
СОДЕРЖАНИЕ

Том 62, номер 4, 2022 год

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

h-, p- и hp-варианты метода коллокации и наименьших квадратов для решения краевых задач для бигармонического уравнения в нерегулярных областях и их приложения

В. А. Беляев, Л. С. Брындин, С. К. Голушко, Б. В. Семисалов, В. П. Шاپеев 531

Общая структура редуцированных базисов для задач агрегационной кинетики различной размерности

С. А. Матвеев, А. П. Смирнов, И. В. Тимохин, Е. Е. Тыртышников 553

Обоснование квадратурно-разностного метода решения интегродифференциальных уравнений с производными переменного порядка

А. И. Федотов 564

О равномерной монотонной аппроксимации непрерывных монотонных функций с помощью сдвигов и сжатий интеграла Лапласа

А. В. Чернов 580

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

Прямо-двойственный метод Ньютона с наискорейшим спуском для линейной задачи полуопределенного программирования. Итерационный процесс

В. Г. Жадан 597

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

Об одной краевой задаче для эллиптического уравнения четвертого порядка на плоскости

А. П. Солдатов 616

МАТЕМАТИЧЕСКАЯ ФИЗИКА

Влияние монотонизации на спектральное разрешение бикомпактных схем в задаче о невязком вихре Тейлора–Грина

М. Д. Брагин 625

Об упрощении численных и аналитических “инструментов” описания “звукового удара”

Х. Ф. Валиев, А. Н. Крайко, Н. И. Тилляева 642

Исследование и применение метода декомпозиции области для моделирования тепловыделяющего элемента

М. П. Галанин, А. С. Родин 659

Моделирование ламинарно-турбулентного перехода с применением гибридных разностных схем

И. В. Егоров, Н. К. Нгуен 677

Сравнительный анализ градиентных методов определения источника диффузионно-логистической модели

Т. А. Звонарева, О. И. Криворотько 694

Дискретная модель уравнения Больцмана для девяти скоростей: решение в виде суммы Вайлда и приложения к моделированию несжимаемых течений

О. В. Ильин 705

**ОБЩИЕ
ЧИСЛЕННЫЕ МЕТОДЫ**

УДК 519.635.1

**Н-, Р- И НР-ВАРИАНТЫ МЕТОДА КОЛЛОКАЦИИ И НАИМЕНЬШИХ
КВАДРАТОВ ДЛЯ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ
ДЛЯ БИГАРМОНИЧЕСКОГО УРАВНЕНИЯ В НЕРЕГУЛЯРНЫХ
ОБЛАСТЯХ И ИХ ПРИЛОЖЕНИЯ¹⁾**

© 2022 г. В. А. Беляев^{1,2,*}, Л. С. Брындин^{1,2,**}, С. К. Голушко^{2,3,***},
Б. В. Семисалов^{2,4,****}, В. П. Шапеев^{1,2,*****}

¹ 630090 Новосибирск, ул. Институтская, 4/1, Институт теоретической и прикладной механики
им. С.А. Христиановича СО РАН, Россия

² 630090 Новосибирск, ул. Пирогова, 2, Новосибирский государственный университет, Россия

³ 630090 Новосибирск, пр-т Акад. Лаврентьева, 6, Федеральный исследовательский центр
информационных и вычислительных технологий, Россия

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

*e-mail: belyaevasily@mail.ru

**e-mail: bryndin-1996@mail.ru

***e-mail: s.k.golushko@gmail.com

****e-mail: vibis@ngs.ru

*****e-mail: shapeev.vasily@mail.ru

Поступила в редакцию 10.02.2020 г.
Переработанный вариант 05.03.2021 г.
Принята к публикации 16.11.2021 г.

Предложены и реализованы новые h-, p- и hp-варианты метода коллокации и наименьших квадратов, позволяющие находить приближенные решения краевых задач для неоднородного бигармонического уравнения в нерегулярных и многосвязных областях. Приведены формулы для операции продолжения при переходе с грубой сетки на более мелкую на многосвязном комплексе в случае применения различных пространств полиномов. Экспериментально показано, что численные решения краевых задач, полученные разработанными вариантами метода, сходятся с повышенным порядком к аналитическим решениям в случаях, когда последние не имеют особенностей. Приведено сравнение полученных результатов с результатами других авторов, использовавших конечно-разностный, конечно-элементный и другие методы, основанные на применении полиномов Чебышёва. Рассмотрены примеры задач с особенностями. Разработанные варианты метода использованы для моделирования изгиба упругой изотропной пластины нерегулярной формы, находящейся под действием поперечной нагрузки. Библ. 30. Фиг. 5. Табл. 12.

Ключевые слова: бигармоническое уравнение, нерегулярная и многосвязная область, краевая задача, метод коллокации и наименьших квадратов, повышенный порядок сходимости, изгиб изотропной пластины.

DOI: 10.31857/S0044466922040020

ВВЕДЕНИЕ

Аналитические решения краевых задач для уравнений с частными производными (УЧП) удается получить в редких случаях и, как правило, в канонических областях. Однако многие физические явления в реальном мире происходят в областях со сложной геометрией, отсюда возникает необходимость разработки эффективных численных методов решения УЧП в нерегулярных областях.

¹⁾ Работа выполнена по теме государственного задания (№ госрегистрации 121030500137-5 и АААА-А19-119051590004-5) и частично поддержана грантом РФФИ № 18-29-18029.

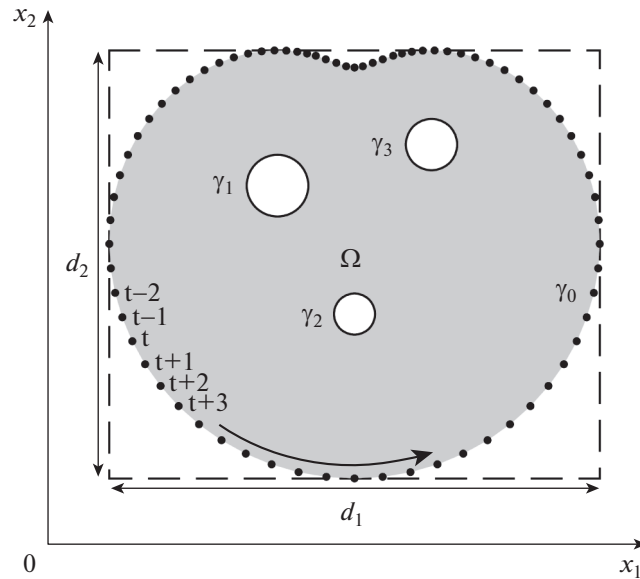
Данная работа посвящена численному решению с повышенной точностью различных краевых задач для бигармонических уравнений, к решению которых сводится анализ поведения многих задач механики сплошной среды (МСС): гидродинамики при малых числах Рейнольдса, линейной теории упругости, теории тонких пластин и др.

При численном решении краевых задач для бигармонического уравнения во многих случаях применяется метод конечных разностей (МКР), главными преимуществами которого являются простота построения расчетной сетки и скорость решения соответствующих задач линейной алгебры. При этом существует ограниченное количество публикаций, посвященных применению МКР для решения бигармонического уравнения в нерегулярных областях (см. [1]–[4]). Другим широко используемым методом для решения таких задач является метод конечных элементов (МКЭ) с использованием неструктурированных сеток (см. [5]–[7]). Как отмечается в [8], несмотря на то, что техника решения УЧП четвертого порядка с помощью МКЭ и МКР хорошо развита, существует не так много результатов, полученных при решении краевых задач в нерегулярных областях со сложными граничными условиями. Кроме того, работа с несколькими тысячами элементов в расчетной области, имеющих разную форму и ориентацию, — достаточно сложная задача (см. [8], [9]).

Условно численные методы решения бигармонического уравнения можно разбить на две основные группы: методы, использующие непосредственную дискретизацию заданного уравнения и краевых условий (см. [8], [10], [11]), и смешанные (см. [3], [4]). Например, в [3], [4] отмечается сложность численного решения бигармонического уравнения из-за наличия производных четвертого порядка в дифференциальном уравнении, поскольку, как известно, переход от системы дифференциальных уравнений к одному уравнению более высокого порядка сопровождается существенным увеличением обусловленности системы линейных алгебраических уравнений (СЛАУ) приближенной задачи. Тем не менее, как сказано в [12], “...построение эффективных методов численного решения бигармонического уравнения может уменьшить влияние этого обстоятельства”. Чтобы избежать проявлений этого недостатка, в [3] вводится дополнительная переменная и решение исходной задачи Дирихле для бигармонического уравнения сводится к последовательному решению двух уравнений Пуассона.

В последнее время при решении задач МСС все более популярными становятся спектральные методы, методы спектральных элементов и методы, основанные на применении полиномов Чебышева достаточно высоких степеней, главным преимуществом которых является экспоненциальный характер уменьшения погрешности в случае достаточно гладких решений (см. [8], [10], [11]). Однако, когда решения не являются достаточно гладкими, такие методы становятся менее эффективными и менее гибкими, например, по сравнению с МКЭ. Существуют и другие методы решения бигармонического уравнения: метод быстрых мультиполей (см. [13]), метод сплайн-коллокации (см. [14]), миметический метод (см. [15]), смешанные методы (см. [9]) и др. Отметим, что большинство работ посвящено решению краевой задачи Дирихле для бигармонического уравнения и существенно меньше задачам, в краевые условия которых входят производные более высоких порядков. Их наличие приводит к дополнительным трудностям и уменьшает возможности получения решений высокой точности.

Данная работа посвящена разработке и реализации h -, p - и hp -вариантов метода коллокации и наименьших квадратов (h -, p - и hp -МКНК) (см. [16]–[24]) для решения бигармонических уравнений в нерегулярных областях, в том числе в многосвязных областях. В проекционно-сеточном МКНК путем проектирования исходной задачи для УЧП в конечномерное линейное функциональное пространство ставится в соответствие приближенная задача, сводящаяся к решению переопределенной СЛАУ, которое определяет приближенное решение дифференциальной задачи. В h -МКНК уточнение решения происходит за счет измельчения шагов сетки при фиксированной степени базисных полиномов используемого полиномиального пространства. В p -МКНК уточнение решения происходит за счет увеличения степени базисных полиномов используемого полиномиального пространства при фиксированной величине шага сетки. В hp -МКНК уточнение решения происходит за счет измельчения шагов сетки и увеличения степени базисных полиномов используемого полиномиального пространства. В данной работе исследованы возможности p -МКНК₁ при решении задачи в пространстве полиномов достаточно высокой степени в одной ячейке, совпадающей со всей расчетной областью, и p -МКНК _{h} с фиксированным количеством ячеек в количестве больше одной.



Фиг. 1. Область решения задачи, где t — номера точек “•”, дискретно задающих ее границу.

1. ПОСТАНОВКА ЗАДАЧИ И СЛУЧАЙ ДИСКРЕТНОГО ЗАДАНИЯ ВНЕШНЕЙ ГРАНИЦЫ ОБЛАСТИ

Рассмотрим краевую задачу Дирихле для неоднородного бигармонического уравнения

$$\Delta^2 u = f(x_1, x_2), \quad (x_1, x_2) \in \Omega, \quad (1.1)$$

$$u|_{\gamma_i} = g_{li}(x_1, x_2), \quad i = \overline{0, N_\gamma}, \quad (1.2)$$

$$u_n|_{\gamma_0} \stackrel{\text{п.в.}}{=} g_{20}(x_1, x_2), \quad (1.3)$$

$$u_n|_{\gamma_i} = g_{2i}(x_1, x_2), \quad i = \overline{1, N_\gamma}, \quad (1.4)$$

в нерегулярной многосвязной области $\Omega \subset \mathbb{R}^2$ с внешней замкнутой границей γ_0 и внутренними границами γ_i (окружности), $i = \overline{1, N_\gamma}$ (фиг. 1), где N_γ — количество круглых отверстий внутри области Ω . Здесь $u(x_1, x_2)$ — искомая функция, $f(x_1, x_2)$, $g_{li}(x_1, x_2)$ и $g_{2i}(x_1, x_2)$ — заданные функции, $\Delta^2 = \frac{\partial^4}{\partial x_1^4} + 2\frac{\partial^4}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4}{\partial x_2^4}$, $u_n = \frac{\partial u}{\partial n}$, $\mathbf{n}$ — внешняя единичная нормаль к соответствующей границе γ_i , $i = \overline{0, N_\gamma}$. Решение задачи (1.1)–(1.4) понимается в обобщенном смысле. Предполагается, что $u \in C^4(\Omega)$, $u|_{\gamma_i} \in C^1(\gamma_i)$, $i = \overline{1, N_\gamma}$, и u — непрерывная кусочно-гладкая функция, дифференцируемая как минимум один раз на каждой из частей γ_0 , отделенных друг от друга точками излома. Условие (1.2) ставится всюду на γ_i , $i = \overline{0, N_\gamma}$, и (1.4) всюду на γ_i , $i = \overline{1, N_\gamma}$. При построении обобщенного решения краевое условие (1.3) требуется выполнить почти всюду (п.в.) на γ_0 (см. [26]): всюду на γ_0 за исключением точек излома в случае их наличия, поскольку направление нормали в них не определено.

На практике зачастую встречаются задачи, граница которых полностью или частично задана дискретным набором точек. Пусть внешняя граница γ_0 области Ω задана дискретно конечной последовательностью точек, расположенных приблизительно равномерно относительно друг друга (фиг. 1). Восполним в этом случае границу области по входным дискретным данным векторной сплайн-функцией, в качестве компонент которой взяты два кубических сплайна, являющиеся функциями параметра t , монотонно меняющегося при монотонном движении точки

вдоль замкнутой границы. Во многих случаях в качестве параметра t , $t = \overline{0, N_p - 1}$, удобно взять номер точки (фиг. 1). Более подробное описание техники восполнения границы области можно найти в [17]. Отметим, что если из входных данных или других каких-то соображений известно, что некоторые точки являются точками излома границы, то следует строить отдельные векторные сплайн-функции на участках, концами которых являются точки излома.

2. ОПИСАНИЕ МКНК

2.1. Построение сеток и выбор базисных элементов

Расчетную область Ω впишем в прямоугольник размера $d_1 \times d_2$ (см. [17]). В нем построим сетку $N_1 \times N_2$ с прямоугольными ячейками Ω_j размера $2h_1 \times 2h_2$, $j = \overline{1, N_{\text{cells}}}$, где N_{cells} — количество расчетных ячеек.

Ячейки сетки, полностью расположенные внутри области, назовем *внутренними* (фиг. 2, ячейки 9 и 11). Прямоугольные ячейки, пересеченные границами области γ_i , $i = \overline{0, N_\gamma}$, назовем *граничными*. Часть граничной ячейки, отсеченную границами γ_i , $i = \overline{0, N_\gamma}$, и лежащую внутри области, назовем *n-ячейкой* (фиг. 2, ячейки 1–8, 10, 12–16). Считаем, что на фиг. 2 номера граничных ячеек совпадают с номерами их внутренних n-ячеек. Часть граничной ячейки, расположенную вне области, назовем *законтурной*. Прямоугольную ячейку сетки, от которой границами области γ_i , $i = \overline{0, N_\gamma}$, отсечена n-ячейка, назовем *материнской*. Внутренние прямоугольные ячейки полностью совпадают со своими материнскими ячейками. Стороны любых ячеек, расположенные внутри области, назовем *внутренними сторонами* ячеек. Части границ γ_i , $i = \overline{0, N_\gamma}$, которые оказались внутри граничной ячейки, назовем ее *внешними сторонами*.

В каждой материнской ячейке области введем локальные координаты

$$y_1 = (x_1 - x_{1j})/h_1, \quad y_2 = (x_2 - x_{2j})/h_2, \quad (2.1)$$

где (x_{1j}, x_{2j}) — центр j -й материнской ячейки, $h_1 = d_1/(2N_1)$, $h_2 = d_2/(2N_2)$. Положим далее $v(y_1, y_2) = u(x_1(y_1), x_2(y_2))$.

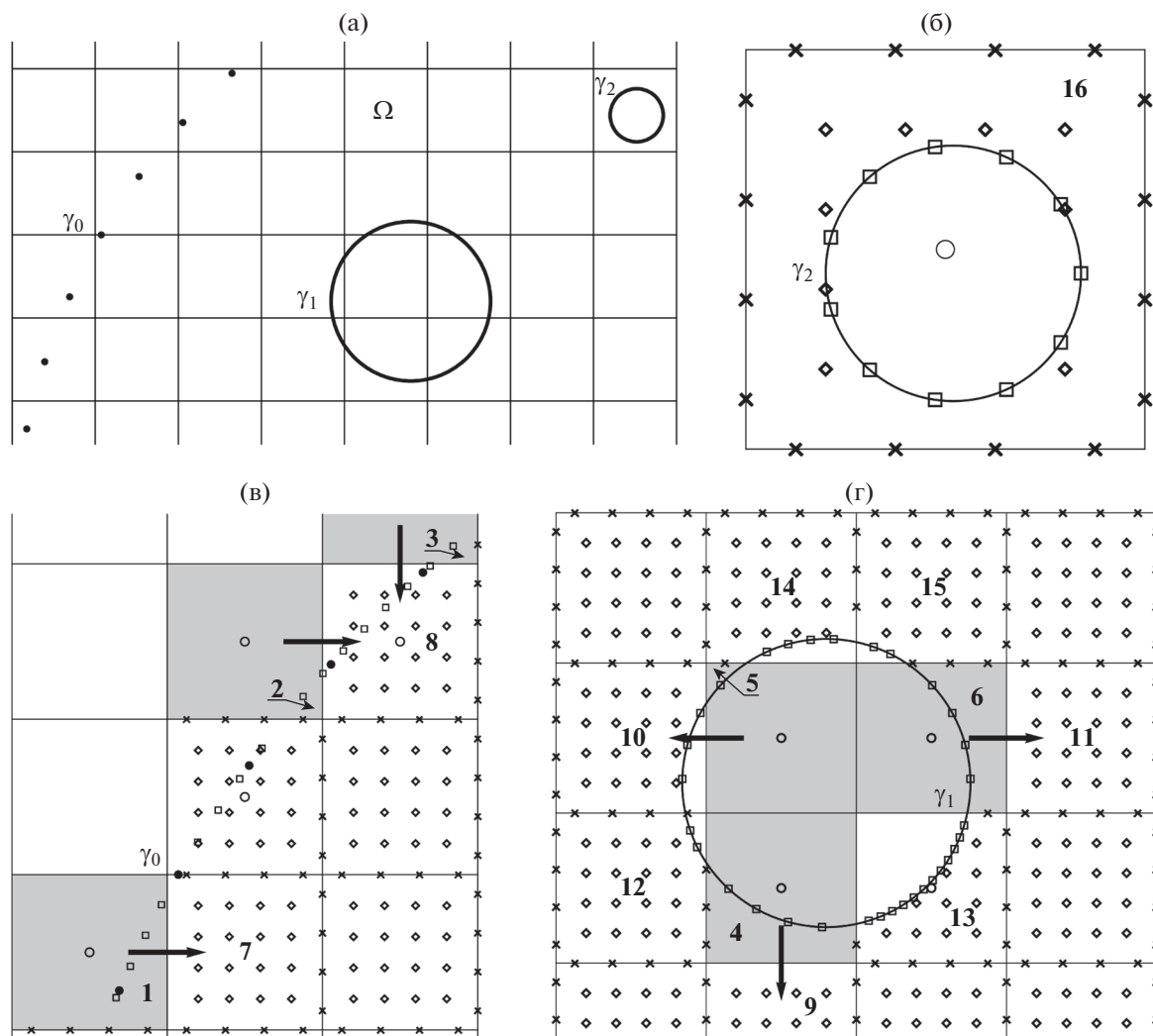
Задача (1.1)–(1.4) после замены (2.1) в локальных переменных примет следующий вид:

$$\begin{aligned} \frac{1}{h_1^4} \frac{\partial^4 v}{\partial y_1^4} + \frac{2}{h_1^2 h_2^2} \frac{\partial^4 v}{\partial y_1^2 \partial y_2^2} + \frac{1}{h_2^4} \frac{\partial^4 v}{\partial y_2^4} &= f(x_1(y_1), x_2(y_2)), \quad (y_1, y_2) \in \Omega, \\ v|_{\gamma_i} &= g_{1i}(x_1(y_1), x_2(y_2)), \quad (y_1, y_2) \in \gamma_i \cap \delta\Omega_j, \quad i = \overline{0, N_\gamma}, \\ \left(\frac{n_1}{h_1} \frac{\partial v}{\partial y_1} + \frac{n_2}{h_2} \frac{\partial v}{\partial y_2} \right) \Big|_{\gamma_0} &\stackrel{\text{п.в.}}{=} g_{20}(x_1(y_1), x_2(y_2)), \quad (y_1, y_2) \in \gamma_0 \cap \delta\Omega_j, \\ \left(\frac{n_1}{h_1} \frac{\partial v}{\partial y_1} + \frac{n_2}{h_2} \frac{\partial v}{\partial y_2} \right) \Big|_{\gamma_i} &= g_{2i}(x_1(y_1), x_2(y_2)), \quad (y_1, y_2) \in \gamma_i \cap \delta\Omega_j, \quad i = \overline{1, N_\gamma}, \end{aligned} \quad (2.2)$$

где (n_1, n_2) — компоненты внешней единичной нормали $\mathbf{n}$ к соответствующей границе γ_i в точке (x_1, x_2) , $i = \overline{0, N_\gamma}$, $\delta\Omega_j$ — внешняя сторона j -й ячейки, $j = \overline{1, N_{\text{cells}}}$.

Приближенное решение v_{hj} задачи (2.2) в каждой j -й ячейке ищется в виде линейной комбинации с неизвестными коэффициентами базисных элементов некоторого функционального пространства. В случае применения полиномов Чебышёва представление приближенного решения, по аналогии с [19], имеет вид

$$\begin{aligned} v_{hj}(y_1, y_2) &= \sum_{i_1=0}^{K_1-1} \sum_{i_2=0}^{K_2-1} c_{i_1 i_2, j} \phi_{i_1}(y_1) \phi_{i_2}(y_2), \quad (y_1, y_2) \in [-1, 1] \times [-1, 1], \\ \phi_{i_1}(y_1) &= \cos(i_1 \arccos(y_1)), \quad \phi_{i_2}(y_2) = \cos(i_2 \arccos(y_2)). \end{aligned} \quad (2.3)$$



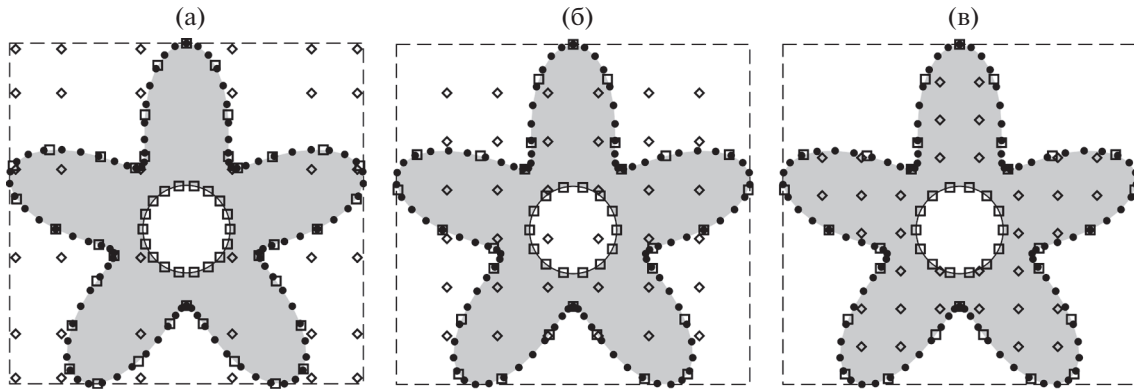
Фиг. 2. Фрагмент расчетной области (а) и его различные части (б)–(г) при $K = 4$; значок “•” обозначает точки, дискретно задающие границу области, “◊” – точки коллокации, “×” – точки согласования, “□” – точки записи краевых условий, “○” – центры некоторых н-ячеек. Темным цветом выделены материнские ячейки, содержащие внутри себя н-ячейки 1–6, которые присоединяются к ячейкам 7–11 соответственно. У граничных ячеек 2, 3, 5 тонкими стрелочками указаны их нерегулярные части, т.е. сами н-ячейки.

Если в качестве базисных элементов взяты мономы, то приближенное решение имеет вид

$$v_{hj}(y_1, y_2) = \sum_{i_1=0}^K \sum_{i_2=0}^{K-i_1} c_{i_1 i_2, j} y_1^{i_1} y_2^{i_2}. \quad (2.4)$$

В МКНК для определения неизвестных коэффициентов $c_{i_1 i_2, j}$ в каждой ячейке выписывается переопределенная “локальная” СЛАУ, состоящая из уравнений коллокации, условий согласования на общих сторонах, принадлежащих двум соседним ячейкам, и краевых условий на γ_i , $i = 0, N_\gamma$, если ячейка является граничной. Совокупность всех “локальных” СЛАУ – глобальная СЛАУ, которая определяет глобальное решение задачи.

Если в расчетной сетке имеется или появилась после измельчения шагов сетки маленькая и/или вытянутая н-ячейка (фиг. 2, ячейки 1–6), то глобальная СЛАУ задачи, как правило, становится хуже обусловленной (см. [16], [17]). Такие ячейки могут быть причиной понижения точности приближенного решения задачи, увеличения количества итераций и времени счета, а также расходимости итерационного процесса (см. [16]). Чтобы справиться с этой проблемой, здесь



Фиг. 3. Двусвязная область. Схема расположения точек записи краевых условий и точек коллокации в трех случаях для p -МКНК₁: (а) – в корнях полиномов Чебышёва в прямоугольнике при $K_1 = K_2 = 6$, (б) – равномерно в прямоугольнике при $K = 6$, (в) – равномерно внутри исходной расчетной области при $K = 6$.

используется идея присоединения таких n -ячеек к соседним ячейкам. Подробную информацию о технике присоединения можно найти, например, в [17]. Здесь кратко обозначим основную идею. Назовем *несамостоятельными* n -ячейки, расположенные в таких граничных ячейках, в которых начало локальной системы координат находится вне расчетной области (фиг. 2, ячейки 1–6). Все остальные ячейки, включая внутренние и некоторые исключительные ячейки, речь о которых пойдет ниже, по определению считаются самостоятельными. Предлагается несамостоятельную n -ячейку присоединять к той соседней самостоятельной ячейке, с которой она имеет наибольшей длины внутреннюю сторону среди ее всех внутренних сторон, общих с другими соседними самостоятельными ячейками (фиг. 2, ячейки 1–6 присоединяются к ячейкам 7–11 соответственно). Через такие стороны в присоединенные n -ячейки продолжается решение из ячейки, к которой их присоединили. Согласно сформулированному выше правилу, к самостоятельной ячейке могут быть присоединены несколько малых/вытянутых несамостоятельных n -ячеек. В этом случае внешняя сторона объединенной ячейки состоит из внешних сторон всех объединенных в ней ячеек. Объединенную ячейку также считаем граничной n -ячейкой. При этом, если круглое отверстие в случае многосвязной области полностью оказалось внутри прямоугольной ячейки, даже если начало локальной системы координат находится вне расчетной области Ω , то такую ячейку считаем самостоятельной (фиг. 2, ячейка 16). Прямоугольная ячейка, которая изначально содержит внутри себя несколько непересекающихся участков границ γ_i , $i = \overline{0, N_\gamma}$, также считается самостоятельной.

В p -МКНК₁ исходная область также включается в прямоугольник (фиг. 3) и в расчетной области строится приближение решения единым полиномом высокой степени. Аналогично по формуле (2.1) вводятся локальные координаты и $N_{\text{cells}} = N_1 = N_2 = j = 1, 2h_1 = d_1, 2h_2 = d_2$.

2.2. Уравнения приближенной задачи и их решение

Будем понимать под степенью переопределенности “локальных” СЛАУ величину η , равную отношению количества уравнений в ней к количеству неизвестных. В работах [16], [17], [22], посвященных решению бигармонического уравнения h -МКНК было установлено, что хорошая обусловленность приближенной задачи достигается при $\eta \approx 3$.

Уравнения коллокации, умноженные на $h_1^2 h_2^2$ в каждой j -й ячейке, $j = \overline{1, N_{\text{cells}}}$, выписываются в N_{col} точках коллокации и имеют вид

$$k_c \left(\frac{h_2^2}{h_1^2} \frac{\partial^4 v_{hj}}{\partial y_1^4} + 2 \frac{\partial^4 v_{hj}}{\partial y_1^2 \partial y_2^2} + \frac{h_1^2}{h_2^2} \frac{\partial^4 v_{hj}}{\partial y_2^4} \right) = k_c h_1^2 h_2^2 f(x_1(y_{1c}), x_2(y_{1c})), \quad (2.5)$$

где (y_{1c}, y_{2c}) – точки коллокации, $c = \overline{1, N_{\text{col}}}$, k_c – положительный весовой множитель уравнения коллокации. В случае применения базисных полиномов Чебышёва количество N_{nu} неизвестных

коэффициентов $c_{i_1 i_2, j}$ в (2.3) будет $N_{\text{nu}} = K_1 K_2$. Аналогично [19], полагаем $N_{\text{col}} = N_{\text{nu}}$ и выписываем уравнения коллокации в точках с координатами $(\alpha_{i_1}^1, \alpha_{i_2}^2)$, $i_1 = \overline{1, K_1}$, $i_2 = \overline{1, K_2}$, $\alpha_{i_1}^1$ и $\alpha_{i_2}^2$ – корни многочленов Чебышева первого рода степеней K_1 и K_2 соответственно. На фиг. 3а изображена расстановка точек коллокации в случае применения полиномов Чебышева при $K_1 = K_2 = 6$ для р-МКНК₁.

В случае применения мономов степени K в качестве базисных элементов количество N_{nu} неизвестных коэффициентов $c_{i_1 i_2, j}$ в (2.4) будет $N_{\text{nu}} = (K+1)(K+2)/2$. Полагаем $N_{\text{col}} = ([\sqrt{N_{\text{nu}}}] + 1)^2$, где “[.]” – целая часть числа. Затем располагаем точки коллокации в каждой ячейке равномерно. На фиг. 3б изображена расстановка точек коллокации в случае применения мономов при $K = 6$ для р-МКНК₁, а на фиг. 2 изображены фрагмент расчетной области и его различные части с сеткой при $K = 4$. При этом во всех вариантах МКНК (кроме р-МКНК₁) в оказавшихся внутри отверстия точках коллокации н-ячеек, к которым не присоединялись н-ячейки, уравнение коллокации (2.5) не выписывалось (см. фиг. 2, ячейки 13 и 16). Авторы этой работы дополнительно рассмотрели р-МКНК₁ без использования законтурной части. На фиг. 3в изображена равномерная расстановка точек коллокации в случае применения мономов при $K = 6$, которая была осуществлена следующим образом. Сначала полагалось $N_{\text{col}} = ([\sqrt{N_{\text{nu}}}] + 1)^2$ и точки равномерно располагались (фиг. 3б) в прямоугольнике, который включает нерегулярную область. Далее компьютерная программа вычисляла количество точек, которые расположены внутри исходной нерегулярной области. Если количество таких точек оказалось меньше, чем значение N_{nu} , то последовательно полагалось $N_{\text{col}} = ([\sqrt{N_{\text{nu}}}] + 2)^2, ([\sqrt{N_{\text{nu}}}] + 3)^2, \dots$ до тех пор, пока очередное N_{col} не оказалось больше, чем N_{nu} .

Во всех вариантах (кроме р-МКНК₁) в каждой j -й ячейке, $j = \overline{1, N_{\text{cells}}}$, в качестве условия согласования решения в $[N_{\text{col}}/4]$ точках согласования (фиг. 2) на каждой общей стороне между соседними ячейками требовалась непрерывность линейной комбинации с весами значений искомого решения и его производных по нормали:

$$k_{m_0} v_{hj} + \frac{k_{m_1}}{h_{nj}} \frac{\partial v_{hj}}{\partial n_j} = k_{m_0} \hat{v}_h + \frac{k_{m_1}}{h_{nj}} \frac{\partial \hat{v}_h}{\partial n_j}, \quad (2.6)$$

$$h_1 h_2 \left(\frac{k_{m_2}}{h_{nj}^2} \frac{\partial^2 v_{hj}}{\partial n_j^2} + \frac{k_{m_3}}{h_{nj}^3} \frac{\partial^3 v_{hj}}{\partial n_j^3} \right) = h_1 h_2 \left(\frac{k_{m_2}}{h_{nj}^2} \frac{\partial^2 \hat{v}_h}{\partial n_j^2} + \frac{k_{m_3}}{h_{nj}^3} \frac{\partial^3 \hat{v}_h}{\partial n_j^3} \right), \quad (2.7)$$

где $\mathbf{n}_j$ – внешняя единичная нормаль к границе j -й ячейки, v_{hj} и $\hat{v}_h$ – пределы приближенного решения задачи при стремлении изнутри и извне к границе j -й ячейки соответственно, $h_{nj} = h_1$, если направление $\mathbf{n}_j$ совпадает с направлением оси y_1 , и $h_{nj} = h_2$, если направление $\mathbf{n}_j$ совпадает с направлением оси y_2 , $k_{m_0}, k_{m_1}, k_{m_2}, k_{m_3}$ – положительные весовые множители условий согласования. Здесь $\frac{\partial v_{hj}}{\partial n_j} = n_{1j} \frac{\partial v_{hj}}{\partial y_1} + n_{2j} \frac{\partial v_{hj}}{\partial y_2}$, где (n_{1j}, n_{2j}) – вектор единичной внешней нормали $\mathbf{n}_j$ к границе j -й ячейки. При отсутствии множителя $h_1 h_2$ в (2.7) k_{m_2}, k_{m_3} нужно согласовывать со значениями шагов сетки (см. п. 2.3). При этом в оказавшихся внутри отверстия точках согласования н-ячеек, к которым не присоединялись н-ячейки, уравнения согласования (2.6), (2.7) не выписывались (фиг. 2, ячейки 12–15).

В р-МКНК₁ ($j = 1$) на внешней границе γ_0 в N_{nu} точках, приблизительно равномерно расположенных на ней (фиг. 3), выписываются краевые условия

$$k_{b_0} v_{hj} = k_{b_0} g_1(x_1, x_2), \quad (2.8)$$

$$k_{b_1} \left(\frac{n_1}{h_1} \frac{\partial v_{hj}}{\partial y_1} + \frac{n_2}{h_2} \frac{\partial v_{hj}}{\partial y_2} \right) = k_{b_1} g_2(x_1, x_2), \quad (2.9)$$

где (n_1, n_2) – вектор единичной внешней нормали $\mathbf{n}$ к границе γ_0 , k_{b_0}, k_{b_1} – положительные весовые множители краевых условий Дирихле. Информацию о технике расстановки точек записи краевых

вых условий на дискретно заданной границе области можно найти в [17]. Если нерегулярная область является многосвязной с внутренними контурами — окружностями радиусов $r_i, i = \overline{1, N_\gamma}$, то на каждом внутреннем контуре дополнительно равномерно располагается $[2r_i N_{\text{nu}} / \sqrt{d_1 d_2}]$ точек для записи краевых условий (фиг. 3).

Во всех вариантах МКНК (кроме р-МКНК₁) на внешней стороне граничной j -й ячейки, в том числе в объединенной граничной n -ячейке, $j = \overline{1, N_{\text{cells}}}$, которая может совпадать с некоторой частью одной или частями нескольких границ $\gamma_i, i = \overline{1, N_\gamma}$, области Ω , в $(4 - N_s)[N_{\text{col}}/4]$ точках на ней (фиг. 2), выписываются краевые условия (2.8), (2.9), где $N_s = \overline{1, 3}$ — количество ячеек, соседних с рассматриваемой. При этом, если к граничной n -ячейке не присоединялась какая-либо n -ячейка, то вместо $(4 - N_s)[N_{\text{col}}/4]$ точек записи краевых условий на ее границе равномерно располагается столько точек записи краевых условий, сколько точек коллокации и согласования располагается внутри отверстия (фиг. 2, ячейки 12–15). Если отверстие оказалось полностью внутри некоторой материнской ячейки (фиг. 2, ячейка 16), то на границе данного отверстия располагается $N_{\text{col}} + [N_{\text{col}}/4] - N_{\text{cic}}$ точек записи краевых условий, где N_{cic} — количество точек коллокации, которые лежат в материнской ячейке, но вне данного отверстия.

Аналогично работам [16], [17], здесь в целях достижения аппроксимации повышенной точности дифференциальной задачи применялись “законтурные” точки коллокации и согласования в граничных n -ячейках около внешней дискретно заданной границы γ_0 (фиг. 2). У n -ячеек, которые являются самостоятельными, имеют четыре соседа и которые пересечены границами $\gamma_i, i = \overline{1, N_\gamma}$, круглых отверстий, “законтурные” части в произвольном случае зачастую малы по сравнению с “законтурными” частями граничных n -ячеек, которые пересечены внешней границей γ_0 . Здесь предложено не использовать у таких ячеек их законтурные части для записи уравнений приближенной задачи. Однако зачастую длина границ $\gamma_i, i = \overline{1, N_\gamma}$, существенно меньше длины внешней границы γ_0 . В таких случаях требование на количество краевых условий становится жестче: необходимо по возможности выписывать краевые условия в каждой ячейке, которая содержит границу $\gamma_i, i = \overline{1, N_\gamma}$, даже если ячейка имеет четыре соседа, т.е. $N_s = 4$.

Объединяя уравнения приближенной задачи (2.5)–(2.9) в соответствии с конкретным вариантом МКНК и типом ячеек, относительно искомым $c_{i_1 i_2, j}$ имеем переопределенную локальную СЛАУ $Ax = b$, где A — прямоугольная вещественная матрица, x, b — векторы неизвестных и правых частей уравнений соответственно. Глобальная СЛАУ здесь решается с помощью метода итераций по подобластям (см. [16]). При построении решения в каждой ячейке делается QR -декомпозиция матрицы A . Итерационный процесс продолжается до тех пор, пока не выполнится условие

$$\max_{i_1 i_2 j} |c_{i_1 i_2, j}^{n+1} - c_{i_1 i_2, j}^n| < \epsilon, \quad (2.10)$$

где $c_{i_1 i_2, j}^n$ — коэффициент полинома в j -й ячейке на n -й итерации, $j = \overline{1, N_{\text{cells}}}$, ϵ — наперед заданная малая константа, называемая псевдопогрешностью решения.

2.3. Ускорение итерационного процесса

Для значительного уменьшения количества итераций и времени, необходимого для проведения расчета, здесь будем использовать: комбинированное применение (см. [24]) операции продолжения на многосеточном комплексе в методе Федоренко (см. [26]); диагонального предобуславливателя (см. [27]); на всех сетках комплекса метода ускорения сходимости итерационного процесса, основанного на подпространствах Крылова (см. [28]); свойства матриц локальных СЛАУ в линейной задаче с постоянными коэффициентами в МКНК (см. [22]). Значения весовых множителей в МКНК влияют на точность приближенного решения задачи, скорость сходимости итерационного процесса, обусловленность локальных СЛАУ (см. [21]). Поэтому в нем одним из эффективных способов уменьшения времени решения задачи является выбор оптимальных значений этих множителей. Авторы представленной работы воспользовались опытом и результатами предыдущих работ, посвященных решению бигармонического уравнения

(см. [17], [22]). Однако в некоторых примерах можно заметить различающиеся значения весовых множителей. Их поиск был осуществлен по алгоритмам из [21], [22].

2.3.1. Операция продолжения на многосеточном комплексе. Приведем формулы для операции продолжения, необходимые для перехода с грубой сетки на более мелкую на многосеточном комплексе. Конкретный вид формул может быть полезен при реализации подобных алгоритмов.

Пусть у нас имеется решение, построенное на сетке размера $N_1/2 \times N_2/2$ с ячейками размера $4h_1 \times 4h_2$, которое в j -й ячейке имеет вид

$$v_{hj}(Y_1, Y_2) = \sum_{i_1=0}^{K_1} \sum_{i_2=0}^{K_2} c_{i_1 i_2, j} \phi_{i_1}(Y_1) \phi_{i_2}(Y_2),$$

где $\phi_{i_1}(Y_1) \phi_{i_2}(Y_2)$ — базисный элемент пространства, в котором ищется решение.

Каждая материнская ячейка сетки размера $N_1/2 \times N_2/2$ (грубая сетка) содержит в себе четыре материнские ячейки сетки размера $N_1 \times N_2$ (мелкая сетка). Пусть для любой точки на плоскости: (Y_1, Y_2) — ее координаты в локальной системе координат, связанной с j -й ячейкой грубой сетки, (y_1, y_2) — в локальной системе координат, связанной с одной из четырех ячеек мелкой сетки, содержащихся в j -й ячейке грубой сетки, (x_1, x_2) — в глобальной системе координат. Обозначим (X_{1j}, X_{2j}) — центр j -й ячейки грубой сетки, (x_{1j}, x_{2j}) — центр одной из четырех ячеек, содержащихся в j -й ячейке грубой сетки. Тогда из формул (2.1) следует:

$$\begin{aligned} x_1 &= 2h_1 Y_1 + X_{1j}, & x_2 &= 2h_2 Y_2 + X_{2j}, \\ y_1 &= (2h_1 Y_1 + X_{1j} - x_{1j})/h_1, & y_2 &= (2h_2 Y_2 + X_{2j} - x_{2j})/h_2. \end{aligned}$$

Отсюда получаем

$$Y_1(y_1) = 0.5y_1 + dy_1, \quad Y_2(y_2) = 0.5y_2 + dy_2,$$

где $dy_1 = (x_{1j} - X_{1j})/2h_1$, $dy_2 = (x_{2j} - X_{2j})/2h_2$.

Запишем решение в ячейке мелкой сетки в виде

$$v_{hj}(y_1, y_2) = \sum_{i_1=0}^{K_1} \sum_{i_2=0}^{K_2} \tilde{c}_{i_1 i_2} \phi_{i_1}(y_1) \phi_{i_2}(y_2) = \sum_{i_1=0}^{K_1} \sum_{i_2=0}^{K_2} c_{i_1 i_2, j} \phi_{i_1}(0.5y_1 + dy_1) \phi_{i_2}(0.5y_2 + dy_2). \quad (2.11)$$

Базис из полиномов Чебышёва. Найдем коэффициенты $\tilde{c}_{i_1 i_2}$ в случае, когда ϕ_{i_1}, ϕ_{i_2} — полиномы Чебышёва степени i_1 и i_2 соответственно.

Для удобства введем матрицы $B_1 = (b_{i,l}^1)_{i=0,l=0}^{K_1,K_1}$ и $B_2 = (b_{i,l}^2)_{i=0,l=0}^{K_2,K_2}$ такие, что

$$\phi_{i_1}(ay_1 + b) = \sum_{l=0}^{i_1} b_{i,l}^1 \phi_l(y_1), \quad \phi_{i_2}(cy_2 + d) = \sum_{l=0}^{i_2} b_{i,l}^2 \phi_l(y_2), \quad (2.12)$$

где a, b, c, d — произвольные константы. Выпишем матрицу B_1 следующим образом. Для $i_1 = 0$ и $i_1 = 1$ имеем

$$\begin{aligned} \phi_0(ay_1 + b) &= 1 = \phi_0(y_1), \\ \phi_1(ay_1 + b) &= ay_1 + b = a\phi_1(y_1) + b\phi_0(y_1). \end{aligned}$$

Для $i_1 = 2$

$$\phi_2(ay_1 + b) = 2(ay_1 + b)^2 - 1 = a^2\phi_2(y_1) + 4ab\phi_1(y_1) + (2b^2 - 1 + a^2)\phi_0(y_1).$$

Следовательно, матрица B_1 имеет вид

$$B_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ b & a & 0 & 0 & \dots \\ 2b^2 - 1 + a^2 & 4ab & a^2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

а матрица B_2 получается аналогичным способом. Заполняя i -ю строку матрицы B_1 ($i \geq 2$) и, используя рекуррентное соотношение для полиномов Чебышёва, будем иметь

$$\phi_{i+1}(ay_1 + b) = 2(ay_1 + b)\phi_i(ay_1 + b) - \phi_{i-1}(ay_1 + b).$$

Подставим в правую часть полученного соотношения выражения для ϕ_i и ϕ_{i-1} из (2.12) и проведем тождественное преобразование

$$\begin{aligned} \phi_{i+1}(ay_1 + b) &= 2(ay_1 + b) \sum_{l=0}^i b_{i,l}^1 \phi_l(y_1) - \sum_{l=0}^{i-1} b_{i-1,l}^1 \phi_l(y_1) = a \sum_{l=1}^i b_{i,l}^1 (2y_1 \phi_l(y_1) - \phi_{l-1}(y_1)) + \\ &+ a \sum_{l=1}^i b_{i,l}^1 \phi_{l-1}(y_1) + 2ay_1 b_{i,0}^1 \phi_0(y_1) + 2b \sum_{l=0}^i b_{i,l}^1 \phi_l(y_1) - \sum_{l=0}^{i-1} b_{i-1,l}^1 \phi_l(y_1). \end{aligned}$$

Используя рекуррентное соотношение для полиномов Чебышёва и тождество $y_1 \phi_0(y_1) = \phi_1(y_1)$, перепишем последнее соотношение в виде

$$\phi_{i+1}(ay_1 + b) = a \sum_{l=1}^i b_{i,l}^1 \phi_{l+1}(y_1) + a \sum_{l=1}^i b_{i,l}^1 \phi_{l-1}(y_1) + 2ab_{i,0}^1 \phi_1(y_1) + 2b \sum_{l=0}^i b_{i,l}^1 \phi_l(y_1) - \sum_{l=0}^{i-1} b_{i-1,l}^1 \phi_l(y_1).$$

Поменяем пределы суммирования у первых двух слагаемых:

$$\phi_{i+1}(ay_1 + b) = a \sum_{l=2}^{i+1} b_{i,l-1}^1 \phi_l(y_1) + a \sum_{l=0}^{i-1} b_{i,l+1}^1 \phi_l(y_1) + 2ab_{i,0}^1 \phi_1(y_1) + 2b \sum_{l=0}^i b_{i,l}^1 \phi_l(y_1) - \sum_{l=0}^{i-1} b_{i-1,l}^1 \phi_l(y_1).$$

Сгруппировав слагаемые в полученном соотношении, для элементов $b_{i+1,l}^1$ матрицы B_1 получим рекуррентные соотношения

$$\begin{aligned} b_{i+1,0}^1 &= ab_{i,1}^1 + 2bb_{i,0}^1 - b_{i-1,0}^1, \\ b_{i+1,1}^1 &= ab_{i,2}^1 + 2ab_{i,0}^1 + 2bb_{i,1}^1 - b_{i-1,1}^1, \\ b_{i+1,l}^1 &= ab_{i,l-1}^1 + ab_{i,l+1}^1 + 2bb_{i,l}^1 - b_{i-1,l}^1, \quad l = \overline{2, i-1}, \\ b_{i+1,i}^1 &= ab_{i,i-1}^1 + 2bb_{i,i}^1, \\ b_{i+1,i+1}^1 &= ab_{i,i}^1. \end{aligned}$$

Воспользуемся равенством (2.11) в представлении решения (2.3), формально заменив K_1 на $K_1 - 1$, K_2 на $K_2 - 1$. В формуле (2.12) положим $a = c = 0.5$, $b = dy_1$, $d = dy_2$ и сконструируем матрицы B_1, B_2 по полученным формулам. В результате будем иметь

$$\begin{aligned} v_{hj}(y_1, y_2) &= \sum_{i_1=0}^{K_1-1} \sum_{i_2=0}^{K_2-1} c_{i_1 i_2, j} \phi_{i_1}(0.5y_1 + dy_1) \phi_{i_2}(0.5y_2 + dy_2) = \\ &= \sum_{i_1=0}^{K_1-1} \sum_{i_2=0}^{K_2-1} c_{i_1 i_2, j} \left(\sum_{k_1=0}^{i_1} b_{i_1, k_1}^1 \phi_{k_1}(y_1) \right) \left(\sum_{k_2=0}^{i_2} b_{i_2, k_2}^2 \phi_{k_2}(y_2) \right) = \sum_{i_1=0}^{K_1-1} \sum_{i_2=0}^{K_2-1} \left[\sum_{k_1=i_1}^{K_1-1} \sum_{k_2=i_2}^{K_2-1} c_{k_1 k_2, j} b_{k_1, i_1}^1 b_{k_2, i_2}^2 \right] \phi_{i_1}(y_1) \phi_{i_2}(y_2). \end{aligned}$$

Следовательно, коэффициенты $\tilde{c}_{i_1 i_2} = \sum_{k_1=i_1}^{K_1-1} \sum_{k_2=i_2}^{K_2-1} c_{k_1 k_2, j} b_{k_1, i_1}^1 b_{k_2, i_2}^2$.

Если $K_1 - 1 = K_2 - 1 = K - 1$, то можно показать, что для осуществления операции продолжения необходимо $O(N_{\text{cells}} K^4)$ арифметических операций, в то время как для выполнения одной итерации требуется $O(N_{\text{cells}} K^6)$ арифметических операций.

Базис из мономов. Воспользуемся соотношением (2.11) в представлении решения (2.4), положив $K_1 = K$, $K_2 = K - i_1$. Применяя формулу бинома Ньютона, получаем

$$\begin{aligned} v_{hj}(y_1, y_2) &= \sum_{i_1=0}^K \sum_{i_2=0}^{K-i_1} c_{i_1 i_2, j} \phi_{i_1}(0.5y_1 + dy_1) \phi_{i_2}(0.5y_2 + dy_2) = \\ &= \sum_{i_1=0}^K \sum_{i_2=0}^{K-i_1} c_{i_1 i_2, j} \left(\sum_{k_1=0}^{i_1} C_{i_1}^{k_1} dy_1^{i_1-k_1} 0.5^{k_1} \phi_{k_1}(y_1) \right) \left(\sum_{k_2=0}^{i_2} C_{i_2}^{k_2} dy_2^{i_2-k_2} 0.5^{k_2} \phi_{k_2}(y_2) \right) = \\ &= \sum_{i_1=0}^K \sum_{i_2=0}^{K-i_1} \left[\sum_{k_1=i_1}^K \sum_{k_2=i_2}^{K-k_1} c_{k_1 k_2, j} C_{i_1}^{k_1} dy_1^{k_1-i_1} 0.5^{i_1} C_{i_2}^{k_2} dy_2^{k_2-i_2} 0.5^{i_2} \right] \phi_{i_1}(y_1) \phi_{i_2}(y_2), \end{aligned}$$

где $C_n^k = \frac{n!}{k!(n-k)!}$. Следовательно, $\tilde{c}_{i_1 i_2} = \sum_{k_1=i_1}^K \sum_{k_2=i_2}^{K-k_1} c_{k_1 k_2, j} C_{i_1}^{k_1} dy_1^{k_1-i_1} 0.5^{i_1} C_{i_2}^{k_2} dy_2^{k_2-i_2} 0.5^{i_2}$.

Полученные формулы позволяют при осуществлении операции продолжения на многосеточном комплексе записать на мелкой сетке решение, полученное на грубой сетке, без потери достигнутой его точности.

3. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

Рассмотрим задачу (1.1)–(1.4) и другие краевые задачи для бигармонического уравнения. В представленных ниже таблицах приведены значения относительной погрешности приближенного решения в дискретном аналоге нормы L_2 :

$$\|E_r\|_2 = \sqrt{\frac{\sum_{j=1}^{N_{\text{cells}}} \sum_{m=1}^{Q_j} (u(x_{1m}, x_{2m}) - u_{hj}(x_{1m}, x_{2m}))^2}{\sum_{j=1}^{N_{\text{cells}}} \sum_{m=1}^{Q_j} (u(x_{1m}, x_{2m}))^2}}, \quad (3.1)$$

относительной погрешности приближенного решения в бесконечной норме

$$\|E_r\|_{\infty} = \frac{\max_{j=1, N_{\text{cells}}} (\max_{m=1, Q_j} |u(x_{1m}, x_{2m}) - u_{hj}(x_{1m}, x_{2m})|)}{\max_{j=1, N_{\text{cells}}} (\max_{m=1, Q_j} |u(x_{1m}, x_{2m})|)}, \quad (3.2)$$

и абсолютной погрешности приближенного решения в бесконечной норме

$$\|E_a\|_{\infty} = \max_{j=1, N_{\text{cells}}} (\max_{m=1, Q_j} |u(x_{1m}, x_{2m}) - u_{hj}(x_{1m}, x_{2m})|), \quad (3.3)$$

где Q_j — количество равномерно распределенных контрольных точек (x_{1m}, x_{2m}) , взятых в j -й материнской ячейке для подсчета в них погрешности, u — точное решение задачи, u_{hj} — приближенное решение в j -й ячейке. Здесь $Q_j = 10\,000$ в р-МКНК₁, а в остальных вариантах МКНК $Q_j = 100$. Для сравнения с МКЭ (см. далее (3.8)) в табл. 11 и 12 приведены значения абсолютной погрешности в норме L_2 , где интеграл считался с помощью прямого произведения одномерных квадратурных формул Гаусса с 20 узлами по каждому из направлений

$$\|E_a\|_{L_2} = \sqrt{\int_{\Omega} (u - u_h)^2 d\Omega}.$$

Порядок сходимости погрешности численного решения в случае h-МКНК определим по формуле

$$R = \log_2 \frac{E_p}{E_c}, \quad (3.4)$$

Таблица 1. Результаты численных экспериментов (пример п. 3.1)

Chen и др., см. [3]			h-МКНК, решение (2.4) при $K = 4$			
$N_1 \times N_2$	$\ E_a\ _\infty$	R	$N_1 \times N_2$	$\ E_a\ _\infty$	R	N_{iter}
64×64	$8.15\text{e}-04$	—	10×10	$1.46\text{e}-05$	—	109
128×128	$2.08\text{e}-04$	1.97	20×20	$1.70\text{e}-06$	3.10	121
256×256	$8.07\text{e}-05$	1.37	40×40	$2.17\text{e}-07$	2.97	205
512×512	$1.83\text{e}-05$	2.13	80×80	$2.96\text{e}-08$	2.87	361

Таблица 2. Значения погрешности $\|E_a\|_\infty$ для численных экспериментов (пример п. 3.1)

K_1-1, K_2-1 или K	5	7	9	11	13
Shao и др., см. [8]	$1.74\text{e}-04$	$1.80\text{e}-06$	$2.69\text{e}-08$	$5.26\text{e}-11$	$3.09\text{e}-12$
p-МКНК ₁ (фиг. 3а), решение (2.3)	$1.04\text{e}-04$	$5.87\text{e}-07$	$2.62\text{e}-09$	$5.72\text{e}-12$	$1.77\text{e}-14$
p-МКНК ₁ (фиг. 3а), решение (2.4)	$4.54\text{e}-04$	$4.08\text{e}-06$	$3.42\text{e}-08$	$1.73\text{e}-10$	$5.16\text{e}-13$
p-МКНК ₁ (фиг. 3б), решение (2.3)	$1.02\text{e}-04$	$2.65\text{e}-07$	$7.04\text{e}-10$	$1.04\text{e}-12$	$1.15\text{e}-14$
p-МКНК ₁ (фиг. 3б), решение (2.4)	$4.67\text{e}-04$	$4.07\text{e}-06$	$3.16\text{e}-08$	$1.80\text{e}-10$	$5.20\text{e}-13$
p-МКНК ₁ (фиг. 3в), решение (2.3)	$9.99\text{e}-05$	$2.52\text{e}-07$	$8.63\text{e}-10$	$2.63\text{e}-12$	$1.42\text{e}-14$
p-МКНК ₁ (фиг. 3в), решение (2.4)	$4.72\text{e}-04$	$3.41\text{e}-06$	$2.45\text{e}-08$	$1.87\text{e}-10$	$3.25\text{e}-13$
p-МКНК _h , решение (2.4)	$2.78\text{e}-04$	$1.47\text{e}-06$	$2.35\text{e}-09$	$5.44\text{e}-12$	$3.6\text{e}-14$

где E_c — значение погрешности ($\|E_r\|_2$, $\|E_r\|_\infty$ или $\|E_a\|_\infty$) на сетке размера $N_1 \times N_2$, E_p — значение погрешности на сетке размера $N_1/2 \times N_2/2$. Обозначим через N_{iter} — количество итераций, при котором останавливался итерационный процесс (2.10).

Во всех численных экспериментах во всех вариантах МКНК (кроме p-МКНК₁) в итерационном процессе на самой грубой сетке здесь в начальном приближении решения взяты $c_{i_1 i_2, j} = 0.4$.

В данной работе во всех рассмотренных примерах известно точное решение задач. Однако в работах по МКНК показана возможность построения сходящегося решения с хорошей точностью и порядком сходимости зачастую не хуже второго при применении полиномов четвертой степени также и в случаях, когда решение различных краевых задач для бигармонического уравнения неизвестно (см. [16], [17]).

3.1. Задача Дирихле в односвязной звездной области

Рассмотрим задачу Дирихле (1.1)–(1.3) для бигармонического уравнения с тестовым решением $u(x_1, x_2) = x_1^2 + x_2^2 + e^{x_1} \cos(x_2)$ в невыпуклой области (фиг. 3, без отверстия), внешняя граница γ_0 которой задана соотношениями

$$\begin{aligned} x_1 &= (0.6 + 0.25 \sin(5t)) \cos(t), \quad t \in [0, 2\pi), \\ x_2 &= (0.6 + 0.25 \sin(5t)) \sin(t). \end{aligned}$$

В табл. 1 приведены результаты численных экспериментов в случае применения h-МКНК, где в качестве базисных элементов взяты мономы при $K = 4$. В качестве весовых множителей здесь и в p-МКНК_h взяты $k_c = k_{m_0} = k_{m_2} = k_{b_0} = k_{b_1} = 1$, $k_{m_1} = 0.01$, $k_{m_3} = 0.015$ из [17], $\epsilon = 10^{-10}$ в h-МКНК, $\epsilon = 10^{-14}$ в p-МКНК_h. В табл. 2 приведены результаты численных экспериментов в случае применения p-МКНК_h (4 ячейки) и p-МКНК₁ при использовании различных вариантов расстановки точек коллокации (фиг. 3). В первом столбце для указания способа расстановки точек коллокации в скобках приведены номера рисунков, на которых изображены соответствующие им расстановки.

В h-МКНК, p-МКНК₁ и p-МКНК_h дополнительно подсчитаны значения чисел обусловленности локальных и глобальных матриц. В первом случае без использования диагонального пре-

Таблица 3. N_{cond} — количество ячеек в h-МКНК, в которых матрицы локальных СЛАУ имеют числа обусловленности порядка 10^2 и 10^3 , cond_{max} — максимальное число обусловленности среди всех матриц локальных СЛАУ и cond_{in} — число обусловленности для матриц СЛАУ внутренних ячеек

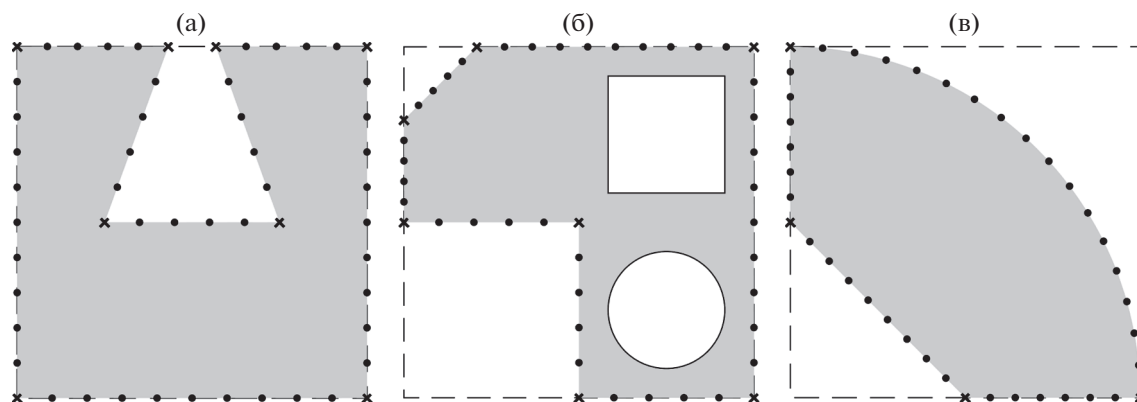
$N_1 \times N_2$	N_{cond}		cond_{max}	cond_{in}
	cond порядка 10^2	cond порядка 10^3		
10×10	50	0	3.34e+2	1.31e+2
20×20	192	0	6.99e+2	1.61e+2
40×40	746	30	1.67e+3	2.26e+2
80×80	2992	110	4.53e+3	3.91e+2

Таблица 4. Значения чисел обусловленности матриц в p-МКНК₁ и их максимальные значения в p-МКНК_h среди всех ячеек области

$K_1 - 1, K_2 - 1$ или K	5	7	9	11	13
Без использования диагонального предобуславливателя					
p-МКНК ₁ (фиг. 3а), решение (2.3)	3.34e+2	8.79e+3	1.30e+5	1.23e+6	1.18e+7
p-МКНК ₁ (фиг. 3а), решение (2.4)	5.91e+1	2.73e+2	1.26e+3	6.93e+3	4.42e+4
p-МКНК ₁ (фиг. 3б), решение (2.3)	2.88e+2	1.60e+3	6.93e+3	6.71e+4	4.51e+5
p-МКНК ₁ (фиг. 3б), решение (2.4)	6.34e+1	2.75e+2	1.33e+3	6.58e+3	3.12e+4
p-МКНК ₁ (фиг. 3в), решение (2.3)	7.76e+2	1.20e+5	6.17e+5	2.81e+7	1.67e+8
p-МКНК ₁ (фиг. 3в), решение (2.4)	6.52e+1	3.58e+2	2.08e+3	1.14e+4	7.83e+4
p-МКНК _h , решение (2.4)	1.12e+3	8.85e+3	9.85e+4	1.14e+6	1.23e+7
С использованием диагонального предобуславливателя					
p-МКНК ₁ (фиг. 3а), решение (2.3)	4.81e+1	1.86e+2	8.54e+2	4.71e+3	2.51e+4
p-МКНК ₁ (фиг. 3а), решение (2.4)	3.51e+1	1.01e+2	3.44e+2	1.76e+3	1.23e+4
p-МКНК ₁ (фиг. 3б), решение (2.3)	8.60e+1	3.30e+2	9.10e+2	4.17e+3	1.90e+4
p-МКНК ₁ (фиг. 3б), решение (2.4)	3.95e+1	1.13e+2	6.98e+2	4.15e+3	1.88e+4
p-МКНК ₁ (фиг. 3в), решение (2.3)	2.60e+2	3.29e+3	1.01e+4	2.54e+6	1.13e+7
p-МКНК ₁ (фиг. 3в), решение (2.4)	4.09e+1	1.59e+2	7.30e+2	2.75e+3	1.70e+4

добуславливателя в зависимости от размера сетки и типов ячеек значения порядка чисел обусловленности варьировались в пределах от 10^2 до 10^3 (см. табл. 3). Во втором случае с представлением решения в виде (2.4) порядок числа обусловленности менялся от 10^1 до 10^4 в зависимости от K , а при представлении решения в виде (7) — от 10^2 до 10^8 в зависимости от $K_1 - 1, K_2 - 1$ (см. табл. 4). Наконец, в последнем случае при увеличении степени K максимальное значение числа обусловленности среди всех матриц локальных СЛАУ достигало порядка 10^7 (см. табл. 4). Здесь число обусловленности вычислялось как отношение максимального сингулярного числа к минимальному, т.е. в спектральной норме. Отметим, что в [8] не приведены значения чисел обусловленности для этого примера, а только для случая прямоугольной области, где они были в диапазоне от 6.29×10^1 до 6.51×10^2 при изменении $K_1 - 1$ и $K_2 - 1$ в пределах от 7 до 13 соответственно.

При дополнительном использовании диагонального предобуславливателя в p-МКНК₁ число обусловленности здесь уменьшалось по-разному. В случае представления решения в виде (2.4) оно уменьшалось в среднем приблизительно в 2–4 раза, в виде (2.3) — на 1–3 десятичных порядка в зависимости от расположения точек коллокации. Например, при $K_1 - 1 = K_2 - 1 = 13$ и расстановкой точек как на фиг. 3а оно изменилось с $1.18e+7$ на $2.51e+4$, как на фиг. 3б — с $4.51e+5$ на $1.90e+4$, как на фиг. 3в — с $1.67e+8$ на $1.13e+7$. Увеличение числа обусловленности при примене-



Фиг. 4. Область решения задачи в примере 3.2 с координатами вершин многоугольника: $(1,1)$, $(0.5 - \text{ctg}70^\circ, 1)$, $(0.5, 0)$, $(-0.5, 0)$, $(-0.5 + \text{ctg}70^\circ, 0)$, $(-1, 1)$, $(-1, -1)$, $(1, -1)$ (а). Область решения задачи в примере 3.3 с координатами вершин многоугольника: $(0, 0)$, $(0, -1)$, $(1, -1)$, $(1, 1)$, $(-7/12, 1)$, $(-1, 7/12)$, $(-1, 0)$. Круглое отверстие радиуса $1/3$ с центром в $(1/2, -1/2)$, квадратное отверстие $[1/6, 1/6] \times [5/6, 5/6]$ (б). Область решения задачи в примере 3.5 с координатами вершин: $(1, 0)$, $(0, 1)$, $(0, 0.5)$, $(0.5, 0)$ (в).

нии базиса из полиномов Чебышёва прежде всего связано с тем, что в этом случае число неизвестных коэффициентов значительно больше, чем при использовании базиса из мономов при фиксированных значениях $K_1 - 1 = K_2 - 1 = K$, вследствие чего точек записи уравнений приближенной задачи берется больше по алгоритму, предложенному в данной работе. К тому же звездная область является невыпуклой, точки коллокации в случае, когда не используется законтурная часть для записи уравнений, будут расположены достаточно близко друг к другу внутри исходной области (а не в прямоугольнике). Это приводит к тому, что обусловленность матрицы СЛАУ ухудшается, хотя точность расчетов остается повышенной благодаря ортогональности используемых полиномов. В р-МКНК₁ в расчетах использованы следующие значения множителей: $k_c = 10^{-3}$, $k_{b_0} = k_{b_1} = 1$.

Из табл. 1 и 2 следует, что результаты, полученные применением МКНК, не только не уступают результатам, полученным другими авторами применением МКР повышенного порядка аппроксимации (см. [3]) и СТММID (англ. Chebyshev tau meshless method based on the integration–differentiation) (см. [8]), но и превосходят их во многих случаях в этом примере. В [8] приближенное решение и его производные искались в виде (2.3). Из табл. 2 видно, что лучшие результаты достигнуты с применением р-МКНК₁.

3.2. Задача Дирихле в односвязной полигональной области

Рассмотрим в невыпуклой области (фиг. 4а) задачу Дирихле (1.1)–(1.3) для бигармонического уравнения с тестовым решением $u(x_1, x_2) = \sin(\pi x_1) \sin(\pi x_2)$. В [11] такая область разбивалась на пять подобластей и решение в каждой подобласти искалось отдельно в виде (2.3). В табл. 5 приведены результаты численных экспериментов, полученных с применением р-МКНК₁ и р-МКНК_h и СТММID-DDM (англ. Chebyshev tau meshless method based on the integration–differentiation with domain decomposition method) (см. [11]). Поскольку в [11] представлены значения погрешности $\|E_a\|_\infty$ в пяти выбранных авторами некоторых подобластях $(\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5)$, то для сравнения с полученными результатами и экономии места приведем в табл. 5 их осредненные значения $\|E_a\|_\infty = (\|E_{a,\Omega_1}\|_\infty + \|E_{a,\Omega_2}\|_\infty + \|E_{a,\Omega_3}\|_\infty + \|E_{a,\Omega_4}\|_\infty + \|E_{a,\Omega_5}\|_\infty) / 5$.

Из табл. 5 видно, что при применении р-МКНК_h и р-МКНК₁ удается получить решение, не уступающее по точности решению по сравнению со случаем применения СТММID-DDM (см. [11]). Кроме того, при применении р-МКНК_h удается достичь значение абсолютной погрешности $\|E_a\|_\infty$ порядка e^{-11} . В р-МКНК_h и р-МКНК₁ при расчетах использовались значения множителей из предыдущего примера.

Таблица 5. Результаты численных экспериментов (пример п. 3.2)

Shao, Wu, см. [11]				
K_1-1, K_2-1	8	10	12	14
$\ E_a\ _\infty$	1.71e-03	4.34e-05	3.91e-06	2.41e-07
p-МКНК ₁ (фиг. 3в), решение (2.3)				
K_1-1, K_2-1	8	10	12	14
$\ E_a\ _\infty$	4.89e-03	1.31e-04	2.92e-06	4.03e-08
p-МКНК _h , решение (2.4), $N_1 \times N_2 = 8 \times 8$				
K	5	7	9	11
$\ E_a\ _\infty$	6.34e-02	5.81e-05	2.02e-06	3.47e-09
p-МКНК _h , решение (2.4), $N_1 \times N_2 = 16 \times 16$				
K	5	7	9	11
$\ E_a\ _\infty$	1.46e-02	2.61e-06	3.30e-08	3.88e-11

Таблица 6. Результаты численных экспериментов (пример п. 3.3)

Mai-Duy и др., см. [10]				
K_1-1, K_2-1	7	9	11	13
$\ E_r\ _2$	4.8e-05	2.5e-07	1.9e-09	8.7e-12
Shao и др., см. [8]				
K_1-1, K_2-1	8	12	16	20
$\ E_r\ _2$	8.06e-04	1.34e-05	2.58e-09	2.45e-13
p-МКНК ₁ (фиг. 3а), решение (2.3)				
K_1-1, K_2-1	8	12	16	20
$\ E_r\ _2$	3.43e-05	2.71e-07	2.42e-11	5.13e-15

3.3. Задача Дирихле в многосвязной области

Рассмотрим в многосвязной области (фиг. 4б) задачу Дирихле (1.1)–(1.4) для бигармонического уравнения с тестовым решением $u(x_1, x_2) = \frac{1}{\pi^4}(1 + \cos(\pi x_1))(1 + \cos(\pi x_2))$. Здесь условие (1.4) выполнено почти всюду на границе квадратного отверстия. В табл. 6 приведены результаты численных экспериментов. В [10] исходная область разбивалась на три подобласти, в каждой из которых приближенное решение искалось отдельно. При этом в [10] не приведены значения погрешности при $K_1, K_2 > 14$. В [8] задача сразу решалась во всей области и там было достигнуто значение погрешности порядка $\sim e-13$. Более того, при увеличении степени полиномов в [8] точность решения падала при $K_1 - 1, K_2 - 1 > 20$. Однако из табл. 6 видно, что с помощью p-МКНК₁ удастся достигнуть погрешности порядка $\sim e-15$. В данном примере использовался p-МКНК₁ с представлением приближенного решения в виде (2.3) при расстановке точек коллокации, изображенной на фиг. 3а. В качестве весовых множителей были взяты $k_c = 10^{-3}$, $k_{b_0} = k_{b_1} = 1$.

3.4. Расчет изгиба шарнирно-закрепленной прямоугольной пластины

Прогиб изотропной упругой пластины в рамках классической теории Кирхгофа–Лява (см. [29]) определяется из решения уравнения (1.1), где $w = u$ – прогиб срединной поверхности пластины, $f = q/D$, q – интенсивность нагрузки, $D = Eh^3/(12(1 - \nu^2))$ – жесткость пластины при цилиндрическом изгибе, E, ν – модуль Юнга и коэффициент Пуассона изотропного мате-

Таблица 7. Результаты численных экспериментов в прямоугольной области (пример п. 3.4)

$N_1 \times N_2$	Операция продолжения			Без операции продолжения		
	$\ E_r\ _\infty$	d_{coeff}	N_{iter}	$\ E_r\ _\infty$	d_{coeff}	N_{iter}
h-МКНК, реш. (2.3) при $K_1 = K_2 = 7$						
8×8	4.76e-06	1.68e-06	90 (41)	4.76e-06	2.73e+3	166
16×16	1.11e-07	3.03e-08	142 (81)	1.11e-07	8.00e+3	329
32×32	1.91e-08	4.23e-10	218 (161)	1.91e-08	1.95e+3	682
h-МКНК, реш. (2.4) при $K = 6$						
8×8	6.91e-05	1.06e-05	99 (41)	6.91e-05	1.55e+2	160
16×16	4.14e-06	2.27e-07	164 (58)	4.14e-06	9.42e+2	322
32×32	2.56e-07	8.18e-09	343 (164)	2.56e-07	5.93e+3	688

риала пластины, h — толщина пластины. Рассмотрим следующие краевые условия закрепления, которые могут быть заданы на каждой части границ γ_i , $i = \overline{0, N_\gamma}$, пластины:

$$w = 0, \quad \frac{\partial w}{\partial n} = 0 \text{ — защемленный край,} \quad (3.5)$$

$$w = 0, \quad M_n = 0 \text{ — шарнирно-закрепленный край,} \quad (3.6)$$

$$M_n = 0, \quad V_n = Q_n - \frac{\partial M_{n\tau}}{\partial \tau} = 0 \text{ — свободный край,} \quad (3.7)$$

где $\mathbf{n}$ и $\boldsymbol{\tau}$ — внешняя нормаль и касательный вектор к границам области γ_i , $i = \overline{0, N_\gamma}$, M_n — изгибающий момент, V_n — обобщенная поперечная сила, Q_n — перерезывающая сила, $M_{n\tau}$ — крутящий момент (см. [29]). Здесь

$$\begin{aligned} M_n &= M_{x_1} \cos^2 \alpha + M_{x_2} \sin^2 \alpha - 2M_{x_1 x_2} \sin \alpha \cos \alpha, \\ M_{n\tau} &= M_{x_1 x_2} (\cos^2 \alpha - \sin^2 \alpha) + (M_{x_1} - M_{x_2}) \sin \alpha \cos \alpha, \\ Q_n &= Q_{x_1} \cos \alpha + Q_{x_2} \sin \alpha, \\ M_{x_1} &= -D \left(\frac{\partial^2 w}{\partial x_1^2} + \nu \frac{\partial^2 w}{\partial x_2^2} \right), \quad M_{x_2} = -D \left(\frac{\partial^2 w}{\partial x_2^2} + \nu \frac{\partial^2 w}{\partial x_1^2} \right), \quad M_{x_1 x_2} = D(1 - \nu) \frac{\partial^2 w}{\partial x_1 \partial x_2}, \\ Q_{x_1} &= -D \frac{\partial}{\partial x_1} (\Delta w), \quad Q_{x_2} = -D \frac{\partial}{\partial x_2} (\Delta w), \end{aligned}$$

где α — угол между вектором нормали и осью x_1 , отсчитываемый от x_1 против хода часовой стрелки, Δ — оператор Лапласа.

Пусть на шарнирно закрепленную (3.6) по всему контуру прямоугольную пластину размера $d_1 \times d_2$ действует синусоидальная нагрузка $q = 10^5 \sin(\pi x_1/d_1) \sin(\pi x_2/d_2)$ [Па]. В этом случае существует точное аналитическое решение краевой задачи (1.1), (3.6) (см. [29])

$$w(x_1, x_2) = \frac{q d_1^4 d_2^4}{\pi^4 D (d_1^2 + d_2^2)^2}. \quad (3.8)$$

В табл. 7 приведены результаты численных экспериментов, полученных h-МКНК с использованием операции продолжения на многосеточном комплексе. В четвертом столбце в скобках указано количество итераций при комбинированном применении операции продолжения и ускорения по Крылову. Здесь в расчетах взяты следующие значения параметров: $d_1 = 10$ м, $d_2 = 10$ м, $h = 0.1$ м, $E = 200$ ГПа, $\nu = 0.28$, $k_c = k_{m_0} = k_{m_2} = k_{b_0} = k_{b_2} = k_{m_1} = k_{m_3} = 1$, $\epsilon = 10^{-14}$ в h-МКНК и p-МКНК_h. Здесь k_{b_2} — весовой множитель условия $M_n = g_2(x_1, x_2)$ в переопределен-

Таблица 8. Результаты численных экспериментов в нерегулярной области (пример п. 3.5)

h-МКНК, решение (2.3) при $K_1 = K_2 = 5$			p-МКНК _h , решение (2.3)		p-МКНК ₁ , решение (2.3)	
$N_1 \times N_2$	$\ E_r\ _\infty$	R	$K_1 = K_2$	$\ E_r\ _\infty$	$K_1 = K_2$	$\ E_r\ _\infty$
12×12	3.92e-04	—	7	1.06e-05	13	7.88e-10
24×24	5.02e-05	2.96	8	2.04e-06	15	4.98e-12
48×48	5.63e-06	3.15	9	1.85e-08	17	3.24e-13
96×96	1.45e-06	1.95	10	7.49e-09	19	5.73e-14

ной СЛАУ приближенной задачи, d_{coeff} — максимальная разница между коэффициентами приближенного решения после первой итерации и коэффициентами начального приближения. Расположение точек коллокации в каждой ячейке соответствовало расположению точек, изображенных на фиг. 3б. Из табл. 7 следует, что применение операции продолжения позволяет сократить количество итераций приблизительно в два раза.

3.5. Краевая задача с граничными условиями, содержащими производные второго порядка

Рассмотрим в нерегулярной области (фиг. 1) краевую задачу для бигармонического уравнения с тестовым решением $u = \sin(\pi x_1/d_1) \sin(\pi x_2/d_2)$, где в численных экспериментах заданы на соответствующих границах значения функции u и действующего на нее дифференциального оператора $M_n = g_{2i}(x_1, x_2)$, $i = 0, 3$, вычисленные из точного решения. Внешней границей γ_0 многосвязной области является плоская кривая “улитка Паскаля” (фиг. 1), заданная соотношениями

$$\begin{aligned} x_1 &= 7 - 4 \sin(t) + \frac{43}{25} \sin(2t), \\ x_2 &= 7.8 + 4 \cos(t) - \frac{43}{25} \cos(2t), \end{aligned}$$

уравнения отверстий $(x_1 - 5.5)^2 + (x_2 - 7)^2 \leq 0.6^2$, $(x_1 - 8.5)^2 + (x_2 - 7.8)^2 \leq 0.5^2$, $(x_1 - 7)^2 + (x_2 - 4.5)^2 \leq 0.4^2$. В табл. 8 приведены результаты численных экспериментов, где в h-МКНК приближенное решение искалось в виде (2.3). Расположение точек коллокации в каждой ячейке соответствовало расположению точек, изображенных на фиг. 3а, т.е. в корнях ортогональных полиномов Чебышёва. В качестве точек согласования брались точки с координатами $(\pm 1, \alpha_{i_2}^2)$, $(\alpha_{i_1}^1, \pm 1)$, $i_1 = \overline{1, K_1}$, $i_2 = \overline{1, K_2}$, $\alpha_{i_1}^1$ и $\alpha_{i_2}^2$ — корни многочленов Чебышёва I рода степеней K_1 и K_2 соответственно. В расчетах взяты следующие значения параметров: $d_1 = 10.6667$, $d_2 = 9.64107$, $k_c = k_{m_0} = k_{m_2} = k_{b_0} = k_{b_2} = 1$, $k_{m_1} = 0.12$, $k_{m_3} = 0.125$, $\epsilon = 10^{-12}$ в h-МКНК и p-МКНК_h, $k_c = 10^{-3}$, $k_{b_0} = k_{b_2} = 1$ в p-МКНК₁.

3.6. Краевая задача с различными граничными условиями

Рассмотрим в нерегулярной области (фиг. 4в) краевую задачу для бигармонического уравнения с тестовым решением $u = \sin(\pi x_1/d_1) \sin(\pi x_2/d_2)$, сформулированную в [8]. На дуге окружности заданы значения действующих на функцию u дифференциальных операторов M_n и V_n , на двух прилежащих к дуге прямолинейных краях заданы значения функции u и действующего на нее дифференциального оператора M_n , на оставшемся прямолинейном крае заданы значения функции u и $\frac{\partial u}{\partial n}$, вычисленные из точного решения.

В табл. 9 приведены результаты численного эксперимента в случае применения p-МКНК₁ с представлением приближенного решения в виде (2.3) при расстановке точек коллокации по схеме, изображенной на фиг. 3б. В этой задаче в краевые условия входят производные второго и третьего порядков. В данной работе использованы следующие значения: $k_{b_0} = k_{b_1} = 1$, k_c , k_{b_2} и k_{b_3} ва-

Таблица 9. Результаты численных экспериментов (пример 3.6)

Shao и др., см. [8]					
K_1-1, K_2-1	4	6	8	10	12
$\ E_r\ _2$	1.75e-02	1.47e-04	2.61e-06	1.29e-07	7.10e-10
p-МКНК ₁ (фиг. 3б), решение (2.3)					
K_1-1, K_2-1	4	6	8	10	12
$\ E_r\ _2$	2.80e-02	1.24e-04	3.09e-06	4.74e-08	2.99e-10

Таблица 10. Результаты численных экспериментов (пример п. 3.7)

$r_0 = r_1/5$		$r_0 = r_1/50$		$r_0 = r_1/500$		$r_0 = r_1/5000$	
h-MKHK, $K = 4$							
$N_1 \times N_2$	$\ E_r\ _\infty$	$N_1 \times N_2$	$\ E_r\ _\infty$	$N_1 \times N_2$	$\ E_r\ _\infty$	$N_1 \times N_2$	$\ E_r\ _\infty$
8×8	3.60e−1	8×8	1.79e−1	8×8	2.78e−1	8×8	2.92e−1
16×16	6.40e−2	16×16	6.93e−2	16×16	7.60e−2	16×16	9.25e−2
32×32	5.20e−3	32×32	3.96e−2	32×32	1.72e−2	32×32	2.77e−2
64×64	1.23e−3	64×64	7.45e−3	64×64	4.24e−3	64×64	5.22e−3
p-MKHK _h , $N_1 \times N_2 = 24 \times 24$, $r_0 = r_1/5$							
K	$\ E_r\ _\infty$	K	$\ E_r\ _\infty$	K	$\ E_r\ _\infty$	K	$\ E_r\ _\infty$
4	1.37e−2	6	2.50e−3	8	2.93e−4	10	2.72e−5

рыировались в диапазоне $1e-3 \sim 1e-2$ при фиксированной степени полинома. Здесь k_{b_3} – весовой множитель перед условием $V_n = g_2(x_1, x_2)$.

3.7. Задача с особенностью. Симметричный изгиб кольцевой пластины

Рассмотрим задачу об изгибе равномерной поперечной нагрузкой q круглой пластины радиуса r_1 , имеющей центральное круглое отверстие радиуса r_0 . Поскольку нагрузка распределена симметрично относительно центра, то прогиб w в полярных координатах имеет вид (см. [29])

$$w = \frac{qr^4}{64D} + C_1 r^2 (\log r - 1) + C_2 \frac{r^2}{4} + C_3 \log r + C_4, \quad (3.9)$$

где $r = \sqrt{x_1^2 + x_2^2}$ – полярный радиус, $C_1, \dots, C_4$ – неизвестные константы, которые определяются из четырех граничных условий.

Рассмотрим сначала кольцевую пластину с защемленными (3.5) внешним и внутренним контурами. В этом случае граничные условия имеют вид

$$w|_{r=r_1} = \frac{\partial w}{\partial n}|_{r=r_1} = w|_{r=r_0} = \frac{\partial w}{\partial n}|_{r=r_0} = 0.$$

Пусть $q = 10^6$ Па, $\nu = 0.28$, $E = 200$ ГПа, $h = 0.1$ м, $r_1 = 5$ м, $r_0 = r_1/k$ м, k – коэффициент отношения, $d_1 = 10$ м, $d_2 = 10$ м. Все значения весовых множителей были взяты равными единице. В табл. 10 приведены результаты численных экспериментов в случае применения h-МКНК при $K = 4$ и p-МКНК_h на сетке 24×24 с представлением приближенного решения в виде (2.4). Отметим, что в случае достаточно малого отверстия все точки коллокации и согласования располагаются вне отверстия (см. п. 2.2). Для учета граничных условий в такой ячейке на ее внешней стороне располагалась одна точка записи краевых условий.

В предложенном здесь p-МКНК₁ приближенные решения (2.3) или (2.4) представляют собой равномерное приближение искомого решения в прямоугольнике, включающем всю исходную

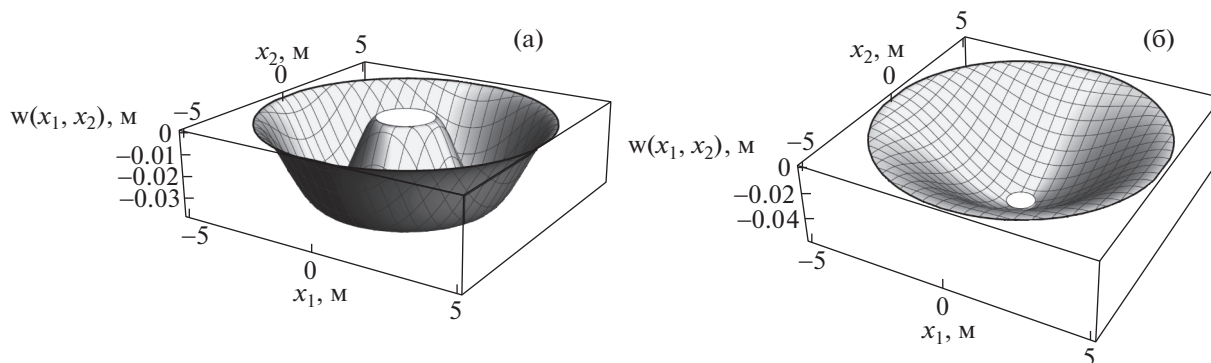
Таблица 11. Результаты численных экспериментов (пример п. 3.8)

$\alpha = 5/3$									
$N_1 \times N_2$	h-МКНК при $K = 4$			h-МКНК при $K = 6$				МКЭ [7]	адапт. МКЭ [7]
	DoF	$\ E_a\ _{L_2}$	R	DoF	$\ E_a\ _{L_2}$	R	DoF	$\ E_a\ _{L_2}$	$\ E_a\ _{L_2}$
4×4	180	6.51e-3	—	336	2.82e-3	—	833	7.15e-3	8.88e-4
8×8	720	3.64e-3	0.84	1344	8.21e-4	1.78	3201	2.72e-3	2.16e-4
16×16	2880	9.93e-4	1.87	5376	1.67e-4	2.29	12545	1.08e-3	5.34e-5
32×32	11 520	3.76e-4	1.40	21 504	1.17e-4	0.51	49665	4.40e-4	1.32e-5
64×64	46080	1.74e-4	1.11	86016	7.52e-5	0.64	197633	1.82e-4	3.28e-6

Таблица 12. Результаты численных экспериментов (пример п. 3.8) при различных α

$N_1 \times N_2$	h-МКНК при $K = 4$			h-МКНК при $K = 6$			h-МКНК при $K = 8$		
	DoF	$\ E_a\ _{L_2}$	R	DoF	$\ E_a\ _{L_2}$	R	DoF	$\ E_a\ _{L_2}$	R
$\alpha = 7/3$									
4×4	180	6.55e-3	—	336	1.42e-3	—	720	1.78e-3	—
8×8	720	1.82e-3	1.85	1344	4.89e-4	1.53	2880	7.45e-4	1.26
16×16	2880	2.77e-4	2.72	5376	9.41e-5	2.37	11 520	1.91e-4	1.96
32×32	11 520	4.23e-5	2.71	21 504	1.52e-5	2.63	46080	5.14e-5	1.89
64×64	46080	6.83e-6	2.63	86016	2.42e-6	2.65	184 320	1.43e-5	1.86
$\alpha = 10/3$									
4×4	180	7.49e-3	—	336	7.70e-4	—	720	2.61e-4	—
8×8	720	6.64e-4	3.50	1344	1.69e-4	2.19	2880	3.62e-5	2.85
16×16	2880	3.49e-5	4.25	5376	2.14e-5	2.98	11 520	5.33e-6	2.76
32×32	11 520	5.88e-7	5.89	21 504	2.39e-6	3.16	46080	6.29e-7	3.08
64×64	46080	4.07e-7	0.53	86016	2.62e-7	3.19	184 320	1.04e-7	2.60
$\alpha = 13/3$									
4×4	180	7.87e-3	—	336	3.63e-4	—	720	7.28e-5	—
8×8	720	7.47e-4	3.40	1344	2.59e-5	3.81	2880	8.91e-6	3.03
16×16	2880	3.46e-5	4.43	5376	1.50e-6	4.11	11 520	5.89e-7	3.92
32×32	11 520	1.62e-6	4.42	21 504	8.05e-8	4.22	46080	3.20e-8	4.18
64×64	46080	1.23e-7	3.72	86016	4.48e-9	4.17	184 320	1.84e-9	4.12
$\alpha = 16/3$									
4×4	180	5.45e-2	—	336	6.16e-4	—	720	7.13e-5	—
8×8	720	7.76e-3	2.81	1344	1.91e-5	5.01	2880	4.12e-6	4.11
16×16	2880	3.39e-4	4.51	5376	9.69e-7	4.30	11 520	1.11e-7	5.21
32×32	11 520	1.32e-5	4.68	21 504	5.76e-8	4.06	46080	2.20e-9	5.65
64×64	46080	6.54e-7	4.33	86016	3.55e-9	4.02	184 320	3.21e-11	6.10

расчетную область и все отверстия в случае многосвязной области (в нашем примере $x_1^2 + x_2^2 \leq r_0^2$). Точное решение (3.9) данной задачи в точке $(0, 0)$ имеет особенность в виде разрыва II рода из-за наличия в нем $\log r$. Поэтому решение (3.9) не является непрерывным в данном прямоугольнике и его не удастся приблизить с помощью р-МКНК₁ (2.3) или (2.4). Однако, зная вид особенности решения и асимптотику его поведения, можно модифицировать базис используемого линейного пространства в р-МКНК₁. С другой стороны, как это видно из табл. 10, при-



Фиг. 5. Величина прогиба w кольцевой пластины с защемленным внешним контуром и защемленным (а) (свободным (б)) внутренним контуром.

менение h -МКНК и p -МКНК $_h$ позволяет построить сходящееся приближенное решение в задаче с особенностью без изменения базиса применяемого пространства. Эта сходимость достигается благодаря использованию кусочно-полиномиального представления решения, что подтверждает универсальность и гибкость h -МКНК и p -МКНК $_h$ при решении краевых задач для УЧП в общем случае.

Для повышения точности аппроксимации и сходимости p -МКНК $_1$ решения рассматриваемой задачи с особенностью необходимо добавить в его представление (2.3) или (2.4) слагаемые вида $\tilde{c} \log r$ и $\tilde{c} r^2 \log r$, где $\tilde{c}$ и $\tilde{c}$ — неизвестные константы. В этом случае уже при $K_1 = K_2 = 5$ или $K = 4$ значение относительной погрешности $\|E_r\|_\infty$ достигается порядка $e-13$ при любых соотношениях r_1/r_2 . На фиг. 5а показана форма прогиба кольцевой пластины с защемленными (3.5) внешним и внутренним контурами с $r_0 = 1$ м.

Рассмотрим теперь кольцевую пластину с защемленным внешним (2.5) и со свободным (3.7) внутренним контурами. На фиг. 5б показана форма прогиба при $r_1 = 5$ м, $r_0 = 0.5$ м, $q = 10^5$ Па. Значение $\|E_r\|_\infty$ здесь достигается также порядка $e-13$ при применении модифицированного p -МКНК $_1$. При применении МКНК с полиномиальным базисом при $K = 10$ на сетке 12×12 значение $\|E_r\|_\infty$ имеет величину порядка $e-3$.

3.8. Задача с особенностью в виде обращения в бесконечность производных решения

Рассмотрим задачу (1.1)–(1.3) в L-образной области $\Omega = (-1, 1) \times (-1, 1) \setminus (0, 1) \times (-1, 0)$ с тестовым решением $u = r^\alpha \sin(\alpha\theta)$, где (r, θ) — полярные координаты, $\alpha > 0$ — const. Здесь $f = 0$, на границе заданы значения u и $\partial u / \partial n$, вычисленные из точного решения.

В табл. 11 приведены результаты численного эксперимента с $\alpha = 53$. В этом случае вторые производные в точке $(0, 0)$ обращаются в бесконечность. Известно, что решение уравнений эллиптического типа в области зависит от его значений на границе. Отсутствие гладкости даже в одной точке на γ_0 может сказываться на характере поведения решения в Ω (см. [30]). Здесь использовался МКНК (2.4) при расстановке точек коллокации по схеме, изображенной на фиг. 2. В табл. 11 DoF обозначает число степеней свободы, что в МКНК эквивалентно общему количеству неизвестных коэффициентов. Отметим, что в МКЭ при решении этой задачи использовалась триангуляция с адаптацией к особенности и без нее (см. табл. 11, столбцы 8–10) (см. [7]). В МКНК были взяты следующие значения весовых множителей $k_c = k_{m_0} = k_{m_2} = k_{b_0} = k_{b_1} = 1$, $k_{m_1} = 0.12$, $k_{m_3} = 0.125$ и $\epsilon = 10^{-12}$.

Из представленных результатов следует, что в МКЭ прежде всего за счет адаптации удается построить более точное решение по сравнению с МКНК и МКЭ с использованием равномерной сетки (более подробно о порядках сходимости при решении рассматриваемой задачи МКЭ см. в [7]). Такая возможность повышения точности решения задач с особенностями в МКНК показана в [20]. Однако отметим, что МКНК здесь применялся к решению краевой за-

дачи для уравнения высокого порядка в исходном виде (1.1)–(1.3), в то время как в МКЭ дополнительно приходится работать со слабыми постановками. При этом надо иметь в виду, что многими численными методами при решении уравнений низкого порядка точность и порядок сходимости будут выше при аналогичном способе аппроксимации уравнений высокого порядка, например, одинаковыми полиномами или конечно-разностными схемами на одинаковых шаблонах. Это было продемонстрировано при решении задачи изгиба шарнирно закрепленной прямоугольной пластины МКНК при $K = 4$ (см. [23]). В этом случае решение краевой задачи для бигармонического уравнения можно легко свести к последовательному решению двух уравнений Пуассона (см. [29]), которые имеют меньший порядок, что приводит к лучше обусловленной задаче по сравнению с исходной. При сравнении с решением бигармонического уравнения в исходном виде этот способ позволяет достичь существенно лучшей точности и более высокого порядка сходимости.

Кроме того, повышение гладкости решения задачи в МКЭ не увеличит порядка сходимости в случае применения линейных элементов (см. [7]) в отличие от МКНК, где он возрастет и, следовательно, погрешность будет уменьшаться быстрее при увеличении степеней базисных элементов полиномиального пространства для приближения искомого решения в ячейке и уменьшении шага сетки. Заметим также, что в МКНК достаточно просто увеличить степень аппроксимирующих полиномов по методике, предложенной в этой и предыдущих работах. В табл. 12 приведены результаты численных экспериментов с различными значениями $\alpha = 7/3, 10/3, 13/3, 16/3$, при этом в точке $(0,0)$ в бесконечность обращаются производные третьего, четвертого, пятого и шестого порядков соответственно. При увеличении гладкости решения его точность и порядок сходимости R возрастают, но R в целом не превышает порядка гладкости решения.

4. ЗАКЛЮЧЕНИЕ

Разработаны новые h -, p - и hp -МКНК повышенной точности для численного решения неоднородных бигармонических уравнений в нерегулярных областях. Возможности предложенных здесь вариантов метода проверены на решении тестовых задач путем вычисления погрешности и порядка сходимости приближенного решения на последовательности сеток. Показано, что новые варианты МКНК во многих случаях не уступают по точности высокоточным известным результатам других авторов, полученным другими методами.

Успешное применение разработанных p -МКНК _{h} и p -МКНК₁ позволило обеспечить экспоненциальный порядок сходимости, поскольку в рассмотренных примерах пп. 3.1–3.6 решения были достаточно гладкими функциями. Однако решения реальных, практически важных задач (пример п. 3.7) зачастую не являются достаточно гладкими (см. также пример п. 3.8). Вследствие этого методы, основанные на полиномиальной аппроксимации высокого порядка, в том числе и предложенный здесь p -МКНК₁, становятся менее эффективными и гибкими, например, по сравнению с h -МКНК и p -МКНК _{h} .

Авторы выражают признательность С.В. Идимешеву, В.И. Исаеву и В.А. Колотилу за полезное обсуждение результатов работы.

СПИСОК ЛИТЕРАТУРЫ

1. Mayo A. The fast solution of Poisson's and the biharmonic equations on irregular regions // SIAM J. Numer. Anal. 1984. V. 21. № 2. P. 265–299.
2. Lai M.C., Liu H.C. Fast direct solver for the biharmonic equation on a disk and its application to incompressible flows // App. Math. Comput. 2005. V. 164. № 3. P. 679–695.
3. Chen G., Li Z., Lin P. A fast finite difference method for biharmonic equations on irregular domains and its application to an incompressible Stokes flow // Adv. Comput. Math. 2007. V. 29. № 2. P. 113–133.
4. Ben-Artzi M., Chorev I., Croisille J.P., Fishelov D. A compact difference scheme for the biharmonic equation in planar irregular domains // SIAM J. Numer. Anal. 2009. V. 47. № 4. P. 3087–3108.
5. Davini C., Pitacco I. An unconstrained mixed method for the biharmonic problem // SIAM J. Numer. Anal. 2000. V. 38. № 3. P. 820–836.
6. Eymard R., Gallouët T., Herbin R., Linke A. Finite volume schemes for the biharmonic problem on general meshes // Math. Comp. 2012. V. 812. № 280. P. 2019–2048.
7. Guo H., Zhang Z., Zou Q. A C^0 Linear Finite Element Method for Biharmonic Problems // J. Sci. Comput. 2018. V. 74. P. 1397–1422.

8. *Shao W., Wu X., Chen S.* Chebyshev tau meshless method based on the integration-differentiation for biharmonic-type equations on irregular domain // *Engin. Anal. Bound. Elem.* 2012. V. 36. № 12. P. 1787–1798.
9. *Jiang Y., Wang B., Yuesheng X.* A fast Fourier–Galerkin method solving a boundary integral equation for the biharmonic equation // *SIAM J. Numer. Anal.* 2014. V. 52. № 5. P. 2530–2554.
10. *Mai-Duy N., See H., Tran-Cong T.* A spectral collocation technique based on integrated Chebyshev polynomials for biharmonic problems in irregular domains // *Appl. Math. Model.* 2009. V. 33. № 1. P. 284–299.
11. *Shao W., Wu X.* An effective Chebyshev tau meshless domain decomposition method based on the integration-differentiation for solving fourth order equations // *Appl. Math. Model.* 2015. V. 39. № 9. P. 2554–2569.
12. *Сорокин С.Б.* Переобусловливание при численном решении задачи Дирихле для бигармонического уравнения // *Сиб. журн. вычисл. матем.* 2011. Т. 14. № 2. С. 205–213.
13. *Greenbaum A., Greengard L., Mayo A.* On the numerical solution of the biharmonic equation in the plane // *Physica D.* 1992. V. 60. № 1–4. P. 216–225.
14. *Bialecki B.* A fast solver for the orthogonal spline collocation solution of the biharmonic Dirichlet problem on rectangles // *J. Comput. Phys.* 2003. V. 191. № 2. P. 601–621.
15. *Gomez-Polanco A., Guevara-Jordan J.M., Molina B.* A mimetic iterative scheme for solving biharmonic equations // *Math. Comput. Model.* 2013. V. 57. № 9–10. P. 2132–2139.
16. *Беляев В.А., Шапеев В.П.* Варианты метода коллокации и наименьших невязок для решения задач математической физики в трапециевидных областях // *Вычисл. технологии.* 2017. Т. 22. № 4. С. 22–42.
17. *Шапеев В.П., Беляев В.А.* Решение с повышенной точностью бигармонического уравнения в нерегулярных областях методом коллокации и наименьших квадратов // *Вычисл. методы и программирование.* 2018. Т. 19. № 4. С. 340–355.
18. *Shapeev V.P., Belyaev V.A., Golushko S.K., Idimeshev S.V.* New possibilities and applications of the least squares collocation method // *EPJ Web of Conferences.* 2018. V. 173. № 01012. P. 01012-1–01012-8.
19. *Идимешев С.В.* Модифицированный метод коллокаций и наименьших невязок и его приложение в механике многослойных композитных балок и пластин: Дис. канд. физ.-мат. наук. Новосибирск: Ин-т вычисл. технологий СО РАН, 2016.
20. *Исаев В.И., Шапеев В.П.* Варианты метода коллокаций и наименьших квадратов повышенной точности для численного решения уравнений Навье–Стокса // *Ж. вычисл. матем. и матем. физ.* 2010. Т. 50. № 10. С. 1758–1770.
21. *Исаев В.И., Шапеев В.П., Еремин С.А.* Исследование свойств метода коллокации и наименьших квадратов решения краевых задач для уравнения Пуассона и уравнений Навье–Стокса // *Вычисл. технологии.* 2007. Т. 12. № 3. С. 53–70.
22. *Голушко С.К., Идимешев С.В., Шапеев В.П.* Метод коллокаций и наименьших невязок в приложении к задачам механики изотропных пластин // *Вычисл. технологии.* 2013. Т. 18. № 6. С. 31–43.
23. *Шапеев В.П., Брындин Л.С., Беляев В.А.* Решение эллиптических уравнений в полигональных областях методом коллокации и наименьших квадратов // *Вестник ЮУрГУ ММП.* 2019. Т. 22. № 3. С. 140–152.
24. *Vorozhtsov E.V., Shapeev V.P.* On the efficiency of combining different methods for acceleration of iterations at the solution of PDEs by the method of collocations and least residuals // *Appl. Math. Comput.* 2019. V. 363. № 124644. P. 1–19.
25. *Лаевский Ю.М.* Метод конечных элементов (основы теории, задачи). Новосибирск: Новосиб. гос. ун-т, 1999. 166 с.
26. *Федоренко Р.П.* Введение в вычислительную физику. М.: МФТИ, 1994. 527 с.
27. *Ramšak M., Škerget L.* A subdomain boundary element method for high-Reynolds laminar flow using stream function–vorticity formulation // *Int. J. Numer. Meth. Fluids.* 2004. V. 46. № 8. P. 815–847.
28. *Saad Y.* Numerical methods for large eigenvalue problems. Manchester Univ. Press, 1992. 346 p.
29. *Тимошенко С.П., Войновский-Кригер С.* Пластины и оболочки. М.: Физматгиз, 1963. 636 с.
30. *Шапеев В.П., Шапеев А.В.* Решение эллиптических задач с особенностями по схемам высокого порядка аппроксимации // *Вычисл. технологии.* 2006. Т. 11. Часть 2. Специальный выпуск. С. 84–91.

