$) равна нулю

$$V_r(\rho, \theta = 0) = 0, \tag{7}$$

а колебания инициализируются электрическим полем E_r следующего вида:

$$E_r(\rho, \theta = 0) = \left(\frac{a_*}{\rho_*}\right)^2 \rho \exp\left\{-2\frac{\rho^2}{\rho_*^2}\right\}, \tag{8}$$

где параметры ρ_* и a_* характеризуют масштаб области локализации и максимальную величину $E_{\max} = a_*^2/\left(\rho_* 2\sqrt{\mathrm{e}}\right) \approx 0.3 a_*^2/\rho_*$ электрического поля (8) соответственно. Здесь и далее $\mathrm{e} = 2.71\ldots$ – основание натурального логарифма. Вид функции (8) выбран из соображений, что подобные колебания могут возбуждаться в разреженной плазме ($\omega_l \gg \omega_p$) лазерным импульсом с частотой ω_l при его фокусировке сферической линзой, когда фокальное пятно имеет форму круга.

Если электрическое поле лазерного излучения имеет аксиальную симметрию и гауссово распределение по пространственным координатам и времени

$$E_L(\rho, z, t) = E_{0L} \exp\left\{-\frac{\rho^2}{\rho_*^2} - \frac{\omega_p^2}{\tau_*^2}\left(t - \frac{z}{c}\right)^2\right\} \cos\left[\omega_l\left(t - \frac{z}{c}\right)\right], \tag{9}$$

где $\tau_* = \omega_p \tau_p$, $\rho_* = k_p R_p$ – безразмерные значения длительности τ_p и радиуса фокального пятна R_p лазерного импульса, то в некоторой точке z, удаленной от заднего фронта импульса на расстояние, превышающее длину плазменной волны ($k_p |z| \gg 1$), справедливо следующее соотношение, связывающее величину a_* с параметрами лазерного импульса [8]

$$a_*^2 = a_0^2 \tau_* \sqrt{\pi/2} \exp\left(-\tau_*^2/8\right), \tag{10}$$

где $a_0 = eE_{0L}/(m\omega_l c)$ – нормированная амплитуда лазерного поля. В условиях оптимального возбуждения кильватерной волны ($\tau_* = 2$), когда ее амплитуда максимальна, соотношение (10) принимает вид $a_*^2 = a_0^2\sqrt{2\pi/\mathrm{e}} \approx 1.52 a_0^2$.

Заметим, что на больших расстояниях от прямой $\rho = 0$, в силу начального условия (8), плазменные колебания не возбуждаются. Поэтому будем считать, что выполнены следующие условия:

$$\begin{gathered} V_r(\rho \to \infty, \theta) = V_\varphi(\rho \to \infty, \theta) = 0, \\ E_r(\rho \to \infty, \theta) = E_\varphi(\rho \to \infty, \theta) = B(\rho \to \infty, \theta) = 0. \end{gathered} \tag{11}$$

Кроме того, аксиальная симметрия решения предусматривает выполнение следующих условий на оси $\rho = 0$:

$$\begin{gathered} V_r(\rho = 0, \theta) = V_\varphi(\rho = 0, \theta) = 0, \\ E_r(\rho = 0, \theta) = E_\varphi(\rho = 0, \theta) = 0. \end{gathered} \tag{12}$$

Таким образом, учитывая специфику цилиндрической системы координат ($\rho \geq 0$), будем изучать в первом квадранте $\{(\rho, \theta) : 0 \leq \rho < \infty,\ \theta > 0\}$ решения системы (5), определяемые начальными и граничными условиями (7), (8), (11), (12) при наличии внешнего магнитного поля B_0, не зависящего ни от времени, ни от пространственной координаты.

В целях моделирования локализованного в пространстве решения системы (13) приме́ним следующий подход ограничения области. Зафиксируем область определения решения по переменной ρ с помощью параметра d, т.е. отрезок $[0, d]$. Предполагается, что в течение времени наблюдения решение не успеет достигнуть границы d и отразиться от нее. Из явного вида началь-

ной функции (8) следует, что если положить $d \approx 4.5\rho_*$, то вне отрезка $[0,d]$ функция $E_r(\rho,\theta=0)$ не будет по порядку превышать величину 10^{-18}. Поэтому указанного значения d достаточно для наблюдения за ленгмюровскими колебаниями, т.е. когда внешнее магнитное поле B_0 отсутствует. В рассматриваемом случае возбуждения волны необходим запас для ее распространения в пространстве, следовательно, параметр d надо брать бóльшим. Как правило, его примерного удвоения бывает достаточно, т.е. параметр d можно выбирать как $d = 10\rho_*$. Конечно, здесь речь идет об ориентировочном значении, так как при необходимости d можно и нужно увеличивать.

2.2. Ряды Фурье–Бесселя

Рассмотрим систему (5) в предположении малости возмущений относительно нулевого фона. После отбрасывания слагаемых второго порядка малости получим

$$\begin{gathered}\frac{\partial V_r^l}{\partial\theta} = -E_r^l - V_\varphi^l B_0, \quad \frac{\partial V_\varphi^l}{\partial\theta} = -E_\varphi^l + V_r^l B_0,\\ \frac{\partial B^l}{\partial\theta} + \frac{1}{\rho}\frac{\partial\left(\rho E_\varphi^l\right)}{\rho} = 0, \quad \frac{\partial E_r^l}{\partial\theta} = V_r^l, \quad \frac{\partial E_\varphi^l}{\partial\theta} + \frac{\partial B^l}{\partial\rho} = V_\varphi^l.\end{gathered} \tag{13}$$

Верхний индекс l здесь обозначает принадлежность к решению линеаризованной задачи.

Отметим, что для (13) имеет место закон сохранения энергии (см. [13]):

$$\frac{1}{2}\frac{\partial}{\partial\theta}\left\{\left(E_r^l\right)^2 + \left(E_\varphi^l\right)^2 + \left(B_z^l\right)^2 + \left(V_r^l\right)^2 + \left(V_\varphi^l\right)^2\right\} + \frac{1}{\rho}\frac{\partial}{\partial\rho}\left\{\rho E_\varphi^l B_z^l\right\} = 0, \tag{14}$$

где, как и ранее, в качестве B_z^l можно понимать не только индуцированное поле $B^l(\rho,\theta)$, но и полное поле $B_0 + B^l(\rho,\theta)$, так как способ получения уравнения (14) допускает оба варианта без изменения формы записи.