**ОБЩИЕ
ЧИСЛЕННЫЕ МЕТОДЫ**

УДК 517.983

**ОБЩАЯ СТРУКТУРА РЕДУЦИРОВАННЫХ БАЗИСОВ ДЛЯ ЗАДАЧ
АГРЕГАЦИОННОЙ КИНЕТИКИ РАЗЛИЧНОЙ РАЗМЕРНОСТИ¹⁾**© 2022 г. С. А. Матвеев^{1,2}, А. П. Смирнов¹, И. В. Тимохин^{1,2,3,*}, Е. Е. Тыртышников^{1,2,3}¹ 119991 Москва, Ленинские горы, стр. 52, МГУ им. М.В. Ломоносова, факультет ВМК, Россия² 119333 Москва, ул. Губкина, 8, Институт вычислительной математики им. Г.И. Марчука
Российской академии наук, Россия³ 119991 Москва, Ленинские горы, МГУ, Московский центр фундаментальной и прикладной математики,
Москва, Россия

*e-mail: m@ivan.timokhin.name

Поступила в редакцию 04.10.2021 г.
Переработанный вариант 04.10.2021 г.
Принята к публикации 16.12.2021 г.

Данная работа посвящена исследованию структуры линейных пространств, позволяющих приближать решения систем дифференциальных уравнений, возникающих в задачах агрегационной кинетики с источником и стоком частиц, на больших интервалах времени с высокой точностью. Основным параметром для семейства рассматриваемых моделей агрегационной кинетики является их размерность N , соответствующая максимальному допустимому размеру частицы в системе. При больших значениях предельного допустимого размера частицы N исследование таких задач оказывается затруднительным, так как начинает требовать значительных вычислительных ресурсов. Для ускорения расчетов мы используем идеологию редукции размерности при помощи поиска базиса искомого маломерного пространства методом снимков. В результате использования такого подхода можно установить существование такого пространства, построить его базис и свести исходную нелинейную задачу высокой размерности к задаче существенно меньшей размерности, позволяющей организовать вычисления более эффективно без существенной потери качества численных решений. Главным результатом данной работы является найденная возможность использования базиса задач большей размерности N для решения задач меньшей размерности $M < N$ без подготовки и построения базиса для размерности M . В численных экспериментах мы демонстрируем, что достаточно просто взять первые $M < N$ компонент векторов базиса задачи размерности N и использовать полученную систему в качестве базиса для новой задачи размерности M . С использованием такого базиса нам удается получать превосходную точность численных решений редуцированной задачи размерности M , хотя сами по себе решения задач размерности M и N не обязаны быть близки. Дополнительно в работе предлагается несколько оценок для норм решений уравнений необратимой агрегации с источником. Из оценок следует, что прямая подстановка векторов базиса N в задачи размерности $M > N$ не приводит к успеху, а расширение базиса для задач большей размерности требует дополнительных усилий, что согласуется с результатами экспериментов. Библ. 28. Фиг. 3.

Ключевые слова: уравнение Смолуховского, редукция модели, агрегация, метод снимков.**DOI:** 10.31857/S004446692204010X**1. ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ**

Процессы агрегации и фрагментации играют важную роль в огромном наборе природных явлений от роста полимерных структур [1]–[3], коагуляции тромбоцитов в крови [4], динамики аэрозолей и примесей в атмосфере [5] до образования звезд и планет [6]–[8]. В случае наличия

¹⁾ Работа выполнена при финансовой поддержке И.В. Тимохина Московским центром фундаментальной и прикладной математики (соглашение с Минобрнауки РФ, проект 075-15-2019-1624); С.А. Матвеева и Е.Е. Тыртышникова РФФИ (проект 19-11-00338).

источника мономеров классические уравнения Смолуховского принимают вид формально бесконечной системы дифференциальных уравнений:

$$\frac{dn_k}{dt} = J\delta_{k1} + \frac{1}{2} \sum_{i+j=k} C_{ij}n_in_j - n_k \sum_{i=1} C_{ik}n_i, \quad k = 1, \dots, \infty, \quad (1.1)$$

где

- n_k — концентрация частиц массы k ;
- C_{ij} — “ядро”, характеризующее вероятность столкновения частиц размеров i и j ;
- J — мощность источников частиц размера 1.

Искомые величинами здесь выступают концентрации n_k частиц размера k на единицу объема среды. При данной записи мы предполагаем, что рассматривается пространственно однородная система, частицы сталкиваются парно и в результате столкновений частиц размеров i и j возникает новая частица размера $i + j$, т.е. выполняется закон сохранения массы. Достаточные условия на коэффициенты C_{ij} для выполнения закона сохранения массы можно найти, например, в работах [9], [10]. Асимптотические свойства стационарных решений уравнений Смолуховского с источником мономеров хорошо изучены в случае коэффициентов

$$C_{ij} = i^v j^\mu + i^\mu j^v \quad (1.2)$$

в работе [11]. В данной работе мы сконцентрируемся на параметрическом семействе моделей агрегации с использованием конечного количества уравнений N , считая, что образующиеся агрегаты большего размера выпадают в осадок под действием внешних сил (например, гравитации), а в качестве основного ядра коагуляции возьмем частный случай (1.2):

$$C_{ij} = i^a j^{-a} + i^{-a} j^a$$

с монодисперсными начальными условиями $n_k(0) = \delta_{k1}$.

В таких условиях исходную систему (1.1) естественно заменить схожей конечной системой

$$\frac{dn_k}{dt} = J\delta_{k1} + \frac{1}{2} \sum_{i+j=k} C_{ij}n_in_j - n_k \sum_{i=1}^N C_{ik}n_i, \quad k = 1, \dots, N. \quad (1.3)$$

Данное семейство моделей вызывает высокий интерес исследователей и позволяет наблюдать ряд интересных эффектов при вариациях параметра размерности системы дифференциальных уравнений N : из асимптотически устойчивого стационарного положения система с ростом N переходит в режим стационарных колебаний по времени [12], [13], что может иметь значение для анализа популяционной динамики в биологии [14]. Свойства смены характера решений на данный момент окончательно не исследованы, но ряд результатов для более простых моделей [15]–[17] позволяет сформировать гипотезу, что колебания в данных системах возникают в результате бифуркации Андронова–Хопфа.

Основной целью данной работы является исследование взаимосвязи решений системы (1.3) при различных N , имея в виду возможность использования информации об одном из них для быстрого построения других. Этот вопрос мы исследуем в контексте идеи редукции модели — аппроксимации многомерной задачи в пространстве меньшей размерности с соответствующим сокращением вычислительных расходов. Конкретное воплощение этой общей идеи следует из работ [18], [19], в которых для уравнения Смолуховского типа (1.3) показана возможность построения маломерной аппроксимации целиком в содержащем решение пространстве малой размерности, без необходимости обращаться к полной системе даже для промежуточных вычислений.

В общем случае редукция модели охватывает широкий спектр различных вычислительных техник с разной степенью привязки к конкретным задачам [20]–[24]; предложенный нами метод (см. алгоритм 1) основан на идее POD (Proper Orthogonal Decomposition) [25], состоящей в простейшем варианте в построении базиса искомого подпространства малой размерности из старших сингулярных векторов матрицы, составленной из некоторых предполагаемых представителей искомого маломерного пространства, которых, следуя [22], мы набираем из “снимков” решения $n(t_k)$ в равноотстоявшие моменты времени. Построение базиса таким методом требует, в общем случае, знания решения заранее (в классическом POD система предполагается зависящей от дополнительных параметров, и “снимки” берутся из решений при нескольких характер-

ных значениях параметра; мы же заинтересованы в применении редукции к отдельно взятой системе). Для обхода этого ограничения в [18] предложен адаптивный алгоритм, строящий по начальному отрезку времени базис для решения в последующие моменты. Эффективность полученного метода проверена на задачах, имеющих циклическое по времени решение — на них показано ускорение вычислений в разы [19] — однако базовый факт наличия подпространства малой размерности, приближенно содержащего в себе решение, экспериментально подтверждается и в других случаях.

В то же время само построение базиса остается достаточно дорогостоящей процедурой, требующей к тому же предварительного построения значительной части искомого решения. Для пространственно неоднородных задач эту проблему можно частично обойти, если ограничиться редукцией по пространству и искать базис одновременно с решением; это было проделано в [26], и показало ускорение в десятки раз. Для пространственно однородной задачи мы предлагаем рассмотреть возможность переноса базиса между системами вида (1.3) с различными N . Для этого естественно задаться вопросом о связи решений систем разного размера, в том числе на достаточно больших промежутках времени (искомое линейное пространство определяется поведением решения в целом, не ограничиваясь каким-либо конечным временным интервалом). В этой связи мы

- экспериментально продемонстрируем существование существенных различий в долгосрочном поведении систем с различным N ;
- получим оценку на величину этого расхождения;
- покажем, что, тем не менее, решения задач с меньшим N с большой точностью лежат в линейной оболочке урезанных решений с большим N .

Алгоритм 1. Общая схема используемого алгоритма построения базиса редукции для (1.3).
 Подробное описание алгоритма см. в [18], [19]

Require: “Ширина” для “окон” $\tau > 0$, параметры точности $\varepsilon' > \varepsilon > 0$, $\delta > 0$

Ensure: V — искомый базис.

$V \leftarrow 0 \in \mathbb{R}^{N \times 0}$

$k \leftarrow 0$

repeat

Решить исходную систему (1.3) полностью на интервале $[k\tau, (k+1)\tau]$ любым способом.

$\hat{V} \leftarrow$ базис POD для $n(t)$ на $[k\tau, (k+1)\tau]$

$\Delta \leftarrow \|(I - VV^T)\hat{V}\|_2$

if $\Delta > \varepsilon'$ then

$U\Sigma W^T \leftarrow (V|\hat{V})$ {сингулярное разложение с сингулярными числами σ_i }

Найти $i : \sigma_i > \delta, \quad \sigma_{i+1} \leq \delta$

$V \leftarrow$ первые i столбцов U

end if

$k \leftarrow k + 1$

until $\Delta \leq \varepsilon$

Последнее означает, что, найдя однажды маломерное подпространство, содержащее решение системы большей размерности N , подпространства для систем меньшей размерности возможно получать совершенно механически с минимальными вычислительными затратами путем “усечения” уже построенного базиса. Полученные редукцией на такой “усеченный” базис решения, как мы демонстрируем в численных экспериментах, с высокой точностью совпадают с полными. Обратный процесс — расширение базиса — не столь прост. Как мы покажем, простое дополнение базиса нулями приводит к существенным отклонениям не только в обнуленных компонентах решения — что в такой ситуации ожидаемо — но и в компонентах, общих для обеих систем. Расши-

рение базиса, таким образом, требует дополнительной априорной информации о структуре решения.

Кроме того, мы получим некоторые общие оценки на отличие решений уравнений (1.3) и (1.1). Полученные оценки и используемая техника отчасти схожи с первыми главами [9], с тем отличием, что мы исследуем именно дискретную формулировку уравнения Смолуховского и фокусируемся больше на конкретных численных оценках, а не качественных результатах.

2. ТЕОРЕТИЧЕСКИЕ ОЦЕНКИ

В данном разделе мы сконцентрируемся на общих оценках отличия решений систем уравнений (1.3) в терминах нормы $\|\cdot\|_1$.

2.1. Решения конечных систем

Прежде всего получим априорную оценку на степень расхождения решений (1.3) при различных N . Для этого введем ряд обозначений:

- $\mathbf{n}(t)$ — решение системы вида (1.3) с M уравнениями; $\mathbf{n}(t) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^M$;
- $\tilde{\mathbf{n}}(t)$ — решение такой же системы, но уже с $N > M$ уравнениями; $\tilde{\mathbf{n}}(t) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^N$;
- $\tilde{\mathbf{n}}_{(1)}(t)$, $\tilde{\mathbf{n}}_{(2)}(t)$ — первые M и последние $N - M$ компонент $\tilde{\mathbf{n}}(t)$ соответственно; $\tilde{\mathbf{n}}_{(1)}(t) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^M$, $\tilde{\mathbf{n}}_{(2)}(t) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{N-M}$;
- C — матрица ядра $M \times M$;
- $\tilde{C}$ — матрица ядра $N \times N$; $\tilde{C} = \begin{pmatrix} C & \tilde{C}_{12} \\ \tilde{C}_{12}^T & \tilde{C}_{22} \end{pmatrix}$.

Кроме того, нам будет полезно ввести в рассмотрение билинейный аналог квадратичного оператора из правой части (1.3):

$$\mathcal{J}_M(x, y) = \frac{1}{2} \sum_{i+j=k} C_{ij} x_i y_j - \frac{1}{2} x_k \sum_{i=1}^M C_{ik} y_i - \frac{1}{2} y_k \sum_{i=1}^M C_{ik} x_i.$$

Тогда уравнения на $\mathbf{n}$ и $\tilde{\mathbf{n}}_{(1)}$ можно переписать в векторном виде:

$$\begin{aligned} \frac{d\mathbf{n}}{dt} &= \mathcal{J} + \mathcal{J}_M(\mathbf{n}, \mathbf{n}), \\ \frac{d\tilde{\mathbf{n}}_{(1)}}{dt} &= \mathcal{J} + \mathcal{J}_M(\tilde{\mathbf{n}}_{(1)}, \tilde{\mathbf{n}}_{(1)}) - \left(\tilde{n}_k \sum_{i=M+1}^N C_{ik} \tilde{n}_i \right)_{k=1}^M, \end{aligned}$$

где под $(f(k))_{k=1}^M$ понимается вектор из значений $f(k)$ при k от 1 до M .

Вычитая теперь одно уравнение из другого, с учетом билинейности и симметрии $\mathcal{J}_M$, получаем

$$\frac{d\mathbf{n}}{dt} - \frac{d\tilde{\mathbf{n}}_{(1)}}{dt} = \mathcal{J}_M(\mathbf{n} - \tilde{\mathbf{n}}_{(1)}, \mathbf{n} + \tilde{\mathbf{n}}_{(1)}) + \left(\tilde{n}_k \sum_{i=M+1}^N C_{ik} \tilde{n}_i \right)_{k=1}^M,$$

откуда уже легко оценить сверху модуль разности

$$\begin{aligned} \left| \frac{dn_k}{dt} - \frac{d\tilde{n}_k}{dt} \right| &\leq \frac{1}{2} \sum_{i+j=k} C_{ij} |n_i - \tilde{n}_i| (n_j + \tilde{n}_j) + \tilde{n}_k \sum_{i=M+1}^N C_{ik} \tilde{n}_i + \\ &+ \frac{1}{2} |n_k - \tilde{n}_k| \sum_{i=1}^M C_{ik} (n_i + \tilde{n}_i) + \frac{1}{2} (n_k + \tilde{n}_k) \sum_{i=1}^M C_{ik} |n_i - \tilde{n}_i|. \end{aligned}$$

Суммируя последнее выражение, получаем оценку на первую норму разности производных:

$$\begin{aligned} \left\| \frac{d\mathbf{n}}{dt} - \frac{d\tilde{\mathbf{n}}_{(1)}}{dt} \right\|_1 &\leq \frac{1}{2} \sum_{i+j \leq M} C_{ij} |n_i - \tilde{n}_i| (n_j + \tilde{n}_j) + \sum_{i,j=1}^M C_{ij} |n_i - \tilde{n}_i| (n_j + \tilde{n}_j) + \\ &+ \sum_{i=1}^N \sum_{j=M+1}^N C_{ij} \tilde{n}_i \tilde{n}_j \leq \frac{3}{2} \|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \|C(\mathbf{n} + \tilde{\mathbf{n}}_{(1)})\| + \tilde{\mathbf{n}}_{(1)}^T \tilde{C}_{12} \tilde{\mathbf{n}}_{(2)}. \end{aligned}$$

Последнее слагаемое в оценке представляет собой скалярное произведение и также легко оценивается:

$$\tilde{\mathbf{n}}_{(1)}^T \tilde{C}_{12} \tilde{\mathbf{n}}_{(2)} \leq \|\tilde{\mathbf{n}}_{(1)}\|_1 \|\tilde{C}_{12} \tilde{\mathbf{n}}_{(2)}\|_\infty.$$

Комбинируя полученные оценки, в области знакопостоянства всех элементов разности получаем для производной нормы

$$\frac{d}{dt} \|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \frac{3}{2} \|C(\mathbf{n} + \tilde{\mathbf{n}}_{(1)})\|_\infty \|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 + \|\tilde{\mathbf{n}}_{(1)}\|_1 \|\tilde{C}_{12} \tilde{\mathbf{n}}_{(2)}\|_\infty,$$

откуда, если элементы разности на отрезке $[t_0, t]$ меняют знак лишь конечное число раз,

$$\|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + \int_{t_0}^t \|\tilde{\mathbf{n}}_{(1)}(\tau)\|_1 \|\tilde{C}_{12} \tilde{\mathbf{n}}_{(2)}\|_\infty d\tau + \frac{3}{2} \int_{t_0}^t \|C(\mathbf{n} + \tilde{\mathbf{n}}_{(1)})\|_\infty \|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 d\tau,$$

а отсюда уже по лемме Гронуолла-Беллмана получаем

$$\|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \exp \left(\|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + \frac{3}{2} \int_{t_0}^t \|C(\mathbf{n} + \tilde{\mathbf{n}}_{(1)})\|_\infty d\tau \right) \int_{t_0}^t \|\tilde{\mathbf{n}}_{(1)}\|_1 \|\tilde{C}_{12} \tilde{\mathbf{n}}_{(2)}\|_\infty d\tau. \quad (2.1)$$

Для дальнейшего оценивания норм в правой части нам пригодится следующая

Лемма 1. В ранее введенных обозначениях,

$$\min \left(\|\mathbf{n}(0)\|_1, \sqrt{\frac{J}{\max_{1 \leq i, j \leq N} C_{ij}}} \right) \leq \|\mathbf{n}(t)\|_1 \leq \max \left(\|\mathbf{n}(0)\|_1, \sqrt{\frac{2J}{\min_{1 \leq i, j \leq N} C_{ij}}} \right). \quad (2.2)$$

Доказательство. В силу неотрицательности $n_k(t)$, $\|\mathbf{n}(t)\|_1 = \sum_{k=1}^N n_k(t)$, откуда

$$\frac{d}{dt} \|\mathbf{n}(t)\|_1 = \sum_{k=1}^N \frac{dn_k}{dt} = \frac{1}{2} \sum_{k=1}^N \sum_{i+j=k} C_{ij} n_i n_j - \sum_{k=1}^N \sum_{i=1}^N C_{ik} n_i n_k + J.$$

Первая из двух сумм равна просто $\sum_{i+j \leq N} C_{ij} n_i n_j$; упрощая, получаем

$$\frac{d}{dt} \|\mathbf{n}(t)\|_1 = J - \frac{1}{2} \sum_{\substack{1 \leq i, j \leq N \\ i+j \leq N}} C_{ij} n_i n_j - \sum_{\substack{1 \leq i, j \leq N \\ i+j > N}} C_{ij} n_i n_j. \quad (2.3)$$

Отсюда уже легко получить двустороннюю оценку на производную $\|\mathbf{n}(t)\|_1$:

$$J - \bar{C} \|\mathbf{n}\|_1^2 \leq J - \sum_{1 \leq i, j \leq N} C_{ij} n_i n_j \leq \frac{d}{dt} \|\mathbf{n}\|_1 \leq J - \frac{1}{2} \sum_{1 \leq i, j \leq N} C_{ij} n_i n_j \leq J - \frac{1}{2} \underline{C} \|\mathbf{n}\|_1^2, \quad (2.4)$$

где $\underline{C} \leq C_{ij} \leq \bar{C}$.

Таким образом, получаем оценки непосредственно на $\|\mathbf{n}(t)\|_1$:

$$\begin{aligned} f(t) &\leq \|\mathbf{n}(t)\|_1 \leq g(t), \\ f'(t) &= J - f^2(t) \bar{C}, \quad f(0) = \|\mathbf{n}(0)\|_1, \\ g'(t) &= J - g^2(t) \frac{1}{2} \underline{C}, \quad g(0) = \|\mathbf{n}(0)\|_1. \end{aligned}$$

Оба уравнения являются частным случаем $h'(t) = a - bh^2(t)$, решением которого является

$$h(t) = \begin{cases} \sqrt{a/b} \operatorname{th}(t\sqrt{ab} + K), & h(0) < \sqrt{a/b}, \\ \sqrt{a/b} \operatorname{cth}(t\sqrt{ab} + K), & h(0) > \sqrt{a/b}, \\ \sqrt{a/b}, & h(0) = \sqrt{a/b} \end{cases}$$

для некоторого K , так что $h(t)$ монотонно стремится к асимптоте $\sqrt{a/b}$.

Подставляя полученные оценки на $f(t)$ и $g(t)$, для $\|\mathbf{n}(t)\|_1$ получаем

$$\min\left(\|\mathbf{n}(0)\|_1, \sqrt{\frac{J}{C}}\right) \leq \|\mathbf{n}(t)\|_1 \leq \max\left(\|\mathbf{n}(0)\|_1, \sqrt{\frac{2J}{C}}\right).$$

Обозначим правую часть (2.2) L и вернемся к уравнению (2.1):

$$\|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \exp\left(\|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + \frac{3}{2} \int_{t_0}^t \|C\|_C (\|\mathbf{n}\|_1 + \|\tilde{\mathbf{n}}_{(1)}\|_1) d\tau\right) \|\tilde{C}_{12}\|_C \int_{t_0}^t \|\tilde{\mathbf{n}}_{(1)}\|_1 \|\tilde{\mathbf{n}}_{(2)}\|_1 d\tau,$$

где $\|\cdot\|_C$ — чебышёвская норма матрицы (максимум модуля элементов).

Поскольку $\|\tilde{\mathbf{n}}_{(1)}\|_1 + \|\tilde{\mathbf{n}}_{(2)}\|_1 \leq L$, $\|\tilde{\mathbf{n}}_{(1)}\|_1 \|\tilde{\mathbf{n}}_{(2)}\|_1 \leq L^2/4$, и окончательно получаем (предполагая, что минимум C_{ij} , а значит, и L , одинаковы для обеих систем) следующее утверждение.

Лемма 2. В предыдущих обозначениях

$$\|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \frac{\|\tilde{C}_{12}\|_C L^2}{4} (t - t_0) \exp\left(\|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + 3L\|C\|_C (t - t_0)\right). \quad (2.5)$$

В частном случае ядра $C_{ij} = i^\mu j^\nu + i^\nu j^\mu$, $C_{ij} \geq 2(ij)^{\frac{\mu+\nu}{2}}$; при $\mu + \nu \geq 0$ получаем $\|\mathbf{n}(t)\|_1 \leq \max(\|\mathbf{n}(0)\|_1, \sqrt{J})$, независимо от N . (Полученная в лемме 1 оценка снизу относительно менее полезна в этом случае: если ν или μ положительны, оценка стремится к 0 при $N \rightarrow \infty$.) Если, кроме того, $J = 1$ и $\sum_k n_k(0) \leq 1$, получаем $L = 1$ и упрощенную оценку:

$$\|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \frac{\|\tilde{C}_{12}\|_C}{4} (t - t_0) \exp\left(\|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + 3\|C\|_C (t - t_0)\right).$$

Полученная оценка достаточно груба, однако ее основная характеристика — экспоненциальный рост нормы — по-видимому, действительно наблюдается в некоторых случаях, что будет продемонстрировано на численных примерах.

2.2. Решения конечной и бесконечной систем

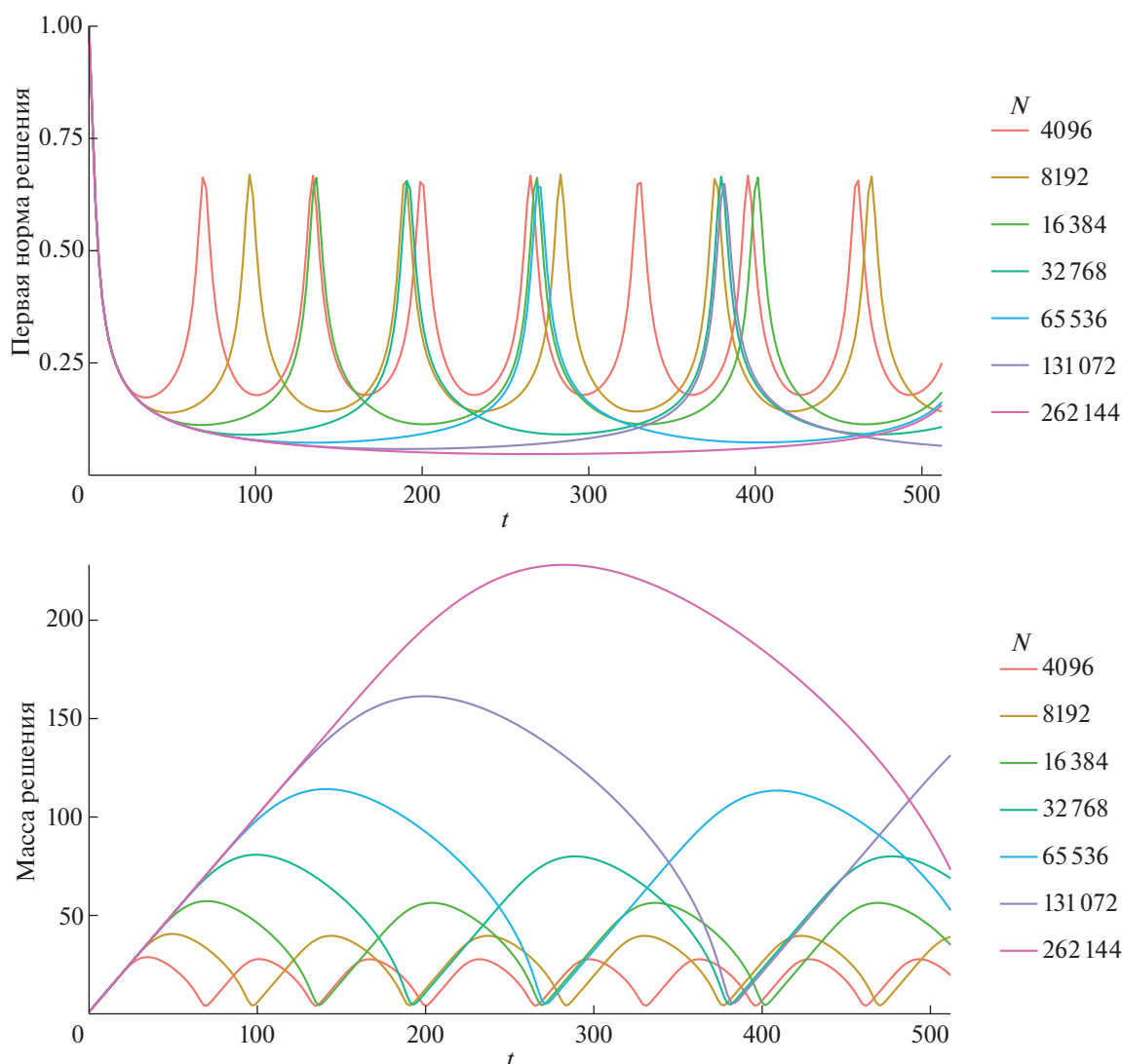
Формально заменяя в выкладках предыдущего пункта N на ∞ (переходить к пределу по N пока, вообще говоря, нельзя, так как не установлена сходимости n), можно перенести полученные оценки на решение исходной системы (1.1), при условии равномерной по t сходимости двойного ряда $\sum_{i,j=1}^{\infty} C_{ij} n_i(t) n_j(t)$ и ограниченности C_{ij} :

$$0 < \underline{C} \leq C_{ij} \leq \bar{C}, \quad 1 \leq i, j.$$

А именно, полагая теперь $\tilde{n}_k(t)$ решением системы (1.1), имеем следующие леммы.

Лемма 3. В вышеприведенных обозначениях,

$$\min\left(\|\tilde{\mathbf{n}}(0)\|_1, \sqrt{\frac{2J}{\sup_{i,j} C_{ij}}}\right) \leq \|\mathbf{n}(t)\|_1 \leq \max\left(\|\tilde{\mathbf{n}}(0)\|_1, \sqrt{\frac{2J}{\inf_{i,j} C_{ij}}}\right).$$



Фиг. 1. Первая норма и общая масса решения (1.3) для разных N . В то время как общая масса решения весьма существенно растет (для бесконечной системы (1.1) она составляет Jt в момент времени t), первая норма (суммарная концентрация частиц всех размеров) остается ограничена в соответствии с леммой 1.

Доказательство. Доказательство точно следует лемме 1, с тем лишь отличием, что в уравнении (2.3) пропадает последний член, а потому (2.4) принимает вид

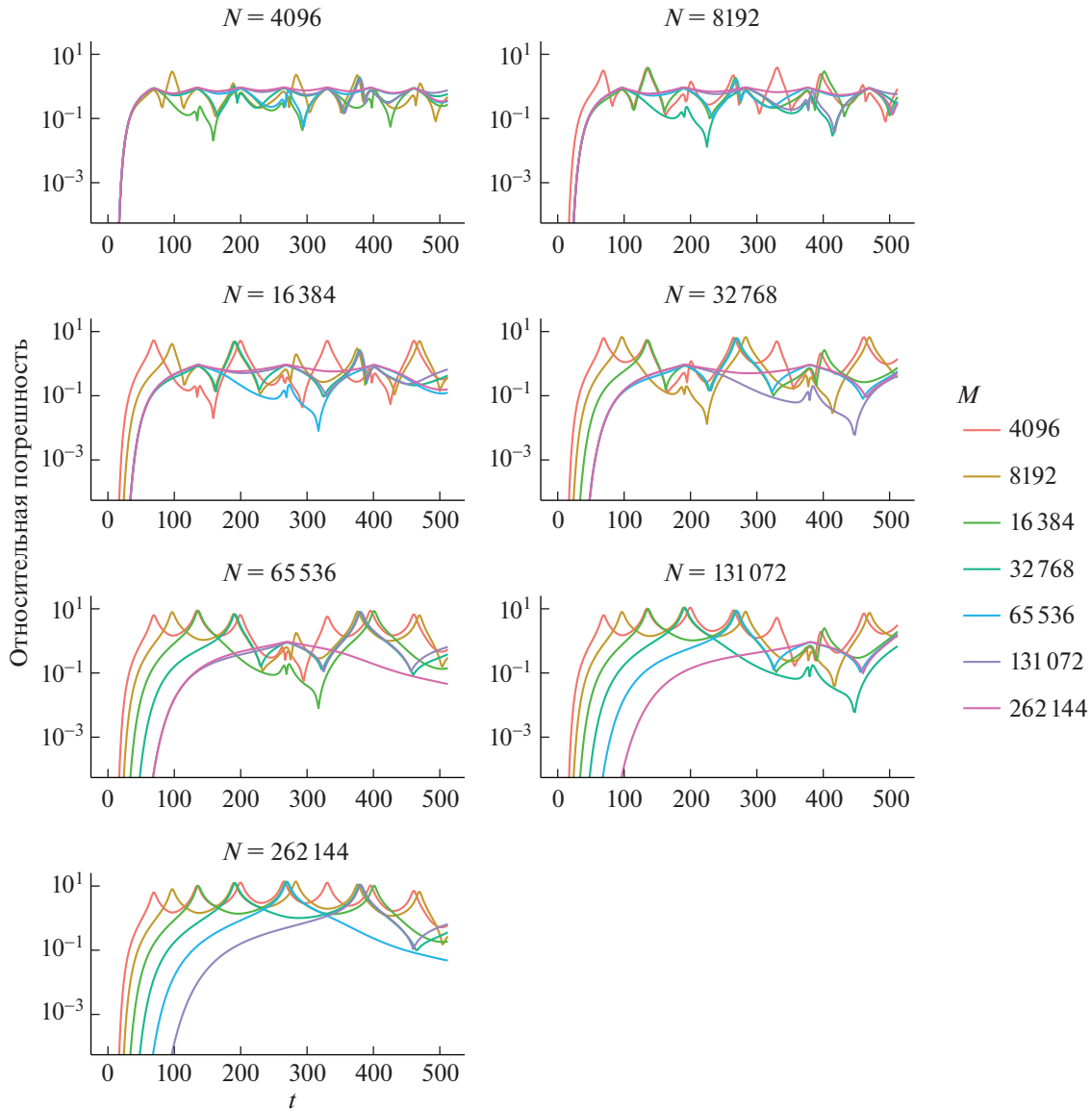
$$J - \frac{1}{2} \bar{C} \|\tilde{\mathbf{n}}\|_1^2 \leq \left\{ J - \frac{1}{2} \sum_{i,j} C_{ij} \tilde{n}_i \tilde{n}_j = \frac{d}{dt} \|\tilde{\mathbf{n}}\|_1 \right\} \leq J - \frac{1}{2} \bar{C} \|\tilde{\mathbf{n}}\|_1^2,$$

что и приводит к отличиям в результате.

Лемма 4. В вышеприведенных обозначениях

$$\begin{aligned} \|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 &\leq \bar{C} \exp(\|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + 3L\bar{C}(t - t_0)) \int_{t_0}^t \|\tilde{\mathbf{n}}_{(1)}\|_1 \|\tilde{\mathbf{n}}_{(2)}\|_1 d\tau \leq \\ &\leq \frac{\bar{C}L^2}{4} (t - t_0) \exp(\|\mathbf{n}(t_0) - \tilde{\mathbf{n}}_{(1)}(t_0)\|_1 + 3L\bar{C}(t - t_0)). \end{aligned}$$

Доказательство. Доказательство (в оговоренном выше предположении о равномерной сходимости двойного ряда) в точности следует доказательству леммы 2, с заменой N на ∞ , так что $\tilde{\mathbf{n}}$ и $\tilde{\mathbf{n}}_{(2)}$ теперь последовательности; соответственно, $\|\tilde{\mathbf{n}}\|_1 = \sum_{i=1}^{\infty} \tilde{n}_i$.

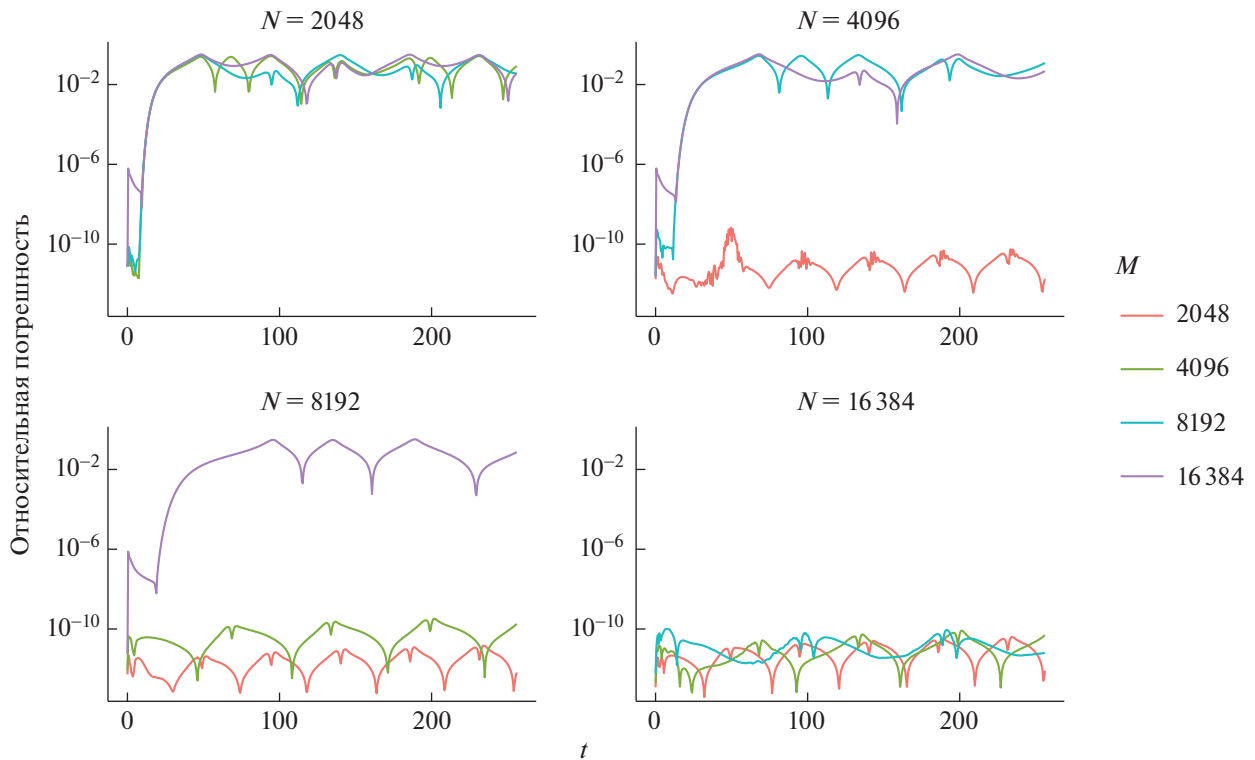


Фиг. 2. Относительная погрешность приближения решения системы размера N решением системы размера M в первой норме, для различных пар M и N ; для каждой пары M и N сравниваются первые $\min(M, N)$ компонент решения (1.3) в один и тот же момент времени. Циклическое поведение решений приводит к циклическому поведению ошибки; рост ошибки вблизи пиков экспоненциальный, что качественно соответствует оценке (2.5).

Если при этом известно поведение $\tilde{\mathbf{n}}(t)$, отсюда можно получить оценки на качество приближения полного решения конечным на конечном интервале времени; например, при $C_{ij} \equiv 1$ известно стационарное решение [11], в котором $\tilde{n}_k = O(k^{-3/2})$. Если это же соотношение сохраняется для всех t , то $\|\tilde{\mathbf{n}}_{(2)}\|_1 \leq KM^{-1/2}$ для некоторого $K > 0$, и из леммы 4 сразу получаем (с учетом $\mathbf{n}(0) = \tilde{\mathbf{n}}_{(1)}(0)$)

$$\|\mathbf{n} - \tilde{\mathbf{n}}_{(1)}\|_1 \leq \frac{1}{\sqrt{M}} t \exp(3t\sqrt{J}) LK,$$

так что при любом фиксированном t норма разности стремится к нулю с ростом M .



Фиг. 3. Погрешность редуцированного решения системы размера M , построенного по базису, полученному для системы размера N . При $N > M$ векторы базиса обрезаются до размера M и реортогонализуются; при $M > N$ они дополняются нулями до требуемого размера. Использование большего базиса для меньшей системы дает превосходную точность решения, несмотря на существенные отличия между самими решениями.

3. ЧИСЛЕННЫЕ ПРИМЕРЫ

В качестве примера рассмотрим систему (1.3) с $C_{ij} = (i/j)^a + (j/i)^a$. При $0 \leq a \leq 1/2$ решение исходной системы (1.1) стремится к стационарному, известному аналитически и воспроизводимому в конечной системе [27], [28]. При $a > 1/2$, однако, в решении конечной системы наблюдаются циклы; как мы вскоре покажем, в этом случае решения при разных N могут существенно отличаться. В качестве конкретного примера такого значения параметра в численных экспериментах использовалось $a = 0.8$.

Прежде всего, в качестве иллюстрации леммы 1, на фиг. 1 приведены первая норма и общая масса решения (1.3) для различных N ; хорошо видно, что, хотя общая масса растет весьма существенно за счет притока частиц размера 1, общее количество частиц, совпадающее с первой нормой, остается ограниченным сверху (лемма 1 дает в нашем случае равномерную по N оценку).

На фиг. 2 показана относительная погрешность приближения $\tilde{\mathbf{n}}_{(1)}$ с помощью n для различных значений M и N от 2^{12} до 2^{18} . Погрешность демонстрирует циклический характер, обусловленный циклическим поведением решений, и экспоненциальный рост по обе стороны от пиков, в соответствии с общей формой теоретической оценки.

Несмотря на экспоненциальный рост ошибки вблизи пиков, и общий рост их высоты (связанный, вероятно, с ростом периода и соответствующим ростом максимума $|t - t_0|$), стоит отметить, что при фиксированных N и t до первого пика ошибка падает с ростом M .

В [18] была показана возможность построения по начальному отрезку решения базиса линейного пространства, в котором лежит решение, и дальнейшего поиска решения в этом пространстве с меньшими вычислительными затратами. Мы можем рассмотреть возможность решения редуцированной таким образом системы размера M в базисе, полученном для системы размера N (должным образом усеченным или дополненным нулями и реортогонализированном) как

непрямое сравнение общего поведения решений систем разного размера, не привязанное к конкретным моментам времени.

Результаты такого эксперимента приведены на фиг. 3 для M и N от 2^{11} до 2^{14} . Как видно на графике, базис, полученный для системы большего размера, позволяет решать меньшие системы с хорошей точностью, т.е. решение меньшего размера с высокой точностью лежит в линейном пространстве, описываемом соответствующими компонентами большего решения, несмотря на высокие поточечные различия. В обратную сторону — от меньшего базиса к большей системе — переход существенно менее успешный; причина здесь, очевидно, в том, что такое редуцированное решение в действительности воспроизводит решение меньшей системы, дополненное нулями — а оно, как мы уже видели, существенно отличается от решения большей.

4. ЗАКЛЮЧЕНИЕ

В данной работе мы продемонстрировали возможности использования редуцированных базисов для решения параметрического семейства моделей процесса необратимой агрегации с источником и стоком частиц. Важным наблюдением является общая структура базисов задач разной размерности, что позволяет использовать базисы большей размерности при решении задач меньшей размерности без дополнительных усилий по их построению.

Для дискретного варианта уравнений Смолуховского получены оценки на расхождение решений систем разных размеров с одним и тем же ядром. Хотя сами по себе оценки достаточно грубы, в отдельных случаях они позволяют тем не менее получить оценки сходимости к решению бесконечной системы, а общее поведение оценки как функции времени — экспоненциальный рост — в численных экспериментах соотносится с поведением нормы решения.

Несмотря на продемонстрированное таким образом существенное расхождение решений систем разного размера, численные эксперименты также показывают, что линейное пространство, описываемое решением большой системы с высокой точностью, содержит в себе решение меньшей, что указывает на общую структуру таких решений, не описываемую обычной оценкой погрешности, и поднимает вопрос о возможности обратного перехода — от меньшей системы к большей.

СПИСОК ЛИТЕРАТУРЫ

1. *Eigen M.* Prionics or the kinetic basis of prion diseases // *Biophysical chemistry*. 1996. V. 63. № 1. P. A1–A18.
2. *Krapivsky P.L., Redner S., Ben-Naim E.* A Kinetic View of Statistical Physics. UK: Cambridge University Press, 2010.
3. *Pöschel T., Brilliantov N.V., Frömmel C.* Kinetics of prion growth // *Biophysical Journal*. 2003. V. 85. № 6. P. 3460–3474.
4. *Garzon Dasgupta A.K., Martyanov A.A., Filkova A.A., Panteleev M.A., Sveshnikova A.N.* Development of a Simple Kinetic Mathematical Model of Aggregation of Particles or Clustering of Receptors // *Life*. 2020. V. 10. № 6. P. 97.
5. *Aloyan A.E., Arutyunyan V.O., Lushnikov A.A., Zagaynov V.A.* Transport of coagulating aerosol in the atmosphere // *J. of Aerosol Sci.* 1997. V. 28. № 1. P. 67–85.
6. *Stoyanovskaya O.P., Snytnikov V.N.* Numerical modeling of formation of high density solitary vortices in a circumstellar disk // *Vychislitel'nye Metody i Programirovanie*. 2012. V. 13. № 3. P. 377–383.
7. *Esposito L.* Planetary Rings. UK: Cambridge University Press, 2006.
8. *Brilliantov N.V., Krapivsky P.L., Bodrova A., Spahn F., Hayakawa H., Stadnichuk V., Schmidt J.* Size distribution of particles in Saturn's rings from aggregation and fragmentation // *PNAS*. 2015. V. 112. № 31. P. 9536–9541.
9. *Галкин В.А.* Уравнение Смолуховского. М.: Физматлит, 2001.
10. *Melzak Z.A.* A scalar transport equation // *Transactions of the American Mathematical Society*. 1957. V. 85. № 2. P. 547–560.
11. *Hayakawa H.* Irreversible kinetic coagulations in the presence of a source // *J. of Physics A: Math. and General*. 1987. V. 20. № 12. P. L801–L805.
12. *Ball R.C., Connaughton C., Jones P.P., Rajesh R., Zaboronski O.* Collective Oscillations in Irreversible Coagulation Driven by Monomer Inputs and Large-Cluster Outputs // *Physical Review Letters*. 2012. V. 109. № 16. P. 168304.
13. *Esenturk E., Connaughton C.* Role of zero clusters in exchange-driven growth with and without input // *Physical Review E*. 2020. V. 101. № 5. P. 052134.

14. *Słomka J., Stocker R.* Bursts Characterize Coagulation of Rods in a Quiescent Fluid // *Physical Review Letters*. 2020. V. 124. № 25. P. 258001.
15. *Pego R.L., Velázquez J.J.L.* Temporal oscillations in Becker–Döring equations with atomization // *Nonlinearity*. 2020. V. 33. № 4. P. 1812.
16. *Budzinskiy S.S., Matveev S.A., Krapivsky P.L.* Hopf bifurcation in addition-shattering kinetics // *Physical Review E*. 2021. V. 103. № 4. P. L040101.
17. *Niethammer B., Pego R.L., Schlichting A., Velázquez J.J.L.* Oscillations in a Becker–Döring model with injection and depletion. 2021. arXiv preprint arXiv:2102.06751.
18. *Тимохин И.В., Матвеев С.А., Тиртышников Е.Е., Смирнов А.П.* Метод поиска редуцированного базиса для нестационарных задач // *Докл. АН. Математика, информатика, процессы управления*. 2021. Т. 497. № 2_21. С. 35–38.
19. *Timokhin I.V., Matveev S.A., Tyrtshnikov E.E., Smirnov A.P.* Model reduction in Smoluchowski-type equations. 2020. arXiv preprint arXiv:2008.09439.
20. *Moore B.* Principal component analysis in linear systems: Controllability, observability, and model reduction // *IEEE Transactions on Automatic Control*. 1981. Т. 26. № 1. С. 17–32.
21. *Glover K.* Optimal Hankel-norm approximations of linear multivariable systems and their l_∞ -error bounds // *Int. J. Control*. 1984. Т. 39. № 6. С. 115–193.
22. *Sirovich L.* Turbulence and the Dynamics of Coherent Structures. I–III // *Quart. Appl. Math.* 1987. Т. 45. № 3. С. 561–590.
23. *Petrov S.* Model Order Reduction Algorithms in the Design of Electric Machines // *Internat. Conference on Large-Scale Sci. Comput.* 2019. P. 140–147.
24. *Gusak J., Daulbaev T., Ponomarev E., Cichocki A., Oseledets I.* Reduced-order modeling of deep neural networks // *Comput. Math. and Math. Phys.* 2021. V. 61. № 5. P. 774–785.
25. *Pinnau R.* Model Reduction via Proper Orthogonal Decomposition. Berlin: Springer, 2008.
26. *Timokhin I.V., Matveev S.A., Tyrtshnikov E.E., Smirnov A.P.* Model reduction for Smoluchowski equations with particle transfer // *Russian Journal of Numerical Analys. and Math. Modelling*. 2021. V. 36. № 3. P. 177–181.
27. *Matveev S.A., Stadnichuk V.I., Tyrtshnikov E.E., Smirnov A.P., Ampilogova N.V., Brilliantov N.V.* Anderson acceleration method of finding steady-state particle size distribution for a wide class of aggregation–fragmentation models // *Comput. Phys. Communicat.* 2018. V. 224. P. 154–163.
28. *Timokhin I.V., Matveev S.A., Siddharth N., Tyrtshnikov E.E., Smirnov A.P., Brilliantov N.V.* Newton method for stationary and quasi-stationary problems for Smoluchowski-type equations // *J. of Comput. Phys.* 2019. V. 382. P. 124–137.

**ОБЩИЕ
ЧИСЛЕННЫЕ МЕТОДЫ**

УДК 519.642

**ОБОСНОВАНИЕ КВАДРАТУРНО-РАЗНОСТНОГО МЕТОДА
РЕШЕНИЯ ИНТЕГРОДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ
С ПРОИЗВОДНЫМИ ПЕРЕМЕННОГО ПОРЯДКА**

© 2022 г. А. И. Федотов

420111 Казань, ул. Карла Маркса, 10, Казанский национальный исследовательский технический университет
им. А.Н. Туполева, Россия

e-mail: fedotovkazan@mail.ru

Поступила в редакцию 08.05.2021 г.
Переработанный вариант 07.09.2021 г.
Принята к публикации 16.12.2021 г.

Дано новое определение производной дробного порядка на основе интерполирования производных натурального порядка. Главным преимуществом нового определения является локальность таких производных. То есть значение производной в точке не зависит от области определения функции, как в случаях производных Римана–Лиувилля и Капуто. Это позволяет строить и обосновывать простые вычислительные методы решения уравнений, содержащих такие производные. Более того, такое определение позволяет обобщить понятие производной на случай переменного порядка дифференцирования. Рассмотрен класс уравнений, содержащих введенные производные. Доказана однозначная разрешимость исходных уравнений и обоснован квадратурно-разностный метод для их решения. Получены эффективные оценки погрешности приближенных решений. Теоретические выводы подтверждены численным решением модельной задачи. Библ. 54. Табл. 1.

Ключевые слова: производные переменного порядка, квадратурно-разностный метод, интегродифференциальные уравнения.

DOI: 10.31857/S0044466922040068**1. ВВЕДЕНИЕ**

Задача обобщения понятия производной на случай производной дробного порядка ставилась с самого начала существования дифференциального исчисления. Так, уже в 1695 г. в письме Лопиталю Лейбниц пишет о возможном обобщении определения введенной им производной на случай дробного порядка и о множественности таких обобщений.

В 1730 г. Эйлер, обобщивший понятие факториала на случай любых действительных чисел, указал, что аналогичным способом, т.е. интерполированием, можно и производную обобщить на случай любого действительного порядка.

В 1812 г. Лаплас, используя свое преобразование, дал определение производной дробного порядка через интеграл.

В 1819 г. Лакруа определил производную дробного порядка от степенных функций. Это определение позволило обобщить понятие дробного дифференцирования на пространство функций, в которых степенные функции образуют полную систему базисных функций. Интересно отметить, что его определение совпадает для таких функций с современным определением производной Римана–Лиувилля.

Фурье в 1822 г. определили дробную производную через свое преобразование. Он впервые указал, что при таком определении порядок производной может быть любым действительным числом как положительным, так и отрицательным.

В дальнейшем производные дробного порядка вводились разными математиками, и к настоящему времени таких определений имеется уже более 20. Исчерпывающую информацию о различных определениях можно найти в монографии [1] и обзорных работах [2], [3].