Далее будем учитывать условия (11) и (12), представляя искомое решение задачи в виде сходящихся рядов Фурье по функциям Бесселя на отрезке $[0,d]$, безусловно, предполагая его достаточную гладкость. С этой целью определим два набора базисных функций $\{Y_k(\rho)\}_{k=1}^\infty$ и $\{Z_k(\rho)\}_{k=1}^\infty$, что связано с различной четностью искомых функций относительно оси симметрии $\rho = 0$. Пусть μ_k, $k = 1, 2, \ldots$ – нули функции Бесселя $J_1(t)$, т.е. $J_1(\mu_k) = 0$. Положим $\kappa_k = \mu_k/d$, тогда на отрезке $[0,d]$ для функций

$$Z_k(\rho) = J_1(\kappa_k\rho)\frac{\sqrt{2}}{d\left|J_0(\mu_k)\right|}, \quad Y_k(\rho) = J_0(\kappa_k\rho)\frac{\sqrt{2}}{d\left|J_0(\mu_k)\right|}$$

справедливы условия ортогональности

$$\int_0^d \rho Z_k(\rho)Z_l(\rho)d\rho = \delta_k^l, \quad \int_0^d \rho\, Y_k(\rho)Y_l(\rho)d\rho = \delta_k^l,$$

где δ_k^l – символ Кронеккера.

Легко проверить полезные соотношения, связывающие функции из обеих наборов:

$$Y_k'(\rho) = -\kappa_k Z_k(\rho), \quad \frac{1}{\rho}\frac{d}{d\rho}(\rho Z_k(\rho)) = \kappa_k Y_k(\rho), \quad k = 1, 2, \ldots.$$

Используя полноту, ортогональность и полезные свойства базисных функций (см., например, [6]), имеем формальное представление решения системы (13) в виде рядов Фурье–Бесселя:

$$E_r^l(\rho,\theta) = \sum_{k=1}^\infty E_r(k)Z_k(\rho)\cos\left(\omega(k)\theta\right),$$

$$V_r^l(\rho,\theta) = \sum_{k=1}^\infty V_r(k)Z_k(\rho)\sin\left(\omega(k)\theta\right),$$

$$E_\varphi^l(\rho,\theta)=\sum_{k=1}^{\infty}E_\varphi(k)Z_k(\rho)\sin\big(\omega(k)\theta\big),\tag{15}$$

$$V_\varphi^l(\rho,\theta)=\sum_{k=1}^{\infty}V_\varphi(k)Z_k(\rho)\cos\big(\omega(k)\theta\big),$$

$$B^l(\rho,\theta)=\sum_{k=1}^{\infty}B(k)Y_k(\rho)\cos\big(\omega(k)\theta\big).$$

Подстановка (15) в (13) порождает формулу зависимости частоты от номера гармоники и напряженности внешнего магнитного поля

$$\omega(k)=\sqrt{1+\frac{\kappa_k^2+B_0^2}{2}\pm\sqrt{B_0^2+\left(\frac{\kappa_k^2-B_0^2}{2}\right)^2}}\tag{16}$$

и явные выражения для коэффициентов Фурье–Бесселя:

$$\begin{aligned}V_r(k)&=-E_r(k),\\ E_\varphi(k)&=-B_0E_r(k)\frac{\omega(k)}{1+\kappa_k^2-\omega^2(k)},\\ V_\varphi(k)&=B_0E_r(k)\frac{\kappa_k^2-\omega^2(k)}{1+\kappa_k^2-\omega^2(k)},\\ B(k)&=-B_0E_r(k)\frac{\kappa_k}{1+\kappa_k^2-\omega^2(k)},\end{aligned}\tag{17}$$

где коэффициенты $E_r(k)$, $k=1,2,\ldots$, определяются равенством

$$E_r^l(\rho,\theta=0)\equiv\left(\frac{a_*}{\rho_*}\right)^2\rho\exp\left\{-2\frac{\rho^2}{\rho_*^2}\right\}=\sum_{k=1}^{\infty}E_r(k)Z_k(\rho).\tag{18}$$

Обратим внимание, что в данном случае объект МНВ [3], [4] соответствует распространению возмущений в направлении, ортогональном вектору магнитного поля, и удовлетворяет только одному из двух соотношений (16), а именно – со знаком минус. В частности, при отсутствии внешнего магнитного поля из (16) следует формула для частоты ленгмюровских колебаний в холодной изотропной плазме $\omega(k)\equiv 1$ (в размерных переменных – $\omega(k)\equiv\omega_p$), т.е. уравнения (13) в этом случае порождают решение хорошо известной задачи. С другой стороны, при $k\to\infty$ справедливо $\omega(k)\to\omega_{up}$, где $\omega_{up}^2=1+B_0^2$, т.е. решение системы (13) может быть в этом случае приближенно описано так называемыми *верхнегибридными колебаниями* (см. [3, с. 112–113]).

Таким образом, построенное аналитическое решение дает возможность определить из (15) необходимые начальные условия для инициализации нелинейной МНВ, которая при достаточно малом внешнем поле B_0 будет близка к линейной. Отметим также, что замена бесконечной области по переменной ρ на конечную, позволяет вместо преобразования Ханкеля (в цилиндрических координатах – аналог интегрального преобразования Фурье) использовать в расчетах ряды Фурье–Бесселя, что может быть реализовано гораздо эффективнее с помощью библиотек стандартных программ численного анализа.