Почему же при наличии такого многообразия различных определений появляются все новые и новые определения? Причина в том, что в каждом определении имеются, при наличии поло-

жительных свойств, недостатки. Именно для их преодоления и вводятся все новые и новые определения. Так, в 2014 г. в [4] определена производная, которая нашла уже много поклонников и последователей. В работе [5] предложено новое определение дробной производной, обобщающей дробную производную Капуто. В мае 2020 г. в [6] дано определение дробной производной с несингулярным ядром. Интересно, что уже в июне того же года появилась работа [7], опровергающая корректность определения дробной производной с несингулярными ядрами. А в [8] изложена даже целая программа построения дробного анализа на основе модификаций оператора дробного дифференцирования Адамара [9]. Известен ряд обобщений определений дробных операторов на случай переменного порядка. Например, в [10]–[12] рассматриваются операторы, у которых порядок дифференцирования является детерминированной функцией времени, а в [13] исследовался случай, когда порядок дифференцирования является случайной величиной. При этом в [10] предложен алгоритм численного расчета и схемотехническая реализация введенного оператора. Следует упомянуть также и работы по дискретному дробному исчислению, введенному в середине XX в. (см. [14]–[16]), и ставшему особенно актуальным в последнее время в связи с разработкой численных методов решения уравнений с дробными операторами. В [17]–[24] даны обобщение и развитие полученных ранее результатов, выведены формулы для дискретного преобразования Лапласа и построен метод решения разностных дробных уравнений на его основе. В [25] предложена процедура дискретизации значений дробной производной Римана–Лиувилля на основе разложения в ряды по полиномам Чебышёва.

Чтобы как-то систематизировать все обилие результатов по дробному дифференцированию, можно условно разделить публикации, посвященные изучению дифференциальных уравнений дробного порядка, на три группы:

а) работы, посвященные изучению математических проблем (существование и единственность решений, зависимость решений от начальных и граничных условий, анализ устойчивости решений, наличия особых точек, форма и смысл начальных и граничных условий, вид и метод построения общего решения для основных типов уравнений и др.);

б) работы по точным методам решения конкретных типов и разновидностей уравнений;

в) работы по разработке, обоснованию и реализации приближенных методов решения уравнений.

Значительный объем материала по математическим вопросам дробных дифференциальных уравнений собран в монографиях [1], [26]–[35]. В частности, в них приводятся теоремы существования и единственности решения начальных и краевых задач для, главным образом, линейных дробных дифференциальных уравнений, содержащих дробные производные в смысле Римана–Лиувилля, Капуто, Грюнвальда–Летникова, и обсуждаются основные методы решения дробных дифференциальных уравнений такие как сведение к интегральному уравнению, метод интегральных преобразований Лапласа, Меллина, Фурье, метод функций Грина и т.д. В обзорной работе [36] обсуждаются введенная Кочубеем [37], [38], [39] дробная производная и вопросы теории, возникающие при ее использовании в дробных дифференциальных уравнениях. В справочнике [40] для решения обыкновенных дробных дифференциальных уравнений, помимо преобразования Лапласа, применяется их сведение к обыкновенным дифференциальным уравнениям целого порядка. Интересным приложением дробных дифференциальных уравнений является построение с их помощью точных решений дифференциальных уравнений целого порядка [1], [30]. Все эти работы позволяют утверждать, что в настоящее время как теория дробных дифференциальных уравнений, так и теория методов их точного решения, хотя и не являются законченными, но представляют собой весьма обширную и хорошо проработанную область исследований. Однако в последнее время все большее значение — это заметно и по количеству появляющихся работ — приобретают работы по разработке, обоснованию и реализации приближенных методов решения дробных дифференциальных уравнений, т.е. работ из группы в).

Число публикаций по последней группе работ растет очень быстрыми темпами, поэтому в данной работе приведем — кроме упомянутых выше — лишь работы, вышедшие в самое последнее время.

В работе [41] обоснован полиномиальный метод решения задачи Коши для уравнения с производными Римана–Лиувилля. Доказаны корректность задачи по Адамару, сходимости метода, получены эффективные оценки погрешности. Описание алгоритма и пример численной реализации метода отсутствуют. В работах [42] и [43] описаны разностные методы решения третьей краевой задачи для нагруженного уравнения теплопроводности с дробной производной Капуто. Получены априорные оценки погрешности приближенного решения в условиях разрешимости исходной задачи. Приведено подробное описание алгоритма решения, но отсутствует его чис-

ленная реализация. Численный метод решения уравнения с дробной производной по времени в смысле Капуто-Фабрицио обоснован в [44]. Теоретические исследования подтверждены численными расчетами. В [45] теоретически обоснован приближенный метод решения дробного дифференциального уравнения с обобщенной производной Капуто. В доказательстве используется теория устойчивости по Уламу. Численные примеры отсутствуют.

Такое обилие определений дробной производной с одной стороны и необходимость дробных производных для прикладных исследований приводят к необходимости очертить рамки допустимости определений, т.е. ввести критерии их корректности. Такие критерии были сформулированы в работе [2] и позже повторены в современной формулировке в работе [46]. Основными критериями являются следующие:

- а) если функция аналитическая, то ее производная тоже должна быть аналитической;
- б) оператор дробного дифференцирования должен быть линейным;
- в) при натуральных значениях порядка дифференцирования оператор должен совпадать с обычным оператором дифференцирования.

Последнее требование означает, что, как указал Эйлер [2], возникает ситуация, похожая на задачу интерполирования. Имеется система операторов целочисленного дифференцирования, требуется связать эти “точки” оператором, допускающим дифференцирование любого действительного порядка. И так же как в случае интерполирования, когда имеется сколь угодно много кривых, соединяющих заданные точки, имеется бесконечно много способов определить оператор дробного дифференцирования, совпадающий при целых значениях порядка дифференцирования с обычным оператором дифференцирования. Все такие операторы должны считаться корректно определенными. При этом важно понимать, что операторы целочисленного дифференцирования являются частными случаями операторов дробного дифференцирования, и поэтому обладают свойствами, которыми операторы дробного дифференцирования могут и не обладать. И вопрос тогда уже не столько в корректности определения оператора дробного дифференцирования, а в выборе такого оператора, который бы наилучшим образом подходил для решения конкретной задачи.

В данной работе предлагается еще одно определение оператора дробного дифференцирования функций, заданных на отрезке, позволяющее, в частности, сделать порядок производной переменным. Это определение удовлетворяет всем критериям корректности определения дробной производной. Кроме того, определенная таким образом производная является, в отличие от большинства предшествующих определений, локальной, что делает ее удобной при использовании в прикладных задачах. Приближенный метод, предлагаемый в статье, обоснован с помощью методики Вайникко [47], [48]. Приведен численный пример, подтверждающий теоретические результаты.

2. ОСНОВНЫЕ ОБОЗНАЧЕНИЯ И ОПРЕДЕЛЕНИЯ

Будем, как обычно, обозначать через $\mathbb{N}$ множество натуральных чисел, $\mathbb{N}_0$ — множество натуральных чисел, дополненных нулем, $\mathbb{Z}$ — множество целых чисел, $\mathbb{R}$ — множество действительных чисел, а $\mathbb{R}_+$ — множество неотрицательных действительных чисел.

Следующие обозначения и определения взяты из монографии [47].

Пусть $X, X_n, n \in \mathbb{N}$, и $Y, Y_n, n \in \mathbb{N}$, две пары банаховых пространств, а $\mathcal{L}(X, Y)$ — множество линейных непрерывных операторов, действующих из X в Y .

Определение 1. Систему $\mathcal{P} = (p_n)_{n \in \mathbb{N}}$ линейных ограниченных операторов $p_n : X \rightarrow X_n$ будем называть *связывающей* для X и $X_n, n \in \mathbb{N}$, если для любого $x \in X$ имеем

$$\|p_n x\|_{X_n} \rightarrow \|x\|_X, \quad n \in \mathbb{N}.$$

Определение 2. Последовательность $(x_n)_{n \in \mathbb{N}}$ элементов $x_n \in X_n$ $\mathcal{P}$ -сходится к $x \in X$, если

$$\|x_n - p_n x\|_{X_n} \rightarrow 0, \quad n \in \mathbb{N}.$$

Определение 3. Последовательность $(x_n)_{n \in \mathbb{N}}$ элементов $x_n \in X_n$ $\mathcal{P}$ -компактна, если любая ее подпоследовательность $(x_n)_{n \in \mathbb{N}' \subseteq \mathbb{N}}$ содержит $\mathcal{P}$ -сходящуюся подпоследовательность $(x_n)_{n \in \mathbb{N}'' \subseteq \mathbb{N}'}$.

Определение 4. Последовательность операторов $T_n : X_n \rightarrow Y_n$, $n \in \mathbb{N}$, $\mathcal{PQ}$ -сходится к оператору $T : X \rightarrow Y$, если для любой $\mathcal{P}$ -сходящейся к $x \in X$ последовательности $(x_n)_{n \in \mathbb{N}}$ последовательность $(T_n x_n)_{n \in \mathbb{N}}$ $\mathcal{Q}$ -сходится к Tx .

Определение 5. Последовательность операторов $T_n \in \mathcal{L}(X_n, Y_n)$, $n \in \mathbb{N}$, компактно $\mathcal{PQ}$ -сходится к вполне непрерывному оператору $T \in \mathcal{L}(X, Y)$, если T_n $\mathcal{PQ}$ -сходится к оператору T и выполнено следующее условие компактности:

$$x_n \in X_n, \quad \|x_n\|_{X_n} \leq 1, \quad n \in \mathbb{N} \Rightarrow (T_n x_n) \mathcal{Q} \text{ компактна.}$$

Обозначим через Y множество непрерывных на $[0, 1]$ функций. С нормой

$$\|y\|_Y = \max_{0 \leq t \leq 1} |y(t)|, \quad y \in Y,$$

множество Y становится банаховым пространством.

Зафиксируем $m \in \mathbb{N}_0$ и обозначим через X множество m раз непрерывно дифференцируемых на $[0, 1]$ функций, удовлетворяющих условиям

$$x^{(v)}(0) = 0, \quad 0 \leq v \leq m-1.$$

С нормой

$$\|x\|_X = \max_{0 \leq t \leq 1} |x^{(m)}(t)|, \quad x \in X,$$

множество X также становится банаховым пространством.

Оператор дифференцирования порядка $v \in \mathbb{N}_0$, $0 \leq v \leq m$, определенный на X , обозначим через D^v . При $v = 0$ оператор D^v будет совпадать с тождественным оператором $I = D^0$.

Из определения пространств X и Y сразу следует, что

$$\|D^m\|_{X \rightarrow Y} = \|D^{-m}\|_{Y \rightarrow X} = 1.$$

Зафиксируем $\alpha \in \mathbb{R}_+$, $\alpha < m$, и определим на X оператор дробного дифференцирования порядка α

$$D^\alpha = (1 - \{\alpha\})D^{[\alpha]} + \{\alpha\}D^{[\alpha]+1}. \quad (1)$$

Здесь $[\alpha]$ обозначает целую часть, а $\{\alpha\}$ дробную часть числа α . При таком определении оператор (1) будет совпадать с обычным оператором дифференцирования при $\alpha \in \mathbb{N}_0$, $\alpha < m$, поэтому это определение должно считаться корректным.

Обозначим теперь через $\alpha(t)$, $t \in [0, 1]$, непрерывную функцию такую, что для некоторого числа $v \in \mathbb{N}_0$, $v < m$, имеем

$$v \leq \alpha(t) < v+1, \quad t \in [0, 1]. \quad (2)$$

Оператором дифференцирования *переменного порядка* $\alpha(t)$ назовем оператор

$$D^{\alpha(t)} = (1 - \{\alpha(t)\})D^v + \{\alpha(t)\}D^{v+1}, \quad t \in [0, 1]. \quad (3)$$

И вновь, при $\alpha(t) = v \in \mathbb{N}_0$ оператор (3) будет совпадать с обычным оператором дифференцирования D^v , поэтому и это определение должно считаться корректным.

Зафиксируем $n \in \mathbb{N}_0$ и обозначим через Y_n множество векторов $y_n = (y_0, y_1, \dots, y_n)$. С нормой

$$\|y_n\|_{Y_n} = \max_{0 \leq i \leq n} |y_i|, \quad y \in Y,$$

множество Y_n становится банаховым пространством.

Пусть X_n — множество векторов $x_n = (x_0, x_1, \dots, x_n)$, а $\bar{D}_n$ имеет вид

$$[\bar{D}_n x_n]_k = \frac{x_k - x_{k-1}}{h}, \quad h = \frac{1}{n}, \quad k \in \mathbb{Z},$$

простейшего оператора численного дифференцирования первого порядка на X_n . Для значений $k \leq 0$ и $k > n$ будем полагать $x_k = 0$. Теперь определим на X_n простейший оператор численного дифференцирования порядка $v \in \mathbb{N}_0$, $0 \leq v \leq m$:

$$\bar{D}_n^v \mathbf{x}_n = \bar{D}_n(\bar{D}_n^{v-1} \mathbf{x}_n), \quad v = 1, 2, \dots.$$

При $v = 0$ оператор $\bar{D}^v$ будем считать тождественным оператором $I_n = \bar{D}_n^0$. Понятно, что при таком определении для всех $\mathbf{x}_n \in X_n$ будут выполняться условия

$$[\bar{D}_n^v \mathbf{x}_n]_0 = 0, \quad 0 \leq v \leq m,$$

а норма

$$\|\mathbf{x}_n\|_{X_n} = \max_{0 \leq k \leq n} |[\bar{D}_n^m \mathbf{x}_n]_k|$$

превращает множество X_n в банахово пространство, причем

$$\|\bar{D}_n^m\|_{X_n \rightarrow Y_n} = \|\bar{D}_n^{-m}\|_{Y_n \rightarrow X_n} = 1.$$

Пусть

$$t_k = \frac{k}{n}, \quad k \in \mathbb{Z}, \quad (4)$$

сетку равноотстоящих узлов на $\mathbb{R}$. Легко видеть, что операторы $p_n \in \mathcal{L}(X, X_n)$ и $q_n \in \mathcal{L}(Y, Y_n)$

$$p_n x = (x(t_0), x(t_2), \dots, x(t_n)), \quad p_n : X \rightarrow X_n,$$

$$q_n x = (x(t_0), x(t_2), \dots, x(t_n)), \quad q_n : Y \rightarrow Y_n,$$

являются связывающими операторами для пар пространств (X, X_n) и (Y, Y_n) .

Операторы $\bar{D}_n^v$, $v = 0, 1, \dots, m$, являются $\mathcal{PQ}$ -сходящимися в том смысле, что для любой функции $x \in X$

$$\|q_n \bar{D}_n^v x - \bar{D}_n^v p_n x\|_{Y_n} \rightarrow 0 \quad \text{при} \quad n \rightarrow \infty,$$

однако скорость их сходимости невелика. В вычислительной схеме предлагаемого метода можно будет для повышения точности использовать любые сходящиеся операторы численного дифференцирования D_n^v

$$[D_n^v \mathbf{x}_n]_k = h^{-v} \sum_{j=-r_v}^0 d_{v,j} x_{k+j}, \quad v = 0, 1, \dots, m, \quad k = 0, 1, \dots, n. \quad (5)$$

Критерием сходимости операторов вида (5) является возможность представить их в виде

$$D_n^v = V_n^v S_n^{-r_v+v} \bar{D}_n^v,$$

где

$$V_n^v = \sum_{j=0}^{r_v-v} \rho_j S_n^j \quad (6)$$

есть полином от оператора сдвига S_n^j :

$$[S_n^j \mathbf{x}_n]_k = x_{j+k},$$

причем

$$\sum_{j=0}^{r_v-v} \rho_j = 1.$$

При этом полином

$$\sum_{j=0}^{r_v-v} \rho_j \zeta^j$$

называется *характеристическим полиномом*, а его корни ζ_j , $j = 1, 2, \dots, r_v - v$, — характеристическими значениями оператора D_n^v .

Характеристические значения сходящегося оператора D_n^m позволяют сформулировать простой необходимый и достаточный критерий его обратимости:

$$|\zeta_j| \neq 1, \quad j = 1, 2, \dots, r_m - m.$$

Действительно, разложение (6) представим в виде произведения

$$V_n^m = \prod_{j=1}^{r_m-m} \frac{S_n - \zeta_j I_n}{1 - \zeta_j}, \quad S_n = S_n^1, \quad I_n - \text{тождественный оператор в } Y_n.$$

Так как оператор $S_n : Y_n \rightarrow Y_n$ изометричен, т.е.

$$\|S_n y_n\|_{Y_n} = \|y_n\|_{Y_n},$$

то справедливы оценки

$$\begin{aligned} \|(S_n - \zeta I_n)^{-1}\|_{Y_n \rightarrow Y_n} &\leq \frac{1}{1 - |\zeta|} \quad \text{при} \quad |\zeta| < 1, \\ \|(S_n - \zeta I_n)^{-1}\|_{Y_n \rightarrow Y_n} &\leq \frac{1}{|\zeta| - 1} \quad \text{при} \quad |\zeta| > 1. \end{aligned}$$

Следовательно, оператор $V_n^m : Y_n \rightarrow Y_n$ имеет ограниченный обратный:

$$\|V_n^{-m}\|_{Y_n \rightarrow Y_n} \leq \prod_{j=1}^{r_m-m} \frac{|1 - \zeta_j|}{|1 - |\zeta_j||}.$$

Это позволяет представить оператор $D_n^{-m} : Y_n \rightarrow X_n$, обратный к оператору D_n^m , в виде композиции операторов

$$D_n^{-m} = \bar{D}_n^{-m} S_n^{r_m-m} V_n^{-m} : Y_n \rightarrow X_n$$

и оценить с учетом изометричности оператора $S_n^{r_m-m}$ и равенства $\|\bar{D}_n^{-m}\|_{Y_n \rightarrow X_n} = 1$ его норму

$$\|D_n^{-m}\|_{Y_n \rightarrow X_n} \leq \|V_n^{-m}\|_{Y_n \rightarrow Y_n}.$$

Эти важные результаты о сходимости и устойчивости разностных схем были получены в работах С.Г. Крейна, Л.Н. Шаблицкой [49], [50] и Н.Н. Гудович [51], [52].

3. ВСПОМОГАТЕЛЬНЫЕ РЕЗУЛЬТАТЫ

В этом разделе приведены три леммы, необходимые для дальнейшего изложения. Доказательство первой леммы имеется, например, в [53], второй и третьей — в [47].

Лемма 1. Пусть A и B — линейные операторы, действующие из банахова пространства X в банахово пространство Y . Предположим, что оператор A обратим и выполнено условие $\|B\|_{X \rightarrow Y} \|A^{-1}\|_{Y \rightarrow X} < 1$. Тогда оператор $A + B : X \rightarrow Y$ также обратим, и справедлива оценка

$$\|(A + B)^{-1}\|_{Y \rightarrow X} \leq \frac{\|A^{-1}\|_{Y \rightarrow X}}{1 - \|B\|_{X \rightarrow Y} \|A^{-1}\|_{Y \rightarrow X}}.$$

Рассмотрим уравнения вида

$$z + Tz = y, \quad T \in \mathcal{L}(Y, Y), \quad y \in Y, \quad (7)$$

$$\mathbf{z}_n + T_n \mathbf{z}_n = \mathbf{y}_n, \quad T_n \in \mathcal{L}(Y_n, Y_n), \quad \mathbf{y}_n \in Y_n, \quad n \in \mathbb{N}. \quad (8)$$

Лемма 2. Пусть выполнены следующие условия:

- 1) операторы $T, T_n, n \in \mathbb{N}$ вполне непрерывны и T_n компактно $\mathcal{Q}\mathcal{Q}$ -сходится к T ;
- 2) последовательность $(\mathbf{y}_n)_{n \in \mathbb{N}}, \mathbf{y}_n \in Y_n, \mathcal{Q}$ -сходится к y ;
- 3) для любого $z \in TY, z \neq 0, \lim_{n \rightarrow \infty} \|q_n z\| > 0$;
- 4) уравнение (7) имеет единственное решение $z^* \in Y$.

Тогда при достаточно больших n уравнение (8) однозначно разрешимо и приближенные решения $\mathbf{z}_n^* \in Y_n$ $\mathcal{Q}$ -сходятся к точному решению $z^* \in Y$ уравнения (7) со скоростью

$$\|\mathbf{z}_n^* - q_n z^*\|_{Y_n} \leq c \|q_n T z^* - T_n q_n z^* - q_n y + \mathbf{y}_n\|_{Y_n}. \quad (9)$$

(Здесь и далее c обозначает вполне определенные константы, не зависящие от n , возможно, разные в разных вхождениях.)

Лемма 3. Если $(q_n)_{n \in \mathbb{N}}$ связывающие операторы и последовательность $(y^{(n)})_{n \in \mathbb{N}}, y^{(n)} \in Y$, компактна, то последовательность $(q_n y^{(n)})_{n \in \mathbb{N}}$ $\mathcal{Q}$ -компактна.

4. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим уравнение

$$\sum_{v=0}^{m-1} \left(\sum_{\mu=0}^{l_v} a_{v,\mu}(t) x^{(\alpha_{v,\mu}(t))}(t) + \sum_{\eta=0}^{k_v} \int_0^1 g_{v,\eta}(t, \tau) x^{(\beta_{v,\eta}(\tau))}(\tau) d\tau \right) = y(t), \quad t \in [0, 1], \quad (10)$$

с начальными условиями

$$x^{(v)}(0) = 0, \quad v = 0, 1, \dots, m-1. \quad (11)$$

Здесь приняты обозначения

$$x^{(\alpha_{v,\mu})} = D^{\alpha_{v,\mu}} x, \quad x^{(\beta_{v,\eta})} = D^{\beta_{v,\eta}} x, \\ \mu = 0, 1, \dots, l_v, \quad \eta = 0, 1, \dots, k_v, \quad v = 0, 1, \dots, m-1, \quad m \in \mathbb{N}.$$

Для индексов v, μ, η и функций $\alpha_{v,\mu}, \beta_{v,\eta}$ на $[0, 1]$ предполагаются выполненными условия

$$v \leq \alpha_{v,\mu}(t) < v+1, \quad \mu = 0, 1, \dots, l_v, \\ v \leq \beta_{v,\eta}(t) < v+1, \quad \eta = 0, 1, \dots, k_v, \quad v = 0, 1, \dots, m-1. \quad (12)$$

5. АНАЛИЗ РАЗРЕШИМОСТИ

По определению (3) с учетом условий (2) в задаче (7), (8) сделаем замену

$$x^{(\alpha_{v,\mu}(t))}(t) = (1 - \{\alpha_{v,\mu}(t)\})x^{(v)}(t) + \{\alpha_{v,\mu}(t)\}x^{(v+1)}(t), \quad \mu = 0, 1, \dots, l_v, \\ x^{(\beta_{v,\eta}(t))}(t) = (1 - \{\beta_{v,\eta}(t)\})x^{(v)}(t) + \{\beta_{v,\eta}(t)\}x^{(v+1)}(t), \quad \eta = 0, 1, \dots, k_v, \\ v = 0, 1, \dots, m-1.$$

Получим уравнение

$$\sum_{v=0}^m (b_v(t) x^{(v)}(t) + \int_0^1 f_v(t, \tau) x^{(v)}(\tau) d\tau) = y(t), \\ b_0(t) = \sum_{\mu=0}^{l_0} (1 - \{\alpha_{0,\mu}(t)\}) a_{v,\mu}(t),$$

$$\begin{aligned}
b_v(t) &= \sum_{\mu=0}^{l_v} (1 - \{\alpha_{v,\mu}(t)\}) a_{v,\mu}(t) + \sum_{\mu=0}^{l_{v-1}} \{\alpha_{v-1,\mu}(t)\} a_{v,\mu}(t), \quad v = 1, 2, \dots, m-1, \\
b_m(t) &= \sum_{\mu=0}^{l_{m-1}} \{\alpha_{m-1,\mu}(t)\} a_{m-1,\mu}(t), \\
f_0(t, \tau) &= \sum_{\eta=0}^{k_0} (1 - \{\beta_{0,\eta}(\tau)\}) g_{0,\eta}(t, \tau), \\
f_v(t, \tau) &= \sum_{\eta=0}^{k_v} (1 - \{\beta_{v,\eta}(\tau)\}) g_{v,\eta}(t, \tau) + \sum_{\eta=0}^{k_{v-1}} \{\beta_{v-1,\eta}(\tau)\} g_{v-1,\eta}(t, \tau), \quad v = 1, 2, \dots, m-1, \\
f_m(t, \tau) &= \sum_{\eta=0}^{k_{m-1}} \{\beta_{m-1,\eta}(\tau)\} g_{m-1,\eta}(t, \tau), \quad t \in [0, 1], \quad \tau \in [0, 1].
\end{aligned} \tag{13}$$

Задачи (10), (11) и (13), (11) равносильны в том смысле, что они обе однозначно разрешимы или нет одновременно, и в случае однозначной разрешимости их решения совпадают.

Теорема 1. Пусть для задачи (10), (11) выполнены следующие условия.

A1. Функции $a_{v,\mu}$, $\mu = 0, 1, \dots, l_v$, $g_{v,\eta}$, $\eta = 0, 1, \dots, k_v$, $v = 0, 1, \dots, m-1$, и u непрерывны, причем $g_{v,\eta}$, $\eta = 0, 1, \dots, k_v$, $v = 0, 1, \dots, m-1$, непрерывны по каждому аргументу равномерно относительно другого аргумента.

A2: $\sum_{v=0}^{l_{m-1}} \{\alpha_{m-1,\mu}(t)\} a_{m-1,\mu}(t) \neq 0$, $t \in [0, 1]$.

A3: $\sum_{v=0}^{m-1} \|b_v/b_m\|_Y + \sum_{v=0}^m \|f_v/b_m\|_Y < 1$.

Тогда задача (13), (11) и, следовательно, задача (10), (11) однозначно разрешимы при любой правой части $y \in Y$.

Доказательство. Задачу (13), (11) запишем в виде операторного уравнения

$$Kx \equiv D^m x + Bx = y/b_m, \quad K : X \rightarrow Y,$$

$$D^m : X \rightarrow Y,$$

$$(Bx)(t) = \left(\sum_{v=0}^{m-1} b_v(t) x^{(v)}(t) + \sum_{v=0}^m \int_0^1 f_v(t, \tau) x^{(v)}(\tau) d\tau \right) / b_m(t), \quad B : X \rightarrow Y,$$

и оценим норму оператора $B : X \rightarrow Y$. Зафиксируем произвольный элемент $x \in X$ и оценим норму Bx в пространстве Y :

$$\|Bx\|_Y = \max_{0 \leq t \leq 1} \left\| \left(\sum_{v=0}^{m-1} b_v(t) x^{(v)}(t) + \sum_{v=0}^m \int_0^1 f_v(t, \tau) x^{(v)}(\tau) d\tau \right) / b_m(t) \right\| \leq \sum_{v=0}^{m-1} \|b_v/b_m\|_Y \|x^{(v)}\|_Y + \sum_{v=0}^m \|f_v/b_m\|_Y \|x^{(v)}\|_Y.$$

По теореме Лагранжа для любого $x \in X$ выполняются равенства

$$\|x^{(v)}\|_Y \leq \|x^{(v+1)}\|_Y, \quad v = 0, 1, \dots, m-1,$$

поэтому

$$\|x^{(v)}\|_Y \leq \|x\|_X, \quad v = 0, 1, \dots, m,$$

и, следовательно,

$$\|Bx\|_Y \leq \left(\sum_{v=0}^{m-1} \|b_v/b_m\|_Y + \sum_{v=0}^m \|f_v/b_m\|_Y \right) \|x\|_X.$$

Это означает, что

$$\|B\|_{X \rightarrow Y} \leq \left(\sum_{v=0}^{m-1} \|b_v/b_m\|_Y + \sum_{v=0}^m \|f_v/b_m\|_m \right).$$

По лемме 1, учитывая, что $\|D^m\|_{X \rightarrow Y} = 1$, получаем, что при

$$u = \sum_{v=0}^{m-1} \|b_v/b_m\|_Y + \sum_{v=0}^m \|b_v/b_m\|_Y < 1$$

оператор $K : X \rightarrow Y$ обратим, т.е. задача (13), (11) и, следовательно, задача (10), (11) однозначно разрешимы при любой правой части $y \in Y$. Более того, это означает, что обратный оператор $\|K^{-1}\|_{Y \rightarrow X}$ ограничен

$$\|K^{-1}\|_{Y \rightarrow X} \leq \frac{1}{1-u},$$

поэтому для решения $x^* \in X$ задач (13), (11) и (10), (11) справедлива оценка

$$\|x^*\|_X \leq \frac{\|y/b_m\|_Y}{1-u}.$$

Теорема 1 доказана.

Условия теоремы 1 являются лишь достаточными для однозначной разрешимости задачи (10), (11). На самом деле класс разрешимых задач вида (10), (11) значительно шире. Тем не менее теорема 1 необходима для того, чтобы предположение о существовании решения этой задачи в теореме 2 не было пустым.

6. ВЫЧИСЛИТЕЛЬНАЯ СХЕМА

Приближенное решение задачи (10), (11) будем искать в виде вектора

$$\mathbf{x}_n = (x_0, x_1, \dots, x_n) \in X_n. \quad (14)$$

Компоненты вектора (14) $[\mathbf{x}_n]_k = x_k$, $k = 0, 1, \dots, n$, приближенные значения искомой функции в узлах t_k , $k = 0, 1, \dots, n$, найдем из системы линейных алгебраических уравнений

$$\sum_{v=0}^m (b_v(t_k) [D_n^v \mathbf{x}_n]_k + \sum_{l=0}^n \xi_{l,v} f_v(t_k, t_l) [D_n^v \mathbf{x}_n]_l) = y(t_k), \quad k = 0, 1, \dots, n. \quad (15)$$

Здесь $D_n^v \mathbf{x}_n$, $v = 0, 1, \dots, m$, – формулы численного дифференцирования вида (5), использующие только значения x_k , $k = \dots, -m+1, -m+2, \dots, n$, причем $x_k = 0$, $k = \dots, -m+1, -m+2, \dots, 0$, а

$$\sum_{l=0}^n \xi_{l,v} f_v(t_k, t_l) [D_n^v \mathbf{x}_n]_l, \quad v = 0, 1, \dots, m,$$

некоторые квадратурные формулы такие, что остаточные члены

$$\rho_{n,v}(w_v) = \int_0^1 w_v(\tau) d\tau - \sum_{l=0}^n \xi_{l,v} w_v(t_l)$$

стремятся к нулю при $n \rightarrow \infty$ для любых непрерывных на $[0, 1]$ функций w_v , $v = 0, 1, \dots, m$.

7. ОБОСНОВАНИЕ КВАДРАТУРНО-РАЗНОСТНОГО МЕТОДА

Теорема 2. Пусть для задачи (10), (11) и вычислительной схемы (14), (15) выполнены условия A1, A2 теоремы 1 и, кроме того, следующие условия.

A4. Задача (10), (11) имеет единственное решение $x^* \in X$.

B1. Формулы численного дифференцирования (5) $\mathcal{P}\mathcal{Q}$ -сходятся.

B2. Характеристические значения оператора D_n^m по модулю отличны от 1.

Тогда при достаточно больших n система уравнений (15) однозначно разрешима и приближенные решения x_n^* сходятся к точному решению x^* задачи (10), (11) при $n \rightarrow \infty$ со скоростью

$$\|x_n^* - p_n x^*\|_{X_n} \leq c(\varepsilon_n + \delta_n), \quad (16)$$

где

$$\varepsilon_n = \max_{0 \leq v \leq m} \|D_n^v p_n x^* - q_n D_n^v x^*\|_{Y_n}, \quad \delta_n = \max_{0 \leq v \leq m} \|q_n \rho_{n,v}(f_v D_n^v x^*)\|_{Y_n}.$$

Доказательство. Пусть $z = D^m x$ и представим задачу (10), (11) в виде равносильного ей операторного уравнения

$$z + Tz = y/b_m, \quad T = BD^{-m} : Y \rightarrow Y. \quad (17)$$

Здесь равносильность понимается в том смысле, что если задача (10), (11) имеет решение $x^* \in X$, то $z^* = D^m x^* \in Y$ будет решением уравнения (17).

Так как условие B2 теоремы 2 означает, что оператор $D_n^m : X_n \rightarrow Y_n$ обратим, то пусть $z_n = D_n^m x_n$ и представим также систему уравнений (15) в виде операторного уравнения

$$\begin{aligned} z_n + T_n z_n &= q_n(y/b_m), \quad T_n = B_n D_n^{-m} : Y_n \rightarrow Y_n, \\ B_n x_n &= \left(\sum_{v=0}^{m-1} (b_v(t_0)/b_m(t_0)) [D_n^v x_n]_0 + \sum_{v=0}^m \sum_{l=0}^n \xi_{l,v}(f_v(t_0, t_l)/b_m(t_0)) [D_n^v x_n]_l, \right. \\ &\quad \sum_{v=0}^{m-1} (b_v(t_1)/b_m(t_1)) [D_n^v x_n]_1 + \sum_{v=0}^m \sum_{l=0}^n \xi_{l,v}(f_v(t_1, t_l)/b_m(t_1)) [D_n^v x_n]_l, \dots, \\ &\quad \left. \sum_{v=0}^{m-1} (b_v(t_n)/b_m(t_n)) [D_n^v x_n]_n + \sum_{v=0}^m \sum_{l=0}^n \xi_{l,v}(f_v(t_n, t_l)/b_m(t_n)) [D_n^v x_n]_l \right). \end{aligned} \quad (18)$$

Покажем, что для уравнений (17) и (18) выполняются все условия леммы 2.

По условию A4 теоремы 2 задача (10), (11) имеет единственное решение $x^* \in X$, тогда, как указано выше, $z^* = D^m x^*$ будет решением уравнения (17), поэтому условие 4) леммы 2 выполнено.

Условие 3) леммы 2 означает, что элемент $z \in TY$ имеет норму $\|z\|_Y > 0$. Оператор q_n является связывающим отображением, поэтому по определению 1 имеем

$$\lim_{n \rightarrow \infty} \|q_n z\|_{Y_n} = \|z\|_Y > 0,$$

и, значит, условие 3) леммы 2 также выполнено.

Условие 2) леммы 2 выполняется тривиально.

Проверим выполнение условия 1).

Так как оператор $D^m : X \rightarrow Y$ ограничен, то все операторы $D^v : X \rightarrow Y$, $v = 0, 1, \dots, m-1$, вполне непрерывны, а вместе с ними, с учетом условия A1, вполне непрерывен и оператор $\sum_{v=0}^{m-1} b_v D^v : X \rightarrow Y$. Интегралы с непрерывными ядрами f_v , $v = 0, 1, \dots, m$, представляют собой (см., например, [54]) вполне непрерывные операторы, поэтому оператор $B : X \rightarrow Y$ вполне непрерывен. Так как оператор $D^{-m} : Y \rightarrow X$ ограничен, то композиция $T = BD^{-m} : Y \rightarrow Y$ является вполне непрерывным оператором.

Проверим теперь $\mathcal{Q}\mathcal{Q}$ -сходимость операторов T_n , $n \in \mathbb{N}$, к оператору T . Покажем вначале, что последовательность операторов $D_n^{-m} : Y_n \rightarrow X_n$, $n \in \mathbb{N}$, $\mathcal{Q}\mathcal{P}$ -сходится к оператору $D^{-m} : Y \rightarrow X$.

Используя равномерную непрерывность и ограниченность на $[0,1]^2$ функций $f_v(t, \tau)/b_m(t)$, $v = 0, 1, \dots, m$, с помощью теоремы Арцела легко убедиться, что последовательности

$$w_{n,v}(t) = \sum_{l=0}^n \xi_{l,v}(f_v(t, t_l)/b_m(t)) [D_n^v \mathbf{x}_n]_l, \quad v = 0, 1, \dots, m,$$

компактны в Y . Вновь применяя лемму 3, получим, что последовательности

$$q_n w_{n,v} = \left(\sum_{l=0}^n \xi_{l,v}(f_v(t_0, t_l)/b_m(t_0)) [D_n^v \mathbf{x}_n]_l, \sum_{l=0}^n \xi_{l,v}(f_v(t_1, t_l)/b_m(t_1)) [D_n^v \mathbf{x}_n]_l, \dots \right. \\ \left. \dots, \sum_{l=0}^n \xi_{l,v}(f_v(t_n, t_l)/b_m(t_n)) [D_n^v \mathbf{x}_n]_l \right), \quad v = 0, 1, \dots, m, \quad (20)$$

а значит, и их сумма $\mathcal{Q}$ -компактны.

Суммируя последовательности (19) и (20), получаем, что последовательность $B_n \mathbf{x}_n$, а с учетом ограниченности операторов D_n^{-m} , и последовательность $T_n \mathbf{z}_n = B_n D_n^{-m} \mathbf{z}_n$ $\mathcal{Q}$ -компактны. Таким образом, условие 1) леммы 2 также выполнено.

По лемме 2 получаем, что для достаточно больших n уравнение (18) имеет единственное решение $\mathbf{z}_n^*$ и верна оценка

$$\|\mathbf{z}_n^* - q_n \mathbf{z}^*\|_{Y_n} \leq c \|q_n T \mathbf{z}^* - T_n q_n \mathbf{z}^*\|_{Y_n}. \quad (21)$$

Для завершения доказательства теоремы 2 осталось получить оценку (16). Используя замены переменных $\mathbf{z}_n^* = D_n^m \mathbf{x}_n^*$ и $\mathbf{z}^* = D^m \mathbf{x}^*$, получаем

$$\|\mathbf{x}_n^* - p_n \mathbf{x}^*\|_{X_n} \leq \|D_n^{-m}\|_{Y_n \rightarrow X_n} \|D_n^m \mathbf{x}_n^* - D_n^m p_n \mathbf{x}^*\|_{Y_n} \leq \|V_n^{-m}\|_{Y_n \rightarrow Y_n} \|\mathbf{z}_n^* - D_n^m p_n \mathbf{x}^*\|_{Y_n} \leq \\ \leq \|V_n^{-m}\|_{Y_n \rightarrow Y_n} \left(\|\mathbf{z}_n^* - q_n \mathbf{z}^*\|_{Y_n} + \|q_n D^m \mathbf{x}^* - D_n^m p_n \mathbf{x}^*\|_{Y_n} \right).$$

Из (21) следует, что

$$\|\mathbf{x}_n^* - p_n \mathbf{x}^*\|_{X_n} \leq c \left(\|q_n T \mathbf{z}^* - T_n q_n \mathbf{z}^*\|_{Y_n} + \|q_n D^m \mathbf{x}^* - D_n^m p_n \mathbf{x}^*\|_{Y_n} \right). \quad (22)$$

Оценим отдельно первое слагаемое правой части оценки (22)

$$\|q_n T \mathbf{z}^* - T_n q_n \mathbf{z}^*\|_{Y_n} = \|q_n B \mathbf{x}^* - B_n D_n^{-m} q_n D^m \mathbf{x}^*\|_{Y_n} \leq \|q_n B \mathbf{x}^* - B_n p_n \mathbf{x}^*\|_{Y_n} + \\ + \|B_n\|_{Y_n \rightarrow Y_n} \|p_n \mathbf{x}^* - D^{-m} q_n D^m \mathbf{x}^*\|_{Y_n} \leq \|q_n B \mathbf{x}^* - B_n p_n \mathbf{x}^*\|_{Y_n} + \|B_n\|_{Y_n \rightarrow Y_n} \|D_n^{-m}\|_{Y_n \rightarrow Y_n} \|D_n^m p_n \mathbf{x}^* - q_n D^m \mathbf{x}^*\|_{Y_n}.$$

Возвращаясь к оценке (22), получаем

$$\|\mathbf{x}_n^* - p_n \mathbf{x}^*\|_{X_n} \leq c \left(\|q_n D \mathbf{x}^* - B_n p_n \mathbf{x}^*\|_{Y_n} + \left(1 + \|B_n\|_{Y_n \rightarrow Y_n} \|D_n^{-m}\|_{Y_n \rightarrow X_n} \right) \|D_n^m p_n \mathbf{x}^* - q_n D^m \mathbf{x}^*\|_{Y_n} \right) \leq \\ \leq c \left(\|D_n^m p_n \mathbf{x}^* - q_n D^m \mathbf{x}^*\|_{Y_n} + \|q_n B \mathbf{x}^* - B_n p_n \mathbf{x}^*\|_{Y_n} \right) \leq c(\varepsilon_n + \delta_n), \\ \varepsilon_n = \max_{0 \leq v \leq m} \|D_n^v p_n \mathbf{x}^* - q_n D^v \mathbf{x}^*\|_{Y_n}, \quad \delta_n = \max_{0 \leq v \leq m} \|q_n p_{n,v}(f_v D^v \mathbf{x}^*)\|_{Y_n}.$$

Теорема 2 доказана полностью.

8. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

Рассмотрим уравнение вида (10)

$$(t^2 + 1)x^{(\alpha_1(t))}(t) + t^3 x^{(\alpha_0(t))}(t) + \int_0^1 (t - \tau) x^{(\beta_1(\tau))}(\tau) d\tau + \int_0^1 (t - \tau^2) x^{(\beta_0(\tau))}(\tau) d\tau = y(t), \quad t \in [0, 1], \quad (23)$$

с производными искомой функции переменных и дробных порядков

$$\alpha_0(t) = \frac{1}{2}t^2, \quad 0 \leq \alpha_0(t) < 1, \quad \alpha_1(t) = \frac{3}{2}, \quad 1 \leq \alpha_1(t) < 2,$$

$$\beta_0(t) = \frac{1}{2}, \quad 0 \leq \beta_0(t) < 1, \quad \beta_1(t) = \frac{3}{2} - \frac{1}{2}t, \quad 1 \leq \beta_1(t) < 2,$$

удовлетворяющих на $[0, 1]$ условиям (2), с начальными условиями

$$x(0) = 0, \quad x'(0) = 0.$$

По определениям (1), (3) производных переменных и дробных порядков найдем

$$x^{(\alpha_1(t))}(t) = \frac{1}{2}x''(t) + \frac{1}{2}x'(t), \quad x^{(\alpha_0(t))}(t) = \frac{1}{2}t^2x'(t) + \left(1 - \frac{1}{2}t^2\right)x(t), \quad t \in [0, 1],$$

$$x^{(\beta_1(\tau))}(\tau) = \left(\frac{1}{2} + \frac{1}{2}\tau\right)x''(\tau) + \left(\frac{1}{2} - \frac{1}{2}\tau\right)x'(\tau), \quad x^{(\beta_0(\tau))}(\tau) = \frac{1}{2}x'(\tau) + \frac{1}{2}x(\tau), \quad \tau \in [0, 1]. \quad (24)$$

Заменяя производные переменных и дробных порядков их выражениями (24), перейдем от уравнения (23) к равносильному ему уравнению

$$(t^2 + 1)x''(t) + (1 + t^5)x'(t) + (2t^3 - t^5)x(t) + \int_0^1 (t - \tau)(\tau + 1)x''(\tau)d\tau +$$

$$+ \int_0^1 (2t - (t + 1)\tau)x'(\tau)d\tau + \int_0^1 (t - \tau^2)x(\tau)d\tau = 2y(t), \quad t \in [0, 1]. \quad (25)$$

Подставляя в уравнение (25) функцию $x(t) = t^2 + \sin^2(t)$, получаем

$$2y(t) = \frac{7}{15} + 5\frac{17}{30}t + 2t^5 + 2t^6 - t^7 + \left(1\frac{1}{2} - 1\frac{5}{8}\right)\sin 2 - \frac{3}{4}\cos 2 +$$

$$+ (4t^2 + 2t^3 - t^5)\sin^2 t + 2\cos 2t + (1 + t^5)\sin 2t. \quad (26)$$

Для уравнения (23) с правой частью (26) функция $x(t) = t^2 + \sin^2 t$ будет точным решением.

Зафиксируем $n \in \mathbb{N}$. По предлагаемому в статье методу приближенное решение уравнения (23) будем искать в виде вектора значений $\mathbf{x}_n = (x_0, x_1, \dots, x_n)$ искомой функции в узлах сетки $t_k = k/n$, $k = 0, 1, \dots, n$. Значения производных искомой функции в узлах сетки будем приближать простейшими разностями

$$x''(t_k) \sim n^2(x_k - 2x_{k-1} + x_{k-2}), \quad x'(t_k) \sim n(x_k - x_{k-1}), \quad k = 1, 2, \dots$$

Интегралы аппроксимируем квадратурными формулами трапеций

$$\int_0^1 (t_k - \tau)(\tau + 1)x''(\tau)d\tau \sim n \sum_{l=1}^{n-1} (t_k - t_l)(t_l + 1)(x_l - 2x_{l-1} + x_{l-2}) + \frac{n}{2}(x_n - 2x_{n-1} + x_{n-2}),$$

$$\int_0^1 (2t_k - (t_k + 1)\tau)x'(\tau)d\tau \sim \sum_{l=1}^{n-1} (2t_k - (t_k + 1)t_l)(x_l - x_{l-1}) + \frac{1}{2}(t_k - 1)(x_n - x_{n-1}),$$

$$\int_0^1 (T_k - \tau^2)x(\tau)d\tau \sim \frac{1}{n} \sum_{l=1}^{n-1} (t_k - t_l^2)x_l + \frac{1}{2n}(t_k - 1)x_n.$$

Заменяя в уравнении (25) производные искомой функции конечными разностями, а интегралы квадратурными суммами, получаем систему уравнений

$$(t_k^2 + 1)n^2(x_k - 2x_{k-1} + x_{k-2}) + (1 + t_k^5)n(x_k - x_{k-1}) + (2t_k^3 - t_k^5)x_k +$$

$$+ n \sum_{l=1}^{n-1} (t_k - t_l)(t_l + 1)(x_l - 2x_{l-1} + x_{l-2}) + \frac{n}{2}(x_n - 2x_{n-1} + x_{n-2}) + \sum_{l=1}^{n-1} (2t_k - (t_k + 1)t_l)(x_l - x_{l-1}) +$$

Таблица 1. Оценки погрешностей приближенного решения задачи (1), (2) разностей и квадратур

n	$\ x_n^* - p_n x^*\ _{X_n}$	ϵ_n	δ_n
10	0.3424108	0.3972079	0.393998
20	0.1654387	0.1997317	0.199488
30	0.1065912	0.1333025	0.133253
40	0.0788687	0.1004065	0.999968
50	0.0666081	0.0804866	0.080014
60	0.0613689	0.0674291	0.066683
70	0.0549078	0.0580792	0.057159

$$\begin{aligned}
& + \frac{1}{2}(t_k - 1)(x_n - x_{n-1}) + \frac{1}{n} \sum_{l=1}^{n-1} (t_k - t_l^2)x_l + \frac{1}{2n}(t_k - 1)x_n = \frac{7}{15} + 5\frac{17}{30}t_k + 2t_k^5 + 2t_k^6 - t_k^7 + \\
& + \left(1\frac{1}{2} - 1\frac{5}{8}\right)\sin 2 - \frac{3}{4}\cos 2 + (4t_k^2 + 2t_k^3 - t_k^5)\sin^2 t_k + 2\cos 2t_k + (1 + t_k^5)\sin 2t_k, \\
& k = 1, 2, \dots, n, \quad x_{-1} = x_0 = 0,
\end{aligned} \quad (27)$$

квадратурно-разностного метода решения задачи (10), (11).

В табл. 1 приведены результаты решений системы уравнений (27). В первой графе указано число узлов n , во второй — погрешность приближенного решения $\|x_n^* - p_n x^*\|_{X_n}$, в третьей и четвертой — погрешности разностей ϵ_n и квадратур δ_n соответственно. Результаты вычислительных экспериментов подтверждают теретические выводы теоремы 2.

ЗАКЛЮЧЕНИЕ

Появление данной работы мотивировано трудностями, возникшими у автора при попытках конструирования и обоснования численных методов решения уравнений с дробными производными. Главным неудобством была нелокальность производных, определенных через интегралы. Действительно, простые разностные схемы решения дифференциальных уравнений используют локальность производных натуральных порядков и при ее отсутствии становятся неприменимыми. Идею определить производные дробного порядка интерполированием производных целого порядка подсказали переписка Лейбница с Лопиталем и замечания Эйлера.

В данной работе определение производных дробного порядка дано самым простым способом: на основании интерполирования сплайнами первого порядка. Очевидно, что возможны и более сложные виды интерполирования. Так же, из соображения простоты и наглядности, для вычислительной схемы были выбраны простейшие формулы численного дифференцирования. И здесь для увеличения скорости сходимости метода можно использовать более точные (но при этом и более сложные) формулы численного дифференцирования. Существенным здесь является то, что при введенном определении дробных производных все результаты по приближенным методам решения уравнений натуральных порядков легко переносятся на случай уравнений с дробными производными.

В дальнейших исследованиях предполагается непосредственное сравнение преимуществ/недостатков введенных производных на задачах, решенных с помощью призводных дробного порядка других авторов.

СПИСОК ЛИТЕРАТУРЫ

1. Самко С.Г., Килбас А.А., Маричев О.И. Интегралы и производные дробного порядка и некоторые их приложения. Минск: Наука и техн., 1987. 688 с.
2. Ross B. The development of fractional calculus 1695–1900 // Historia math. 1977. № 4. P. 75–89.

3. Бутковский А.Г., Постнов С.С., Постнова Е.А. Дробное интегро-дифференциальное исчисление и его приложения в теории управления. I. Математические основы и проблема интерпретации // Автомат. и телемехан. 2013. Вып. 4. С. 3–42.
4. Khalil R., Al Horani M., Yousef A., Sababhehb M. A new definition of fractional derivative // J. of Comput. and Appl. Math. 2014. № 264. P. 65–70.
5. Rezapour M., Sijuwade A., Asaki T.J. A new sigmoidal fractional derivative for regularization // <https://arxiv.org/pdf/2001.01610.pdf>
6. Khalid Hattaf. A new generalized definition of fractional derivative with non-singular kernel // Comput. 2020. 8. 49. P. 1–9. www.mdpi.com/journal/computation
7. Diethelm K., Garrappa R., Giusti A., Stynes M. Why fractional derivatives with nonsingular kernels should not be used // Fract. Calc. Appl. Anal. 2020. 23. № 3. P. 610–634.
8. Чуриков В.А. Программа и принципы построения дробного анализа // Известия Томского политехнического университета. 2009. Т. 314. № 2. С. 9–12.
9. Hadamar J. Essai sur l'étude des fonctions donnees par leur development de Taylor // J. Math. Pures et Appl. 1892. Ser. 4. V. VIII. P. 101–186.
10. Samko S.G. Fractional integration and differentiation of variable order // Anal. Math. 1995. V. 218. P. 213–236.
11. Lorenzo C.F., Hartley T.T. Variable order and distributed order fractional operators // Nonlin. Dyn. 2002. V. 29. P. 57–98.
12. Valerio D., da Costa J.S. Variable-order fractional derivatives and their numerical approximations // Signal Proc. 2011. V. 91. P. 470–483.
13. Sun H., Chen Y., Chen W. Time fractional differential equation model with random derivative order // Proc. ASME int.design engin. tecnical conf & computes and inform in engin. Conf. IDETC/CIE 2009. San Diego, 2009. Paper ID DETC2009-87483 (6 pages)
14. Al-Salam W.A., Verma A. A fractional Leibniz q-formula // Pac. J. Math. 1975. V. 60. P. 1–9.
15. Al-Salam W.A. Some fractional q-integrals and q-derivatives // Proc. Edin. Math. Soc. 1969. V. 15. P. 135–140.
16. Agarwal R.P. Certain fractional q-integrals and q-derivatives // Proc. Camb. Phil. Soc. 1969. V. 66. P. 365–370.
17. Predrag M.R., Sladana D.M., Miomir S.S. Fractional integrals and derivatives in q-calculus // Appl. Anal. Discr. Math. 2007. V. 1. P. 311–323.
18. Atici F.M., Eloe P.W. A transform method in discrete fractional calculus // Int. J. Differ. Equat. 2007. V. 2. № 2. P. 165–176.
19. Atici F.M., Eloe P.W. Initial value problems in discrete fractional calculus // Proc. Amer. Math Soc. 2009. V. 137. P. 981–989.
20. Atici F.M., Eloe P.W. Fractional q-calculus on a time scale // J. Nonlin. Math. Phys. 2007. V. 14. № 3. P. 341–352.
21. Holm M.T. The Laplace transform in discrete fractional calculus // Comput. Math. Appl. 2011. V. 62. P. 1591–1601.
22. Abdeljawad T., Baleanu D. Fractional differences and integration by parts // J. Comput. Anal. Appl. 2011. V. 13. № 3. P. 574–582.
23. Abdeljawad T., Baleanu D. Caputo q-fractional initial value problems and a q-analogue Mittag-Leffler function // Commun. Nonlin.Sci. Numer. Simulat. 2011. V. 16. P. 4682–4688.
24. Abdeljawad T. On Riemann and Caputo fractional differences // Comput. Math. Appl. 2011. V. 62. P. 1602–1611.
25. Miyakoda T. Direct discretization of the fractional-order differential by using Chebyshev series expansion // Proc. Appl. Math. Mech. 2007. V. 7. P. 2020011–2020012.
26. Miller K.S., Ross B. An introduction to the fractional calculus and fractional differential equations. New York: John Wiley & Sons, 1993.
27. Учайкин В.В. Метод дробных производных. Ульяновск: Артишок, 2008.
28. Nishimoto K. An essence of Nishimoto's fractional calculus (Calculus of the 21st century), Integrals and differentiations of arbitrary order. Koriyama: Descartes Press, 1991.
29. Нахушев А.М. Дробное исчисление и его применение М.: Физматлит, 2003.
30. Kilbas A.A., Srivastava H.M., Trujillo J.J. Theory and applications of fractional differential equations. Amsterdam: Elsevier, 2006.
31. Podlubny I. Fractional differential equations. San Diego: Academic Press, 1999.
32. Diethelm K. The analysis of fractional differential equations. Berlin: Springer, 2010.
33. Псху А.В. Уравнения в частных производных дробного порядка. М.: Наука, 2005.
34. Lakshmikantham V., Leela S., Vasundhara D.J. Theory of fractional Dynamic systems. Cambridge: Cambridge Academic Publishers, 2009.
35. Caponetto R., Dongola G., Fortuna L., Petras I. Fractional Order systems. Modeling and control applications. Singapore: World Scientific, 2010.

36. *Luchko Y., Yamamoto M.* The general fractional derivative and related fractional differential equations // *Math.* 2020. 8. 2115.
37. *Kochubei A.N.* General fractional calculus, evolution equations, and renewal processes // *Integr. Equa. Oper. Theory.* 2011. V. 71. P. 583–600.
38. *Kochubei A.N.* General fractional calculus // *Handbook of Fractional Calculus with Applications; Volume 1: Basic Theory*, Berlin, Germany; Boston, MA, USA. 2019. P. 111–126.
39. *Kochubei A.N.* Equations with general fractional time derivatives. Cauchy problem // *Handbook of Fractional Calculus with Applications; Volume 2: Fractional Differential Equations*; Berlin, Germany; Boston, MA, USA. 2019. P. 223–234.
40. *Zwillinger D.* Handbook of differential equations. New York: Academic Press, 1997.
41. *Агачев Ю.Р., Гуськова А.В.* Обобщенный полиномиальный метод решения задачи типа Коши для одного дробно-дифференциального уравнения // *Итоги науки и техн. Сер. Современ. мат. и ее прил. Темат. обз.* 2020. Т. 176. С. 80–90.
42. *Бештоков М.Х., Худалов М.З.* Третья краевая задача для нагруженного уравнения теплопроводности с дробной производной Капуто // *Матем. и матем. моделирование.* 2020. № 3. С. 52–64.
43. *Бештоков М.Х.* Нелокальные краевые задачи в дифференциальной и разностной трактовках для обобщенного нагруженного уравнения влагопереноса // *Дифференц. ур-ния и процессы управления.* 2020. № 4. С. 1–27. <http://diffjournal.spbu.ru/>
44. *Алимбеков Н.Б., Байгереев Д.Р., Мадияров М.Н.* Исследование численного метода решения краевой задачи для дифференциального уравнения с дробной производной по времени // *Известия АлтГУ. Матем. и механ.* 2020. № 4 (114) С. 64–69.
45. *Minh Duc Tran, Vu Ho, Hoa Ngo Van.* On the stability of Fractional differential equations involving generalized Caputo fractional derivative // *Hindawi Mathematical Problems in Engineering Volume.* 2020. Article ID 1680761, 14 pages
<https://doi.org/10.1155/2020/1680761>
46. *Hilfer R., Luchko Yu.* Disiderata for fractional derivatives and integrals // *Mathematics.* 2019. 7. 149.
<https://doi.org/10.3390/math7020149>
47. *Вайникко Г.М.* Анализ дискретизационных методов. Тарту: Изд-во Тартуского ун-та, 1976. 161 с.
48. *Вайникко Г.М.* О сходимости квадратурно-разностного метода для линейных интегро-дифференциальных уравнений // *Ж. вычисл. матем. и матем. физ.* 1971. Т. 11. № 3. 1971. С. 770–776.
49. *Крейн С.Г., Шаблицкая Л.Н.* Об устойчивости разностных схем для задачи Коши // *Ж. вычисл. матем. и матем. физ.* 1966. Т. 6. № 4. С. 648–664.
50. *Крейн С.Г., Шаблицкая Л.Н.* Необходимые условия устойчивости разностных схем и собственные значения разностных операторов // *Ж. вычисл. матем. и матем. физ.* 1973. Т. 13. № 3. С. 647–657.
51. *Гудович Н.Н.* Об абстрактной схеме разностного метода // *Ж. вычисл. матем. и матем. физ.* 1966. Т. 6. № 5. С. 916–921.
52. *Гудович Н.Н.* О построении устойчивых разностных схем любого наперед заданного порядка для линейных обыкновенных дифференциальных уравнений // *Докл. АН СССР.* 1974. Т. 217. № 2. С. 264–267.
53. *Красносельский М.А., Вайникко Г.М., Забрейко П.П., Рунтцкий Я.Б., Стеценко В.Я.* Приближенное решение операторных уравнений. М.: Наука, 1969. 418 с.
54. *Люстерник Л.А., Соболев В.И.* Элементы функционального анализа. 2-е издание. М.: Наука, 1965. 520 с.

ОБЩИЕ
ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.651

**О РАВНОМЕРНОЙ МОНОТОННОЙ АППРОКСИМАЦИИ
НЕПРЕРЫВНЫХ МОНОТОННЫХ ФУНКЦИЙ С ПОМОЩЬЮ
СДВИГОВ И СЖАТИЙ ИНТЕГРАЛА ЛАПЛАСА**© 2022 г. А. В. Чернов^{1,2}¹ 603950 Нижний Новгород, пр-т Гагарина, 23, Нижегородский гос. ун-т им. Н.И. Лобачевского, Россия² 603950 Нижний Новгород, ул. Минина, 24, Нижегородский гос. технический ун-т им. Р.Е. Алексеева, Россия

e-mail: chavnn@mail.ru

Поступила в редакцию 14.10.2020 г.

Переработанный вариант 05.11.2021 г.

Принята к публикации 16.12.2021 г.

Для непрерывных монотонных функций, заданных на конечном отрезке $[-b; b]$, строится монотонная аппроксимация $Q(x)$ с любой заранее заданной точностью в метрике пространства $C[-b; b]$ с помощью сдвигов и сжатий функции (интеграла) Лапласа. В свою очередь, для функции Лапласа строится высокоточная аппроксимация в той же метрике суммой линейной функции и линейной комбинации квадратичных экспонент. Исследуется вопрос об устойчивости свойства монотонности функции $Q(x)$ при замене в ней интеграла Лапласа его аппроксимацией. Задача об аппроксимации непрерывной монотонной функции возникает, в частности, в рамках подхода к аппроксимации непрерывных функций многих переменных, основанного на сочетании теоремы А.Н. Колмогорова об их представлении функциями одного переменного (несколькими внешними и одной внутренней — монотонной), которые, собственно, и подвергаются аппроксимации. Соответствующий подход, в котором указанные внешние и внутренняя функции аппроксимировались линейными комбинациями квадратичных экспонент, исследовался автором ранее. Поскольку в качестве внутренней функции представления А.Н. Колмогорова выступает всегда одна и та же монотонная непрерывная функция Ψ одного переменного, то здесь возникает вопрос: как эффективным образом производить ее аппроксимацию с сохранением условия монотонности? Данная статья как раз и дает ответ на этот вопрос. Библ. 24. Фиг. 4.

Ключевые слова: непрерывные монотонные функции, равномерная аппроксимация, интеграл Лапласа, квадратичные экспоненты.

DOI: 10.31857/S0044466922040044**ВВЕДЕНИЕ**

Пусть $C[a; b]$ — пространство функций, непрерывных на отрезке $[a; b]$. Классическим подходом к аппроксимации функций из $C[a; b]$ является использование многочленов. Этот подход опирается на известную теорему Вейерштрасса о том, что всякую такую функцию можно с любой заданной точностью в метрике пространства $C[a; b]$ аппроксимировать многочленом. Получить наилучшее представление непрерывной функции многочленом позволяет теория П.Л. Чебышева. Уместно упомянуть здесь теоремы типа С.М. Никольского, которые позволяют при фиксированном $s > 0$ указать класс непрерывных функций, каждая из которых для всякого $n \in \mathbb{N}$ допускает равномерное на $[a; b]$ приближение с точностью порядка n^{-s} полиномами степени n (подробнее см. [1, гл. VI]). В случае, когда аппроксимируемая функция является достаточно гладкой всюду, за исключением одной или нескольких точек, более эффективно использование рациональных многочленов вида $\mathcal{P}_n(x)/\mathcal{Q}_n(x)$, где $\mathcal{P}_n(x)$, $\mathcal{Q}_n(x)$ — алгебраические многочлены степени n (см. [1, гл. VI]). (О других способах аппроксимации см. в [2, гл. 2, § 2.7, с. 66], [3, гл. 10].)

В теории обработки сигналов, изображений и т.д. используются также подходы, основанные, в частности, на теории вейвлетов (всплесков). Эти подходы оказались весьма эффективными в смысле возможности сжатия и преобразования информации (см. [4], [5]). Конкретное семейство вейвлетов порождается сдвигами и сжатиями из так называемого материнского вейвлета — неко-

торой функции, удовлетворяющей специальным условиям. Типичным примером является материнский вейвлет “мексиканская шляпа” (подробнее см. [4, гл. 3]). Сдвиги и сжатия квадратичной экспоненты (т.е. функции Гаусса) – это тоже квадратичные экспоненты. В [6] было показано, что материнский вейвлет “мексиканская шляпа” на любом конечном отрезке можно сколь угодно точно в равномерной метрике приблизить линейной комбинацией в точности двух квадратичных экспонент. Поэтому на фиксированном отрезке аппроксимация квадратичными экспонентами сопоставима по эффективности с аппроксимацией вейвлетом “мексиканская шляпа”. В [7] доказано, что любой многочлен степени n на любом конечном отрезке можно сколь угодно точно в равномерной метрике приблизить линейной комбинацией не более $(n + 1)$ квадратичной экспоненты. Отметим, что аппроксимация квадратичными экспонентами строится неоднозначно и допускает бесконечное количество вариантов и, вообще, имеет ряд общих черт с представлением фреймами (см. [5, § 1.8]). Что касается указанной неоднозначности, то для инженерных приложений она бывает полезна (см. [5, § 1.8]).

В представленном выше обзоре речь шла об аппроксимации без дополнительных условий. Однако бывают ситуации (например, при решении некоторых задач математической статистики, см. [8]), когда требуется строить монотонную аппроксимацию для монотонной непрерывной функции. В качестве содержательного примера такой ситуации укажем следующую. Прежде всего, приведем формулировку модификации теоремы А.Н. Колмогорова в стиле работы [9, разд. “Введение”, “Построение внутренних функций”], опирающейся в свою очередь на [10].

Лемма 1. *При произвольно заданных числах: натуральном $n \geq 2$, алгебраическом $\lambda > 0$ степени $\deg \lambda \geq n$, натуральном $\gamma \geq 2(n + 1)$ и $\sigma = 1/(\gamma - 1)$, существует действительная непрерывная монотонно возрастающая функция одного переменного $\Psi(y)$, $y \geq 0$, такая, что каждая определенная на n -мерном единичном кубе E^n непрерывная действительная функция $f(x_1, \dots, x_n)$ представима в виде*

$$f(x) = \sum_{q=0}^{2n} F_q[G_q(x)], \quad G_q(x) = \sum_{p=1}^n \lambda^p \Psi(x_p + \sigma q), \quad x = (x_1, \dots, x_n) \in E^n,$$

где семейство действительных непрерывных функций одного переменного $\{F_q(y), q = \overline{0, 2n}\}$ определяется по функции f .

Способ из [9] построения функции Ψ (в виде бесконечного ряда) сложный и многоэтапный. Поэтому его практическое использование затруднительно. При этом функции Ψ , F_q определяются неоднозначно.

В [11] было предложено аппроксимировать непрерывные функции многих переменных на основе леммы 1 и аппроксимации функций $\Psi(y)$, $F_q(y)$ линейными комбинациями квадратичных экспонент (там же см. теоретическое обоснование этого подхода, описание его преимуществ и в случае размерности 2 результаты численных экспериментов). Как видно из этих экспериментов, для различных функций $f(x)$ аппроксимация функции, играющей роль $\Psi(y)$ в выражении $G_q(x)$ из леммы 1, может выглядеть по-разному, быть не только возрастающей, но и убывающей (что соответствует умножению аргумента функций $F_q(y)$ на (-1) ; возможен сдвиг или растяжение этой функции); условие монотонности может оказаться нарушенным на некоторых участках. Указанное нарушение монотонности, вообще говоря, не удивительно, поскольку аппроксимируемая функция $\Psi(y)$ входит в аргумент внешних функций $F_q(y)$, которые одновременно с ней подвергаются аппроксимации. Может оказаться и так, что поведение функции Ψ на некоторых участках не оказывает влияния на суперпозиции внешних и внутренних функций. С другой стороны, нарушение монотонности, по крайней мере, по общим соображениям может приводить к замедлению или затруднению процесса сходимости при оптимизации квадратичной невязки, либо минимизации функционала в задаче бесконечномерной оптимизации, при конечномерной аппроксимации искомой функции, по параметрам аппроксимации. Естественно возникает вопрос о разработке эффективного способа монотонной аппроксимации функции $\Psi(y)$. Желание получить ответ на этот вопрос как раз и послужило исходной мотивацией написания данной статьи.

В настоящей работе для непрерывных монотонных функций, заданных на конечном отрезке $[-b; b]$, строится монотонная аппроксимация с любой заранее заданной точностью в метрике пространства $C[-b; b]$ с помощью сдвигов и сжатий функции (интеграла) Лапласа. Этот способ построения аппроксимации основан, в общих чертах, на следующих соображениях.

1. Всякая гладкая монотонная (неубывающая) функция может быть представлена с точностью до постоянного слагаемого в виде интеграла с переменным верхним пределом от квадрата непрерывной функции.

2. В свою очередь, непрерывная функция эффективно аппроксимируется линейной комбинацией квадратичных экспонент.

3. Интеграл от квадрата линейной комбинации квадратичных экспонент можно представить в виде линейной комбинации (некоторого специального шаблона) сдвигов и сжатий интеграла Лапласа. Отсюда следует вывод, что гладкую монотонную функцию можно эффективно аппроксимировать специальной линейной комбинацией сдвигов и сжатий интеграла Лапласа, обладающей свойством монотонности.

4. Для непрерывной монотонной функции всегда можно построить бесконечно дифференцируемую монотонную аппроксимацию (в виде многочлена). Таким образом, тот же вывод можно сделать и в отношении непрерывных монотонных функций. Для краткости построенную таким образом аппроксимацию будем обозначать $Q(x)$.

По сравнению с применением для аппроксимации непосредственно квадратичных экспонент, здесь можно усмотреть лишь некоторое неудобство, связанное с тем, что интеграл Лапласа не выражается в элементарных функциях. Но с практической точки зрения, при использовании любого математического пакета, квадратичная экспонента (так же, как и интеграл Лапласа) вычисляется все равно приближенно. Кроме того, существуют математические пакеты (в частности, MATLAB), в которых реализована *функция ошибки (error function)*: $\operatorname{erf}(x) = 2/\sqrt{\pi} \int_0^x e^{-t^2} dt$. Ясно,

что интеграл Лапласа $\hat{\Phi}_0(x) = 1/\sqrt{2\pi} \int_0^x e^{-t^2/2} dt = 1/2 \operatorname{erf}(x/\sqrt{2})$. Аналогично функции Гаусса, интеграл Лапласа обладает свойством локализации: достаточно знать его значения на $[0; 5]$, поскольку это нечетная функция, а при $x > 5$ значения ее практически не отличаются от $1/2$; на $[0; 5]$ из математической статистики известны таблицы значений, которые можно использовать в готовом виде, не тратя времени на вычисления (в точках между узлами сетки можно использовать линейную интерполяцию). Наконец, в разд. 3 мы показываем, что интеграл Лапласа можно с высокой степенью точности аппроксимировать суммой линейной функции и линейной комбинации двух-трех квадратичных экспонент, и указываем конкретные значения ее коэффициентов. Представляется, что этот последний результат интересен и сам по себе, поскольку интеграл Лапласа играет важную роль при решении многих прикладных задач, связанных с теорией вероятностей, математической статистикой и вычислительной математикой. При повышении количества квадратичных экспонент в линейной комбинации точность такой аппроксимации (обозначим ее $\hat{\Phi}_{0,\varepsilon}(x)$), разумеется, возрастает. При этом точность аппроксимации оценивается по разности вторых производных функций $\hat{\Phi}_0(x)$ и $\hat{\Phi}_{0,\varepsilon}(x)$ в пространстве L_2 и, соответственно, первых производных в пространстве $\mathbf{C}$. Более того, мы показываем, что при достаточной малости отклонения $\left| \hat{\Phi}'_0(x) - \hat{\Phi}'_{0,\varepsilon}(x) \right|$ в метрике $\mathbf{C}$ имеет место устойчивость свойства монотонности аппроксимации $Q(x)$ при замене в ней интеграла Лапласа на $\hat{\Phi}_{0,\varepsilon}(x)$. Кроме того, мы представляем результаты численных экспериментов по аппроксимации изучаемым способом различных монотонных функций: четырех гладких и одной непрерывной кусочно линейной. Во всех указанных случаях эффективность изучаемого способа аппроксимации оказалась достаточно высокой, а устойчивость свойства монотонности аппроксимации подтвердилась экспериментально.

Сделаем краткий обзор по работам, посвященным аппроксимации монотонных функций. По-видимому, впервые эта проблема была поставлена в [12], [13]. В [14] был получен результат об аппроксимации непрерывной возрастающей функции возрастающими многочленами. В дальнейшем этот результат уточнялся, в основном, на случай монотонных функций различной степени гладкости (см., например, [15]–[17]). Вопрос о достижимости точности аппроксимации непрерывной неубывающей функции $f(x)$ неубывающим многочленом $\mathcal{P}_n(x)$ на уровне порядка $\omega_k(f, n^{-1})$, $k = 2, 3, 4$, где ω_k — модуль непрерывности порядка k , исследовался в [18]. Отметим, что процедура построения аппроксимирующих многочленов была достаточно сложной и многоэтапной. Аппроксимация достаточно гладких s -монотонных функций изучалась в [19]. Функция $f: I = (-1; 1) \rightarrow \mathbb{R}$ называется s -монотонной для целого $s \geq 0$, если $\Delta_t^s f(t) \geq 0$ для всех $t \in I$,

$\tau > 0$. Здесь $\Delta_\tau^s f(t) = \sum_{k=0}^s (-1)^{s-k} C_s^k f(t + k\tau)$ при $(t, t + s\tau) \subset I$ (иначе равная нулю) — s -я конечная разность с шагом τ . В [20] изучалась проблема L_1 -аппроксимации монотонных функций d переменных по n вычислениям функции методом Монте-Карло.

Упомянутые работы [14]–[20] носили, между тем, чисто теоретический характер; результаты численных экспериментов в них представлены не были (строго говоря, в [20] есть один пример с численными расчетами, но, как там сказано, он имеет лишь теоретическое значение).

Подчеркнем, что автора проблема аппроксимации функций интересует с точки зрения приближения бесконечномерных задач оптимизации конечномерными. В такой ситуации значения искомой функции в контрольных точках нам неизвестны, поскольку заранее неизвестна сама функция (и в этом состоит принципиальное отличие от обычной задачи аппроксимации). В этом случае неизвестные параметры аппроксимации определяются путем минимизации значения целевого функционала, вычисляемого на конечномерном множестве аппроксимаций Ω , если они являются допустимыми по смыслу задачи либо на суперпозициях, содержащих функции класса Ω и принимающих допустимые значения. Указанное принципиальное отличие следует учитывать при сравнении различных подходов к аппроксимации (подробнее об этом уже было сказано в [11]): здесь главное — это достижение приемлемой точности аппроксимации произвольной функции уже при небольшом количестве параметров. Как показано в [7], [11] (там же см. результаты численных экспериментов), достаточно эффективными (в указанном смысле) оказываются аппроксимации с помощью квадратичных экспонент. Однако вопрос сохранения монотонности аппроксимации там не исследовался.

1. МОНОТОННАЯ АППРОКСИМАЦИЯ МОНОТОННЫХ НЕПРЕРЫВНЫХ ФУНКЦИЙ СДВИГАМИ И СЖАТИЯМИ ИНТЕГРАЛА ЛАПЛАСА

Для теории вероятностей и математической статистики важное значение имеет так называемый *интеграл Лапласа*

$$\widehat{\Phi}_0(x) = \frac{1}{\sqrt{2\pi}} \Phi_0(x), \quad \Phi_0(x) = \int_0^x \exp\left(-\frac{t^2}{2}\right) dt, \quad x \in (-\infty; \infty).$$

Функцию $\Phi_0(x)$ мы, соответственно, называем *функцией Лапласа*. Эта функция будет играть существенную роль в наших построениях.

Далее, называя непрерывную функцию возрастающей на отрезке, мы имеем в виду, что она допускает продолжение до непрерывной строго возрастающей функции на некотором объемлющем интервале. Называя гладкую функцию возрастающей на отрезке, мы имеем в виду, что ее производная строго положительна на всем отрезке. Следующее утверждение реализует соображения 1–3 построения монотонной аппроксимации гладких монотонных функций, описанные во Введении.

Теорема 1. Пусть $b > 0$, $\Psi \in C^1[-b; b]$ — неубывающая функция. Тогда для любого $\varepsilon > 0$ найдутся константа C и непрерывная функция $r(t)$, $t \in [-b; b]$, а также параметры $v \in \mathbb{N}$, $(\alpha, \beta, \gamma) \in \mathbb{R}^{3v}$ такие, что

$$\Psi(x) = C + Q(x) + R(x), \quad R(x) = \int_0^x r(t) dt, \quad |r(x)| \leq \varepsilon, \quad \forall x \in [-b; b],$$

$$Q(x) = \frac{1}{2} \sum_{j=1}^v \alpha_j^2 \gamma_j \Phi_0\left(\frac{x - \beta_j}{\gamma_j / 2}\right) + 2 \sum_{i < j} \alpha_i \alpha_j \gamma_{ij} e^{\sigma_{ij}} \Phi_0\left(\frac{x - \beta_{ij}}{\gamma_{ij}}\right), \quad (1)$$

$$\beta_{ij} = \frac{\beta_i \gamma_j^2 + \beta_j \gamma_i^2}{\gamma_i^2 + \gamma_j^2}, \quad \gamma_{ij} = \frac{\gamma_i \gamma_j}{\sqrt{2(\gamma_i^2 + \gamma_j^2)}}, \quad \sigma_{ij} = -\frac{(\beta_i - \beta_j)^2}{\gamma_i^2 + \gamma_j^2}. \quad (2)$$

Всякая функция вида (1) не убывает на $\mathbb{R}$. В случае возрастающей функции $\Psi(x)$ функцию $Q(x)$ тоже можно считать возрастающей на $[-b; b]$ (при некотором наборе параметров).

Следующее утверждение реализует соображение 4 построения монотонной аппроксимации непрерывных монотонных функций из тех, которые были описаны во Введении.

Теорема 2. Пусть $b > 0$, $\Psi \in C[-b; b]$ — возрастающая функция. Тогда для любого $\varepsilon > 0$ найдутся константа C и непрерывная функция $R(x)$, а также параметры $v \in \mathbb{N}$, $(\alpha, \beta, \gamma) \in \mathbb{R}^{3v}$ такие, что

$$\Psi(x) = C + Q(x) + R(x),$$

где функция $Q(x)$ имеет вид (1) с параметрами (2) и является возрастающей на $[-b; b]$, $|R(x)| \leq \varepsilon$ $\forall x \in [-b; b]$. Всякая функция вида (1) не убывает на $\mathbb{R}$.

Доказательства теорем 1, 2 представлены в следующем разделе. Аппроксимация интеграла Лапласа с помощью квадратичных экспонент обсуждается в разд. 3. Как видно из теоремы 1, выражение (1) можно заменить следующим:

$$\begin{aligned} Q(x) &= \frac{1}{2} \sum_{j=1}^v \alpha_j^2 \gamma_j \left[\Phi_0 \left(\frac{x - \beta_j}{\gamma_j/2} \right) + \Phi_0 \left(\frac{\beta_j}{\gamma_j/2} \right) \right] + 2 \sum_{i < j} \alpha_i \alpha_j \gamma_{ij} e^{\sigma_{ij}} \left[\Phi_0 \left(\frac{x - \beta_{ij}}{\gamma_{ij}} \right) + \Phi_0 \left(\frac{\beta_{ij}}{\gamma_{ij}} \right) \right] = \\ &= \int_0^x \left(\sum_{j=1}^v \alpha_j \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] \right)^2 dt. \end{aligned} \quad (3)$$

В этом случае $C = \Psi(0)$. Выражение (3) более удобно, если предполагается использовать производные по параметрам. Примем обозначения

$$\begin{aligned} b_{ij}(x) &= \exp \left[-\frac{(x - \beta_{ij})^2}{2\gamma_{ij}^2} \right], \quad c_{ij}(x) = \gamma_{ij}^2 b_{ij}(x) [-(x - \beta_{ij}) + 2(\beta_j - \beta_{ij})], \\ a_{ij}(x) &= \int_0^x b_{ij}(t) dt = \gamma_{ij} \left[\Phi_0 \left(\frac{x - \beta_{ij}}{\gamma_{ij}} \right) + \Phi_0 \left(\frac{\beta_{ij}}{\gamma_{ij}} \right) \right]. \end{aligned}$$

Нетрудно заметить, что $\beta_{ii} = \beta_i$, $\gamma_{ii} = \gamma_i/2$, $\sigma_{ii} = 0$, откуда

$$Q(x) = \sum_{i,j=1}^v \alpha_i \alpha_j e^{\sigma_{ij}} a_{ij}(x).$$

Следующее утверждение необходимо при использовании методов первого порядка для минимизации квадратичной невязки между аппроксимацией $Q(x)$ вида (3) и аппроксимируемой функцией.

Теорема 3. Для функции (3) имеем

$$\frac{\partial}{\partial \alpha_j} Q(x) = 2 \sum_{i=1}^v \alpha_i e^{\sigma_{ij}} a_{ij}(x), \quad (4)$$

$$\frac{\partial}{\partial \beta_j} Q(x) = -\frac{4\alpha_j}{\gamma_j^2} \sum_{i=1}^v \alpha_i e^{\sigma_{ij}} \{(\beta_j - \beta_{ij}) a_{ij}(x) + \gamma_{ij}^2 [b_{ij}(x) - b_{ij}(0)]\}, \quad (5)$$

$$\frac{\partial}{\partial \gamma_j} Q(x) = \frac{4\alpha_j}{\gamma_j^3} \sum_{i=1}^v \alpha_i e^{\sigma_{ij}} \{c_{ij}(x) - c_{ij}(0) + [\gamma_{ij}^2 + (\beta_{ij} - \beta_j)^2] a_{ij}(x)\}. \quad (6)$$

Доказательство теоремы 3 представлено в следующем разделе.

Теоремы 1–3 являются теоретической основой алгоритма построения монотонной аппроксимации непрерывных монотонных функций и состоящего в минимизации квадратичной невязки (по заданному набору контрольных точек) между шаблоном вида (1) или, что эквивалентно, вида (3) и аппроксимируемой функцией. При его практической реализации на компьютере возникает вопрос о конкретном способе вычисления значений интеграла Лапласа в формулах (1) или (3). Как уже было отмечено во Введении, интеграл Лапласа не выражается в элементарных функциях. Поэтому при расчетах на компьютере приходится использовать его аппроксимацию. При этом возможно нарушение монотонности полученной компьютерной реализации $P(x)$ аппроксимации $Q(x)$. Из общих соображений ясно, что это нарушение будет достаточно незначительным (и в этом смысле приемлемым) при условии, что аппроксимация интеграла Лапласа достаточно точна. Говоря о приемлемости, можно иметь в виду, в частности, что монотонность аппроксимации требуется нам для того, чтобы облегчить процесс минимизации невязки и

повысить шансы на получение более точной аппроксимации непрерывной монотонной функции. Соответственно, если отклонения от монотонности аппроксимации достаточно невелики, то это существенным образом отсекает посторонние “рыскания” процесса минимизации невязки, как бы направляя его “в нужное русло”, и в практическом плане это уже достаточно ценное достижение. Вопросы построения высокоточной аппроксимации интеграла Лапласа обсуждаются далее в разд. 3. Естественно, что при построении аппроксимации интеграла Лапласа необходимо иметь такие способы, которые позволили бы нам обеспечить сохранение свойства монотонности для реализации $P(x)$ аппроксимации $Q(x)$.

Следующее утверждение естественно понимать как теорему об устойчивой монотонности аппроксимирующей функции.

Теорема 4. Пусть (при заданном наборе параметров) функция $Q(x)$, определяемая формулой (1), является возрастающей на $[-b; b]$. Тогда при всех достаточно малых $\varepsilon > 0$ функция $P_\varepsilon(x)$, определяемая аналогичной формулой при замене интеграла Лапласа $\Phi_0(x)$ его гладкой аппроксимацией $\Phi_{0,\varepsilon}(x)$, удовлетворяющей условиям

$$\|\Phi'_0 - \Phi'_{0,\varepsilon}\|_{C[0;5]} < \varepsilon; \quad \Phi_{0,\varepsilon}(-x) = -\Phi_{0,\varepsilon}(x), \quad x \geq 0; \quad \Phi_{0,\varepsilon}(x) = \Phi_0(x), \quad x > 5,$$

остаётся возрастающей на $[-b; b]$.

Доказательство теоремы 4 представлено в следующем разделе.

Описываемый в разд. 3 способ IV аппроксимации интеграла Лапласа (основанный на аппроксимации $\Phi''_0(x)$ в пространстве $L_2[0; 5]$ линейной комбинацией вторых производных квадратичных экспонент) позволяет обеспечить выполнение условий теоремы 4. На самом деле при расчетах на компьютере вместо числовой оси мы всегда имеем дело с ее дискретизацией в виде конечного набора точек. Да и кроме того, нет смысла повышать точность расчетов выше некоторой неустраиваемой (или диктуемой какими-то иными причинами) погрешности вычислений. В этом смысле всегда присутствует некий уровень неразличимости точек на числовой оси. В соответствии с этим будем говорить, что функция $P(x)$ приближенно возрастает на $[a; b]$ на уровне неразличимости $\delta > 0$, если для любого разбиения $a = x_0 < x_1 < \dots < x_n = b$ при условии $\Delta x_i = x_i - x_{i-1} \geq \delta$ имеем $P(x_i) > P(x_{i-1})$ для всех $i = \overline{1, n}$.

Теорема 5. Пусть (при заданном наборе параметров) функция $Q(x)$, определяемая формулой (1), является возрастающей на $[-b; b]$. Тогда при заданном уровне неразличимости δ и всех достаточно малых $\varepsilon > 0$ функция $P_\varepsilon(x)$, определяемая аналогичной формулой при замене интеграла Лапласа $\Phi_0(x)$ его гладкой аппроксимацией $\Phi_{0,\varepsilon}(x)$, удовлетворяющей условиям

$$\|\Phi'_0 - \Phi'_{0,\varepsilon}\|_{L_1[0;5]} < \varepsilon; \quad \Phi_{0,\varepsilon}(-x) = -\Phi_{0,\varepsilon}(x), \quad x \geq 0; \quad \Phi_{0,\varepsilon}(x) = \Phi_0(x), \quad x > 5,$$

будет приближенно возрастающей на $[-b; b]$.

Доказательство теоремы 5 представлено в следующем разделе.

Описываемый в разд. 3 способ III аппроксимации интеграла Лапласа (основанный на аппроксимации $\Phi'_0(x)$ в пространстве $L_2[0; 5]$ линейной комбинацией первых производных квадратичных экспонент) позволяет обеспечить выполнение условий теоремы 5.

2. ДОКАЗАТЕЛЬСТВО РЕЗУЛЬТАТОВ ОБ АППРОКСИМАЦИИ С ПОМОЩЬЮ ИНТЕГРАЛА ЛАПЛАСА

Лемма 2. Имеем $\left(\sum_{i=1}^n a_i\right)^2 = \sum_{i=1}^n a_i^2 + 2\sum_{i<j} a_i a_j$ для любых $a_i \in \mathbb{R}$, $i = \overline{1, n}$.

Доказательство легко проводится математической индукцией.

Доказательство теоремы 1. Очевидно, что

$$\psi_0 = \Psi' \in C[-b; b], \quad \psi_0 \geq 0; \quad \Psi(x) = \Psi(0) + \int_0^x \psi_0(t) dt, \quad x \in [-b; b].$$

В свою очередь, $\psi_0(t) = \psi^2(t)$, где $\psi(t) = \sqrt{\psi_0(t)}$ — непрерывная функция. Согласно [7], функция $\psi(t)$ с любой степенью точности в метрике $C[-b; b]$ представима линейной комбинацией квадратичных экспонент

$$\psi(t) = \sum_{j=1}^v \alpha_j \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] + \sigma(t), \quad |\sigma(t)| \leq \varepsilon_1 \quad \forall t \in [-b; b],$$

где $\varepsilon_1 > 0$ — произвольно фиксированное число. Положим

$$K(x) = \int_0^x \left(\sum_{j=1}^v \alpha_j \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] \right)^2 dt, \quad R(x) = \int_0^x r(t) dt,$$

$$r(t) = 2\{\psi(t) - \sigma(t)\}\sigma(t) + \sigma^2(t). \text{ Тогда } \int_0^x \psi^2(t) dt = K(x) + R(x).$$

Далее будем считать, что ε_1 выбрано так, чтобы

$$|r(t)| \leq u(\varepsilon_1) \equiv 2(\bar{\Psi} + \varepsilon_1)\varepsilon_1 + \varepsilon_1^2 \leq \varepsilon, \quad \bar{\Psi} = \max_{t \in [-b; b]} |\psi(t)|.$$

Пользуясь леммой 2, рассмотрим

$$K(x) = \sum_{j=1}^v \alpha_j^2 \int_0^x \exp \left[-\frac{(t - \beta_j)^2}{(\gamma_j/\sqrt{2})^2} \right] dt + 2 \sum_{i < j} \alpha_i \alpha_j \int_0^x \exp \left[-\frac{(t - \beta_i)^2}{\gamma_i^2} - \frac{(t - \beta_j)^2}{\gamma_j^2} \right] dt.$$

Преобразуем сумму

$$\begin{aligned} \frac{(t - \beta_i)^2}{\gamma_i^2} + \frac{(t - \beta_j)^2}{\gamma_j^2} &= \left(\frac{1}{\gamma_i^2} + \frac{1}{\gamma_j^2} \right) t^2 - 2t \left(\frac{\beta_i}{\gamma_i^2} + \frac{\beta_j}{\gamma_j^2} \right) + \frac{\beta_i^2}{\gamma_i^2} + \frac{\beta_j^2}{\gamma_j^2} = \\ &= \frac{\gamma_i^2 + \gamma_j^2}{\gamma_i^2 \gamma_j^2} [t^2 - 2t\beta_{ij} + \beta_{ij}^2 - \beta_{ij}^2] + \frac{\beta_i^2}{\gamma_i^2} + \frac{\beta_j^2}{\gamma_j^2} \end{aligned}$$

(см. (2)). Положим для краткости $\ell_{ij} = \sqrt{2}\gamma_{ij}$. Заметим, что

$$\begin{aligned} \frac{\beta_{ij}^2}{\ell_{ij}^2} - \frac{\beta_i^2}{\gamma_i^2} - \frac{\beta_j^2}{\gamma_j^2} &= \beta_{ij} \left(\frac{\beta_i}{\gamma_i^2} + \frac{\beta_j}{\gamma_j^2} \right) - \frac{\beta_i^2}{\gamma_i^2} - \frac{\beta_j^2}{\gamma_j^2} = \frac{\beta_i \gamma_j^2 + \beta_j \gamma_i^2}{\gamma_i^2 + \gamma_j^2} \left(\frac{\beta_i}{\gamma_i^2} + \frac{\beta_j}{\gamma_j^2} \right) - \frac{\beta_i^2}{\gamma_i^2} - \frac{\beta_j^2}{\gamma_j^2} = \\ &= \beta_i^2 \left[\frac{\gamma_j^2}{\gamma_i^2(\gamma_i^2 + \gamma_j^2)} - \frac{1}{\gamma_i^2} \right] + \beta_j^2 \left[\frac{\gamma_i^2}{\gamma_j^2(\gamma_i^2 + \gamma_j^2)} - \frac{1}{\gamma_j^2} \right] + \frac{2\beta_i \beta_j}{\gamma_i^2 + \gamma_j^2} = \\ &= \frac{\beta_i^2}{\gamma_i^2} \left[\frac{\gamma_j^2}{\gamma_i^2 + \gamma_j^2} - 1 \right] + \frac{\beta_j^2}{\gamma_j^2} \left[\frac{\gamma_i^2}{\gamma_i^2 + \gamma_j^2} - 1 \right] + \frac{2\beta_i \beta_j}{\gamma_i^2 + \gamma_j^2} = \frac{-\beta_i^2 - \beta_j^2 + 2\beta_i \beta_j}{\gamma_i^2 + \gamma_j^2} = \sigma_{ij}. \end{aligned}$$

Таким образом,

$$\frac{(t - \beta_i)^2}{\gamma_i^2} + \frac{(t - \beta_j)^2}{\gamma_j^2} = \frac{(t - \beta_{ij})^2}{\ell_{ij}^2} - \sigma_{ij}.$$

Стало быть,

$$K[x] = \sum_{j=1}^v \alpha_j^2 \int_0^x \exp \left[-\frac{(t - \beta_j)^2}{(\gamma_j/\sqrt{2})^2} \right] dt + \sum_{i < j} 2\alpha_i \alpha_j e^{\sigma_{ij}} \int_0^x \exp \left[-\frac{(t - \beta_{ij})^2}{\ell_{ij}^2} \right] dt.$$

Делая замену $t - \beta_{ij} = \tau/\sqrt{2}$, $\gamma = \ell/\sqrt{2}$, получаем

$$\int_0^x \exp \left[-\frac{(t - \beta)^2}{\ell^2} \right] \frac{dt}{\gamma} = \int_{-\beta/\gamma}^{(x-\beta)/\gamma} \exp \left[-\frac{\tau^2}{2} \right] d\tau = \left[\Phi_0 \left(\frac{x - \beta}{\gamma} \right) + \Phi_0 \left(\frac{\beta}{\gamma} \right) \right].$$

Следовательно, $K(x) = c + Q(x)$, где $Q(x)$ определяется формулой (1),

$$c = \frac{1}{2} \sum_{i=1}^v \alpha_i^2 \gamma_i \Phi_0 \left(\frac{2\beta_i}{\gamma_i} \right) + \sum_{i < j} 2\alpha_i \alpha_j e^{\sigma_{ij}} \gamma_{ij} \Phi_0 \left(\frac{\beta_{ij}}{\gamma_{ij}} \right).$$

Таким образом, $\Psi(x) = C + Q(x) + R(x) \quad \forall x \in [-b; b]; C = c + \Psi(0)$.

По построению $Q'(x) = K'(x) \geq 0, x \in \mathbb{R}$. Следовательно, функция $Q(x)$ не убывает. Предположим теперь, что функция $\Psi(x)$ возрастает, т.е. $\Psi'(x) > 0, x \in [-b; b]$. По теореме Вейерштрасса существует минимум $\omega = \min_{x \in [-b; b]} \Psi'(x)$, и он достигается в некоторой точке $x_0 \in [-b; b]$. Таким образом, $\omega = \Psi'(x_0) > 0$. Имеем

$$Q'(x) = \Psi'(x) - R'(x) = \Psi'(x) - r(x) \geq \omega - r(x) > 0 \quad \forall x \in [-b; b],$$

если $|r(x)| < \omega$, т.е. достаточно, чтобы $u(\varepsilon_1) < \omega$. Теорема 1 доказана.

Пусть $\omega(f, \delta) = \sup\{|f(x) - f(y)| : |x - y| \leq \delta, x, y \in X\}$ — модуль непрерывности. Следующее утверждение доказано в [14, theorem 2].

Лемма 3. Для любой возрастающей функции $f \in C(X)$, $X = [-1; 1]$, и для всякого $n \in \mathbb{N}$ существует возрастающий многочлен $P_n(x)$ степени n такой, что $\|f - P_n\|_{C(X)} \leq c\omega(f, n^{-1})$.

Доказательство теоремы 2. В силу теоремы Кантора для компакта $X \subset \mathbb{R}$ и функции $f \in C(X)$ получаем, что $\omega(f, \delta) \rightarrow 0$ при $\delta \rightarrow +0$. Таким образом, согласно лемме 3, непрерывную возрастающую функцию можно с любой степенью точности приблизить возрастающим многочленом (т.е. гладкой возрастающей функцией). Отсюда и непосредственно из теоремы 1 вытекает теорема 2. Теорема 2 доказана.

Доказательство теоремы 3. Непосредственно из (3) получаем

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_j} &= 2 \int_0^x \left(\sum_{i=1}^v \alpha_i \exp \left[-\frac{(t - \beta_i)^2}{\gamma_i^2} \right] \right) \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] dt = \\ &= 2 \sum_{i=1}^v \alpha_i \int_0^x \exp \left[-\frac{(t - \beta_i)^2}{\gamma_i^2} \right] \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] dt = 2 \sum_{i=1}^v \alpha_i e^{\sigma_{ij}} a_{ij}(x), \end{aligned}$$

т.е. (4). Аналогично,

$$\begin{aligned} \frac{\partial Q}{\partial \beta_j} &= 2 \int_0^x \left(\sum_{i=1}^v \alpha_i \exp \left[-\frac{(t - \beta_i)^2}{\gamma_i^2} \right] \right) \alpha_j \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] \left[\frac{2(t - \beta_j)}{\gamma_j^2} \right] dt = \\ &= \frac{4\alpha_j}{\gamma_j^2} \sum_{i=1}^v \alpha_i e^{\sigma_{ij}} \int_0^x b_{ij}(t) dt - \frac{2\alpha_j \beta_j}{\gamma_j^2} \frac{\partial}{\partial \alpha_j} Q(x). \end{aligned}$$

Рассмотрим интеграл

$$\begin{aligned} \int_0^x b_{ij}(t) dt &= \int_0^x b_{ij}(t)(t - \beta_{ij}) dt + \beta_{ij} a_{ij}(x) = \\ &= -\gamma_{ij}^2 \int_0^x b_{ij}(t) d \left[-\frac{(t - \beta_{ij})^2}{2\gamma_{ij}^2} \right] + \beta_{ij} a_{ij}(x) = -\gamma_{ij}^2 b_{ij}(t) \Big|_0^x + \beta_{ij} a_{ij}(x). \end{aligned}$$

Отсюда очевидным образом получаем (5). Вычислим теперь

$$\begin{aligned} \frac{\partial Q}{\partial \gamma_j} &= 2 \int_0^x \left(\sum_{i=1}^v \alpha_i \exp \left[-\frac{(t - \beta_i)^2}{\gamma_i^2} \right] \right) \alpha_j \exp \left[-\frac{(t - \beta_j)^2}{\gamma_j^2} \right] \frac{2(t - \beta_j)^2}{\gamma_j^3} dt = \\ &= \frac{4\alpha_j}{\gamma_j^3} \sum_{i=1}^v \alpha_i e^{\sigma_{ij}} \int_0^x b_{ij}(t)(t - \beta_j)^2 dt. \end{aligned}$$

Заметим, что

$$(t - \beta_j \pm \beta_{ij})^2 = (t - \beta_{ij})^2 + 2(\beta_{ij} - \beta_j)(t - \beta_{ij}) + (\beta_{ij} - \beta_j)^2.$$

Беря по частям при $u = (t - \beta_{ij})$, $v = -\gamma_{ij}^2 b_{ij}(t)$, получаем

$$\int_0^x b_{ij}(t)(t - \beta_{ij})^2 dt = -\gamma_{ij}^2 b_{ij}(t)(t - \beta_{ij}) \Big|_0^x + \gamma_{ij}^2 a_{ij}(x).$$

Далее,

$$2(\beta_{ij} - \beta_j) \int_0^x b_{ij}(t)(t - \beta_{ij}) dt = 2\gamma_{ij}^2 (\beta_j - \beta_{ij}) b_{ij}(t) \Big|_0^x.$$

Таким образом,

$$\begin{aligned} \int_0^x b_{ij}(t)(t - \beta_j)^2 dt &= -\gamma_{ij}^2 b_{ij}(t)(t - \beta_{ij}) \Big|_0^x + 2\gamma_{ij}^2 (\beta_j - \beta_{ij}) b_{ij}(t) \Big|_0^x + \\ &+ \{\gamma_{ij}^2 + (\beta_{ij} - \beta_j)^2\} a_{ij}(x) = c_{ij}(x) - c_{ij}(0) + \{\gamma_{ij}^2 + (\beta_{ij} - \beta_j)^2\} a_{ij}(x). \end{aligned}$$

Отсюда сразу следует (6). Теорема 3 доказана.

Доказательство теоремы 4. Поскольку функция $Q(x)$ гладкая и возрастающая, согласно теореме Вейерштрасса, найдется число $\omega > 0$ такое, что $Q'(x) \geq \omega$ на $[-b; b]$. В соответствии с формулой (1) для всех $x \in [-b; b]$ имеем

$$\left| Q'(x) - P'_\varepsilon(x) \right| \leq \kappa \left\| \Phi'_0 - \Phi'_{0,\varepsilon} \right\|_{C[0;5]}, \quad \kappa = \frac{1}{2} \sum_{j=1}^v \alpha_j^2 |\gamma_j| + 2 \sum_{i < j} |\alpha_i \alpha_j \gamma_{ij}| e^{\sigma_{ij}}.$$

Таким образом, если $2\kappa\varepsilon \leq \omega$, то

$$P'_\varepsilon(x) = Q'(x) + P'_\varepsilon(x) - Q'(x) \geq \omega - \kappa\varepsilon \geq \frac{\omega}{2}.$$

Теорема 4 доказана.

Доказательство теоремы 5. Поскольку функция $Q(x)$ гладкая и возрастающая, согласно теореме Вейерштрасса, найдется число $\omega > 0$ такое, что $Q'(x) \geq \omega$ на $[-b; b]$. Выберем произвольное разбиение отрезка $-b = x_0 < x_1 < \dots < x_n = b$, $\Delta x_i \geq \delta$, $i = \overline{1, n}$. Для произвольных $\beta \in \mathbb{R}$, $\gamma > 0$ рассмотрим точки $\tilde{x}_i = (x_i - \beta)/\gamma$, $i = \overline{0, n}$. Имеем

$$\begin{aligned} \Phi_0(\tilde{x}_i) - \Phi_0(\tilde{x}_{i-1}) &= \int_{\tilde{x}_{i-1}}^{\tilde{x}_i} \Phi'_0(t) dt = \Phi_{0,\varepsilon}(\tilde{x}_i) - \Phi_{0,\varepsilon}(\tilde{x}_{i-1}) + r, \\ r &= \int_{\tilde{x}_{i-1}}^{\tilde{x}_i} \{\Phi'_0(t) - \Phi'_{0,\varepsilon}(t)\} dt. \end{aligned}$$

Очевидно, что $|r| \leq 2 \left\| \Phi'_0 - \Phi'_{0,\varepsilon} \right\|_{L_1[0;5]} \leq 2\varepsilon$. В соответствии с (1)

$$P'_\varepsilon(x_i) - P'_\varepsilon(x_{i-1}) = Q(x_i) - Q(x_{i-1}) + R \geq \omega \Delta x_i + R \geq \omega \delta + R,$$

где $|R| \leq \kappa |r| \leq 2\kappa\varepsilon$. Таким образом, при условии $2\kappa\varepsilon < \omega\delta$ имеем $P'_\varepsilon(x_i) > P'_\varepsilon(x_{i-1})$, $i = \overline{1, n}$. Теорема 5 доказана.

3. АППРОКСИМАЦИЯ ИНТЕГРАЛА ЛАПЛАСА С ПОМОЩЬЮ КВАДРАТИЧНЫХ ЭКСПОНЕНТ

Рассмотрим четыре подхода к аппроксимации интеграла Лапласа с помощью квадратичных экспонент. Первые два из этих четырех подходов применимы, вообще говоря, к аппроксимации произвольной непрерывной функции; третий — к аппроксимации произвольной непрерывной

функции с первой производной из класса L_2 ; четвертый — к аппроксимации произвольной функции со второй производной из класса L_2 . Эти подходы и сами по себе представляют определенный интерес. Третий подход соответствует теореме 5, четвертый — теореме 4.

Применительно к интегралу Лапласа, прежде всего, заметим, что он обладает рядом замечательных свойств. Для нас важно, главным образом, то, что он, во-первых, является нечетной функцией, и, во-вторых, значения, принимаемые им при $x > 5$, практически неотличимы от числа $1/2$. Поэтому аппроксимацию интеграла Лапласа достаточно строить лишь на отрезке $[0; 5]$. Именно этим отрезком мы и ограничимся. Подходы к аппроксимации мы ранжируем по степени возрастания сложности. Что касается иных характеристик, сделаем следующие замечания.

Наша цель состоит в том, чтобы при возможно меньшем количестве μ квадратичных экспонент получить высокоточную аппроксимацию интеграла Лапласа (вполне достаточную для практических приложений). Упомянутые четыре подхода — это четыре возможности добиться поставленной цели. Первый — наиболее простой и быстрый в смысле реализации. Однако он носит эвристический характер, поскольку сводится к минимизации квадратичной невязки по фиксированному набору контрольных точек. Таким образом, малость невязки по норме $L_p[a; b]$ или $C[a; b]$ строго теоретически не гарантируется. Второй, третий и четвертый имеют теоретически обоснованную оценку точности.