3. ЧИСЛЕННЫЙ АЛГОРИТМ ВТОРОГО ПОРЯДКА ТОЧНОСТИ

Пусть исходное уравнение имеет вид

$$\frac{\partial\mathbf{U}}{\partial\theta}+\frac{\partial\mathbf{F}(\mathbf{U})}{\partial\rho}=\mathbf{S}(\mathbf{U},\rho,\theta),\tag{19}$$

где $\mathbf{U}$, $\mathbf{F}$, $\mathbf{S}$ – вектор-функции, рассматриваемые в полуплоскости $\{(\rho,\theta):\theta\ge 0,\ \rho\in\mathbb{R}\}$, и в момент времени $\theta=0$ заданы начальные условия

$$\mathbf{U}(\rho,\theta=0)=\mathbf{U}^0(\rho),\quad \rho\in\mathbb{R}.\tag{20}$$

Будем считать, что нас интересует приближенное решение задачи Коши, определенной соотношениями (19), (20), про которое известно, что оно существует, единственно и обладает достаточной гладкостью.

Определим дискретизацию независимых переменных с помощью постоянных параметров τ и h так, что

$$\theta^n = n\tau, \quad n \geq 0, \quad \rho_i = ih, \quad i = 0, 1, 2, \ldots,$$

и будем обозначать зависимую переменную $\mathbf{U}(\rho, \theta)$ в узле сетки (ρ_i, θ^n) как $\mathbf{U}_i^n$. Стандартная схема Мак-Кормака [14] состоит из двух этапов вычислений, известных как предиктор и корректор, и может быть представлена в виде

$$\mathbf{U}_i^p = \mathbf{U}_i^n - \frac{\tau}{h}\left(\mathbf{F}_{i+1}^n - \mathbf{F}_i^n\right) + \tau \mathbf{S}_i^n, \tag{21}$$

$$\mathbf{U}_i^c = \mathbf{U}_i^n - \frac{\tau}{h}\left(\mathbf{F}_i^p - \mathbf{F}_{i-1}^p\right) + \tau \mathbf{S}_i^p, \tag{22}$$

где верхний индекс p обозначает шаг предиктор (c – корректор) или n – временной слой θ^n.

Окончательная формула, формирующая решение на следующем временном слое $(n+1)$, имеет вид:

$$\mathbf{U}_i^{n+1} = \frac{\mathbf{U}_i^p + \mathbf{U}_i^c}{2}. \tag{23}$$

Приведенная схема хорошо известна и давно используется в вычислительной практике. Ее свойства представлены в популярных монографиях, посвященных численному анализу и математическому моделированию (см., например, [15], [16]). Напомним, что на гладких решениях схема Мак-Кормака имеет порядок аппроксимации $O\left(\tau^2 + h^2\right)$ и условие устойчивости $\tau = O(h)$, поэтому ее традиционно относят к схемам второго порядка точности.

Важным отличием системы (5) от традиционной формы записи уравнений (19) являются недивергентные слагаемые следующего вида:

$$V_r \frac{\partial E_r}{\partial \rho}, \quad V_r \frac{\partial V_\varphi}{\partial \rho}, \quad V_\varphi \frac{\partial E_r}{\partial \rho}.$$

Например, одно из них присутствует в уравнении, описывающем динамику r – компоненты электрического поля

$$\frac{\partial E_r}{\partial \theta} + V_r \frac{\partial E_r}{\partial \rho} = V_r - \frac{V_r E_r}{\rho}.$$

В работе [7] был предложен подход для устранения трудности с недивергентной формой слагаемых при сохранении прежнего (второго) порядка точности разностной схемы Мак-Кормака. В дальнейшем он хорошо зарекомендовал себя при расчетах плоских МНВ [5], [17], поэтому на его деталях здесь останавливаться не будем.

Существенным отличием обсуждаемой схемы от плоского случая является необходимость расчетов на оси симметрии $\rho = 0$. В частности, речь идет об уравнении для индуцированного магнитного поля в системе (5):

$$\frac{\partial B}{\partial \theta} + \frac{1}{\rho}\frac{\partial\left(\rho E_\varphi\right)}{\rho} = 0. \tag{24}$$

Функция $B(\rho, \theta)$ является единственной функцией, которая не обращается в нуль на оси симметрии, поэтому для вычислений особенность в нуле требует модификации уравнения. Учитывая условие из (12), т.е. $E_\varphi(\rho = 0, \theta) = 0$, в пределе при $\rho \to 0$ из (24) имеем

$$\frac{\partial B}{\partial \theta} + 2\frac{\partial E_\varphi}{\partial \rho} = 0.$$

В свою очередь, соотношения предиктор и корректор для этого уравнения примут вид

$$(B)_0^p = (B)_0^n - 2\frac{\tau}{h}(E_\varphi)_1^n, \quad (B)_0^c = (B)_0^n - 2\frac{\tau}{h}(E_\varphi)_1^p. \tag{25}$$

Нижние индексы 0 и 1 обозначают дискретные значения функций на оси ($\rho = 0$) и в соседней точке ($\rho = h$).

Отметим также, что при $\rho = d$ вычисления не производятся совсем: просто функции E_r, V_r, E_φ, V_φ полагаются равными нулю, как и производная $\partial B/\partial\rho$, в силу достаточной удаленности границы d.

4. РЕЗУЛЬТАТЫ РАСЧЕТОВ

Процесс опрокидывания нерелятивистских ленгмюровских колебаний при учете аксиальной симметрии рассматривался в [8], [18], в свою очередь, трансформация плоских колебаний в МНВ – в работе [5]. Поэтому в целях сохранения преемственности с приведенными в них результатами расчетов зафиксируем значения параметров в начальном условии (8): $a_* = 0.365$, $\rho_* = 0.6$. Выберем параметр d, характеризующий искусственную границу, равным 7. При этом характерное значение параметра дискретизации по пространственной переменной, использовавшееся в расчетах, равно $h = 0.25\times10^{-3}$. Шаг интегрирования по времени τ из соображений устойчивости выбирался равным $h/2$, а в целях контроля точности регулярно проводились расчеты с сеточными параметрами в два раза меньшими, чем основные (рабочие).