Подход I. При заданном $\mu \in \mathbb{N}$ и наборе параметров $(\xi, \eta, \zeta) \in \mathbb{R}^{3\mu}$ в качестве аппроксимирующих функций (аппроксимаций) будем использовать функции вида

$$\Phi_\mu[\xi, \eta, \zeta](x) = \sum_{j=1}^{\mu} \xi_j \phi_j[\eta_j, \zeta_j](x), \quad \phi_j[\eta_j, \zeta_j](x) = \exp \left[-\frac{(x - \eta_j)^2}{\varepsilon + \zeta_j^2} \right],$$

где $\varepsilon > 0$ — некоторое достаточно малое число (нужное для того только, чтобы избежать нуля в знаменателе; в численных расчетах мы везде брали $\varepsilon = 10^{-10}$). Пусть $a = z_1 < \dots < z_m = b$ — некоторое разбиение отрезка $[a; b] = [0; 5]$ с помощью достаточно большого количества контрольных точек. Мы брали $m = 1000$. Подбор оптимальных параметров аппроксимации производится путем минимизации квадратичной невязки

$$W[\xi, \eta, \zeta] = \sum_{i=1}^m \left\{ \Phi_\mu[\xi, \eta, \zeta](z_i) - \hat{\Phi}_0(z_i) \right\}^2$$

за счет варьирования всей тройки $(\xi, \eta, \zeta) \in \mathbb{R}^{3\mu}$. Задачу минимизации

$$W[\xi, \eta, \zeta] \rightarrow \min_{(\xi, \eta, \zeta) \in \mathbb{R}^{3\mu}}$$

мы решали различными методами: Левенберга—Марквардта, Гаусса—Ньютона, Хука—Дживса, DFP, BFGS, BFS. Для одномерного поиска использовались метод квадратичной интерполяции, Брента и More—Thuente (см. [21]). Наилучшие результаты показал метод Левенберга—Марквардта в сочетании с методом More—Thuente.

Для случая $\mu = 2$ были получены следующие результаты:

$\xi = (0.50102, -0.74041)$, $\eta = (0.16707, -1.0234)$, $\zeta = (91.92, 1.6321)$, $W = 7.548e-5$ (запись $e-5$

означает умножение на 10^{-5}), максимальное отклонение в контрольных точках аппроксимации от аппроксимируемой функции $\delta_{\max} = 0.0013038$, максимальное отклонение в контрольных точках производной аппроксимации от производной аппроксимируемой функции $\delta'_{\max} = 0.014958$. Графики аппроксимации и аппроксимируемой функции не приводим, поскольку различия в них можно увидеть только при очень большом увеличении (на малом участке между точками 3 и 4). Таким образом, достаточно хорошее качество аппроксимации удается получить уже этим простейшим способом при $\mu = 2$. Для сравнения укажем, что при использовании классического способа (взять частичную сумму ряда, полученного почленным интегрированием разложения подынтегральной функции в ряд Маклорена) приходится учитывать порядка сорока членов разложения для достижения сопоставимой точности.

Приведем результаты для случая $\mu = 3$:

$\xi = (0.89868, -0.80742, -0.012592)$, $\eta = (-110.5396, -1.1287, 2.6466)$, $\zeta = (151.7562, 1.7226, 2.094)$, $W = 1.0964e-5$, $\delta_{\max} = 0.00052134$, $\delta'_{\max} = 0.0072675$. Графики не приводим, поскольку различия в них можно увидеть только при очень большом увеличении.

Подход II. Этот подход основан на следующем общем утверждении об аппроксимации.

Теорема 6. Пусть E — линейное пространство со скалярным произведением $\langle \cdot, \cdot \rangle$, $x_i \in E$, $i = \overline{1, m}$, $\zeta_{ij} = \langle x_i, x_j \rangle$, $i, j = \overline{1, m}$, $\Gamma = (\zeta_{ij})_{i,j=\overline{1,m}}$ — матрица Грама указанной системы векторов, $\Gamma_1 = (\zeta_{ij})_{i,j=\overline{1,m-1}}$, $\Delta = \det \Gamma$, $\Delta_1 = \det \Gamma_1$. Предположим, что $\Delta_1 \neq 0$ (отсюда по свойствам определителя Грама вытекает, что $\Delta_1 > 0$). Тогда найдутся вектор $\tilde{\xi} \in \mathbb{R}^{m-1}$ и соответствующая ему линейная комбинация $\tilde{x} = \sum_{i=1}^{m-1} \tilde{\xi}_i x_i$ такие, что для разности $h = \tilde{x} - x_m$ имеем $\|h\|^2 = \Delta/\Delta_1$. Более того, вектор $\tilde{x}$ минимизирует расстояние от вектора x_m до линейного подпространства L , натянутого на векторы x_i , $i = \overline{1, m-1}$, причем вектор $\tilde{\xi}$ определяется однозначно как решение системы

$$\Gamma_1 \tilde{\xi} = (\zeta_{im})_{i=\overline{1,m-1}}. \quad (7)$$

Доказательство. Согласно [22, § XVIII.1, пример 9, с. 601], существует единственный вектор $\tilde{\xi}$ (определяемый как решение системы (7)) такой, что

$$\|x_m - \tilde{x}\| = \inf_{y \in L} \|x_m - y\|, \quad \tilde{x} = \sum_{i=1}^{m-1} \tilde{\xi}_i x_i.$$

По построению вектор $h = \tilde{x} - x_m$ ортогонален пространству L , причем $\tilde{x} \in L$. Согласно второму следствию из [23, гл. 12, теорема 98.2, с. 349–350], справедливо равенство

$$|\Gamma(x_1, \dots, x_{m-1}, h)| = |\Gamma(x_1, \dots, x_{m-1})| \cdot |\Gamma(h)|,$$

где $|\Gamma(x_1, \dots, x_{m-1}, h)|$ — определитель Грама системы $x_1, \dots, x_{m-1}, h$, и соответственно, указанное равенство переписывается как

$$|\Gamma(x_1, \dots, x_{m-1}, h)| = \Delta_1 \|h\|^2.$$

Известно (см. [23, гл. 12, свойства 2, 3, с. 348]), что определитель Грама не изменяется от прибавления к любому вектору системы линейной комбинации остальных векторов, и при умножении любого вектора системы на число α умножается на α^2 . С учетом принадлежности $\tilde{x} \in L$ отсюда ясно, что $|\Gamma(x_1, \dots, x_{m-1}, h)| = \Delta$. Теорема доказана.

Замечание 1. В частности, при $\Delta = 0$ получаем в качестве следствия классический факт: из равенства нулю определителя Грама вытекает линейная зависимость системы векторов. Вообще, как известно, равенство нулю определителя Грама необходимо и достаточно для линейной зависимости системы векторов.

Замечание 2. Для аппроксимации $x_m \approx \sum_{i=1}^{m-1} \tilde{\xi}_i x_i$ точность оценивается как $\sqrt{\Delta/\Delta_1}$.

Обозначим для краткости $\varphi_j(x) = \varphi_j[\eta_j, \zeta_j](x)$, $j = \overline{1, \mu}$. Рассмотрим систему функций $\{\varphi_j, j = \overline{1, \mu}; \hat{\Phi}_0\}$ как систему из $(\mu + 1)$ векторов в пространстве $L_2[a; b]$. Положим Γ — матрица Грама этой системы, Γ_1 — матрица Грама системы $\{\varphi_j, j = \overline{1, \mu}\}$, $\Delta = \det \Gamma$, $\Delta_1 = \det \Gamma_1$. Непосредственно из теоремы 6 вытекает

Теорема 7. Пусть $\Delta_1 \neq 0$ (следовательно, $\Delta_1 > 0$). Тогда существует вектор $\xi \in \mathbb{R}^\mu$ такой, что для аппроксимации $\Phi_\mu(x) = \sum_{j=1}^\mu \xi_j \varphi_j(x)$ имеем $\|\Phi_\mu - \hat{\Phi}_0\|_{L_2[a; b]} = \sqrt{\Delta/\Delta_1}$. При этом в качестве ξ можно взять решение системы

$$\Gamma_1 \xi = (\sigma_i)_{i=\overline{1, \mu}}, \quad (8)$$

где

$$\sigma_i = \langle \varphi_i, \hat{\Phi}_0 \rangle_{L_2} = \int_a^b \varphi_i(x) \hat{\Phi}_0(x) dx; \quad \Gamma_1 = (\langle \varphi_i, \varphi_j \rangle_{L_2})_{i,j=\overline{1, \mu}}.$$

Из [24, гл. 4, упражнение 15, с. 92] видно, что в теореме 7 вектор ξ как раз и обеспечивает минимум расстояния аппроксимируемой функции $\widehat{\Phi}_0$ до подпространства, натянутого на систему $\{\varphi_j, j = \overline{1, \mu}\}$.

Исходя из теоремы 7, при заданном $\mu \in \mathbb{N}$ возникает идея подобрать параметры $\eta, \zeta \in \mathbb{R}^\mu$ путем минимизации отношения $R(\eta, \zeta) = \Delta(\eta, \zeta)/\Delta_1(\eta, \zeta)$, скажем, методом Хука–Дживса (нулевого порядка), а затем взять в качестве вектора коэффициентов $\xi \in \mathbb{R}^\mu$ решение системы (8). Впрочем, здесь следует соблюдать осторожность: при неудачном выборе начального приближения минимизируемая величина может повести себя как неопределенность типа $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$. С другой стороны, эту трудность можно обойти, если ввести штраф за слишком малые значения Δ_1 (но, как показывает опыт, в таком случае происходит сильное замедление процесса численной минимизации). Отметим, наконец, что отношение $R(\eta, \zeta)$ нельзя вычислить точно (аналитически). Поэтому приходится его вычислять приближенно с помощью численного интегрирования. Тем не менее близость к оптимуму полученных в итоге параметров ξ, η, ζ можно оценить с помощью квадратичной невязки $W[\xi, \eta, \zeta]$, а также оценкой максимального отклонения аппроксимации от аппроксимируемой функции в контрольных точках.

Описанный способ позволил нам получить следующие результаты.

Для $\mu = 2$: $\xi = (-0.74267, 1.6952)$, $\eta = (-1.0268, -2798.7635)$, $\zeta = (1.6348, 2536.7604)$, $W = 6.6945e-6$, $\delta_{\max} = 0.0012945$, $\delta'_{\max} = 0.014745$. Графики не приводим, поскольку различия в них можно увидеть только при очень большом увеличении.

Для $\mu = 3$: $\xi = (-0.32242, -0.74678, 0.82234)$, $\eta = (4.2016, -1.05, 4.1338)$, $\zeta = (8.2595, 1.65, 13.411)$, $W = 7.4844e-5$, $\delta_{\max} = 0.00079101$, $\delta'_{\max} = 0.011011$. Графики не приводим, поскольку различия в них можно увидеть только при очень большом увеличении.

Подход III аналогичен подходу II, с тем лишь отличием, что вместо аппроксимации самого интеграла Лапласа $\widehat{\Phi}_0(x)$ точно тем же способом строится аппроксимация (в пространстве $L_2[a; b]$) по функциям $\varphi'_j(x)$ его производной $\widehat{\Phi}'_0(x) = 1/\sqrt{2\pi} \exp\left[-\frac{x^2}{2}\right]$. После этого в качестве аппроксимации функции $\widehat{\Phi}_0(x)$ принимается $Q_\mu[\xi, \eta, \zeta](x) = C + \Phi_\mu[\xi, \eta, \zeta](x)$, где константа C определяется из условия $Q_\mu(a) = \widehat{\Phi}_0(a)$.

Итак, рассмотрим систему функций

$$\left\{ \varphi'_j(x) = -\frac{2(x - \eta_j)}{\varepsilon + \zeta_j^2} \varphi_j(x), j = \overline{1, \mu}; \widehat{\Phi}'_0(x) \right\}$$

как систему из $(\mu + 1)$ векторов в пространстве $L_2[a; b]$. Положим Γ' — матрица Грама этой системы, Γ'_1 — матрица Грама системы

$$\{\varphi'_j, j = \overline{1, \mu}\},$$

$\Delta' = \det \Gamma'$, $\Delta'_1 = \det \Gamma'_1$. Непосредственно из теоремы 6 вытекает

Теорема 8. Пусть $\Delta'_1 \neq 0$ (следовательно, $\Delta'_1 > 0$). Тогда существует вектор $\xi \in \mathbb{R}^\mu$ такой, что для аппроксимации $Q_\mu(x) = C + \Phi_\mu(x)$, $C = \widehat{\Phi}_0(a) - \Phi_\mu(a)$, имеем $\|Q_\mu - \widehat{\Phi}_0\|_{C[a; b]} \leq \sqrt{(b - a)\Delta'/\Delta'_1}$. При этом в качестве ξ можно взять решение системы

$$\Gamma'_1 \xi = (\sigma'_i)_{i=\overline{1, \mu}}, \quad (9)$$

где

$$\sigma'_i = \langle \varphi'_i, \widehat{\Phi}'_0 \rangle_{L_2} = \int_a^b \varphi'_i(x) \widehat{\Phi}'_0(x) dx; \quad \Gamma'_1 = (\langle \varphi'_i, \varphi'_j \rangle_{L_2})_{i, j=\overline{1, \mu}}.$$

Доказательство. Используя неравенство Гёльдера, получаем

$$\begin{aligned} |Q_\mu(x) - \widehat{\Phi}_0(x)| &= \left| Q_\mu(a) + \int_a^x Q'_\mu(t) dt - \widehat{\Phi}_0(a) - \int_a^x \widehat{\Phi}'_0(t) dt \right| = \\ &= \left| \int_a^x \{ \Phi'_\mu(t) - \widehat{\Phi}'_0(t) \} dt \right| \leq \int_a^b \left| \Phi'_\mu(t) - \widehat{\Phi}'_0(t) \right| dt \leq \sqrt{(b-a)} \left\| \Phi'_\mu - \widehat{\Phi}'_0 \right\|_{L_2[a;b]}. \end{aligned}$$

Дальнейшее очевидно. Теорема доказана.

Исходя из теоремы 8, при заданном $\mu \in \mathbb{N}$ возникает идея подобрать параметры $\eta, \zeta \in \mathbb{R}^\mu$ путем минимизации отношения $R'(\eta, \zeta) = \Delta'(\eta, \zeta)/\Delta'_1(\eta, \zeta)$, скажем, методом Хука–Дживса (нулевого порядка), а затем взять в качестве вектора коэффициентов $\xi \in \mathbb{R}^\mu$ решение системы (9). Отношение $R'(\eta, \zeta)$ опять же нельзя вычислить точно (аналитически). Поэтому приходится его вычислять приближенно с помощью численного интегрирования. Тем не менее близость к оптимуму полученных в итоге параметров ξ, η, ζ можно оценить с помощью квадратичной невязки $\mathcal{W}[\xi, \eta, \zeta]$, а также оценкой максимального отклонения аппроксимации от аппроксимируемой функции в контрольных точках.

Описанный способ позволил нам получить следующие результаты.

Для $\mu = 2$: $C = 0.50001$, $\xi = (-0.30633, -0.7523)$, $\eta = (-0.43757, -1.9128)$, $\zeta = (1.453, 1.7258)$, квадратичное отклонение аппроксимации от аппроксимируемой функции в контрольных точках $\widetilde{\mathcal{W}} = 6.3837e-8$, $\delta_{\max} = 1.4717e-5$, $\delta'_{\max} = 7.9605e-5$. Рисунков не приводим, поскольку визуально график аппроксимации полностью совпадает с графиком аппроксимируемой функции. Для практических приложений такой точности уже вполне достаточно.

Для $\mu = 3$: $C = 0.5$, $\xi = (-0.025654, -0.27286, -0.76869)$, $\eta = (0.47302, -0.40261, -1.824)$, $\zeta = (1.3121, 1.3683, 1.6522)$, $\widetilde{\mathcal{W}} = 3.5108e-9$, $\delta_{\max} = 3.2107e-6$, $\delta'_{\max} = 4.8288e-5$. Рисунков не приводим, поскольку визуально график аппроксимации полностью совпадает с графиком аппроксимируемой функции.

Подход IV совершенно аналогичен подходу III, с тем лишь отличием, что вместо аппроксимации производной интеграла Лапласа $\widehat{\Phi}'_0(x)$ точно тем же способом строится аппроксимация (в пространстве $L_2[a; b]$) по функциям $\varphi'_i(x)$ его второй производной $\widehat{\Phi}''_0(x)$. После этого в качестве аппроксимации $\widehat{\Phi}_0(x)$ принимается $Q_\mu[\xi, \eta, \zeta](x) = C_1 + C_2x + \Phi_\mu[\xi, \eta, \zeta](x)$, где C_1, C_2 определяются из условий $Q_\mu(0) = \widehat{\Phi}_0(0)$, $Q'_\mu(0) = \widehat{\Phi}'_0(0)$.

Описанный способ позволил нам получить следующие результаты.

Для $\mu = 2$: $C_1 = 0.50023$, $C_2 = -4.8295e-005$, $\xi = (-0.50169, -0.60577)$, $\eta = (-0.64487, -2.0532)$, $\zeta = (1.5027, 1.456)$, по функции ($\widetilde{\mathcal{W}}$ – квадратичное отклонение, $\delta_{\max}$ – максимальное отклонение) $\widetilde{\mathcal{W}} = 6.125e-7$, $\delta_{\max} = 4.9302e-5$; по первой производной $\widetilde{\mathcal{W}}' = 1.4818e-6$, $\delta'_{\max} = 7.5314e-5$; по второй производной $\widetilde{\mathcal{W}}'' = 9.9166e-6$, $\delta''_{\max} = 0.00056108$. Рисунков не приводим, поскольку визуально график аппроксимации полностью совпадает с графиком аппроксимируемой функции. Как видим, разница с подходом III не очень большая, из чего можно заключить, что и подход III будет достаточно работоспособен. То, что по функции отклонение получилось даже несколько больше, чем для подхода III, не должно нас смущать, поскольку отклонение вычислялось не от “идеального” интеграла Лапласа, а от его приближения, полученного численным интегрированием. Поэтому отклонение по производной здесь более информативно.

Для $\mu = 3$ результаты следующие: $C_1 = 0.49988$, $C_2 = 2.5236e-5$, $\xi = (10.9656, -0.28145, -11.7683)$, $\eta = (-1.9261, -0.46093, -1.9349)$, $\zeta = (1.9636, 1.4548, 1.9586)$, по функции $\widetilde{\mathcal{W}} = 7.5652e-9$, $\delta_{\max} = 9.2179e-6$; по первой производной $\widetilde{\mathcal{W}}' = 8.9872e-8$, $\delta'_{\max} = 2.202e-5$; по второй производной $\widetilde{\mathcal{W}}'' = 1.222e-6$, $\delta''_{\max} = 0.00021175$. Рисунков не приводим, поскольку визуально график аппроксимации полностью совпадает с графиком аппроксимируемой функции.

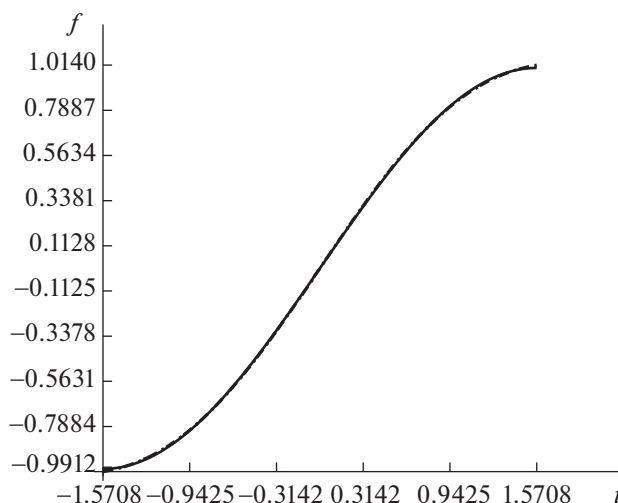
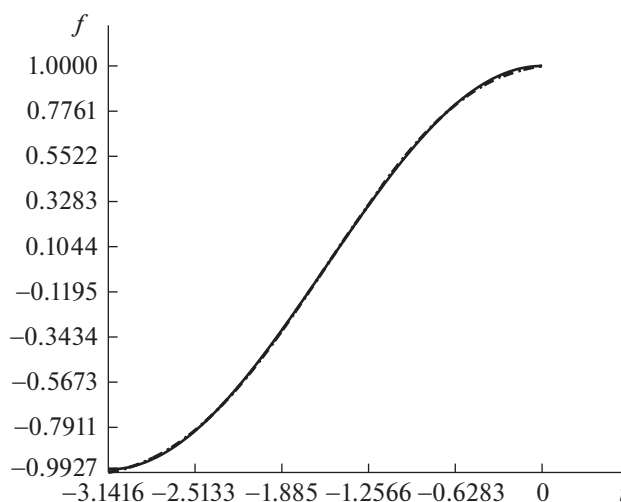
В заключение этого раздела приведем также следующие данные. Нам было интересно сравнить время счета при использовании нашего способа аппроксимации интеграла Лапласа с помощью линейной функции и двух квадратичных экспонент (подход IV), с одной стороны, и при использовании функции ошибки erf , реализованной в системе MATLAB (седьмой версии), с другой стороны. Оценивалось совокупное время, необходимое для вычисления интеграла Лапласа (точнее, его аппроксимации) на массиве узловых точек сетки отрезка $[0; 5]$ с шагом 10^{-6} (т.е. на массиве из порядка пяти миллионов точек). Разумеется, время счета определенным образом зависит от других процессов, функционирующих в системе (а также от типа и конфигурации компьютера и операционной системы). Поэтому ради чистоты эксперимента мы повторили его 20 раз. Во всех 20 из них вычисления по нашему способу оказались несколько быстрее. Приведем типичный результат. Для нашего способа время счета 0.317164 с; при использовании функции erf — 0.384069 с. Максимум отклонений в узлах сетки: 4.4487×10^{-5} . Графики визуально полностью совпадают. При этом следует понимать, что MATLAB — это коммерческий продукт и алгоритм, реализованный в программе erf — это коммерческая тайна. Можно предположить, что там реализован табличный способ в сочетании с линейной интерполяцией, но автору об этом ничего не известно. Остается лишь отметить, что подобный способ требует хранения в памяти компьютера соответствующего массива данных, а при собственной реализации такой массив еще надо создать.

4. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

Для неубывающих функций $f(t)$, $t \in [a; b]$, различного вида мы строили аппроксимацию по формуле (3), $C = f(a)$, при значениях $v \geq 2$ — до достижения достаточно высокой точности. Параметры $(\alpha, \beta, \gamma) \in \mathbb{R}^{3v}$ определялись, исходя из минимизации квадратичной невязки по тысяче контрольных точек. Вместо интеграла Лапласа использовалась его аппроксимация в рамках подхода IV при $\mu = 2$. При использовании подхода III и встроенной функции MATLAB $\text{erf}(t)$ для вычисления интеграла Лапласа отличия практически неразличимы. Например, для теста № 1 при $v = 2$ для подхода III получаем $\delta = 0.00085953$, $\delta_{\max} = 0.0026705$; $\alpha = (0.36985, 0.62644)$, $\beta = (1.6809e-5, 8.4071e-11)$, $\gamma = (0.97662, 3.437)$. При использовании функции $\text{erf}(t)$ $\delta = 0.0008385$, $\delta_{\max} = 0.0023415$; $\alpha = (0.37384, 0.62218)$, $\beta = (-9.5407e-12, -3.7628e-12)$, $\gamma = (0.98342, 3.4653)$. Здесь (и далее) используются следующие обозначения: δ — значение квадратичной невязки, $\delta_{\max}$ — модуль максимального отклонения аппроксимации от аппроксимируемой функции по набору контрольных точек. Графики по сравнению с подходом IV визуально никак не различаются (см. далее соответствующие результаты для подхода IV). Кроме того, проводилась проверка полученной аппроксимации на монотонность. При использовании функции $\text{erf}(t)$ аппроксимация во всех тестах оказалась неубывающей (разумеется, на конечной сетке — из 1000 контрольных точек). При использовании подхода III наблюдалось незначительное отклонение от монотонности в одной-двух контрольных точках: тест № 3, $v = 2$ — порядка 10^{-3} ; тест № 4, $v = 2$, $v = 3$ — порядка 10^{-3} ; тест № 5, $v = 5$ — порядка 10^{-4} . Отметим, что нарушение положительности (отрицательности) на величину порядка 10^{-5} и меньше может возникать вследствие накопленной погрешности вычислений (в том числе и достаточно простых) даже там, где этого теоретически не должно быть. Автор сталкивался с подобным явлением, например, при программной реализации симплекс-метода. Итак, приведем полученные результаты для подхода IV. Здесь $\omega = \min_{i=1,n} \{Q(x_i) - Q(x_{i-1})\}$ — сеточный показатель монотонности.

Тест № 1. Функция $f(t) = \arctg t$, $t \in [-3; 3]$. При $v = 2$: $\delta = 0.00085827$, $\delta_{\max} = 0.0027058$, $\omega = 0.00051423$; $\alpha = (0.36984, 0.62644)$, $\beta = (2.4837e-5, 2.0265e-10)$, $\gamma = (0.97661, 3.437)$. При $v = 3$: $\delta = 3.1132e-5$, $\delta_{\max} = 0.00050259$, $\omega = 0.00062426$; $\alpha = (0.20861, 0.46246, 0.328)$, $\beta = (3.2261e-5, 1.2145e-5, 1.7191e-11)$, $\gamma = (0.79389, 1.7984, 9.0161)$. Графики не приводим, поскольку визуально они полностью совпадают.

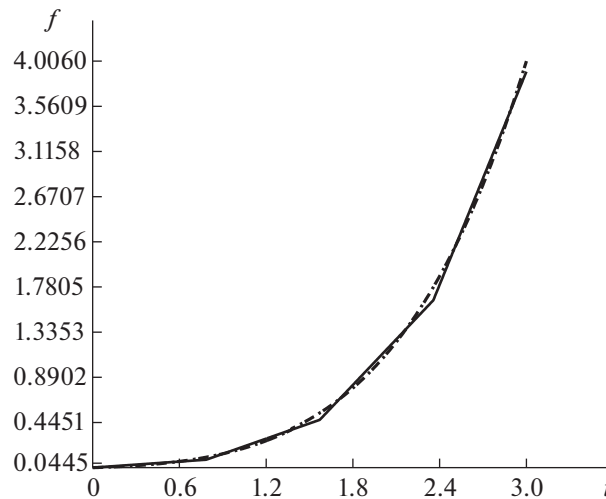
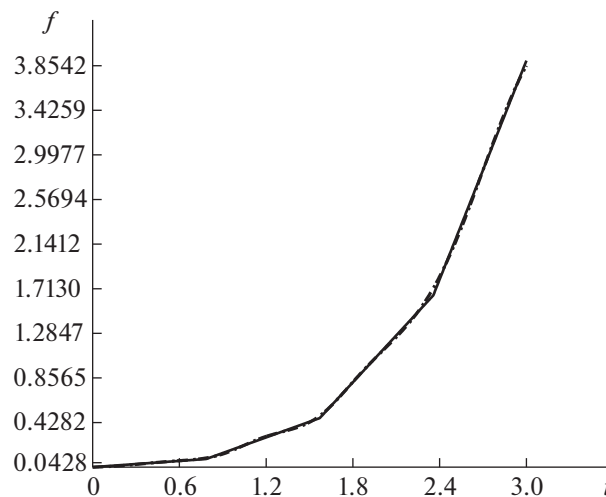
Тест № 2. Функция $f(t) = \sin t$, $t \in [-\pi/2; \pi/2]$. При $v = 2$: $\delta = 0.015425$, $\delta_{\max} = 0.013986$, $\omega = 0.00036968$, $\alpha = (5.4264, -4.4197)$, $\beta = (-0.023435, -0.14461)$, $\gamma = (3.9465, 9.0997)$. Графики см. на фиг. 1: аппроксимация изображена пунктирной линией, аппроксимируемая функция — сплошной линией. При $v = 3$: $\delta = 0.0010741$, $\delta_{\max} = 0.0049779$, $\omega = 0.00025331$;

Фиг. 1. Тест № 2, $v = 2$.Фиг. 2. Тест № 3, $v = 2$.

$\alpha = (-0.38012, -0.941, -0.38081)$, $\beta = (1.0384, 0.00052416, -1.038)$, $\gamma = (-0.6554, 1.0154, 0.65592)$.
Графики не приводим, поскольку визуально они полностью совпадают.

Тест № 3. Функция $f(t) = \cos t$, $t \in [-\pi; 0]$. При $v = 2$: $\delta = 0.027529$, $\delta_{\max} = 0.015041$, $\omega = -0.0011892$: нарушение монотонности лишь в одной контрольной точке с индексом 905; $\alpha = (-4.5679, 5.4914)$, $\beta = (-0.29867, -1.3424)$, $\gamma = (8.3551, 3.8693)$. При использовании для аппроксимации интеграла Лапласа подхода IV при $\mu = 3$ нарушение монотонности устраняется: $\delta = 0.028168$, $\delta_{\max} = 0.015552$, $\omega = 8.1668e-5$. Графики см. на фиг. 2. При $v = 3$: $\delta = 0.0020035$, $\delta_{\max} = 0.0085406$, $\omega = 4.0173e-5$; $\alpha = (0.28284, -1.6716, 0.95752)$, $\beta = (-2.4766, 0.57694, -1.3151)$, $\gamma = (0.76281, 0.44424, 1.4416)$. Графики не приводим, поскольку визуально они полностью совпадают.

Тест № 4. Функция $f(t) = \exp t$, $t \in [-3; 3]$. При $v = 2$: $\delta = 0.010551$, $\delta_{\max} = 0.015452$, $\omega = 2.173e-4$; $\alpha = (13.0709, 2.7665)$, $\beta = (7.5862, 5.8611)$, $\gamma = (3.5002, 5.5059)$. При $v = 3$: $\delta = 0.0051879$, $\delta_{\max} = 0.013392$, $\omega = 0.00030832$; $\alpha = (6.5988, 4.0538, 0.13998)$, $\beta = (6.1664, 5.9167, 0.72595)$, $\gamma = (2.6424, 4.7098, 8.6292)$. Графики не приводим, поскольку визуально они полностью совпадают.

Фиг. 3. Тест № 5, $v = 3$.Фиг. 4. Тест № 5, $v = 4$.

Тест № 5. Непрерывная кусочно-линейная функция, см. фиг. 3–4, при $t \in [0; 3]$. При $v = 3$: $\delta = 1.376$, $\delta_{\max} = 0.12811$, $\omega = 0.00012769$; $\alpha = (13.8495, 3.6601, -0.22289)$, $\beta = (11.0969, 5.8107, 2.2459)$, $\gamma = (3.8892, 3.9442, 9.1205)$. Графики см. на фиг. 3. При $v = 4$: $\delta = 0.21604$, $\delta_{\max} = 0.068932$, $\omega = 0.00011881$; $\alpha = (-6.7454, 5.2712, 2.4614, -1.9459)$, $\beta = (1.0425, 0.89296, 1.3388, 2.7021)$, $\gamma = (0.61874, 0.49289, 0.3392, 0.66608)$. Графики см. на фиг. 4. При $v = 5$: $\delta = 0.049677$, $\delta_{\max} = 0.030583$, $\omega = -0.00024438$; — нарушение монотонности лишь в одной контрольной точке с индексом 256; $\alpha = (-0.33219, -7.8355, 6.5769, -0.49263, -0.58114)$, $\beta = (0.22402, 2.3705, 2.337, 2.4544, 0.9735)$, $\gamma = (0.54247, 0.70341, 0.58191, 0.13697, -0.27281)$. При использовании для аппроксимации интеграла Лапласа подхода IV при $\mu = 3$ нарушение монотонности существенно уменьшается (остается в одной контрольной точке с индексом 204): $\delta = 0.049884$, $\delta_{\max} = 0.030384$, $\omega = -8.6895e-5$. При использовании для аппроксимации интеграла Лапласа функции erf нарушение монотонности полностью устраняется: $\delta = 0.049594$, $\delta_{\max} = 0.030323$, $\omega = 0.00023876$. Графики не приводим, поскольку визуально они практически совпадают. При $v = 6$: $\delta = 0.027565$, $\delta_{\max} = 0.031594$, $\omega = 2.3785e-4$; $\alpha =$

$= -(0.46309, 0.43636, 1.8492, 1.1983, 0.33155, -0.37226)$, $\beta = (0.92851, 2.462, 2.8463, 1.7451, 0.21825, 1.4669)$, $\gamma = (0.22713, 0.12356, -0.56682, 0.62113, 0.53356, -0.086174)$.

Графики не приводим, поскольку визуально они полностью совпадают.

Сделаем некоторые комментарии. Функции тестов № 1–4 гладкие, поэтому подпадают под ситуацию, описанную в теореме 1. Функция теста № 5 негладкая, но непрерывная и строго возрастающая, следовательно, подпадает под ситуацию, описанную в теореме 2. В численных экспериментах этого раздела для минимизации невязки мы использовали численные методы первого порядка (в основном, метод Левенберга–Марквардта и метод DFP – в зависимости от того, какой из них показывал лучшие результаты, с одномерным поиском по методу More–Thuente). Поэтому требовалось вычисление производных невязки по параметрам аппроксимации, и тут использовались формулы из теоремы 3.

СПИСОК ЛИТЕРАТУРЫ

1. Дзядык В.К. Введение в теорию равномерного приближения функций полиномами. М.: Наука, 1977. 512 с.
2. Зюко А.Г., Кловский Д.Д., Назаров М.В., Финк Л.М. Теория передачи сигналов. М.: Связь, 1980. 288 с.
3. Голубов Б.И., Ефимов А.В., Скворцов В.А. Ряды и преобразования Уолша. Теория и применения. М.: ЛКИ, 2008. 352 с.
4. Добеши И. Десять лекций по вейвлетам. Ижевск: НИЦ “Регулярная и хаотическая динамика”, 2001. 464 с.
5. Новиков И.Я., Протасов В.Ю., Скопина М.А. Теория всплесков. М.: Физматлит, 2006. 616 с.
6. Чернов А.В. О применении квадратичных экспонент для дискретизации задач оптимального управления // Вестн. Удмурт. ун-та. Математика. Механика. Компьютерные науки. 2017. Т. 27. Вып. 4. С. 558–575.
7. Чернов А.В. О применении функций Гаусса для численного решения задач оптимального управления // Автоматика и телемехан. 2019. 6. С. 51–69.
8. Robbins H., Monro S. A stochastic approximation method // Ann. Math. Stat. 1951. V. 22. 3. P. 400–407.
9. Голубков А.Ю. Построение внешних и внутренних функций представления непрерывных функций многих переменных суперпозицией непрерывных функций одного переменного // Фундамент. и прикл. матем. 2002. Т. 8. Вып. 1. С. 27–38.
10. Sprecher D.A. On the structure of continuous functions of several variables // Trans. Amer. Math. Soc. 1965. V. 115. P. 340–355.
11. Чернов А.В. О применении функций Гаусса в сочетании с теоремой Колмогорова для аппроксимации функций многих переменных // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. 5. С. 784–801.
12. Shisha O. Monotone approximation // Pacific J. of Math. 1965. V. 15. P. 667–671.
13. Roulier J.A. Monotone approximation of certain classes of function // J. of Approximat. Theory. 1968. V. 1. P. 319–324.
14. Lorentz G.G., Zeller K.L. Degree of approximation by monotone polynomials. I // J. of Approximat. Theory. 1968. V. 1. P. 501–504.
15. DeVore R.A. Monotone approximation by polynomials // SIAM J. of Math. Anal. 1977. V. 8. 5. P. 906–921.
16. DeVore R.A., Yu X.M. Pointwise estimates for monotone polynomial approximation // Constructive Approximat. 1985. V. 1. P. 323–331.
17. Шевчук И.А. Приближение монотонных функций монотонными многочленами // Матем. сборник. 1992. Т. 183. 5. С. 63–78.
18. DeVore R.A., Leviatan D., Shevchuk I.A. Approximation of monotone functions: a counter example // Proc. of Chamonix, 1996 / A. Le Mehaute, C. Rabut, L.L. Schumaker (eds.). P. 95–102.
19. Gilewicz J., Konovalov V.N., Leviatan D. Widths and shape-preserving widths of sobolev-type classes of s-monotone functions // J. of Approximat. Theory. 2006. V. 140. P. 101–126.
20. Kunsch R.J. The difficulty of Monte Carlo approximation of multivariate monotone functions // J. of Approximat. Theory. 2019. V. 241. P. 33–56.
21. More J.J., Thuente D.J. Line search algorithms with guaranteed sufficient decrease // ACM Transact. on Math. Software. 1994. V. 20. P. 286–307.
22. Зорич В.А. Математический анализ. Ч. II. М.: МЦНМО, 2002. XIV+794 с.
23. Воеводин В.В. Линейная алгебра. М.: Наука, 1980. 400 с.
24. Беллман Р. Введение в теорию матриц. М.: Наука, 1969. 368 с.

ОПТИМАЛЬНОЕ
УПРАВЛЕНИЕ

УДК 519.658

ПРЯМО-ДВОЙСТВЕННЫЙ МЕТОД НЬЮТОНА С НАИСКОРЕЙШИМ
СПУСКОМ ДЛЯ ЛИНЕЙНОЙ ЗАДАЧИ ПОЛУОПРЕДЕЛЕННОГО
ПРОГРАММИРОВАНИЯ. ИТЕРАЦИОННЫЙ ПРОЦЕСС¹⁾© 2022 г. В. Г. Жадан^{1,2}¹ 117333 Москва, ул. Вавилова, 40, ФИЦ ИУ РАН, Россия² 141700 Долгопрудный, М. о., Институтский пер., 9, МФТИ, Россия

e-mail: zhadan@ccas.ru

Поступила в редакцию 02.04.2021 г.
Переработанный вариант 02.04.2021 г.
Принята к публикации 16.12.2021 г.

Рассматривается прямо-двойственный метод Ньютона для решения линейной задачи полуопределенного программирования. При этом как прямые переменные, так и слабые двойственные переменные могут иметь неполный ранг и принадлежать границам допустимых множеств. Приводятся формулы для определения направлений перемещения, а также исследуются свойства этих направлений. Описывается способ выбора шагов перемещения, приводящий к уменьшению ранга симметризованного произведения прямой и слабой двойственной переменных. Библ. 16.

Ключевые слова: линейная задача полуопределенного программирования, прямо-двойственный метод Ньютона, итерационный процесс, наискорейший спуск.

DOI: 10.31857/S0044466922040135

ВВЕДЕНИЕ

Настоящая статья является продолжением работы [1], в которой рассматривался прямо-двойственный метод Ньютона для решения линейной задачи полуопределенного программирования и были выведены уравнения для определения направлений перемещений. Основная особенность при получении данных уравнений заключалась в том, что точки, в которых они выводились, могли принадлежать границам допустимых множеств. В таких точках ньютоновская система линейных уравнений для определения направлений перемещений принимает специальный вид. Было показано, что в пространстве симметричных матриц можно выделить четыре подпространства, которым должны принадлежать данные перемещения.

Цель настоящей работы заключается в получении точных аналитических формул для ньютоновских направлений как в пространстве прямых переменных, так и в пространстве слабых двойственных переменных, а также в построении итерационного процесса. Исследуются свойства ньютоновских направлений, рассматривается вопрос о выборе шага перемещения. Показывается, что в ходе итерационного процесса ранг симметризованного произведения прямой переменной и слабой двойственной переменной может только уменьшаться. Данный ранг характеризует выполнение условия комплементарности, которое входит в условия оптимальности для линейных задач полуопределенного программирования.

Приведем некоторые обозначения, введенные ранее в [1] и используемые в данной работе. Ниже $\mathbb{S}^n$ — пространство симметричных матриц порядка n , и $\mathbb{S}_+^n$ — конус положительно-полуопределенных матриц из $\mathbb{S}^n$. Конус $\mathbb{S}_+^n$ самосопряженный и задает в $\mathbb{S}^n$ частичный порядок $\succeq$, используя который будем писать $X \succeq 0$, если $X \in \mathbb{S}_+^n$. Размерность пространства $\mathbb{S}^n$ равняется “треугольному” числу $n^\Delta = \frac{1}{2}n(n+1)$.

¹⁾ Работа выполнена при финансовой поддержке РФФИ (проект 21-71-30005).

Скалярное (внутреннее) произведение между матрицами $X \in \mathbb{S}^n$ и $Y \in \mathbb{S}^n$ определяется по Фробениусу: $X \bullet Y = \text{tr}(X^T Y)$. Обычное евклидово скалярное произведение в $\mathbb{R}_n$ обозначается с помощью угловых скобок. Символ J^n обозначает множество индексов от 1 до n . Через $X \circ Y$ обозначается симметризованное произведение между матрицами из $\mathbb{S}^n$, определяемое как $X \circ Y = \frac{1}{2}(XY + YX)$.

Линейная задача полуопределенного программирования ставится стандартным образом

$$\min C \bullet X, \quad A_i \bullet X = b^i, \quad i \in J^m, \quad X \succeq 0. \quad (1)$$

Матрицы C , X и A_i , $i \in J^m$, принадлежат пространству $\mathbb{S}^n$. Двойственной к (1) является задача

$$\max \langle b, u \rangle, \quad Y = C - \sum_{i=1}^m u^i A_i, \quad Y \succeq 0, \quad (2)$$

в которой $b = (b^1; \dots; b^m) \in \mathbb{R}^m$. Предполагается, что решения задач (1) и (2) существуют и что матрицы A_i , $i \in J^m$, линейно независимы. Предполагается также, что $C \neq 0_{nn}$ и $b = [b^1; \dots; b^m] \neq 0_n$. Здесь и ниже 0_k есть k -мерный нулевой вектор, 0_{kl} — нулевая матрица размером $k \times l$. Точка с запятой указывает на то, что компоненты вектора располагаются в виде столбца, т.е. одна компонента под другой.

Пусть $\mathcal{F}_p$ и $\mathcal{F}_d$ — допустимые множества в задачах (1) и (2) соответственно. Если $X \in \mathcal{F}_p$ и $[u, Y] \in \mathcal{F}_d$ — решения (1) и (2), то они удовлетворяют системе равенств

$$X \circ Y = 0, \quad A_i \bullet X = b^i, \quad i \in J^m, \quad Y = C - \sum_{i=1}^m u^i A_i. \quad (3)$$

Последнее равенство в (3) можно представить в другом виде, аналогичном равенствам $A_i \bullet X = b^i$. Пусть $l = n^\Delta - m$. Обозначим через $K_1, \dots, K_l$ матрицы из $\mathbb{S}^n$, задающий базис в ортогональном дополнении пространства, порождаемого матрицами A_i , $i \in J^m$. Тогда, умножая равенство $Y = C - \sum_{i=1}^m u^i A_i$ скалярно на K_j , $j \in J^l$, получаем $K_j \bullet Y = c^j$, где $c^j = K_j \bullet C$. Условия оптимальности (3) заменяются на следующие:

$$X \circ Y = 0, \quad A_i \bullet X = b^i, \quad i \in J^m, \quad K_j \bullet Y = c^j, \quad j \in J^l. \quad (4)$$

Предлагаемый в [1] метод является вариантом метода Ньютона для решения системы (4).

1. НЬЮТОНОВСКАЯ СИСТЕМА В ВЕКТОРНОМ ВИДЕ

Пусть имеется допустимая пара $[X, Y] \in \mathbb{S}^n \times \mathbb{S}^n$. Это означает, что $X \succeq 0$, $Y \succeq 0$ и выполнены последние два равенства из (4). Считаем также, что пара $[X, Y]$ является согласованной, т.е. $X \circ Y \succeq 0$. Если для $[X, Y]$ первое равенство в (4) не выполняется, другими словами, пара не является *комплементарной*, то она неоптимальна, т.е. X и Y вместе с соответствующей двойственной переменной u не являются решениями задач (1) и (2). В этом случае желательно перейти в новую допустимую согласованную пару $[\bar{X}, \bar{Y}]$, используя для этой цели направления перемещения ΔX и ΔY . В качестве ΔX и ΔY могут быть взяты ньютоновские направления, получающиеся из применения метода Ньютона для решения системы (4). В [1] построена специальная система линейных уравнений для нахождения таких ΔX и ΔY . Дадим ее краткое описание.

Для симметричной матрицы $X \circ Y$ имеет место разложение: $X \circ Y = UD(\lambda)U^T$, где U — ортогональная матрица порядка n , $D(\lambda)$ — диагональная матрица с вектором собственных значений $X \circ Y$ на диагонали. Пусть n_p — ранг матрицы $X \circ Y$ и пусть $\lambda = [\lambda_p; \lambda_o]$, где λ_p — ненулевые собственные значения, λ_o — нулевые собственные значения. Тогда $U = [U_p, U_o]$, где подматрица U_p состоит из n_p столбцов U .

Введем в рассмотрение еще три матрицы U_B , U_N и U_Z , зависящие от образов матриц $X \in \mathbb{S}_+^n$ и $Y \in \mathbb{S}_+^n$, а также от их нуль-пространств. Матрица U_B задает ортонормированный базис в пересечении образа матрицы X с нуль-пространством матрицы Y . Соответственно, матрица U_N задает ортонормированный базис в пересечении образа матрицы Y с нуль-пространством матрицы X . Если пересечение нуль-пространств матриц X и Y имеет положительную размерность, то такие пары $[X, Y]$ называются *особыми*, в противном случае они называются *неособыми*. В неособых парах матрица U_Z отсутствует. Неособая регулярная пара $[X, Y] \in \mathbb{S}_+^n \times \mathbb{S}_+^n$ — это такая пара, когда $U_O = [U_B, U_N]$. В особой регулярной паре $U_O = [U_B, U_N, U_Z]$.

Предположим сначала для простоты, что неоптимальная пара $[X, Y] \in \mathbb{S}_+^n \times \mathbb{S}_+^n$ регулярная и неособая. В этом случае $\lambda_p > 0_{n_p}$. При этих предположениях, как показано в [1], матрицы $X_U = U^T X U$ и $Y_U = U^T Y U$ имеют блочно-диагональный вид

$$X_U = \begin{bmatrix} X_{U_p} & 0 & 0 \\ 0 & X_{U_B} & 0 \\ 0 & 0 & X_{U_N} \end{bmatrix}, \quad Y_U = \begin{bmatrix} Y_{U_p} & 0 & 0 \\ 0 & Y_{U_B} & 0 \\ 0 & 0 & Y_{U_N} \end{bmatrix} \quad (5)$$

с диагональными блоками

$$\begin{aligned} X_{U_p} &= U_p^T X U_p, & X_{U_B} &= U_B^T X U_B, & X_{U_N} &= U_N^T X U_N, \\ Y_{U_p} &= U_p^T Y U_p, & Y_{U_B} &= U_B^T Y U_B, & Y_{U_N} &= U_N^T Y U_N. \end{aligned}$$

Тогда в качестве направлений ΔX и ΔY также целесообразно брать блочно-диагональные матрицы с соответственно блоками ΔX_{U_p} , ΔX_{U_B} , ΔX_{U_N} и ΔY_{U_p} , ΔY_{U_B} , ΔY_{U_N} , т.е. полагать

$$\Delta X_U = \begin{bmatrix} \Delta X_{U_p} & 0 & 0 \\ 0 & \Delta X_{U_B} & 0 \\ 0 & 0 & \Delta X_{U_N} \end{bmatrix}, \quad \Delta Y_U = \begin{bmatrix} \Delta Y_{U_p} & 0 & 0 \\ 0 & \Delta Y_{U_B} & 0 \\ 0 & 0 & \Delta Y_{U_N} \end{bmatrix}. \quad (6)$$

Симметричные матрицы X_{U_p} , Y_{U_p} и ΔX_{U_p} , ΔY_{U_p} имеют порядок n_p (ранг матрицы $X \circ Y$). У матриц X_{U_B} , Y_{U_B} и ΔX_{U_B} , ΔY_{U_B} ранг n_B , у матриц X_{U_N} , Y_{U_N} и ΔX_{U_N} , ΔY_{U_N} ранг n_N . Пространство блочно-диагональных симметричных матриц с блоками соответственно порядков n_p , n_B и n_N обозначим через $\mathbb{S}_U^n$. Его размерность равняется $n_U = n_p^\Delta + n_B^\Delta + n_N^\Delta$.

В случае блочно-диагональных приращений (6) из ньютоновской системы уравнений для равенств (4) следует, что $\Delta X_{U_N} = 0_{n_N}$ и $\Delta Y_{U_B} = 0_{n_B}$, а сама система сводится к следующей:

$$\begin{aligned} Y_{U_p} \circ \Delta X_{U_p} + X_{U_p} \circ \Delta Y_{U_p} &= -D(\lambda_p), \\ A_{i,U_p} \bullet \Delta X_{U_p} + A_{i,U_B} \bullet \Delta X_{U_B} &= 0, \quad i \in J^m, \\ K_{j,U_p} \bullet \Delta Y_{U_p} + K_{j,U_N} \bullet \Delta Y_{U_N} &= 0, \quad j \in J^{l_U}. \end{aligned} \quad (7)$$

Матрицы A_{i,U_p} , A_{i,U_B} и K_{j,U_p} , K_{j,U_N} следующие:

$$A_{i,U_p} = U_p^T A_i U_p, \quad A_{i,U_B} = U_B^T A_i U_B, \quad K_{j,U_p} = U_p^T K_j U_p, \quad K_{j,U_N} = U_N^T K_j U_N,$$

где $i \in J^m$, $j \in J^{l_U}$, $l_U = n_U - m$. Блочно-диагональные матрицы с блоками K_{j,U_p} , K_{j,U_B} и K_{j,U_N} , $j \in J^{l_U}$, задают некоторый базис в ортогональном дополнении подпространства из $\mathbb{S}_U^n$, определяемым блочно-диагональными матрицами с блоками A_{i,U_p} , A_{i,U_B} , A_{i,U_N} , $i \in J^m$.

Чтобы решить систему линейных матричных уравнений (7), представим ее в векторном виде. Для этого введем дополнительные обозначения.

Пусть вес M обозначает *вектор длины* n^2 , являющийся прямой суммой столбцов квадратной матрицы M порядка n . В случае симметричной матрицы M будем пользоваться также вектором

$\text{svec } M$ длины n^Δ . В данный вектор в отличие от $\text{vec } M$ помещаются столбцы не полностью, а только их части, начинающиеся с главного элемента, причем все элементы, не стоящие на диагонали матрицы M , при помещении в $\text{svec } M$ умножаются на $\sqrt{2}$.

Для перехода от $\text{vec } M$ к $\text{svec } M$ и для обратного перехода нам потребуются специальные *элиминационные* и *дублирующие* матрицы (см. [2], [3]), которые обозначим как $\tilde{\mathcal{L}}_n$ и $\tilde{\mathcal{D}}_n$ соответственно. Обе матрицы $\tilde{\mathcal{L}}_n$ и $\tilde{\mathcal{D}}_n$ являются матрицами полного ранга и имеют соответственно размеры $n^\Delta \times n^2$ и $n^2 \times n^\Delta$. Если M – симметричная матрица порядка n , то

$$\text{svec } M = \tilde{\mathcal{L}}_n \text{vec } M, \quad \text{vec } M = \tilde{\mathcal{D}}_n \text{svec } M.$$

Для матриц $\tilde{\mathcal{L}}_n$ и $\tilde{\mathcal{D}}_n$ справедливо равенство $\tilde{\mathcal{L}}_n \tilde{\mathcal{D}}_n = I_{n^\Delta}$. Кроме того, $\tilde{\mathcal{D}}_n^\top \tilde{\mathcal{D}}_n = I_{n^\Delta}$. Здесь и ниже I_k – единичная матрица порядка k .

Матрице $M \in \mathbb{S}^n$ можно поставить в соответствие ее кронекеровскую сумму $M^\otimes$, определяемую как

$$M^\otimes = \frac{1}{2}(M \otimes I_n + I_n \otimes M),$$

где знак $\otimes$ между матрицами обозначает их произведение по Кронекеру. Матрица $M^\otimes$ симметричная и имеет порядок n^2 . Ниже через $\tilde{M}^\otimes$ обозначается квадратная матрица $\tilde{\mathcal{L}}_n M^\otimes \tilde{\mathcal{D}}_n$, имеющая уже порядок n^Δ . Если M – неособая матрица, то матрицы $M^\otimes$ и $\tilde{M}^\otimes$ также неособые. Более того, неособой оказывается и матрица $\tilde{\mathcal{L}}_n(M \otimes M)\tilde{\mathcal{D}}_n$, причем для обратной матрицы справедлива формула (см. [2])

$$[\tilde{\mathcal{L}}_n(M \otimes M)\tilde{\mathcal{D}}_n]^{-1} = \tilde{\mathcal{L}}_n(M^{-1} \otimes M^{-1})\tilde{\mathcal{D}}_n. \quad (8)$$

Для любых матриц M_1 , M_2 и M_3 , если определено их произведение $M_1 M_2 M_3$, имеет место равенство

$$\text{vec } M_1 M_2 M_3 = (M_3^\top \otimes M_1) \text{vec } M_2. \quad (9)$$

С помощью формулы (9) получаем, что равенство $X_{U_p} \circ Y_{U_p} = D(\lambda_p)$ может быть записано следующим образом:

$$\text{svec}(X_{U_p} \circ Y_{U_p}) = \tilde{X}_{U_p}^\otimes \text{svec } Y_{U_p} \tilde{Y}_{U_p}^\otimes \text{svec } X_{U_p} = \text{svec } D(\lambda_p). \quad (10)$$

Здесь $\tilde{X}_{U_p}^\otimes = \tilde{\mathcal{L}}_{n_p} X_{U_p}^\otimes \tilde{\mathcal{D}}_{n_p}$, $\tilde{Y}_{U_p}^\otimes = L_{n_p} Y_{U_p}^\otimes \tilde{\mathcal{D}}_{n_p}$.

Если матрицы $\tilde{X}_{U_p}^\otimes$ и $\tilde{Y}_{U_p}^\otimes$ неособые, то неформально через $\tilde{X}_{U_p}^{-\otimes}$ и $\tilde{Y}_{U_p}^{-\otimes}$ будем обозначать их обратные матрицы, т.е. $(\tilde{X}_{U_p}^\otimes)^{-1}$ и $(\tilde{Y}_{U_p}^\otimes)^{-1}$ соответственно. Обозначим также для упрощения записи $\mathcal{G}_p = \tilde{Y}_{U_p}^{-\otimes} \tilde{X}_{U_p}^\otimes$. Тогда на основании (10) получаем

$$\text{svec } X_{U_p} = \mathcal{G}_p \text{svec } Y_{U_p} = \tilde{Y}_{U_p}^{-\otimes} \text{svec } D(\lambda_p), \quad (11)$$

$$\text{svec } Y_{U_p} = \mathcal{G}_p^{-1} \text{svec } X_{U_p} = \tilde{X}_{U_p}^\otimes \text{svec } D(\lambda_p). \quad (12)$$

Пусть $\mathcal{A}_{\text{svec}}^{U_p}$ и $\mathcal{A}_{\text{svec}}^{U_B}$ – матрицы, строками которых являются соответственно векторы $\text{svec } A_{i,U_p}$ и $\text{svec } A_{i,U_B}$, $i \in J^m$. Аналогично, $\mathcal{K}_{\text{svec}}^{U_p}$ и $\mathcal{K}_{\text{svec}}^{U_N}$ – матрицы, строками которых являются соответственно векторы $\text{svec } K_{j,U_p}$ и $\text{svec } K_{j,U_N}$, $j \in J^{l_U}$. Тогда, используя равенства

$$\text{svec}(Y_{U_p} \circ \Delta X_{U_p}) = \tilde{Y}_{U_p}^\otimes \text{svec } \Delta X_{U_p}^{U_p}, \quad \text{svec}(X_{U_p} \circ \Delta Y_{U_p}) = \tilde{X}_{U_p}^\otimes \text{svec } \Delta Y_{U_p}, \quad (13)$$

получающиеся аналогично (11) и (12), приходим к тому, что ньютоновская система (7) может быть представлена в следующем виде:

$$\tilde{Y}_{U_P}^{\otimes} \text{svec} \Delta X_{U_P} + \tilde{X}_{U_P}^{\otimes} \text{svec} \Delta Y_{U_P} = -\text{svec} D(\lambda_P), \quad (14)$$

$$\mathcal{A}_{\text{svec}}^{U_P} \text{svec} \Delta X_{U_P} + \mathcal{A}_{\text{svec}}^{U_B} \text{svec} \Delta X_{U_B} = 0_m, \quad (15)$$

$$\mathcal{H}_{\text{svec}}^{U_P} \text{svec} \Delta Y_{U_P} + \mathcal{H}_{\text{svec}}^{U_N} \text{svec} \Delta Y_{U_N} = 0_{l_U}. \quad (16)$$

Обозначим через Ψ матрицу системы (14)–(16), через f – ее правую часть, т.е. положим

$$\Psi = \begin{bmatrix} \tilde{Y}_{U_P}^{\otimes} & 0 & \tilde{X}_{U_P}^{\otimes} & 0 \\ \mathcal{A}_{\text{svec}}^{U_P} & \mathcal{A}_{\text{svec}}^{U_B} & 0 & 0 \\ 0 & 0 & \mathcal{H}_{\text{svec}}^{U_P} & \mathcal{H}_{\text{svec}}^{U_N} \end{bmatrix}, \quad f = \begin{bmatrix} -\text{svec} D(\lambda_P) \\ 0_m \\ 0_{l_U} \end{bmatrix}.$$

Тогда система (14)–(16) перепишется в виде $\Psi z = f$, где

$$z = [\text{svec} \Delta X_{U_P}; \text{svec} \Delta X_{U_B}; \text{svec} \Delta Y_{U_P}; \text{svec} \Delta Y_{U_N}].$$

В [1] дано определение невырожденной пары $[X, Y]$, где $X \in \mathbb{S}_U^n$, $Y \in \mathbb{S}_U^n$. В невырожденной паре $[X, Y]$ строки матриц $[\mathcal{A}_{\text{svec}}^{U_P}, \mathcal{A}_{\text{svec}}^{U_B}]$ и $[\mathcal{H}_{\text{svec}}^{U_P}, \mathcal{H}_{\text{svec}}^{U_N}]$ линейно независимы. Более того, пара $[X, Y]$ является невырожденной в том и только в том случае, когда выполняются неравенства $m \geq n_B^\Delta \geq m - n_P^\Delta$. Если $n_B > 0$, то матрицы A_{i, U_B} , $i \in J^m$, порождают все пространство $\mathbb{S}^{n_B}$. Аналогично, если $n_N > 0$, то матрицы K_{j, U_N} , $j \in J^{l_U}$, порождают все пространство $\mathbb{S}^{n_N}$.

Утверждение 1. Пусть допустимая пара $[X, Y] \in \mathbb{S}_U^n \times \mathbb{S}_U^n$ регулярная и согласованная. Пусть, кроме того, она является невырожденной парой. Тогда матрица Ψ неособая.

Доказательство. Так как при сделанных предположениях матрицы X_{U_P} и Y_{U_P} , имеющие порядок n_P , неособые, то матрицы $\tilde{X}_{U_P}^{\otimes}$ и $\tilde{Y}_{U_P}^{\otimes}$ также неособые порядка n_P^Δ . Учитывая это, умножим первую строку матрицы Ψ на матрицу $\mathcal{A}_{\text{svec}}^{U_P} \tilde{Y}_{U_P}^{-\otimes}$ и вычтем ее из второй строки. В результате приходим к матрице

$$\Psi_1 = \begin{bmatrix} \tilde{Y}_{U_P}^{\otimes} & 0 & \tilde{X}_{U_P}^{\otimes} & 0 \\ 0 & \mathcal{A}_{\text{svec}}^{U_B} & -\mathcal{A}_{\text{svec}}^{U_P} G_P & 0 \\ 0 & 0 & \mathcal{H}_{\text{svec}}^{U_P} & \mathcal{H}_{\text{svec}}^{U_N} \end{bmatrix}.$$

Определитель получившейся матрицы Ψ_1 совпадает с определителем матрицы Ψ . Отсюда следует, что $\det \Psi = \det \tilde{Y}_{U_P}^{\otimes} \det \Psi_2$, где

$$\Psi_2 = \begin{bmatrix} \mathcal{A}_{\text{svec}}^{U_B} & -\mathcal{A}_{\text{svec}}^{U_P} G_P & 0 \\ 0 & \mathcal{H}_{\text{svec}}^{U_P} & \mathcal{H}_{\text{svec}}^{U_N} \end{bmatrix}.$$

Следовательно, $\det \Psi \neq 0$, если $\det \Psi_2 \neq 0$.

В случае, когда $n_B = 0$ и, следовательно, $n_B^\Delta = 0$, матрица Ψ_2 имеет вид

$$\Psi_3 = \begin{bmatrix} -\mathcal{A}_{\text{svec}}^{U_P} G_P & 0 \\ \mathcal{H}_{\text{svec}}^{U_P} & \mathcal{H}_{\text{svec}}^{U_N} \end{bmatrix}.$$

Более того, поскольку $[X, Y]$ – невырожденная пара и $n_B = 0$, то $n_P^\Delta \geq m$. Поэтому матрица $\mathcal{A}_{\text{svec}}^{U_P}$ имеет полный ранг, равный m . Матрица $\mathcal{A}_{\text{svec}}^{U_P} G_P$ также имеет полный ранг, равный m .

Обратимся к линейной однородной системе

$$\Psi_3 \mu = 0_{n_U}. \quad (17)$$

Пусть $\mu = [\mu_P; \mu_N]$. Тогда, как следует из уравнения $\mathcal{H}_{\text{svec}}^{U_P} \mu_P + \mathcal{H}_{\text{svec}}^{U_N} \mu_N = 0$, вектор μ должен принадлежать ортогональному дополнению к подпространству, порожденному строками матрицы $[K_{\text{svec}}^{U_P}, K_{\text{svec}}^{U_N}]$. Но данное ортогональное дополнение согласно определению матриц K_{j,U_P} и K_{j,U_N} , $j \in J^{l_U}$, порождается строками матрицы $[A_{\text{svec}}^{U_P}, A_{\text{svec}}^{U_N}]$. Поэтому решение μ должно иметь вид

$$\mu = \begin{bmatrix} \mu_P \\ \mu_N \end{bmatrix} \begin{bmatrix} (A_{\text{svec}}^{U_P})^T \\ (A_{\text{svec}}^{U_N})^T \end{bmatrix} v$$

для некоторого вектора $v \in \mathbb{R}_m$. Отсюда, в частности, получаем $\mu_P = (A_{\text{svec}}^{U_P})^T v$. После подстановки данного μ_P в уравнение в (17), приходим к заключению, что должно выполняться равенство $A_{\text{svec}}^{U_P} \mathcal{G}_P (A_{\text{svec}}^{U_P})^T v = 0_m$. Так как матрица $A_{\text{svec}}^{U_P} \mathcal{G}_P (A_{\text{svec}}^{U_P})^T$ неособая, то $v = 0_m$. Таким образом, однородная система (17) имеет только нулевое решение. Следовательно, матрица этой системы неособая, ее ранг отличен от нулевого.

Перейдем теперь к рассмотрению общего случая, когда $n_B > 0$. Так как пара $[X, Y]$ невырожденная, то выполняется неравенство $n_B^\Delta \leq m$.

Если $n_B^\Delta = m$, то матрица $A_{\text{svec}}^{U_B}$ квадратная порядка m . Опять же в силу невырожденности пары $[X, Y]$ матрицы A_{i,U_B} , $i \in J^m$, порождают все пространство $\mathbb{S}^{n_B}$. Поэтому матрица $A_{\text{svec}}^{U_B}$, стоящая в левом верхнем углу Ψ_2 , обязательно должна быть неособой. Кроме того, матрица $[\mathcal{H}_{\text{svec}}^{U_P}, \mathcal{H}_{\text{svec}}^{U_N}]$, стоящая в нижней строке Ψ_2 , также оказывается квадратной порядка $l_U = n_P^\Delta + n_N^\Delta$, причем в силу линейной независимости матриц $K_j^{U_N}$, $j \in J^{l_U}$, следующей из невырожденности пары $[X, Y]$, она неособая. Поэтому вся матрица Ψ_2 , будучи квадратной и имеющей порядок $n_P^\Delta + n_B^\Delta + n_N^\Delta$, оказывается неособой.

Рассмотрим наконец случай, когда $0 < n_B^\Delta < m$, и покажем, что однородная система уравнений $\Psi_2 \mu = 0_{n_U}$ с квадратной матрицей Ψ_2 порядка n_U также может иметь только тривиальное решение.

Пусть $d_B = m - n_B^\Delta > 0$. Разобьем вектор μ на три подвектора, а именно, $\mu = [\mu_B; \mu_P; \mu_N] \in \mathbb{R}^{n_U}$. Система $\Psi_2 \mu = 0$ распадается на две подсистемы:

$$A_{\text{svec}}^{U_B} \mu_B - A_{\text{svec}}^{U_P} \mathcal{G}_P \mu_P = 0_m, \quad (18)$$

$$\mathcal{H}_{\text{svec}}^{U_P} \mu_P + \mathcal{H}_{\text{svec}}^{U_N} \mu_N = 0_{l_U}. \quad (19)$$

Так как пара $[X, Y]$ невырожденная, то матрица $[\mathcal{H}_{\text{svec}}^{U_P}, \mathcal{H}_{\text{svec}}^{U_N}]$ подсистемы (19) размером $l_U \times (l_U + d_B)$ имеет полный ранг, равный l_U . Следовательно, общее решение этой подсистемы есть d_B -мерное линейное подпространство.

Пусть $\mathcal{L}_B$ — подпространство в $\mathbb{R}^m$, порожденное столбцами матрицы $A_{\text{svec}}^{U_B}$ и пусть E_B — матрица некоторого базиса в подпространстве $\mathcal{L}_B^\perp$, являющемся ортогональным дополнением $\mathcal{L}_B$. Поскольку опять в силу невырожденности пары $[X, Y]$ матрицы A_{i,U_B} , $i \in J^m$, порождают пространство $\mathbb{S}^{n_B}$, то матрица $A_{\text{svec}}^{U_B}$ имеет полный ранг, равный n_B^Δ . Поэтому у матрицы E_B размер $m \times d_B$.

Покажем, что общее решение подсистемы (19) имеет вид

$$\begin{bmatrix} \mu_P \\ \mu_N \end{bmatrix} = \begin{bmatrix} (A_{\text{svec}}^{U_P})^T \\ (A_{\text{svec}}^{U_N})^T \end{bmatrix} E_B v,$$

где $\mathbf{v} \in \mathbb{R}^{d_B}$. Действительно, подставляя данное решение в (19), в силу ортогональности матриц $\tilde{A}_i^U$ и $\tilde{K}_j^U$, $i \in J^m$, $j \in J^{l_U}$, получаем

$$\begin{aligned} & \left[\mathcal{H}_{\text{svec}}^{U_P} \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T + \mathcal{H}_{\text{svec}}^{U_N} \left(\mathcal{A}_{\text{svec}}^{U_N} \right)^T \right] E_B \mathbf{v} = \\ & = \left[\mathcal{H}_{\text{svec}}^{U_B} \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^T + \mathcal{H}_{\text{svec}}^{U_P} \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T + \mathcal{H}_{\text{svec}}^{U_N} \left(\mathcal{A}_{\text{svec}}^{U_N} \right)^T \right] E_B \mathbf{v} = 0. \end{aligned}$$

При этом матрица $\left[\mathcal{A}_{\text{svec}}^{U_P}, \mathcal{A}_{\text{svec}}^{U_N} \right]^T E_B$ имеет полный ранг, равный d_B . Чтобы в этом убедиться, достаточно показать, что полный ранг, равный d_B , имеет подматрица $\left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B$.

Число столбцов матрицы $\left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B$ не больше числа строк. Действительно, иначе выполнялось бы неравенство $d_B = m - n_B^\Delta > n_P^\Delta$ или $m > n_B^\Delta + n_P^\Delta$, которое противоречит неравенству $m \leq n_B^\Delta + n_P^\Delta$, следующему из предположения о невырожденности пары $[X, Y]$. Таким образом, ранг матрицы $\left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B$ не превышает d_B . На самом деле, он равен d_B . Пусть это не так, тогда $\left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B \mathbf{h} = 0_{n_P^\Delta}$ для некоторого ненулевого вектора $\mathbf{h} \in \mathbb{R}^{d_B}$. Но это означает, что ненулевой вектор $E_B \mathbf{h}$, принадлежащий нуль-пространству матрицы $\left(\mathcal{A}_{\text{svec}}^{U_B} \right)^T$, принадлежит одновременно нуль-пространству матрицы $\left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T$. Последнее невозможно, из-за того, что в силу предположения о невырожденности пары $[X, Y]$ строки матрицы $\left[\mathcal{A}_{\text{svec}}^{U_P}, \mathcal{A}_{\text{svec}}^{U_B} \right]$ линейно независимы.

Подставим найденное выражение для μ_P в подсистему (18), в результате получаем следующее уравнение:

$$\mathcal{A}_{\text{svec}}^{U_B} \mu_B - \mathcal{A}_{\text{svec}}^{U_P} \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B \mathbf{v} = 0_m, \quad (20)$$

которое после умножения слева на матрицу E_B^T принимает вид

$$E_B^T \mathcal{A}_{\text{svec}}^{U_P} \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B \mathbf{v} = 0_m.$$

Матрица этой системы, как следует из вышесказанного, неособая, поэтому $\mathbf{v} = 0_{d_B}$.

Мы установили, что $\mu_P = 0_{n_P^\Delta}$ и $\mu_N = 0_{n_N^\Delta}$. Кроме того, из (20) в силу линейной независимости столбцов матрицы $\mathcal{A}_{\text{svec}}^{U_B}$ получаем, что $\mu_B = 0_{n_B^\Delta}$. Таким образом, однородная система $\Psi_2 \mu = 0_{n_U}$ имеет только нулевое решение. Следовательно, матрица Ψ_2 неособая. Утверждение доказано.

2. РЕШЕНИЕ НЬЮТОНОВСКОЙ СИСТЕМЫ

Рассмотрим разные случаи решения системы (14)–(16) в зависимости от ранга n_B матрицы U_B . Предполагаем, что текущая допустимая пара $[X, Y] \in \mathbb{S}_U^n \times \mathbb{S}_U^n$ является невырожденной, а также регулярной и согласованной.

Случай 1: $n_B^\Delta = m$. При выполнении этого равенства матрица $\mathcal{A}_{\text{svec}}^{U_B}$ квадратная порядка m , причем из-за предположения о невырожденности пары $[X, Y]$ данная матрица неособая. Имеем

$$\Delta Y_U = - \sum_{i=1}^m \Delta u^i \begin{bmatrix} A_{i,U_P} & 0 & 0 \\ 0 & A_{i,U_B} & 0 \\ 0 & 0 & A_{i,U_N} \end{bmatrix} \quad (21)$$

для некоторого $\Delta u = [\Delta u^1; \dots, \Delta u^m] \in \mathbb{R}^m$. Кроме того, как уже отмечалось, $\Delta Y_{U_B} = 0_{n_B n_B}$. Отсюда, поскольку квадратные симметричные матрицы A_{i,U_B} порядка n_B линейно независимы и их ровно m штук, приходим к выводу, что $\Delta u = 0_m$. В этом случае согласно (21) выполняется следую-

щее: $\Delta Y_U = 0$ и, в частности, $\Delta Y_{U_p} = 0_{n_p n_p}$. Таким образом, первое уравнение (14) сводится к $\tilde{Y}_{U_p}^{\otimes} \text{svec} \Delta X_{U_p} = -\text{svec} D(\lambda_p)$, разрешая которое, получаем $\text{svec} \Delta X_{U_p} = -\tilde{Y}_{U_p}^{-\otimes} \text{svec} D(\lambda_p)$.

Так как $D(\lambda_p) = X_{U_p} \circ Y_{U_p}$, то в силу (10) $\text{svec} D(\lambda_p) = \tilde{Y}_{U_p}^{\otimes} \text{svec} X_{U_p}$. Поэтому выражение для $\text{svec} \Delta X_{U_p}^{U_p}$ упрощается

$$\text{svec} \Delta X_{U_p} = -\text{svec} X_{U_p}. \quad (22)$$

Зная $\text{svec} \Delta X_{U_p}$, из второго уравнения (15) находим

$$\text{svec} \Delta X_{U_B} = \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^{-1} \mathcal{A}_{\text{svec}}^{U_p} \text{svec} X_{U_p} = \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^{-1} b - \text{svec} X_{U_B}. \quad (23)$$

Случай 2: $0 < n_B^{\Delta} < m$. При этом предположении матрица $\mathcal{A}_{\text{svec}}^{U_B}$ становится прямоугольной, причем число столбцов в ней меньше числа строк. Обозначим через $\mathcal{M}$ пространство столбцов матрицы $\mathcal{A}_{\text{svec}}^{U_B}$, через $\mathcal{M}^{\perp}$ — его ортогональное дополнение. Пусть E_B — матрица некоторого базиса в подпространстве $\mathcal{M}^{\perp}$, она имеет размер $m \times d_B$, где $d_B = m - n_B^{\Delta}$. Так как, по-прежнему, $\Delta Y_{U_B} = 0_{n_B n_B}$, то

$$\text{svec} \Delta Y_{U_B} = -\left(\mathcal{A}_{\text{svec}}^{U_B} \right)^{\top} \Delta u = 0_{n_B^{\Delta}}.$$

Отсюда следует, что $\Delta u = E_B q$ для некоторого $q \in \mathbb{R}^{d_B}$. Поэтому

$$\text{svec} \Delta Y_{U_p} = -\left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top} E_B q, \quad \text{svec} \Delta Y_{U_N} = -\left(\mathcal{A}_{\text{svec}}^{U_N} \right)^{\top} E_B q. \quad (24)$$

Умножим обе части уравнения (14) слева на матрицу $\tilde{Y}_{U_p}^{-\otimes}$ и подставим в получившееся уравнение $\text{svec} \Delta Y_{U_p}$ из (24). В результате приходим к следующему выражению для $\text{svec} \Delta X_{U_p}$:

$$\text{svec} \Delta X_{U_p} = \tilde{Y}_{U_p}^{-\otimes} \left[\tilde{X}_{U_p}^{\otimes} \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top} E_B q - \text{svec} D(\lambda_p) \right].$$

Если учесть (11), то

$$\text{svec} \Delta X_{U_p} = \mathcal{G}_p \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top} E_B q - \text{svec} X_{U_p}.$$

Подставляя найденное $\text{svec} \Delta X_{U_p}$ в уравнение (15) и умножая далее (15) слева на матрицу $E_B^{\top}$, с учетом того, что $E_B^{\top} \mathcal{A}_{\text{svec}}^{U_B} = 0_{d_B}$, получаем

$$\left(E_B^{\top} \mathcal{A}_{\text{svec}}^{U_p} \mathcal{G}_p \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top} E_B \right) q = E_B^{\top} \mathcal{A}_{\text{svec}}^{U_p} \text{svec} X_{U_p}.$$

Отсюда находим $q = \Phi_p^{-1} E_B^{\top} \mathcal{A}_{\text{svec}}^{U_p} \text{svec} X_{U_p}$, где $\Phi_p = E_B^{\top} \mathcal{H}_p E_B$, $\mathcal{H}_p = \mathcal{A}_{\text{svec}}^{U_p} \mathcal{G}_p \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top}$. Таким образом, вводя еще одно обозначение $\tilde{\Phi}_p^{-1} = E_B \Phi_p^{-1} E_B^{\top}$, приходим к следующей формуле для приращения $\text{svec} \Delta X_{U_p}$:

$$\text{svec} \Delta X_{U_p} = \left[\mathcal{G}_p \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top} \tilde{\Phi}_p^{-1} \mathcal{A}_{\text{svec}}^{U_p} - I_{n_p^{\Delta}} \right] \text{svec} X_{U_p}. \quad (25)$$

Но $\mathcal{A}_{\text{svec}}^{U_p} \text{svec} X_{U_p} = b - \mathcal{A}_{\text{svec}}^{U_B} \text{svec} X_{U_B}$, поэтому

$$\text{svec} \Delta X_{U_p} = \mathcal{G}_p \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^{\top} \tilde{\Phi}_p^{-1} b - \text{svec} X_{U_p}. \quad (26)$$

Кроме того, из (15) получаем

$$\text{svec} \Delta X_{U_B} = -\left[\left(\mathcal{A}_{\text{svec}}^{U_B} \right)^{\top} \mathcal{A}_{\text{svec}}^{U_B} \right]^{-1} \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^{\top} \mathcal{A}_{\text{svec}}^{U_p} \text{svec} \Delta X_{U_p}.$$

Следовательно, согласно (26),

$$\mathcal{A}_{\text{svec}}^{U_P} \text{svec} \Delta X_{U_P} = \mathcal{H}_P \tilde{\Phi}_P^{-1} b + \mathcal{A}_{\text{svec}}^{U_B} \text{svec} X_{U_B} - b,$$

что влечет

$$\text{svec} \Delta X_{U_B} = \left[\left(\mathcal{A}_{\text{svec}}^{U_B} \right)^T \mathcal{A}_{\text{svec}}^{U_B} \right]^{-1} \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^T \left(I_m - \mathcal{H}_P \tilde{\Phi}_P^{-1} \right) b - \text{svec} X_{U_B}. \quad (27)$$

Далее, используя соотношение $\text{svec} \Delta Y_U = - \left(\mathcal{A}_{\text{svec}}^U \right)^T \Delta u$, вычисляем соответствующие направления в Y -пространстве:

$$\text{svec} \Delta Y_{U_P} = - \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} b, \quad \text{svec} \Delta Y_{U_N} = - \left(\mathcal{A}_{\text{svec}}^{U_N} \right)^T \tilde{\Phi}_P^{-1} b. \quad (28)$$

Как уже отмечалось, $\text{svec} \Delta X_{U_N} = 0_{n_N^\Delta}$ и $\text{svec} \Delta Y_{U_B} = 0_{n_B^\Delta}$.

Случай 3: $n_B = 0$. При этом предположении $\mathcal{A}_{\text{svec}}^{U_P} \text{svec} X_{U_P} = b$, а из уравнения (15) следует, что $\mathcal{A}_{\text{svec}}^{U_P} \text{svec} \Delta X_{U_P} = 0_m$. Имеем $\text{svec} \Delta Y_{U_P} = - \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \Delta u$ для некоторого m -мерного вектора Δu . Тогда уравнение (14) с учетом (11) сводится к

$$\text{svec} \Delta X_{U_P} - \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \Delta u = - \text{svec} X_{U_P}. \quad (29)$$

Умножим уравнение (29) слева на матрицу $\mathcal{A}_{\text{svec}}^{U_P}$, в результате находим $\Delta u = \mathcal{H}_P^{-1} b$. После подстановки найденного Δu в (29) получаем

$$\text{svec} \Delta X_{U_P} = \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \mathcal{H}_P^{-1} b - \text{svec} X_{U_P}. \quad (30)$$

Отсюда, в частности, следует, что $\text{svec} \Delta X_{U_P} = 0_{n_P^\Delta}$, когда $n_P^\Delta = m$. Для $\text{svec} \Delta X_{U_N}$ по-прежнему имеем $\text{svec} \Delta X_{U_N} = 0_{n_N^\Delta}$.

С помощью соотношения $\text{svec} \Delta Y_U = - \left(\mathcal{A}_{\text{svec}}^U \right)^T \Delta u$ определяем дополнительно

$$\text{svec} \Delta Y_{U_P} = - \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \mathcal{H}_P^{-1} b, \quad \text{svec} \Delta Y_{U_N} = - \left(\mathcal{A}_{\text{svec}}^{U_N} \right)^T \mathcal{H}_P^{-1} b.$$

Обратим внимание, что матрица $\mathcal{H}_P^{-1}$ переходит в введенную ранее матрицу $\tilde{\Phi}_P^{-1}$, если в последней взять $E_B = I_m$.

Нами получены формулы для приращений ΔX и ΔY , в которых присутствуют матрицы A_{i,U_P} и A_{i,U_B} , $i \in J^m$. Можно также вывести формулы для приращений ΔX и ΔY , в которых присутствуют только матрицы K_{j,U_P} и K_{j,U_N} , $j \in J^{l_U}$. При нахождении таких приращений вместо (15) используется уравнение (16).

3. СВОЙСТВА НЬЮТОНОВСКИХ НАПРАВЛЕНИЙ

Рассмотрим некоторые простейшие свойства, которыми обладают ньютоновские направления ΔX_U и ΔY_U . Эти свойства являются характерными для метода Ньютона, когда он используется для решения оптимизационных задач.

Утверждение 2. Направления ΔX_U и ΔY_U ортогональны друг другу, т.е.

$$\Delta X_U \bullet \Delta Y_U = 0. \quad (31)$$

Доказательство. Равенство (31) в эквивалентной форме записывается в виде

$$\langle \text{svec} \Delta X_U, \text{svec} \Delta Y_U \rangle = 0. \quad (32)$$

Учтем далее, что $\text{svec} \Delta X_{U_N} = 0_{n_N^\Delta}$ и $\text{svec} \Delta Y_{U_B} = 0_{n_B^\Delta}$. Тогда (32) выполняется, если

$$\langle \text{svec} \Delta X_{U_P}, \text{svec} \Delta Y_{U_P} \rangle = 0. \quad (33)$$

Покажем, что равенство (33) действительно имеет место, причем считаем, что направления $\text{svec } \Delta X_{U_p}$ и $\text{svec } \Delta Y_{U_p}$ берутся равными соответственно (26) и (28). Подставляя их, получаем

$$\begin{aligned} \langle \text{svec } \Delta X_{U_p}, \text{svec } \Delta Y_{U_p} \rangle &= \\ &= - \left\langle \left[\mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_p} - I_{n_p^\Delta} \right] \text{svec } X_{U_p}, \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^T \tilde{\Phi}_P^{-1} b \right\rangle = \\ &= \left\langle \left[\tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_p} \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_p} - \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_p} \right] \text{svec } X_{U_p}, b \right\rangle = 0, \end{aligned}$$

так как $\tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_p} \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_p} \right)^T \tilde{\Phi}_P^{-1} = \tilde{\Phi}_P^{-1}$. Мы пришли к (33) и, следовательно, к (31). Утверждение доказано.

Введем в рассмотрение функцию

$$\varphi(\alpha) = (X + \alpha \Delta X) \bullet (Y + \alpha \Delta Y),$$

где $\alpha \geq 0$.

Утверждение 3. Пусть $[X, Y]$ – регулярная допустимая согласованная пара. Тогда

$$\varphi(\alpha) = (1 - \alpha) \varphi(0). \quad (34)$$

Доказательство. Поскольку $X \succeq 0$, $Y \succeq 0$, то $\varphi(0) = X \bullet Y \geq 0$, причем равенство возможно только в том случае, когда пара $[X, Y]$ оптимальная. Имеем

$$\varphi(\alpha) = X_U \bullet Y_U + \alpha(Y_U \bullet \Delta X_U + X_U \bullet \Delta Y_U) + \alpha^2 \Delta X_U \bullet \Delta Y_U,$$

или, если учесть (31),

$$\varphi(\alpha) = X_U \bullet Y_U + \alpha(Y_U \bullet \Delta X_U + X_U \bullet \Delta Y_U).$$

Примем далее во внимание, что $X_{U_N} = 0$, $\Delta X_{U_N} = 0$ и $Y_{U_B} = 0$, $\Delta Y_{U_B} = 0$. Тогда $\varphi(\alpha) = X_{U_p} \bullet Y_{U_p} + \alpha h$, где $h = Y_{U_p} \bullet \Delta X_{U_p} + X_{U_p} \bullet \Delta Y_{U_p}$.

Так как

$$Y_{U_p} \bullet \Delta X_{U_p} = \text{tr } Y_{U_p} \Delta X_{U_p} = \text{tr } \Delta X_{U_p} Y_{U_p}$$

и

$$X_{U_p} \bullet \Delta Y_{U_p} = \text{tr } X_{U_p} \Delta Y_{U_p} = \text{tr } \Delta Y_{U_p} X_{U_p},$$

то $h = \text{tr}(Y_{U_p} \circ \Delta X_{U_p}) + \text{tr}(X_{U_p} \circ \Delta Y_{U_p})$. Поэтому в силу первого уравнения из (7) справедливо равенство $h = -\text{tr } D(\lambda_p) = -\sum_{i=1}^{n_p} \lambda_p^i$. Поскольку $\text{tr } D(\lambda_p) = \text{tr}(X_{U_p} \circ Y_{U_p}) = \varphi(0)$, то выполняется (34). Утверждение доказано.

Приращение $\text{svec } \Delta Y_U$ связано с приращением двойственной переменной Δu формулой: $\text{svec } \Delta Y_U = -\left(A_{\text{svec}}^U\right)^T \Delta u$, причем, как следует из вышесказанного, имеем

$$\Delta u = \begin{cases} 0_m, & n_B^\Delta = m, \\ E_B \Phi_P^{-1} E_B^T b, & 0 < n_B^\Delta < m, \\ \mathcal{H}_P^{-1} b, & n_B = 0. \end{cases} \quad (35)$$

Покажем, что направление Δu приводит к увеличению значения целевой функции в двойственной задаче (2). Для этого сделаем дополнительное предположение относительно задачи (1). Пусть

$$\mathcal{F}_{\mathcal{A}} = \{Z \in \mathbb{S}^n : A_i \bullet Z = b^i, 1 \leq i \leq m\}.$$

Условие A_1 . Не существует матрицы $Z \in \mathcal{F}_{\mathcal{A}}$ ранга r такой, что $r^\Delta < m$.

Утверждение 4. Пусть $[X, Y] \in \mathbb{S}_U^n \times \mathbb{S}_U^n$ – регулярная допустимая согласованная пара. Пусть, кроме того, выполнено условие A_1 . Тогда $\langle b, \Delta u \rangle \geq 0$, причем равенство возможно только при $\Delta u = 0_m$.

Доказательство. Рассмотрим только общий случай формулы (35) для приращения Δu , когда $0 < n_B^\Delta < m$. Как уже отмечалось, третий случай $n_B^\Delta = m$ сводится ко второму, если в качестве матрицы E_B взять единичную матрицу I_m . Имеем

$$\langle b, \Delta u \rangle = \langle E_B^\top b, \Phi_P^{-1} E_B^\top b \rangle \geq 0, \quad (36)$$

поскольку матрица Φ_P положительно определена. Неравенство (36) переходит в равенство, когда $E_B^\top b = 0_{d_B}$. Но выполнение данного равенства означает, что вектор b принадлежит нуль-пространству матрицы $E_B^\top$, совпадающему с пространством столбцов матрицы $\mathcal{A}_{\text{svec}}^{U_B}$. Но тогда для некоторого ненулевого $z \in \mathbb{R}^{n_B^\Delta}$ имеет место равенство

$$\mathcal{A}_{\text{svec}}^{U_B} z = b. \quad (37)$$

Пусть $Z \in \mathbb{S}^n$ – матрица, у которой соответствующая матрица $Z_U = U^\top Z U$ принадлежит пространству $\mathbb{S}_U^n$, является блочно-диагональной, причем с единственным ненулевым диагональным блоком Z_{U_B} . Считаем также, что $\text{svec } Z_{U_B} = z$. Ранг r матрицы Z не больше, чем n_B , и на основании (37) получаем

$$A_i \bullet Z = A_i^U \bullet Z_U = \langle \text{svec } A_i^{U_B}, \text{svec } Z_{U_B} \rangle = b^i, \quad 1 \leq i \leq m,$$

т.е. $Z \in \mathcal{F}_{\mathcal{A}}$. Но $r^\Delta \leq n_B^\Delta < m$, что противоречит условию A_1 . Таким образом, возможен только случай, когда $n_B = m$. Но согласно (35) тогда выполняется равенство $\Delta u = 0_m$. Утверждение доказано.

Утверждение 5. Пусть $[X, Y]$ – регулярная допустимая согласованная пара. Тогда верно

$$C \bullet \Delta X \leq 0. \quad (38)$$

причем равенство возможно только при $\Delta X_U = 0$.

Доказательство. Рассмотрим общий случай, когда направления $\text{svec } \Delta X_{U_P}$ и $\text{svec } \Delta Y_{U_P}$ вычисляются согласно (26) и (28). Неравенство (38) перепишем в виде

$$\langle \text{svec } C_U, \text{svec } \Delta X_U \rangle \leq 0. \quad (39)$$

Так как $\Delta X_{U_N} = 0_{n_N n_N}$, то

$$\langle \text{svec } C_U, \text{svec } \Delta X_U \rangle = \langle \text{svec } C_{U_P}, \text{svec } \Delta X_{U_P} \rangle + \langle \text{svec } C_{U_B}, \text{svec } \Delta X_{U_B} \rangle. \quad (40)$$

Имеем с учетом того, что $Y_{U_B} = 0_{n_B n_B}$,

$$\text{svec } C_{U_P} = \text{svec } Y_{U_P} + \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^\top u, \quad \text{svec } C_{U_B} = \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^\top u. \quad (41)$$

Следовательно,

$$\langle \text{svec } C_{U_P}, \text{svec } \Delta X_{U_P} \rangle = \langle \text{svec } Y_{U_P}, \text{svec } \Delta X_{U_P} \rangle + \langle u, \mathcal{A}_{\text{svec}}^{U_P} \text{svec } \Delta X_{U_P} \rangle, \quad (42)$$

причем $\text{svec } Y_{U_P}$ удовлетворяют равенству (12).

Для второго слагаемого в (40) выполняется

$$\langle \text{svec } C_{U_B}, \text{svec } \Delta X_{U_B} \rangle = \langle u, \mathcal{A}_{\text{svec}}^{U_B} \text{svec } \Delta X_{U_B} \rangle. \quad (43)$$

Подставляя (42), (43) в (40) и учитывая (15), получаем

$$\begin{aligned} \langle \text{svec } C_U, \text{svec } \Delta X_U \rangle &= \langle \text{svec } Y_{U_P}, \text{svec } \Delta X_{U_P} \rangle + \langle u, \mathcal{A}_{\text{svec}}^{U_P} \text{svec } \Delta X_{U_P} + \mathcal{A}_{\text{svec}}^{U_B} \text{svec } \Delta X_{U_B} \rangle = \\ &= \langle \text{svec } Y_{U_P}, \text{svec } \Delta X_{U_P} \rangle. \end{aligned}$$

На основании (26) имеем

$$\begin{aligned} \langle \text{svec } Y_{U_P}, \text{svec } \Delta X_{U_P} \rangle &= \left\langle \text{svec } Y_{U_P}, \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} b \right\rangle - \langle \text{svec } Y_{U_P}, \text{svec } X_{U_P} \rangle = \\ &= \left\langle \text{svec } Y_{U_P}, \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_P} \text{svec } X_{U_P} \right\rangle - \langle \text{svec } Y_{U_P}, \text{svec } X_{U_P} \rangle. \end{aligned}$$

Здесь мы учли, что $b = \mathcal{A}_{\text{svec}}^{U_P} \text{svec } X_{U_P} + \mathcal{A}_{\text{svec}}^{U_B} \text{svec } X_{U_B}$.

Из (12) и (26) следует, что

$$\begin{aligned} \langle \text{svec } Y_{U_P}, \text{svec } \Delta X_{U_P} \rangle &= \left\langle \mathcal{G}_P^{-1} \text{svec } X_{U_P}, \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_P} \mathcal{G}_P \mathcal{G}_P^{-1} \text{svec } X_{U_P} \right\rangle - \\ &- \left\langle \mathcal{G}_P^{-1} \text{svec } X_{U_P}, \mathcal{G}_P \mathcal{G}_P^{-1} \text{svec } X_{U_P} \right\rangle, \end{aligned}$$

или, обозначая $z = \mathcal{G}_P^{-1} \text{svec } X_{U_P}$, $Z = Z_1 - \mathcal{G}_P$, $Z_1 = \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_P} \mathcal{G}_P$,

$$\langle \text{svec } Y_{U_P}, \text{svec } \Delta X_{U_P} \rangle = \langle z, Zz \rangle.$$

Найдем собственные значения ξ матрицы Z . Они удовлетворяют уравнению

$$\left(\mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_P} \mathcal{G}_P - \mathcal{G}_P \right) y = \xi y, \quad (44)$$

где y — ненулевой собственный вектор Z . Умножая равенство (44) слева на матрицу $E_B^T \mathcal{A}_{\text{svec}}^{U_P}$, приходим к равенству $\xi E_B^T \mathcal{A}_{\text{svec}}^{U_P} y = 0$. Следовательно, если $E_B^T \mathcal{A}_{\text{svec}}^{U_P} y \neq 0$, т.е. вектор y не принадлежит нуль-пространству матрицы $E_B^T \mathcal{A}_{\text{svec}}^{U_P}$, то $\xi = 0$.

Предположим далее, что $E_B^T \mathcal{A}_{\text{svec}}^{U_P} y = 0$. В этом случае вектор $\mathcal{A}_{\text{svec}}^{U_P} y$ принадлежит подпространству $\mathcal{M}$, являющемуся пространством столбцов матрицы $\mathcal{A}_{\text{svec}}^{U_B}$, т.е. $\mathcal{A}_{\text{svec}}^{U_P} y = \mathcal{A}_{\text{svec}}^{U_B} \rho$ для некоторого ρ .

Умножим скалярно равенство (44) на вектор $\mathcal{G}_P^{-T} y$. Тогда

$$\left\langle \mathcal{G}_P^{-T} y, \left(\mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_P} \mathcal{G}_P - \mathcal{G}_P \right) y \right\rangle = \xi \langle \mathcal{G}_P^{-T} y, y \rangle.$$

Так как $\mathcal{A}_{\text{svec}}^{U_P} y = \mathcal{A}_{\text{svec}}^{U_B} \rho$, то данное равенство сводится к $\langle y, y \rangle + \xi \langle \mathcal{G}_P^{-T} y, y \rangle = 0$. Откуда $\xi = -\|y\|^2 / \langle \mathcal{G}_P^{-T} y, y \rangle < 0$.

Таким образом, у матрицы Z все собственные значения действительные и неположительные. Следовательно, неравенство $C \bullet \Delta X \leq 0$ действительно имеет место.

Равенство $C \bullet \Delta X = 0$ выполняется тогда и только тогда, когда $Zz = 0$. В этом случае

$$\mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \tilde{\Phi}_P^{-1} \mathcal{A}_{\text{svec}}^{U_P} \text{svec } X_{U_P} = \text{svec } X_{U_P}.$$

Отсюда согласно (25) $\text{svec } \Delta X_{U_P} = 0$. Поэтому согласно (15) имеем

$$\mathcal{A}_{\text{svec}}^{U_B} \text{svec } \Delta X_{U_B} = 0_m.$$

Поскольку в невырожденной паре $[X, Y]$ столбцы матрицы $\mathcal{A}_{\text{svec}}^{U_B}$ линейно независимы, то $\text{svec } \Delta X_{U_B} = 0$. Таким образом, $\Delta X_U = 0$. Утверждение доказано.

4. ИТЕРАЦИОННЫЙ ПРОЦЕСС И ВЫБОР ШАГА

Предположим, что задана начальная допустимая регулярная согласованная пара $[X^0, Y^0]$. Пусть, кроме того, на k -м шаге итерационного процесса найдена пара $[X^k, Y^k]$, также являющаяся допустимой регулярной согласованной парой. Обозначим через U ортогональную матрицу порядка n такую, что $X^k \circ Y^k = UD(\lambda)U^T$. Считаем, для определенности, что $U = [U_P, U_B, U_N]$, где

ранги матриц U_P , U_B и U_N равны соответственно n_P , n_B и n_N . Для вектора собственных значений λ матрицы U выполняется: $\lambda = [\lambda_P; 0]$, $\lambda_P \in \mathbb{R}^{n_P}$. Для X^k и Y^k справедливы представления

$$X_U^k = \begin{bmatrix} X_{U_P}^k & 0 & 0 \\ 0 & X_{U_B}^k & 0 \\ 0 & 0 & X_{U_N}^k \end{bmatrix}, \quad Y_U^k = \begin{bmatrix} Y_{U_P}^k & 0 & 0 \\ 0 & Y_{U_B}^k & 0 \\ 0 & 0 & Y_{U_N}^k \end{bmatrix}, \quad (45)$$

где

$$X_{U_P}^k = U_P^T X^k U_P, \quad X_{U_B}^k = U_B^T X^k U_B, \quad X_{U_N}^k = U_N^T X^k U_N$$

и

$$Y_{U_P}^k = U_P^T Y^k U_P, \quad Y_{U_B}^k = U_B^T Y^k U_B, \quad Y_{U_N}^k = U_N^T Y^k U_N.$$

Возьмем симметричные блочно-диагональные матрицы

$$\Delta X_U^k = \begin{bmatrix} \Delta X_{U_P}^k & 0 & 0 \\ 0 & \Delta X_{U_B}^k & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Delta Y_U^k = \begin{bmatrix} \Delta Y_{U_P}^k & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta Y_{U_N}^k \end{bmatrix},$$

которым соответствуют векторы $\text{svec } \Delta X_{U_P}^k$, $\text{svec } \Delta X_{U_B}^k$ и $\text{svec } \Delta Y_{U_P}^k$, $\text{svec } \Delta Y_{U_N}^k$.

Имея формулы для ньютоновских направлений ΔX_U^k и ΔY_U^k , построим итерационный процесс

$$X_U^{k+1} = X_U^k + \alpha_k \Delta X_U^k, \quad Y_U^{k+1} = Y_U^k + \alpha_k \Delta Y_U^k, \quad (46)$$

где шаг $\alpha_k \geq 0$ подбирается таким образом, чтобы он был максимально возможным при выполнении следующих условий: $X_U^{k+1} \succeq 0$, $Y_U^{k+1} \succeq 0$ и $X_U^{k+1} \circ Y_U^{k+1} \succeq 0$. Тогда, если все пары $[X^k, Y^k]$, $k \geq 0$, оказываются допустимыми, регулярными, согласованными и неособыми, то итерационный процесс (46) полностью определен.

Для того, чтобы $X_U^{k+1} \succeq 0$, должны выполняться неравенства

$$X_{U_P}^{k+1} = X_{U_P}^k + \alpha_k \Delta X_{U_P}^k \succeq 0, \quad X_{U_B}^{k+1} = X_{U_B}^k + \alpha_k \Delta X_{U_B}^k \succeq 0. \quad (47)$$

Для матриц $X_{U_P}^k$ и $X_{U_B}^k$ справедливы представления (см. [1]):

$$X_{U_P}^k = F_{X_P} D(\eta_{BN}) F_{X_P}^T, \quad X_{U_B}^k = D(\eta_{BB}), \quad (48)$$

где F_{X_P} — ортогональная матрица порядка n_P , η_{BN} и η_{BB} — векторы собственных значений матрицы X^k со всеми строго положительными компонентами.

Пусть

$$\Delta \hat{X}_{U_P}^k = \Delta X_{U_P}^k + X_{U_P}^k, \quad \Delta \hat{X}_{U_B}^k = \Delta X_{U_B}^k + X_{U_B}^k.$$

Тогда на основании (22), (26) и (30) имеем

$$\text{svec } \Delta \hat{X}_{U_P}^k = \begin{cases} 0_m, & n_B^\Delta = m, \\ \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T E_B \left(E_B^T \mathcal{H}_P E_B \right)^{-1} E_B^T b, & 0 < n_B^\Delta < m, \\ \mathcal{G}_P \left(\mathcal{A}_{\text{svec}}^{U_P} \right)^T \mathcal{H}_P^{-1} b, & n_B = 0, \end{cases}$$

а на основании (23) и (27), получаем

$$\text{svec} \Delta \hat{X}_{U_B}^k = \begin{cases} \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^{-1} b, & n_B^\Delta = m, \\ \left[\left(\mathcal{A}_{\text{svec}}^{U_B} \right)^T \mathcal{A}_{\text{svec}}^{U_B} \right]^{-1} \left(\mathcal{A}_{\text{svec}}^{U_B} \right)^T \left(I_m - \mathcal{H}_P \tilde{\Phi}_P^{-1} \right) b, & 0 < n_B^\Delta < m. \end{cases}$$

В введенных обозначениях первое рекуррентное соотношение в итерационном процессе (46) расщепляется на следующие два:

$$X_{U_P}^{k+1} = (1 - \alpha_k) X_{U_P}^k + \alpha_k \hat{\Delta} X_{U_P}^k, \quad X_{U_B}^{k+1} = (1 - \alpha_k) X_{U_B}^k + \alpha_k \hat{\Delta} X_{U_B}^k.$$

Тогда, как следует из (48), имеем

$$X_{U_P}^{k+1} = F_{X_P} \left[(1 - \alpha_k) D(\eta_{BN}) + \alpha_k W_{X_P} \right] F_{X_P}^T,$$

где через W_{X_P} обозначена матрица $W_{X_P} = F_{X_P}^T \hat{\Delta} X_{U_P}^k F_{X_P}$.

Обе матрицы $D(\eta_{BN})$ и W_{X_P} являются симметричными, причем первая матрица положительно-определенная. Поэтому с помощью конгруэнтного преобразования обе они могут быть приведены к диагональному виду [16]. Пусть $\mathcal{T}_{X_P}$ — неособая матрица, осуществляющая такое преобразование:

$$\mathcal{T}_{X_P} D(\eta_{BN}) \mathcal{T}_{X_P}^{-1} = D(\bar{e}), \quad \mathcal{T}_{X_P} W_{X_P} \mathcal{T}_{X_P}^{-1} = D(\sigma_P),$$

где $\bar{e}$ есть n_P -мерный вектор, состоящий из единиц. Тогда

$$X_{U_P}^{k+1} = F_{X_P} \mathcal{T}_{X_P}^{-1} \left[(1 - \alpha_k) D(\bar{e}) + \alpha_k D(\sigma_P) \right] \mathcal{T}_{X_P} F_{X_P}^T. \quad (49)$$

Схожим образом получаем

$$X_{U_B}^{k+1} = \mathcal{T}_{X_B}^{-1} \left[(1 - \alpha_k) D(\bar{e}) + \alpha_k D(\sigma_B) \right] \mathcal{T}_{X_B}, \quad (50)$$

где $\mathcal{T}_{X_B}$ — неособая матрица, осуществляющая конгруэнтное преобразование:

$$\mathcal{T}_{X_B} D(\eta_{BB}) \mathcal{T}_{X_B}^{-1} = D(\bar{e}), \quad \mathcal{T}_{X_B} \Delta \hat{X}_B^k \mathcal{T}_{X_B}^{-1} = D(\sigma_B).$$

Теперь $\bar{e}$ есть n_B -мерный вектор, состоящий из единиц. Из (49) и (50) видно, что если минимальная из компонент векторов σ_P и σ_B меньше единицы, то шаг α_k конечный. Более того, если среди компонент векторов σ_P и σ_B имеются отрицательные, то обязательно шаг α_k удовлетворяет неравенству: $0 < \alpha_k < 1$.