4.1. Расчет начальных данных

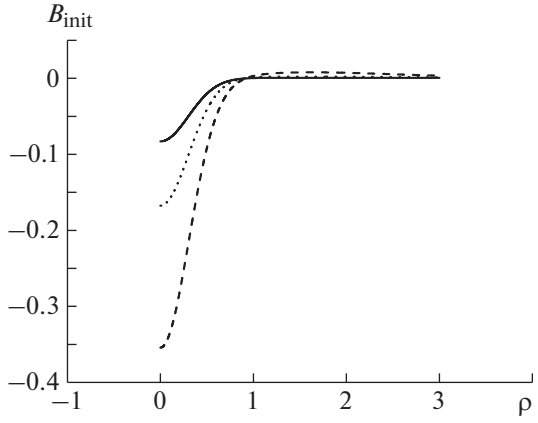
Выбор начальных данных для инициализации МНВ представляет определенные трудности. Дело в том, что в соответствии с формулами (15) начальные условия (7), (8) порождают ненулевые функции V_φ и B при $\theta = 0$. Причем определяются они в терминах коэффициентов рядов Фурье–Бесселя, зависящих от функции E_r в начальный момент времени. Как правило, чтобы избежать проблем с недостающими начальными условиями, волны в магнитоактивной плазме рассматривают либо в электростатическом приближении [1], [19], либо полагают недостающие начальные функции V_φ и B тождественно нулевыми. При использовании слабых магнитных полей $B_0 \ll 1$ последнее вполне допустимо, однако в общем случае пространственная форма волны может быть сильно искажена тривиальными начальными данными (см. детали в [5]).

Заметим, что вышесказанное порождает определенную новизну в постановке задачи, так как при отсутствии внешнего магнитного поля начальных условий (7), (8) вполне достаточно для однозначного определения ленгмюровских колебаний в произвольный момент времени.

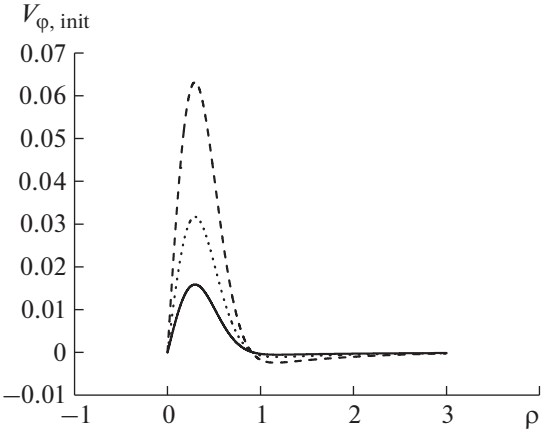
Приведем пример, иллюстрирующий согласование начальных данных в медленной необыкновенной волне. Для одного и того же пространственного распределения функции $E_r(\rho, \theta = 0)$, инициирующего волну в соответствии с (8) на фиг. 2 и фиг. 3 изображены начальные распределения функций $B(\rho, \theta = 0)$ и $V_\varphi(\rho, \theta = 0)$, определяемые решением (15) при различных внешних полях $B_0 \in \{0.25, 0.5, 1\}$.

Процедура их расчета базировалась на использовании пакетов QUADPACK [20] (библиотека SLATEC) и FUNPACK [21], реализующих вычисление определенных интегралов и работу с функциями Бесселя. На отрезке $[0, d]$ вычислялись коэффициенты Фурье–Бесселя $E_r(k)$ из разложения (18). Далее с помощью формул (15) при $\theta = 0$ и (16) вычислялись искомые коэффициенты функций $B^l(\rho, \theta = 0)$ и $V_\varphi^l(\rho, \theta = 0)$, которые использовались для определения на сетке недостающих начальных данных. Количество коэффициентов определялось параметром MCOEF, характерное значение которого полагалось 40. Для контрольных расчетов, как правило, значение параметра удваивалось.

Легко заметить, что диапазон изменения этих начальных функций вполне соизмерим с диапазоном $E_r(\rho, \theta = 0) \in [0, 0.068]$. Это дает основание для вывода, что как полагать обе начальные функции нулевыми, так и совсем пренебрегать индуцированным магнитным полем (случай электростатического приближения или верхнегибридных колебаний), при рассмотрении МНВ с начальным возмущением (7), (8) было бы не вполне оправдано.



Фиг. 2. Начальное индуцированное магнитное поле $B_{\text{init}}(\rho) = B(\rho, \theta = 0)$ для различных значений B_0: сплошная линия – $B_0 = 0.25$, пунктирная – $B_0 = 0.5$, штриховая – $B_0 = 1$.



Фиг. 3. Начальная φ-компонента скорости электронов $V_{\varphi,\text{init}}(\rho) = V_\varphi(\rho, \theta = 0)$ для различных значений B_0: сплошная линия – $B_0 = 0.25$, пунктирная – $B_0 = 0.5$, штриховая – $B_0 = 1$.

Отметим, что речь о согласовании начальных данных идет только в смысле сохранения физического смысла решения МНВ, так как, в силу гиперболичности исходной системы (5), с математической точки зрения все наборы достаточно гладких начальных условий являются равноправными.

4.2. Линейная МНВ

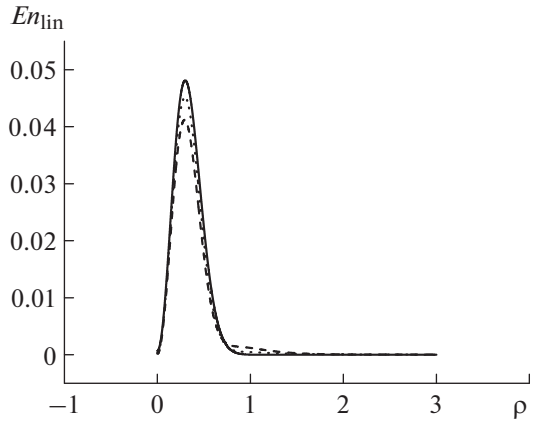
Для иллюстраций динамики линейной волны выберем функцию, характеризующую ее энергию с точностью до множителя 1/2,

$$En_{\text{lin}}(\rho, \theta) = E_r^2(\rho, \theta) + E_\varphi^2(\rho, \theta) + B^2(\rho, \theta) + V_r^2(\rho, \theta) + V_\varphi^2(\rho, \theta). \tag{26}$$

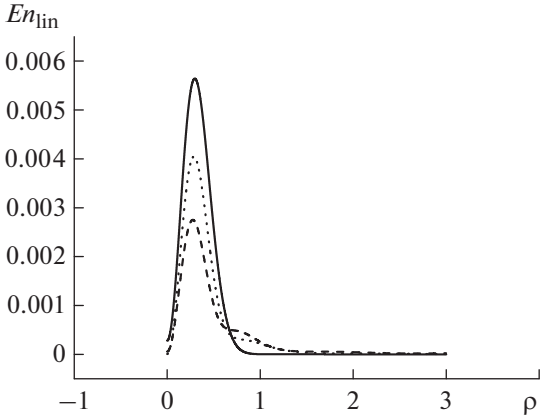
Из соотношения (14) следует, что интеграл по отрезку $[0, d]$ с весом ρ от функции $En_{\text{lin}}(\rho, \theta)$ равен константе, поэтому ее пространственные распределения в различные моменты времени могут иллюстрировать волновой характер переноса энергии в пространстве.

Сравним поведение волны для различных внешних полей $B_0 \in \{0.25, 0.5, 1\}$. Выберем с этой целью характерные моменты времени – $\theta \in \{0, 40\pi, 80\pi\}$ и рассмотрим для них функцию En_{lin}.

На фиг. 4 приведены пространственные распределения энергии для слабого внешнего магнитного поля $B_0 = 0.25$. Легко заметить, что волна очень похожа на неподвижную, т.е. переме-



Фиг. 4. Энергия линейной волны во внешнем поле $B_0 = 0.25$ для различных значений θ: сплошная линия – $\theta = 0$, пунктирная – $\theta = 40\pi$, штриховая – $\theta = 80\pi$.

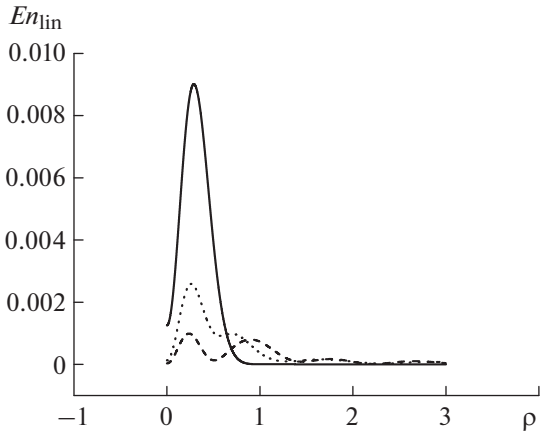


Фиг. 5. Энергия линейной волны во внешнем поле $B_0 = 0.5$ для различных значений θ: сплошная линия – $\theta = 0$, пунктирная – $\theta = 40\pi$, штриховая – $\theta = 80\pi$.

щение экстремальных значений энергии практически незаметно. Однако пусть медленно, но энергия перемещается в пространстве, о чем свидетельствует ее монотонное увеличение на периферии волны в зависимости от времени. Конечно, в силу закона сохранения, эта монотонность полностью компенсируется соответствующим убыванием максимальных значений. Поэтому можно сказать, что для слабых внешних полей происходит медленное “расплывание” энергии практически без изменения ее начальной формы.

Влияние значимого увеличения внешнего поля представлено на фиг. 5. При $B_0 = 0.5$ со временем происходит изменение формы волны, связанное с заметным перемещением энергии на периферии волны. Легко заметить, что скорость этого изменения при этом увеличивается. Другими словами, в умеренных внешних полях скорость переноса энергии линейной волны со временем возрастает, хотя максимальные значения энергии при этом монотонно убывают. Отметим, что фиг. 4 и 5 являются замечательной иллюстрацией “медленности” рассматриваемой необыкновенной волны.

Наконец, динамика энергии волны в сильном магнитном поле $B_0 = 1$ представлена на фиг. 6. Представленные пространственные распределения функции En_{lin} наглядно иллюстрируют волновой характер – скорость переноса энергии достаточно велика, а амплитуда уменьшается при выполнении закона сохранения, т.е. наблюдаются одновременно две тенденции: ускорение переноса энергии и немонотонное “расплывание” формы волны.



Фиг. 6. Энергия линейной волны во внешнем поле $B_0 = 1$ для различных значений θ: сплошная линия – $\theta = 0$, пунктирная – $\theta = 40\pi$, штриховая – $\theta = 80\pi$.

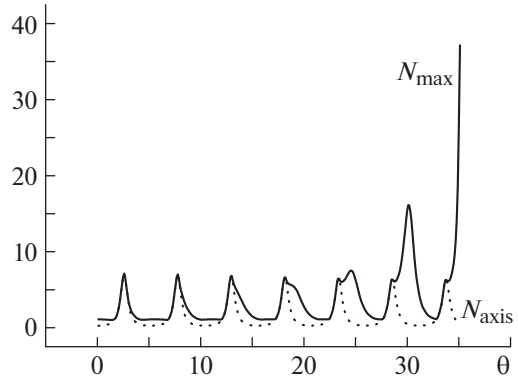
Уточним, что представленные выше графики получены одновременно как с помощью формул (15), так и при использовании численного алгоритма из разд. 3, упрощенного для линейных уравнений (13). В силу второго порядка точности приближенного метода и малости параметров дискретизации, заметить визуально различия в графиках функции En_{lin}, полученных различными способами, не представляется возможным.

4.3. Нелинейная МНВ

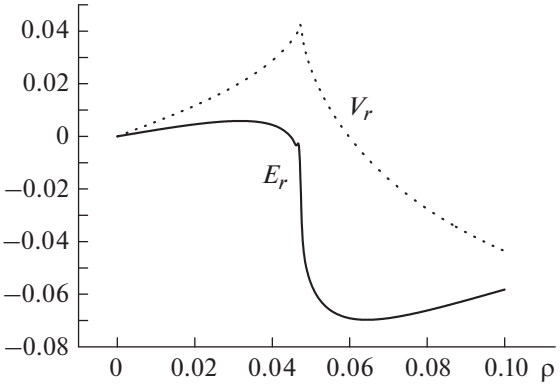
Наиболее интересным свойством нелинейной МНВ, в отличие от линейной, является ее опрокидывание по прошествии некоторого количества периодов. Аналогичный эффект имеет место и в случае отсутствия внешнего магнитного поля, т.е. в случае аксиально симметричных ленгмюровских колебаний [8], [18].

Краткое изложение эффекта опрокидывания в рамках рассматриваемой модели холодной плазмы (без внешнего магнитного поля) состоит в следующем. Начальное пространственное распределение электронной плотности N, как следствие формул (4) и (8), приводит к избытку положительного заряда в начале координат (при $\rho = 0$). По этой причине начинается движение электронов в направлении центра области, что через половину периода колебаний порождает распределение плотности с глобальным максимумом также при $\rho = 0$. Если бы нелинейные плазменные колебания сохраняли во времени свою пространственную форму, то описанные распределения плотности электронов регулярно меняли бы друг друга через каждую половину периода, порождая в центре области строго периодическую последовательность экстремумов с неизменными амплитудами. Однако с течением времени происходит постепенное формирование абсолютного максимума плотности, расположенного вне оси и сравнимого по величине с осевыми. После возникновения этого внеосевого максимума наблюдается резкое возрастание его по величине и достаточно быстро (часто через период – другой) на его месте возникает сингулярность электронной плотности. Численному моделированию этого процесса посвящена монография [2], кроме того, графики для аксиально-симметричного случая и используемых в настоящей статье параметров имеются в [18]. Уточним, что здесь речь идет о колебаниях, которые существуют на протяжении нескольких периодов во времени.

Приведем иллюстрации влияния умеренного по напряженности внешнего магнитного поля $B_0 = 0.5$ на эффект опрокидывания, в данном случае, уже МНВ. На фиг. 7 пунктиром изображено для электронной плотности изменение во времени в начале координат, а сплошной линией – динамика максимального по области значения. Влияние внешнего поля приводит к значительному (практически в два раза) уменьшению амплитуды колебаний на оси симметрии по сравнению с его отсутствием. Сначала колебания носят регулярный характер, т.е. глобальные по области максимумы и минимумы плотности сменяют друг друга через половину периода и располагаются в начале координат. После четвертого регулярного (центрального) максимума в момент времени $\theta \approx 19$ возникает новая структура – внеосевой максимум электронной плотности, по ве-



Фиг. 7. Динамика плотности плазмы N при $B_0 = 0.5$: сплошная линия – максимум по области, пунктирная линия – ось симметрии.



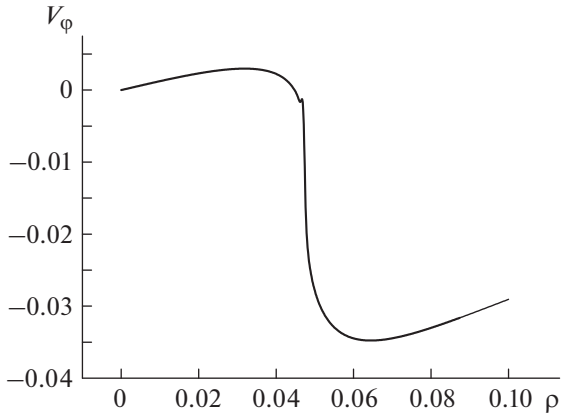
Фиг. 8. Пространственные распределения r-компонент скорости и электрического поля в момент опрокидывания для $B_0 = 0.5$: сплошная линия – E_r, пунктирная линия – V_r.

личине сравнимый с ближайшим регулярным. Далее в течение двух следующих периодов он монотонно возрастает, а затем в $\theta_{br} \approx 35.1$ на его месте возникает сингулярность электронной плотности.

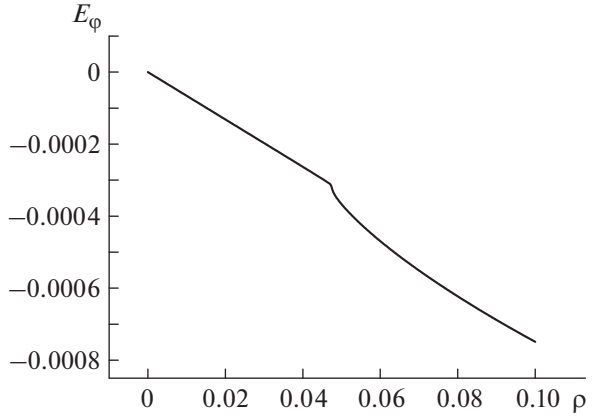
Распределения компонент скорости V_r и электрического поля E_r в момент опрокидывания изображены на фиг. 8. Отметим, что в окрестности сингулярности плотности функция скорости имеет скачок производной, а функция электрического поля принимает ступенчатый характер. Именно такие качественные характеристики и обеспечивают опрокидывание колебаний в момент $\theta_{br} \approx 35.1$. Важно отметить, что опрокидывание носит характер “градиентной катастрофы”, т.е. сами функции E_r и V_r при этом остаются ограниченными. Для внеосевого опрокидывания подобные распределения являются характерными и в большом количестве представлены в [2].

Распределения φ-компонент скорости и электрического поля, а также индуцированное магнитное поле в момент опрокидывания изображены на фиг. 9–11. В окрестности сингулярности плотности их поведение качественно сходно с r-компонентами: функция скорости V_φ принимает ступенчатый характер, а функция электрического поля E_φ имеет скачок производной. В свою очередь, функция B имеет скачок производной, а в окрестности начала координат $\rho = 0$ практически совпадает с константой. Отметим, что все представленные на фиг. 9–11 функции также ограничены, а их изображения ранее в научной литературе не встречались.

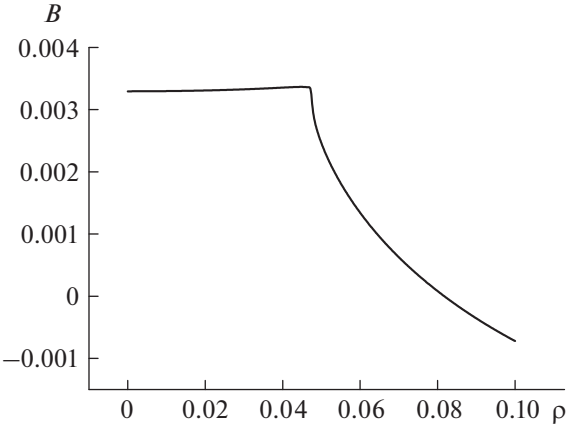
Следует обратить внимание, что для колебаний и волн, порождаемых узким импульсом (в данном случае $\rho_* = 0.6$), абсолютные значения φ-компоненты скорости и индуцированного магнитного поля по порядку меньше абсолютных значений скорости V_r и электрического



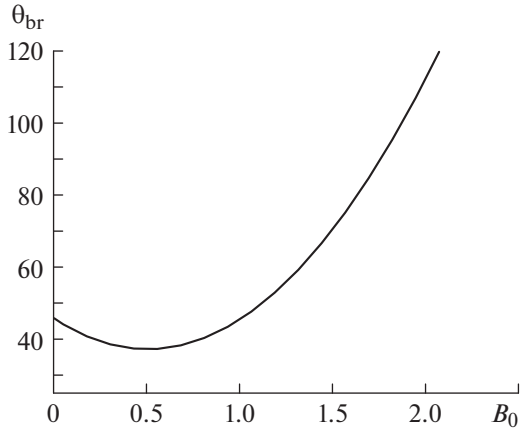
Фиг. 9. Пространственное распределение φ-компоненты скорости в момент опрокидывания для $B_0 = 0.5$.



Фиг. 10. Пространственное распределение φ-компоненты электрического поля в момент опрокидывания для $B_0 = 0.5$.



Фиг. 11. Пространственное распределение индуцированного магнитного поля в момент опрокидывания для $B_0 = 0.5$.



Фиг. 12. Зависимость времени опрокидывания колебаний θ_{br} от внешнего поля B_0.

поля E_r. В свою очередь, абсолютные значения электрического поля E_φ еще примерно на порядок меньше, чем V_φ. Тем не менее такие малые возмущения, сформировавшиеся при $B_0 = 0.5$, привели к опрокидыванию волны примерно на 10 безразмерных единиц времени (около 20%) быстрее, чем при $B_0 = 0$, т.е. при тождественно нулевых V_φ, E_φ, B.

Ожидаемое влияние внешнего магнитного поля на ленгмюровские колебания плазмы сводится к двум процессам: возбуждение бегущей волны при любом B_0 и пресечение ее опрокидывания при достаточно большом B_0. Аналогичное влияние оказывало увеличение температуры электронов на релятивистские колебания, рассмотренное в [22]. Причиной является схожесть физических факторов: увеличение внешнего магнитного поля, как и температуры, приводит к увеличению групповой скорости плазменных волн, что способствует выносу энергии из первоначальной области локализации колебаний. В случае нагревания плазмы в слабо-нелинейном приближении удалось в явной форме получить условие, при выполнении которого может наблюдаться эффект опрокидывания. Это было связано с монотонным увеличением времени опрокидывания в зависимости от температуры электронов. В рассматриваемом случае магнитоактивной плазмы подобного условия вывести не удалось, и это имеет вескую причину. Дело в том, что внешнее поле влияет на время опрокидывания МНВ немонотонным образом: в слабых полях процесс опрокидывания ускоряется, а затем – при увеличении поля – процесс замедляется, вплоть до полного прекращения, которое достаточно сложно отследить по причине необходимости моделирования волны на больших временах.

Асимптотическое запаздывание по времени эффекта опрокидывания легко заметить при увеличении внешнего поля. Качественно процесс опрокидывания, если он имеет место, практически не изменяется. Сначала наблюдаются регулярные максимумы плотности на оси симметрии, которые сопровождаются небольшим затуханием. Затем формируется практически неподвижный внеосевой максимум, который располагается вблизи оси симметрии, монотонно возрастает и приводит к сингулярности. Увеличение внешнего магнитного поля B_0 “растягивает” этот процесс во времени. Качественные отличия самих пространственных распределений в момент опрокидывания от представленных на фиг. 8–11 не обнаружены. Как и выше, опрокидывание носит характер “градиентной катастрофы”, т.е. все функции, кроме электронной плотности, при этом остаются ограниченными.

На фиг. 12 изображена зависимость времени опрокидывания от величины B_0 при взятых параметрах a_* и ρ_*. Спадание времени опрокидывания при малых значениях напряженности магнитного поля связано с ненулевым значением φ-компоненты скорости электронов, которая увеличивается с ростом магнитного поля. Последующее нарастание времени опрокидывания с ростом напряженности магнитного поля связано с увеличением групповой скорости медленной необыкновенной волны, которая выносит энергию из области первоначальной локализации колебаний. Подобный эффект уже наблюдался в процессе опрокидывания плоской релятивистской МНВ [17], однако формальное аналитическое обоснование его пока получить не удалось. Дополнительное увеличение внешнего поля приводит к резкому увеличению времени опроки-

дывания. Например, при $B_0 = 2.5$ опрокидывание наблюдается при $\theta_{br} \approx 193.3$, а при $B_0 = 3$ имеем $\theta_{br} \approx 375.9$. Отображение этих данных на фиг. 12 сильно бы ухудшило наглядность представленного графика.

Следует отметить, что расчеты эффекта опрокидывания подвергались строгому контролю, так как в этом процессе из гладкого решения формировалась так называемая градиентная катастрофа. При уменьшении/увеличении параметров дискретизации, размеров области и количества слагаемых в рядах в два раза (контрольные расчеты) пространственно-временные координаты опрокидывания изменялись на доли процентов.

ЗАКЛЮЧЕНИЕ

В работе исследовано влияние внешнего магнитного поля на цилиндрические нерелятивистские нелинейные плазменные колебания и волны, порождаемые мощным лазерным импульсом. В результате согласованного взаимодействия электромагнитных полей и частиц в магнитоактивной плазме формируется медленная необыкновенная волна. Для ее инициализации использован метод построения недостающих начальных условий на основе решения линейной задачи в рядах Фурье–Бесселя, которые несложно адаптировать к их дискретному аналогу по пространству. Реализация дискретных преобразований осуществляется с помощью стандартных пакетов программ численного анализа.

С целью численного моделирования нерелятивистской волны построена схема метода конечных разностей второго порядка точности типа Мак-Кормака на основе эйлеровых переменных. В силу новизны постановки задачи результаты расчетов подвергались тщательному контролю. Если внешнее магнитное поле отсутствует, то в соответствии с начальными условиями колебания сосредоточены в окрестности начала координат и продолжаются там до их опрокидывания. Эффект опрокидывания характеризуется формированием внеосевого максимума плотности по истечении нескольких периодов колебаний, затем – быстрым ростом этого максимума, и, наконец, сингулярностью плотности, что связано с пересечением траекторий соседних частиц. В случае учета внешнего магнитного поля ленгмюровские колебания трансформируются в медленную необыкновенную волну. При этом групповая скорость волны увеличивается с ростом внешнего постоянного поля, что способствует выносу энергии из первоначальной области локализации колебаний. По этой причине, начиная с некоторого критического значения внешнего поля, эффект опрокидывания уже перестает наблюдаться, т.е. формируется глобальное по времени гладкое решение. Представленная динамика электронной плотности и пространственные распределения энергиии хорошо иллюстрируют опрокидывание медленной необыкновенной волны.

Полученные результаты можно использовать для моделирования более сложных по структуре волн в магнитоактивной плазме, например, при релятивистских скоростях электронов или при дополнительном учете температуры электронов. Учитывая малость угловой компоненты электрического поля в момент опрокидывания, представляют определенный интерес моделирование электростатических верхнегибридных колебаний [1] и сравнение их опрокидывания с представленным в работе. Кроме того, возможно обобщение полученных данных на случай намагниченной квантовой плазмы, а также они могут быть полезны при обсуждении различных физических эффектов, связанных с плазменными колебаниями и волнами.

СПИСОК ЛИТЕРАТУРЫ

1. *Davidson R.C.* Methods in nonlinear plasma theory. New York: Academic Press, 1972. P. 33–53.
2. *Чижонков Е.В.* Математические аспекты моделирования колебаний и кильватерных волн в плазме. М.: Физматлит, 2018. С. 12–240.
3. *Александров А.Ф., Богданкевич Л.С., Рухадзе А.А.* Основы электродинамики плазмы. М.: Высшая школа, 1988. С. 102–113.
4. *Гинзбург В.Л., Рухадзе А.А.* Волны в магнитоактивной плазме. М.: Наука, 1975. С. 112–124.
5. *Фролов А.А., Чижонков Е.В.* О численном моделировании медленной необыкновенной волны в магнитоактивной плазме // Вычисл. методы и программирование. 2020. Т. 21. С. 420.
6. *Джексон Д.* Ряды Фурье и ортогональные полиномы. М.: ГИТТЛ, 1948. С. 124–129.
7. *Чижонков Е.В.* О схемах второго порядка точности для моделирования плазменных колебаний // Вычисл. методы и программирование. 2020. Т. 21. С. 115.

8. *Горбунов Л.М., Фролов А.А., Чижонков Е.В., Андреев Н.Е.* Опрокидывание нелинейных цилиндрических колебаний плазмы // Физ. плазмы. 2010. Т. 36. № 4. С. 375.
9. *Borhanian J.* Extraordinary electromagnetic localized structures in plasmas: Modulational instability, envelope solitons, and rogue waves // Physics Letters A. 2015. V. 379. № 6. P. 595.
10. *Moradi A.* Energy behaviour of extraordinary waves in magnetized quantum plasmas // Phys. of Plasmas. 2018. V. 25. P. 052123.
11. *Силин В.П.* Введение в кинетическую теорию газов. М.: Наука, 1971. С. 119.
12. *Силин В.П., Рухадзе А.А.* Электромагнитные свойства плазмы и плазмоподобных сред. М.: Книжный дом "ЛИБРОКОМ", 2012. С. 104–110.
13. *Фролов А.А., Чижонков Е.В.* О применении закона сохранения энергии в модели холодной плазмы // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. № 3. С. 503.
14. *MacCormack R.W.* The effect of viscosity in hypervelocity impact cratering // J. Spacecr. Rockets. 2003. V. 40. № 5. P. 757.
15. *Шокин Ю.И., Яненко Н.Н.* Метод дифференциального приближения. Применение к газовой динамике. Новосибирск: Наука, 1985. С. 251–252.
16. *Андерсон Д., Таннехилл Дж., Плетчер Р.* Вычислительная гидромеханика и теплообмен. Т. 1. М.: Мир, 1990. С. 179–180.
17. *Фролов А.А., Чижонков Е.В.* Об опрокидывании медленной необыкновенной волны в холодной магнитоактивной плазме // Матем. моделирование. 2021. Т. 33. № 6. С. 3.
18. *Фролов А.А., Чижонков Е.В.* Влияние электрон-ионных соударений на опрокидывание цилиндрических плазменных колебаний // Матем. моделирование. 2018. Т. 30. № 10. С. 86.
19. *Maity C.* Lagrangian Fluid Technique to Study Nonlinear Plasma Dynamics. PHD Thesis. Kolkata, India: Saha Institute of Nuclear Physics, 2013.
20. *de Doncker E.* An adaptive extrapolation algorithm for automatic integration // SIGNUM Newsletter. 1978. № 13. P. 12.
21. *Cody W.J.* The FUNPACK package of special function subroutines // ACM Trans. Math. Soft. 1975. № 1. P. 13.
22. *Chizhonkov E.V., Frolov A.A.* Influence of electron temperature on breaking of plasma oscillations // Russ. J. Numer. Anal. Math. Modelling. 2019. V. 34. № 2. P. 71.