Обратимся ко второму рекуррентному соотношению из (46). Наряду с (47) должны выполняться соотношения

$$Y_{U_P}^{k+1} = Y_{U_P}^k + \alpha_k \Delta Y_{U_P}^k \succeq 0, \quad Y_{U_N}^{k+1} = Y_{U_N}^k + \alpha_k \Delta Y_{U_N}^k \succeq 0,$$

причем для матриц $Y_{U_P}^k$ и $Y_{U_N}^k$ справедливы представления, аналогичные (48):

$$Y_{U_P}^k = F_{Y_P} D(\theta_{NB}) F_{Y_P}^T, \quad Y_{U_N}^k = D(\theta_{NN}),$$

где F_{Y_P} — ортогональная матрица порядка n_P , θ_{NB} и θ_{NN} — векторы собственных значений матрицы Y^k со всеми строго положительными компонентами.

Проведем с помощью неособых матриц $\mathcal{T}_{Y_P}$ и $\mathcal{T}_{Y_N}$ конгруэнтные преобразования:

$$\mathcal{T}_{Y_P} D(\theta_{NB}) \mathcal{T}_{Y_P}^{-1} = D(\bar{e}), \quad \mathcal{T}_{Y_P} W_{Y_P} \mathcal{T}_{Y_P}^{-1} = D(\omega_P),$$

$$\mathcal{T}_{Y_N} D(\theta_{NN}) \mathcal{T}_{Y_N}^{-1} = D(\bar{e}), \quad \mathcal{T}_{Y_N} \Delta Y_N^k \mathcal{T}_{Y_N}^{-1} = D(\omega_N),$$

где $W_{Y_P} = F_{Y_P}^T \Delta Y_{U_P}^k F_{Y_P}$. Тогда получаем

$$Y_{U_P}^{k+1} = F_{Y_P} \mathcal{T}_{Y_P}^{-1} \left[D(\bar{e}) + \alpha_k D(\omega_P) \right] \mathcal{T}_{Y_P} F_{Y_P}^T, \quad Y_{U_N}^{k+1} = \mathcal{T}_{Y_N}^{-1} \left[D(\bar{e}) + \alpha_k D(\omega_N) \right] \mathcal{T}_{Y_N}.$$

Таким образом, шаг α_k должен быть ограничен, если среди компонент векторов ω_p или ω_N имеются отрицательные. Более того, если какая-либо компонента ω_p или ω_N окажется меньше -1 , то $0 < \alpha_k < 1$. Так как $\Delta Y = 0_{nn}$ при $n_B^\Delta = m$, то в данном случае ограничение на шаг α_k становится несущественным.

Пусть $Z^k = X^k \circ Y^k$. Из разложения $Z^k = U_p D(\lambda_p) U_p^T$ и равенства

$$U_p^T Z^k U_p + U_p^T \Delta Z^k U_p = 0$$

следует, что $\Delta Z_{U_p}^k = U_p^T \Delta Z^k U_p = -D(\lambda_p)$. Поэтому $Z^k + \alpha_k \Delta Z^k \succeq 0$ для $0 \leq \alpha_k \leq 1$.

На основании вышесказанного приходим к следующему результату.

Утверждение 6. Пусть $0 \leq \alpha_k \leq 1$. Тогда

$$\text{rank } X_{U_p}^{k+1} \leq \text{rank } X_{U_p}^k, \quad \text{rank } Y_{U_p}^{k+1} \leq \text{rank } Y_{U_p}^k.$$

Более того, $\text{rank } Z^{k+1} \leq \text{rank } Z^k$, т.е. ранг матрицы $Z = X \circ Y$ не увеличивается в ходе итераций.

Определение 1. k -ю итерацию назовем *активной*, если $\text{rank } Z^{k+1} < \text{rank } Z^k$.

Так как допустимая пара $[X, Y]$ является оптимальной, если $\text{rank } X \circ Y = 0$, то в случае, когда метод находит решение задач (1) и (2), то он находит это решение за не более чем n активных итераций.

Согласно утверждению 7, если пара $[X^k, Y^k]$ была допустимой, регулярной, согласованной и неособой, то следующая пара $[X^{k+1}, Y^{k+1}]$ является допустимой, регулярной и согласованной. Однако она может оказаться особой. В этом случае, следует решить вспомогательную линейную полуопределенную задачу дополненности, о которой пойдет речь в следующем разделе.

В заключение данного раздела рассмотрим вопрос о матрице $\tilde{\Phi}_p^{-1}$. Чтобы ее вычислить, надо знать матрицу $\mathcal{G}_p = \tilde{Y}_{U_p}^{-\otimes} \tilde{X}_{U_p}^{\otimes}$. Найдем выражение для $Y_{U_p}^{-\otimes}$.

Матрица $\tilde{Y}_{U_p}^{\otimes}$ представима в виде

$$\tilde{Y}_{U_p}^{\otimes} = \tilde{\mathcal{L}}_{n_p} \left(F_{Y_p} D(\theta_{NB}) F_{Y_p}^T \right)^{\otimes} \tilde{\mathcal{D}}_{n_p},$$

причем, так как F_{Y_p} — ортогональная матрица, то справедлива следующая формула:

$$\left(F_{Y_p} D(\theta_{NB}) F_{Y_p}^T \right)^{\otimes} = (F_{Y_p} \otimes F_{Y_p}) D^{\otimes}(\theta_{NB}) (F_{Y_p}^T \otimes F_{Y_p}^T), \quad (51)$$

где матрица $F_{Y_p} \otimes F_{Y_p}$ также ортогональная, $(F_{Y_p} \otimes F_{Y_p})^{-1} = F_{Y_p}^T \otimes F_{Y_p}^T$.

Воспользуемся далее формулами, имеющими место для любой квадратной матрицы M порядка n_p (см. [2]):

$$\tilde{\mathcal{D}}_{n_p} \tilde{\mathcal{L}}_{n_p} (M \otimes M) \tilde{\mathcal{D}}_{n_p} = (M \otimes M) \tilde{\mathcal{D}}_{n_p}, \quad (52)$$

$$\tilde{\mathcal{D}}_{n_p} \tilde{\mathcal{L}}_{n_p} M^{\otimes} \tilde{\mathcal{D}}_{n_p} = M^{\otimes} \tilde{\mathcal{D}}_{n_p}. \quad (53)$$

Тогда согласно (51), (52) и (53) получаем

$$\begin{aligned} \tilde{Y}_{U_p}^{\otimes} &= \tilde{\mathcal{L}}_{n_p} (F_{Y_p} \otimes F_{Y_p}) D^{\otimes}(\theta_{NB}) (F_{Y_p}^T \otimes F_{Y_p}^T) \tilde{\mathcal{D}}_{n_p} = \\ &= \tilde{\mathcal{L}}_{n_p} (F_{Y_p} \otimes F_{Y_p}) D^{\otimes}(\theta_{NB}) \tilde{\mathcal{D}}_{n_p} \tilde{\mathcal{L}}_{n_p} (F_{Y_p}^T \otimes F_{Y_p}^T) \tilde{\mathcal{D}}_{n_p} = \\ &= \tilde{\mathcal{L}}_{n_p} (F_{Y_p} \otimes F_{Y_p}) \tilde{\mathcal{D}}_{n_p} \tilde{\mathcal{L}}_{n_p} D^{\otimes}(\theta_{NB}) \tilde{\mathcal{D}}_{n_p} \tilde{\mathcal{L}}_{n_p} (F_{Y_p}^T \otimes F_{Y_p}^T) \tilde{\mathcal{D}}_{n_p}. \end{aligned}$$

Отсюда, используя формулу (8), приходим к следующему выражению для обратной матрицы:

$$\tilde{Y}_{U_p}^{-\otimes} = \mathcal{W} \left(\tilde{\mathcal{L}}_{n_p} D^{\otimes}(\theta_{NB}) \tilde{\mathcal{D}}_{n_p} \right)^{-1} \mathcal{W}^{-1},$$

где введено обозначение $\mathcal{W} = \tilde{\mathcal{L}}_{n_p}(F_{Y_p} \otimes F_{Y_p})\tilde{\mathcal{D}}_{n_p}$, причем для обратной матрицы $\mathcal{W}^{-1}$ выполняется следующее: $\mathcal{W}^{-1} = \tilde{\mathcal{L}}_{n_p}(F_{Y_p}^T \otimes F_{Y_p}^T)\tilde{\mathcal{D}}_{n_p} = \mathcal{W}^T$. Матрица $\tilde{\mathcal{L}}_{n_p}D^{\otimes}(\Theta_{NB})\tilde{\mathcal{D}}_{n_p}$ — диагональная порядка n_p^{Δ} со строго положительными диагональными элементами.

5. ВСПОМОГАТЕЛЬНАЯ ЗАДАЧА ДОПОЛНИТЕЛЬНОСТИ

Полученные в разд. 2 ньютоновские направления ΔX_U и ΔY_U вычислялись в парах $[X, Y]$, которые являлись неособыми, т.е. пересечение нуль-пространств матриц X и Y имело нулевую размерность. Предположим, теперь, что пара $[X, Y]$ особая, в таких парах размерность указанного подпространства положительная. Предполагаем также, что $U = [U_p, U_B, U_N, U_Z]$ и ранг матрицы U_Z равен $n_Z > 0$.

Ньютоновская система (7) в особой паре видоизменяется

$$\begin{aligned} Y_{U_p} \circ \Delta X_{U_p} + X_{U_p} \circ \Delta Y_{U_p} &= -D(\lambda_p), \\ A_{i,U_p} \bullet \Delta X_{U_p} + A_{i,U_B} \bullet \Delta X_{U_B} + A_{i,U_Z} \bullet \Delta X_{U_Z} &= 0, \quad i \in J^m, \\ K_{j,U_p} \bullet \Delta Y_{U_p} + K_{j,U_N} \bullet \Delta Y_{U_N} + K_{j,U_Z} \bullet \Delta Y_{U_Z} &= 0, \quad j \in J^{l_v}. \end{aligned} \quad (54)$$

Здесь $A_{i,U_Z} = U_Z^T A_i U_Z$, $K_{j,U_Z} = U_Z^T K_j U_Z$, $\Delta X_{U_Z} = U_Z^T \Delta X U_Z$, $\Delta Y_{U_Z} = U_Z^T \Delta Y U_Z$. Векторное представление системы следующее:

$$\tilde{Y}_{U_p}^{\otimes} \text{svec} \Delta X_{U_p} + \tilde{X}_{U_p}^{\otimes} \text{svec} \Delta Y_{U_p} = -\text{svec} D(\lambda_p), \quad (55)$$

$$\mathcal{A}_{\text{svec}}^{U_p} \text{svec} \Delta X_{U_p} + \mathcal{A}_{\text{svec}}^{U_B} \text{svec} \Delta X_{U_B} + \mathcal{A}_{\text{svec}}^{U_Z} \text{svec} \Delta X_{U_Z} = 0_m, \quad (56)$$

$$\mathcal{K}_{\text{svec}}^{U_p} \text{svec} \Delta Y_{U_p} + \mathcal{K}_{\text{svec}}^{U_N} \text{svec} \Delta Y_{U_N} + \mathcal{K}_{\text{svec}}^{U_Z} \text{svec} \Delta Y_{U_Z} = 0_{l_v}, \quad (57)$$

где $\mathcal{A}_{\text{svec}}^{U_Z}$ и $\mathcal{K}_{\text{svec}}^{U_Z}$ — матрицы, строками которых являются соответственно векторы $\text{svec} A_i^{U_Z}$ и $\text{svec} K_j^{U_Z}$.

Система (55)–(57) недоопределена и, следовательно, имеет целое множество решений. Зафиксируем $\text{svec} \Delta X_{U_Z}$ и $\text{svec} \Delta Y_{U_Z}$. Тогда решение системы относительно остальных переменных единственное и может быть получено аналогично тому, как это делается в разд. 2. При этом выражения для ΔX_{U_p} , ΔX_{U_B} , ΔY_{U_p} и ΔY_{U_N} полностью сохраняются с единственной заменой b на $b_1 = b - \mathcal{A}_{\text{svec}}^{U_Z} \text{svec} \Delta X_{U_Z}$, т.е. становятся зависимыми от $\text{svec} \Delta X_{U_Z}$.

Пусть ΔX_{U_Z} и ΔY_{U_Z} — матрицы, которым соответствуют векторы $\text{svec} \Delta X_{U_Z}$ и $\text{svec} \Delta Y_{U_Z}$. Распорядимся приращением ΔX_{U_Z} таким образом, чтобы

$$\Delta X_{U_Z} \succeq 0, \quad \Delta Y_{U_Z} \succeq 0, \quad \Delta X_{U_Z} \bullet \Delta Y_{U_Z} = 0.$$

Обозначим через $\mathcal{A}_{U_Z}$ матрицу $\mathcal{A}_{U_Z} = [A_{1,U_Z}; \dots; A_{m,U_Z}]$. Здесь, как и в случае с векторами, знак точки с запятой между матрицами означает, что эти матрицы помещаются одна под другой в виде матрицы-столбца. Под $\mathcal{A}_{U_Z} \bullet \Delta X_{U_Z}$ будем понимать m -мерный вектор

$$\mathcal{A}_{U_Z} \bullet \Delta X_{U_Z} = [A_{1,U_Z} \bullet \Delta X_{U_Z}; \dots; A_{m,U_Z} \bullet \Delta X_{U_Z}].$$

Согласно своему определению

$$\Delta Y_{U_Z} = -\sum_{i=1}^m A_{i,U_Z} \Delta u^i.$$

Теперь для Δu при $0 \leq n_B^{\Delta} < m$ выполняется

$$\Delta u = \tilde{\Phi}_p^{-1} \left(b - \mathcal{A}_{\text{svec}}^{U_Z} \text{svec} \Delta X_{U_Z} \right),$$

причем матрица $\tilde{\Phi}_p^{-1}$ заменяется на $\mathcal{H}_p^{-1}$, когда $n_B = 0$. Поэтому верно

$$\Delta Y_{U_Z} = -\sum_{i=1}^m (\tilde{\Phi}_p^{-1} b)^i A_{i,U_Z} + \sum_{i=1}^m [\tilde{\Phi}_p^{-1} (\mathcal{A}_{U_Z} \bullet \Delta X_{U_Z})]^i A_{i,U_Z}. \quad (58)$$

Обозначим далее

$$\lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) = \tilde{\Phi}_p^{-1} [\mathcal{A}_{U_Z} \bullet \Delta X_{U_Z}] \in \mathbb{R}^m, \quad \Theta = -\sum_{i=1}^m (\tilde{\Phi}_p^{-1} b)^i A_{i,U_Z}$$

и определим линейное относительно ΔX_{U_Z} отображение $\Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) : S^{n_Z} \rightarrow S^{n_Z}$, положив

$$\Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) = \sum_{i=1}^m \lambda_{\mathcal{A}, \tilde{\Phi}_p}^i(\Delta X_{U_Z}) A_{i,U_Z}.$$

Тогда согласно (58) имеем $\Delta Y_{U_Z} = \Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) + \Theta$.

Введем в рассмотрение вспомогательную задачу, которая является линейной полуопределенной задачей дополнителности

$$\begin{aligned} \Delta X_{U_Z} \bullet \Delta Y_{U_Z} &= 0, \\ \Delta Y_{U_Z} &= \Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) + \Theta, \\ \Delta X_{U_Z} &\succeq 0, \quad \Delta Y_{U_Z} \succeq 0. \end{aligned} \quad (59)$$

Утверждение 7. Пусть допустимая особая пара $[X, Y]$ такова, что матрицы

$$\begin{bmatrix} A_{i,U_B} & 0 \\ 0 & A_{i,U_Z} \end{bmatrix}, \quad i \in J^m, \quad (60)$$

линейно независимы. Тогда решение задачи (59) существует, причем единственное.

Доказательство. Поставим в соответствие задаче дополнителности (59) задачу квадратичного программирования

$$\begin{aligned} \Delta X_{U_Z} \bullet (\Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) + \Theta) &\rightarrow \min, \\ \Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) + \Theta &\succeq 0, \quad \Delta X_{U_Z} \succeq 0. \end{aligned} \quad (61)$$

Если взять $q = \mathcal{A}_{U_Z} \bullet \Delta X_{U_Z} \in \mathbb{R}^m$, то целевую функцию в задаче (61) можно записать как $\tilde{f}(q) = \langle q, \tilde{\Phi}_p^{-1} q \rangle - \langle \tilde{\Phi}_p^{-1} b, q \rangle$.

Покажем, что матрица $\tilde{\Phi}_p^{-1}$ такова, что $\langle q, \tilde{\Phi}_p^{-1} q \rangle > 0$ для любых ненулевых q . Действительно, так как матрица Φ_p^{-1} положительно определена, то $\langle q, \tilde{\Phi}_p^{-1} q \rangle = 0$ в том и только в том случае, когда $E_B^T q = 0$, т.е. вектор $q = \mathcal{A}_{U_Z} \bullet \Delta X_{U_Z} = \mathcal{A}_{\text{svec}}^{U_Z} \text{svec} \Delta X_{U_Z}$ принадлежит ядру матрицы E_B^T . Но данное ядро совпадает с образом матрицы $\mathcal{A}_{\text{svec}}^{U_B}$. Поэтому столбцы матрицы $[\mathcal{A}_{\text{svec}}^{U_B}, \mathcal{A}_{\text{svec}}^{U_Z}]$ оказываются линейно зависимыми, что противоречит предположению о линейной независимости матриц (60).

Матрица Φ_p^{-1} является положительно-определенной, поэтому квадратичная функция $\langle w, \Phi_p^{-1} w \rangle$, где $w = E_B^T q$, является сильно выпуклой и, следовательно, достигает своего минимального значения по w , причем единственного, на любом выпуклом замкнутом множестве. Отсюда следует достижение минимального значения функции

$$f(\Delta X_{U_Z}) = \Delta X_{U_Z} \bullet (\Lambda_{\mathcal{A}, \tilde{\Phi}_p}(\Delta X_{U_Z}) + \Theta)$$

на допустимом множестве в задаче (61), т.е. задача (61) также имеет решение.

Согласно необходимым условиям минимума

$$f_{\Delta X_{U_Z}}(\Delta X_{U_Z}) \succeq 0, \quad X \bullet f_{\Delta X_{U_Z}}(\Delta X_{U_Z}) = 0,$$

где $f_{\Delta X_{U_Z}}(\Delta X_{U_Z})$ — градиент функции $f(\Delta X_{U_Z})$ по ΔX_{U_Z} . Данные условия можно переписать в виде

$$\Delta Y_{U_Z} = f_{\Delta X_{U_Z}}(\Delta X_{U_Z}) = \Lambda_{\mathcal{A}, \Phi_p}(\Delta X_{U_Z}) + \Theta \succeq 0, \quad \Delta X_{U_Z} \bullet \Delta Y_{U_Z} = 0.$$

Таким образом, решение задачи квадратичного программирования (61) является одновременно решением задачи дополнителности (59). Утверждение доказано.

Замечание 1. Пусть $\tilde{\Phi}_p^{-1} = [\phi^{kl}]$, $1 \leq k, l \leq m$. Тогда производная компоненты $\lambda_{\mathcal{A}, \Phi_p}^i(\Delta X_{U_Z})$ по $\Delta X_{U_Z}^{pq}$, $1 \leq p, 1 \leq p, q \leq n_Z$, равна

$$\frac{d\lambda_{\mathcal{A}, \Phi_p}^i(\Delta X_{U_Z})}{dX_{U_Z}^{pq}} = \sum_{j=1}^m \phi^{ij} A_j^{pq}.$$

Отсюда следует, что (k, l) -й элемент градиента линейной функции $\Lambda_{\mathcal{A}, \Phi_p}(\Delta X_{U_Z}) \bullet \Delta X_{U_Z}$ по $\Delta X_{U_Z}^{pq}$ равняется $\sum_{i,j=1}^m \phi^{ij} A_i^{kl} A_j^{pq}$.

Замечание 2. Если пара $[X, Y]$ невырожденная, то $n_p^\Delta + n_B^\Delta \geq m$. При выполнении предположения (60) должно выполняться неравенство $n_B^\Delta + n_Z^\Delta \leq m$. Отсюда следует оценка сверху на число n_Z , определяющее размерность задачи дополнителности (59), а именно, $n_Z \leq n_p$.

Решение задачи дополнителности (59) единственное. Подставляя найденные ΔX_{U_Z} и ΔY_{U_Z} в систему (54), получаем, что она становится зависимой только от остальных переменных и однозначным образом разрешима относительно них. Как уже отмечалось, формулы для определения этих направлений отличаются от полученных в разд. 2 только заменой вектора правых частей b на вектор b_1 .

ЗАКЛЮЧЕНИЕ

Предложенный алгоритм предназначен для решения линейных задач полуопределенного программирования в стандартной постановке. Вычисления ведутся в пространствах прямых переменных и слабых двойственных переменных, хотя несложно переписать формулы для вычислений в пространстве двойственных переменных. Все текущие приближения в итерационном процессе принадлежат допустимым множествам, поэтому метод можно трактовать как метод внутренней точки, но не строго внутренней точки, так как точки могут принадлежать границам допустимых множеств. Это принципиальное отличие от методов внутренней точки в традиционном понимании позволяет уменьшать вычислительные затраты при приближении к решениям прямой и двойственной задач. Делается это за счет более точного выполнения условия комплементарности между ними. Поэтому особенно эффективно алгоритм должен работать на конечной стадии вычислений, когда текущие точки находятся вблизи решений задач.

СПИСОК ЛИТЕРАТУРЫ

1. Жадан В.Г. Прямо-двойственный метод Ньютона с наискорейшим спуском для линейной задачи полуопределенного программирования. Ньютоновская система уравнений // Ж. вычисл. матем. и матем. физ. 2022. Т. 62. № 2. С. 22–38.
2. Magnus J.R., Neudecker H. The elimination matrix: some lemmas and applications // SIAM J. Alg. Disc. Meth. 1980. V. 1. № 4. P. 422–449.
3. Магнус Я.Р., Нейдеккер Ч. Матричное дифференциальное исчисление с приложениями к статистике и эконометрике. М.: Физматлит, 2002.
4. Nesterov Yu.E., Nemirovski A.S. Interior Point Polynomial Algorithms in Convex Programming // Philadelphia: SIAM Publications, SIAM, 1994.
5. Alizadeh F. Interior point methods in semidefinite programming with applications to combinatorial optimization // SIAM J. Optim. 1995. V. 5. P. 13–51.
6. Vandenberghe L., Boyd S. Semidefinite programming // SIAM Rev. 1996. V. 38. P. 49–95.
7. Todd M.J. Semidefinite optimization // Acta Numer. 2001. V. 10. P. 515–560.
8. Wolkowicz H., Saigal R., Vandenberghe L. (eds). Handbook of Semidefinite Programming. Boston: Kluwer Academic Publishers, 2000.

9. *Laurent M., Rendle F.* Semidefinite Programming and Integer Programming. Centrum voor Wiskunde en Informatica, 2002.
10. *de Klerk E.* Aspects of Semidefinite Programming. Interior Point Algorithms and Selected Applications. Kluwer Academic Publishers, 2004.
11. *Monteiro R.D.C.* First- and second-order methods for semidefinite programming // Mathematical programming, Series B. 2003. V. 97. № 1–2. P. 209–244.
12. *Жадан В.Г.* Прямо-двойственный метод Ньютона для задач линейного программирования // Ж. вычисл. матем. и матем. физ. 1999. Т. 39. № 1. С. 17–32.
13. *Жадан В.Г.* О сходимости прямо-двойственного метода Ньютона для задачи линейного программирования // Ж. вычисл. матем. и матем. физ. 1999. Т. 39. № 5. С. 431–445.
14. *Арнольд В.И.* О матрицах, зависящих от параметров // Успехи матем. наук. 1971. Т. 26. вып. 2 (158). С. 101–114.
15. *Alizadeh F., Haeblerly J.-P.F., Overton M.L.* Complementarity and nondegeneracy in semidefinite programming // Math. Program., Series B. 1997. V. 77. № 2. P. 129–162.
16. *Тыртышников Е.Е.* Матричный анализ и линейная алгебра. М.: Физматлит, 2007.

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.958

ОБ ОДНОЙ КРАЕВОЙ ЗАДАЧЕ ДЛЯ ЭЛЛИПТИЧЕСКОГО УРАВНЕНИЯ ЧЕТВЕРТОГО ПОРЯДКА НА ПЛОСКОСТИ

© 2022 г. А. П. Солдатов^{1,2,3}¹ 119333 Москва, Вавилова, 44, корп. 2, ФИЦ ИУ РАН, Россия² 119992 Москва, Воробьевы горы, Московский центр фундаментальной и прикладной математики, Россия³ 111250 Москва, Красноказарменная ул, 14, НИУ МЭИ, Россия

e-mail: soldatov48@gmail.com

Поступила в редакцию 12.07.2021 г.

Переработанный вариант 12.07.2021 г.

Принята к публикации 16.12.2021 г.

Для эллиптического уравнения четвертого порядка с постоянными вещественными коэффициентами в многосвязной области рассмотрена краевая задача, заключающаяся в задании на границе этой области самой функции и ее нормальной производной третьего порядка. В работе дан критерий фредгольмовости, удобный для использования, и приведена формула индекса этой задачи. Выделены классы уравнений, для которых критерий фредгольмовости особенно упрощается, и подсчитаны точные значения индекса. Библ. 8.

Ключевые слова: эллиптические уравнения, нормальные производные, условие фредгольмовости задачи, формула индекса.

DOI: 10.31857/S0044466922040111

В области D , ограниченной гладким контуром Γ , рассмотрим для эллиптического уравнения

$$\frac{\partial^4 u}{\partial y^4} - \sum_{r=0}^3 a_r \frac{\partial^4 u}{\partial x^{4-r} \partial y^r} + \sum_{i+j \leq 3} a_{ij} \frac{\partial^{i+j} u}{\partial x^i \partial y^j} = f \quad (1)$$

с коэффициентами $a_r \in \mathbb{R}$, $a_{ij} \in C^\mu(\bar{D})$, $0 < \mu < 1$, краевую задачу

$$u|_\Gamma = f_1, \quad \frac{\partial^3 u}{\partial n^3}|_\Gamma = f_2, \quad (2)$$

где $n = n_1 + in_2$ означает единичную внешнюю нормаль. В предположении $\Gamma \in C^{4,\mu}$ оператор задачи ограничен: $C^{4,\mu}(\bar{D}) \rightarrow C^\mu(\bar{D}) \times C^{4,\mu}(\Gamma) \times C^{1,\mu}(\Gamma)$. Соответственно, решение ищется в классе $C^{4,\mu}(\bar{D})$ с правыми частями $f \in C^\mu(\bar{D})$, $f_1 \in C^{4,\mu}(\Gamma)$ и $f_2 \in C^{1,\mu}(\Gamma)$. Фредгольмовость и индекс задачи понимаются по отношению к этому оператору.

Эта задача занимает особое место в классе задач этого типа

$$\frac{\partial^{k_1} u}{\partial n^{k_1}}|_\Gamma = f_1, \quad \frac{\partial^{k_2} u}{\partial n^{k_2}}|_\Gamma = f_2, \quad 1 \leq k_1 < k_2 \leq 3,$$

поскольку, как установлено в [1], все остальные задачи фредгольмовы индекса нуль.

В силу эллиптичности характеристический многочлен $z^4 - a_3 z^3 - a_2 z^2 - a_1 z - a_0$ уравнения (1) не имеет вещественных корней. Обозначим через v_1, v_2 его корни в верхней полуплоскости, при этом случаи (i) $v_1 \neq v_2$ различных корней и (ii) $v_1 = v_2 = v$ кратного корня выделим особо.

С точки зрения общей эллиптической теории (см. [2]) задача (1), (2) фредгольмова в пространстве $C^{4,\mu}(\bar{D})$ тогда и только тогда, когда ее краевые условия удовлетворяют так называемому условию дополненности (или условию Шапиро–Лопатинского). В этом случае говорят

также (см. [3]), что краевые условия (2) накрывают дифференциальный оператор, отвечающий главной части (1). Как показано в [1], в обозначениях

$$\omega(e, v) = \frac{e_2 - v e_1}{e_1 + v e_2}, \quad e = e_1 + i e_2,$$

это условие равносильно тому, что функция

$$H(e) = \begin{cases} [\omega(e, v_2)]^3 - [\omega(e, v_1)]^3, & \text{(i)} \\ [\omega(e, v)]^2, & \text{(ii)} \end{cases} \quad (3)$$

всюду отлична от нуля на единичной окружности $\mathbb{T}$.

Очевидно, в случае кратного корня это условие всегда выполнено и индекс задачи равен нулю. В случае (i) его можно выразить следующим образом.

Лемма 1. Условие

$$[\omega(e, v_2)]^3 \neq [\omega(e, v_1)]^3, \quad e \in \mathbb{T}, \quad (4)$$

равносильно тому, что

$$v_1 \neq e^{\pm 2\pi i} v_2, \quad a_k^\pm = \delta_k^{-1} [\sqrt{3}(1 - v_1 v_2) \pm \sqrt{\Delta}] \notin \mathbb{R}, \quad k = 1, 2, \quad (5)$$

где положено

$$\delta_1 = 2i(e^{\pi i/3} v_1 - e^{-\pi i/3} v_2), \quad \delta_2 = 2i(e^{\pi i/3} v_2 - e^{-\pi i/3} v_1), \\ \Delta = 3(v_1 v_2 - 1)^2 + 4(v_1^2 + v_2^2 + v_1 v_2).$$

При выполнении этого условия индекс Коши

$$\frac{1}{2\pi} [\arg H]_{\mathbb{T}} = -2(n+1), \quad (6)$$

где n равно числу точек $a_1^\pm, a_2^\pm$, лежащих в нижней полуплоскости (с учетом их возможных совпадений), и приращение на окружности берется против часовой стрелки.

Доказательство. Запишем

$$H(e) = \prod [\omega(e, v_2) - q \omega(e, v_1)],$$

где произведение берется по трем корням $q = 1, e^{\pm 2\pi i/3}$ уравнения $q^3 = 1$. В явном виде:

$$H(e) = \frac{(v_1 - v_2)(e_1^2 + e_2^2)h_1(e)h_2(e)}{(e_1 + v_1 e_2)^3 (e_1 + v_2 e_2)^3},$$

где $h_k(e) = (v_1 - q v_2)e_2^2 + (1 - q)(1 - v_1 v_2)e_1 e_2 - (v_2 - q v_1)e_1^2$ и $q = e^{2\pi i/3}$ для $k = 1$, $q = e^{-2\pi i/3}$ для $k = 2$.

В частности, первое условие в (5) непосредственно следует из (4) при $e_2 = 0$. Поскольку $q^2 + q + 1 = 0$, то можно записать

$$h_k(e) = (v_1 - q v_2)(e_2 - z_k^+ e_1)(e_2 - z_k^- e_1), \quad k = 1, 2,$$

с корнями

$$z_k^\pm = \frac{(q-1)(1 - v_1 v_2) \pm \sqrt{-q\Delta}}{2(v_1 - q v_2)}.$$

Так как $e^{\pm 2\pi i/3} - 1 = \pm i\sqrt{3}e^{\pm \pi i/3}$, откуда $z_1^\pm = a_2^\pm$, $z_2^\pm = a_1^\mp$. Итак, при выполнении первого условия в (5) имеем равенство

$$H(e) = c \frac{(e_1^+ e_2^2)(e_2 - a_1^+ e_1)(e_2 - a_2^+ e_1)(e_2 - a_1^- e_1)(e_2 - a_2^- e_1)}{(e_1 + v_1 e_2)^3 (e_1 + v_2 e_2)^3} \quad (7)$$

с некоторым $c \neq 0$, которое приводит к эквивалентности условий (4) и (5).

Обратимся ко второму утверждению леммы. Функция H четна, и потому

$$\frac{1}{2\pi}[\arg H]_{\mathbb{T}} = \frac{1}{\pi}[\arg H]_{\mathbb{T}^+}, \quad (8)$$

где $\mathbb{T}^+$ есть правая полуокружность. С другой стороны, отображение $e = e_1 + ie_2 \rightarrow t = e_2/e_1$ осуществляет гомеоморфизм этой полуокружности на расширенную вещественную прямую $\overline{\mathbb{R}}$, причем обход ее от точки $e = -i$ к $e = i$ против часовой стрелки соответствует движению на прямой в положительном направлении. Согласно (7) имеем соотношение $H(e) = cR(e_2/e_1)$ с некоторой ненулевой постоянной c и рациональной функцией

$$R(\zeta) = \frac{(\zeta^2 + 1)(\zeta - a_1^+)(\zeta - a_2^+)(\zeta - a_1^-)(\zeta - a_2^-)}{(1 + v_1\zeta)^3(1 + v_2\zeta)^3},$$

полюса $-1/v_k$ которой лежат в верхней полуплоскости. Поэтому на основании принципа аргумента, примененного к аналитической функции R в нижней полуплоскости, приходим к равенству

$$\frac{1}{2\pi}[\arg H]_{\mathbb{T}^+} = \frac{1}{2\pi}[\arg R]_{\mathbb{R}} = -(n+1),$$

где учтено, что положительное направление вещественной прямой оставляет эту полуплоскость слева. Совместно с (8) отсюда следует формула (6), что завершает доказательство леммы.

Теорема 1. Пусть контур $\Gamma \in C^{4,\mu}$ состоит из связных компонент $\Gamma_0, \Gamma_1, \dots, \Gamma_s$, причем Γ_0 охватывает все остальные контуры, и условие (4) фредгольмовости задачи (1), (2) выполнено.

Тогда ее индекс $\varkappa$ дается формулой

$$\varkappa = 4(1-s)n, \quad (9)$$

где $n = 0$ в случае (ii), и n определяется леммой 1 в случае (i).

Доказательство опирается на общие результаты (см. [1], [4], [5]) о фредгольмовости и индексе задач для эллиптических уравнений $2l$ -го порядка с l краевыми условиями типа (2). Для односвязных областей D эти вопросы подробно исследованы в [1], [4], в частности, равенство (9) согласуется с формулой индекса, данной в этих статьях. Случаю многосвязной области посвящена работа [5], однако доказательство ее основной теоремы содержит пробел, который привел к ошибочной формуле индекса. Поэтому, следуя рассуждениям указанных работ, приведем полное доказательство применительно к рассматриваемой задаче (1), (2), восполняющее этот пробел.

В соответствии с леммой 1 достаточно убедиться, что

$$\varkappa = \begin{cases} 2(s-1)(\varkappa_0 + 2), & \text{(i)} \\ 0, & \text{(ii)} \end{cases} \quad (10)$$

где $\varkappa_0$ означает левую часть (6).

Для функции $\varphi \in C^1(\Gamma)$ положим

$$d\varphi = \varphi' + m^0\varphi, \quad (m^0\varphi)|_{\Gamma_i} = \frac{1}{|\Gamma_i|} \int_{\Gamma_i} \varphi(t) d_1t,$$

где штрих указывает на дифференцирование по параметру длины дуги, $|\Gamma_i|$ есть длина контура Γ_i , и d_1t означает элемент длины дуги. Видно, что оператор d обратим: $C^k(\Gamma) \rightarrow C^{k-1}(\Gamma)$, причем $d^k\varphi = \varphi^{(k)} + m^0\varphi$. Следовательно, краевое условие (2) можно заменить равносильным ему

$$d^3u^+ = d^3f_1, \quad \left. \frac{\partial^3 u}{\partial n^3} \right|_{\Gamma} = f_2, \quad (2)'$$

где знак $+$ указывает на граничное значение функции u . Очевидно,

$$d^3 u^+ = \left[\left(e_1 \frac{\partial}{\partial x} + e_2 \frac{\partial}{\partial y} \right)^3 u + \sum_{i+j=1,2} b_{ij} \frac{\partial^{i+j} u}{\partial x^i \partial y^j} \right]_{\Gamma} + m^0(u^+)$$

с некоторыми $b_{ij} \in C^{1,\mu}(\Gamma)$, где $e(t) = in(t)$ означает единичный касательный вектор к контуру Γ в точке t . В явном виде $2b_{20} = 3(e_1^2)'$, $2b_{02} = 3(e_2^2)'$, $b_{11} = 3(e_1 e_2)'$ и $b_{10} = e_1^{(2)}$, $b_{01} = e_2^{(2)}$.

Очевидно, операторы

$$u \rightarrow \frac{\partial^{i+j} u}{\partial x^i \partial y^j}, \quad i+j \leq 3,$$

компактны $C^{4,\mu}(\bar{D}) \rightarrow C^{\mu}(\bar{D})$, а операторы

$$u \rightarrow \left(\frac{\partial^{i+j} u}{\partial x^i \partial y^j} \right)^+, \quad i+j = 1, 2,$$

компактны $C^{4,\mu}(\bar{D}) \rightarrow C^{1,\mu}(\Gamma)$.

Поэтому вместе с (1), (2)' задача

$$\frac{\partial^4 u}{\partial y^4} - \sum_{r=0}^3 a_r \frac{\partial^4 u}{\partial x^{4-r} \partial y^r} = f, \quad (1)''$$

$$\left[\left(e_1 \frac{\partial}{\partial x} + e_2 \frac{\partial}{\partial y} \right)^3 u \right]_{\Gamma} = f_1^0, \quad \left[\left(n_1 \frac{\partial}{\partial x} + n_2 \frac{\partial}{\partial y} \right)^3 u \right]_{\Gamma} = f_2^0 \quad (2)''$$

также фредгольмова и ее индекс равен $\mathfrak{x}$.

Зафиксируем точку $z_0 \in D$ и обозначим $\tilde{C}^{4,\mu}(\bar{D})$ подпространство всех функций $u \in C^{4,\mu}(\bar{D})$, обращающихся в точке в нуль вместе с частными производными до порядка 3 включительно. Очевидно, оно замкнуто и его коразмерность, т.е. размерность фактор-пространства $C^{4,\mu}/\tilde{C}^{4,\mu}$, равна 6. Поэтому в силу известных свойств фредгольмовых операторов (см. [6]) задача (1)'', (2)'' фредгольмова в классе $\tilde{C}^{4,\mu}(\bar{D})$ и ее индекс

$$\tilde{\mathfrak{x}} = \mathfrak{x} - 6. \quad (11)$$

Каждой функции $u \in \tilde{C}^{4,\mu}(\bar{D})$ поставим в соответствие вектор $\mathcal{D}u = U$, составленный из частных производных третьего порядка:

$$U_j = \frac{\partial^3 u}{\partial^{4-j} x \partial^{j-1} y}, \quad 1 \leq j \leq 4.$$

Очевидно, оператор $\mathcal{D}$ взаимно однозначен на $\tilde{C}^{4,\mu}(\bar{D})$ и его образ $\tilde{X}$ содержится в замкнутом подпространстве $X \subseteq [C^{1,\mu}(\bar{D})]^4$, выделяемом условием

$$\frac{\partial U_j}{\partial y} = \frac{\partial U_{j+1}}{\partial x}, \quad 1 \leq j \leq 4. \quad (12)$$

Утверждается, что подпространство $\tilde{X}$ замкнуто в X и имеет конечную коразмерность, равную

$$\dim(X/\tilde{X}) = 6s. \quad (13)$$

Действительно, в любой односвязной подобласти $D_0 \subseteq D$, содержащей точку z_0 , уравнение $\mathcal{D}u = U$ с правой частью $U \in X$ всегда разрешимо, и его решение определяется последователь-

ным интегрированием вдоль дуги, соединяющей произвольную точку $z \in D_0$ с z_0 . Например, для $l = 1$

$$u(z) = \int_{z_0}^z U_1 dx + U_2 dy.$$

Аналогично, для $l = 2$ следует положить

$$u_{1,j}(z) = \int_{z_0}^z U_j dx + U_{j+1} dy, \quad 1 \leq j \leq 3,$$

$$u_{2,j}(z) = \int_{z_0}^z u_{1,j} dx + u_{2,j+1} dy, \quad j = 1, 2; \quad (\mathcal{D}^{(-1)}U)(z) = \int_{z_0}^z u_{2,1} dx + u_{2,2} dy$$

и т.д.

По отношению к исходной области D решением уравнения $\mathcal{D}u = U$ с правой частью $U \in X$ служит многозначная функция u , принадлежащая классу $C^{4,\mu}$ в односвязных подобластях D_0 . При обходе контура Γ_j , $1 \leq j \leq s$, она получает приращение в виде некоторого многочлена $p_j \in P_2$ степени не выше двух. Этот факт можно выразить следующим образом. Соединим внутри $D \setminus \{z_0\}$ контура Γ_0 и Γ_j , $1 \leq j \leq s$, дугой L_j , считая эти дуги попарно непересекающимися. Тогда в области $D_0 = D \setminus (L_1 \cup \dots \cup L_n)$ решение u рассматриваемого уравнения однозначно. Предполагая дуги L_j ориентируемыми, для односторонних предельных значений $u_j^\pm$ на L_j функции u будем иметь соотношения

$$u_j^+ - u_j^- = p_j|_{L_j}, \quad 1 \leq j \leq n,$$

с некоторыми p_j , причем аналогичные соотношения выполняются и для частных производных функций u и p_j до порядка $k - 1$ включительно. Очевидно, отображение $U \rightarrow (p_1, \dots, p_n)$ переводит пространство $\tilde{X}$ на все P_2^s , и класс $\mathcal{X}_0$ описывается условиями $p_1 = \dots = p_n = 0$ в этих соотношениях. Поскольку $\dim P_2 = 6$, отсюда следует равенство (13).

По отношению к вектору $U = \mathcal{D}u$ задача (1)", (2)" может быть переписана в эквивалентной форме

$$\frac{\partial U}{\partial y} - A \frac{\partial U}{\partial x} U = f^1, \quad CU^+ = f^0, \quad (14)$$

с матрицами

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a_0 & a_1 & a_2 & a_3 \end{pmatrix}, \quad C = \begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \end{pmatrix},$$

где элементы $C_{ij}(t)$, $t \in \Gamma$, определяются из соотношений

$$\sum_{k=1}^4 C_{1k}(t) z^{k-1} = [e_1(t) + e_2(t)z]^3, \quad \sum_{k=1}^4 C_{2k}(t) z^{k-1} = [e_2(t) - e_1(t)z]^3,$$

и положено $f^1 = (0, 0, 0, f)$, $f^0 = (f_1^0, f_1^0)$. В этой связи введем подпространство $Y \subseteq [C^\mu(\bar{D})]^4$ векторов f^1 , первые три компоненты которых равны нулю. Заметим, что прообраз

$$L^{-1}(Y) = X, \quad (15)$$

где для краткости

$$L = \frac{\partial}{\partial y} - A \frac{\partial U}{\partial x}.$$

Задачу Римана–Гильберта (14) для эллиптической системы первого порядка, рассматриваемую во всем классе $[C^{1,\mu}(\bar{D})]^4$, обозначим через R , символ R сохраняем и для ее оператора, действующего $[C^{1,\mu}(\bar{D})]^4 \rightarrow [C^\mu(\bar{D})]^4 \times [C^{1,\mu}(\Gamma)]^2$. Символы $R_{\tilde{X}}$ и R_X указывают на эту задачу в соответствующих классах, причем ее операторы действуют из этих классов в $Y \times [C^{1,\mu}(\Gamma)]^2$.

Поскольку задача $R_{\tilde{X}}$ равносильна задаче (9), (10), то она фредгольмова и ее индекс равен $\tilde{\alpha}$. Поэтому на основании (13) задача R_X также фредгольмова и ее индекс

$$\text{ind } R_X = \text{ind } R_{\tilde{X}} + 6s.$$

Совместно с (11) отсюда

$$\alpha = \text{ind } R_X + 6(1-s). \quad (16)$$

К задаче R , рассматриваемой во всем классе $[C^{1,\mu}(\bar{D})]^4$, можно применить результаты [7]. С этой целью введем 4×2 -матрицу B с элементами

$$\begin{aligned} B_{k,1} &= v_1^{k-1}, & B_{k,2} &= v_2^{k-1}, & \text{(i)} \\ B_{k,1} &= v^{k-1}, & B_{k,2} &= (k-1)v^{k-2}, & \text{(ii)} \end{aligned} \quad 1 \leq k \leq 4.$$

Составленная с помощью нее квадратная матрица $\tilde{B} = (B\bar{B})$ обратима и приводит матрицу A к жордановой форме

$$\tilde{B}^{-1}A\tilde{B} = \begin{pmatrix} J & 0 \\ 0 & \bar{J} \end{pmatrix},$$

где

$$\text{(i)} \quad J = \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix}, \quad \text{(ii)} \quad J = \begin{pmatrix} v & 1 \\ 0 & v \end{pmatrix}.$$

В этих обозначениях на основании теоремы 2 из [7] задача R фредгольмова в классе $C^{1,\mu}(\bar{D})$ тогда и только тогда, когда

$$\det C(t)B \neq 0, \quad t \in \Gamma, \quad (18)$$

и ее индекс дается формулой

$$\text{ind } R = -\sum_{j=0}^s \frac{1}{\pi} [\arg \det(CB)]_{\Gamma_j} + 2(1-s), \quad (19)$$

где приращение вдоль Γ_j берется в направлении, оставляющем область D слева.

Из определений матриц C и B видно, что

$$CB = \begin{pmatrix} (e_1 + e_2 v_1)^3 & (e_1 + e_2 v_2)^3 \\ (e_2 - e_1 v_1)^3 & (e_2 - e_1 v_2)^3 \end{pmatrix}, \quad \text{(i)}$$

$$CB = \begin{pmatrix} (e_1 + e_2 v)^3 & 3e_2(e_1 + e_2 v)^2 \\ (e_2 - e_1 v)^3 & -3e_1(e_2 - e_1 v)^2 \end{pmatrix}, \quad \text{(ii)}$$

так что в обозначениях (3) имеем соотношение

$$\det[C(t)B] = \begin{cases} (e_1 + e_2 v_1)^3 (e_1 + e_2 v_2)^3 H[e(t)], & \text{(i)} \\ -3(e_1 + e_2 v)^4 H[e(t)], & \text{(ii)} \end{cases} \quad t \in \Gamma.$$

В частности, на основании (5) условие (18) выполнено, так что задача R фредгольмова.

Когда точка t обходит простой контур Γ_j в положительном направлении, единичный вектор $e(t)$ обходит окружность $\mathbb{T}$ против часовой стрелки при $j = 0$ и по часовой стрелке при $1 \leq j \leq s$. Поэтому при $\operatorname{Im} z > 0$ имеем равенство

$$\frac{1}{2\pi} [\arg(e_1 + ze_2)]_{\Gamma} = 1 - s$$

и в обозначениях (8) аналогичным образом

$$\frac{1}{2\pi} \arg H[e(t)]_{\Gamma} = 1 - s,$$

так что формула (19) принимает следующий вид:

$$\operatorname{ind} R = \begin{cases} -2(1-s)(\alpha_0 + 5), & \text{(i)} \\ -2(1-s)(\alpha_0 + 3). & \text{(ii)} \end{cases} \quad (20)$$

Утверждается, что индексы задач R и R_X совпадают:

$$\operatorname{ind} R = \operatorname{ind} R_X. \quad (21)$$

В самом деле, в силу (15) пространство решений однородных задач R_X и R совпадают. С другой стороны, условия разрешимости неоднородной задачи R достаточны и для разрешимости неоднородной задачи R_X . Поэтому остается убедиться, что в действительности они и необходимы. Запишем необходимые и достаточные условия разрешимости неоднородной задачи R в виде $p_i^1 f^1 + p_i^0 f^0 = 0$, $1 \leq i \leq s$, где линейные функционалы p_i^0 и p_i^1 непрерывны на $[C^{1,\mu}(\Gamma)]^2$ и $[C^{\mu}(\bar{D})]^4$ соответственно, и пары $p_i = (p_i^1, p_i^0)$ линейно независимы. Необходимо показать, что они линейно независимы и как функционалы на $Y \times [C^{1,\mu}(\Gamma)]^2$.

Предположим противное, т.е. найдется некоторая ненулевая линейная комбинация этих функционалов с коэффициентами α_i , равная нулю на $Y \times [C^{1,\mu}(\Gamma)]^2$. В частности,

$$\sum_{i=1}^s \alpha_i p_i^0 = 0. \quad (22)$$

Пусть вектор-функция $f^1 \in [C^{\mu}(\bar{D})]^4$, продолжим ее до вектор-функции φ с компактным носителем из того же класса C^{μ} и положим

$$U(z) = \frac{1}{2\pi i} \int_{\mathbb{C}} (t-z)_A^{-1} \varphi(t) d_2 t, \quad z \in \mathbb{C},$$

где $d_2 t$ означает элемент площади. Легко проверяется, что при дополнительном условии $\varphi \in C^1$, функция $U = T\varphi$ также непрерывно дифференцируема и удовлетворяет неоднородной системе $Lu = \varphi$. В действительности, этот факт справедлив и для функций φ , локально удовлетворяющих условию Гельдера (см., например, [8, лемма 3.4.2]). Таким образом, в области D вектор-функция U является решением задачи R с правой частью f^1 и $f^0 = CU^+$. Поэтому $p_i^1(f^1) + p_i^0(f^0) = 0$, $1 \leq i \leq s$, так что с учетом (22)

$$\left(\sum_{i=1}^s \alpha_i p_i^1 \right) (f^1) = 0.$$

Поскольку вектор-функция f^1 была выбрана произвольно, совместно с (22) отсюда заключаем, что система функционалов p_i линейно зависима, что невозможно. Тем самым равенство (21) установлено.

Объединяя равенства (20), (21) с (16), приходим к справедливости формулы (10), завершающей доказательство теоремы.

Для некоторых классов уравнений (1) в условиях теоремы 1 критерий фредгольмовости задачи (2) упрощается, и в формуле (9) ее индекса число n подсчитывается явно.

Теорема 2. (а) При $|v_1| = |v_2| = 1$ условие (4) равносильно

$$|v_1 - v_2| \neq \sqrt{3}, \quad (23a)$$

и в формуле (9) число $n = 0$ при $|v_1 - v_2| < \sqrt{3}$ и $n = 2$ при $|v_1 - v_2| > \sqrt{3}$.

(б) При $v_1 = i\rho_1$, $v_2 = i\rho_2$ условие (4) равносильно

$$|\rho_1 - \rho_2| \neq \sqrt{3}(1 + \rho_1\rho_2), \quad (23б)$$

и в формуле (9) число $n = 0$ при $|\rho_1 - \rho_2| < \sqrt{3}(1 + \rho_1\rho_2)$ и $n = 2$ при $|\rho_1 - \rho_2| > \sqrt{3}(1 + \rho_1\rho_2)$.

(в) При $v_1 = v$, $v_2 = -1/v$ условие (4) равносильно

$$4 \operatorname{Im} v \neq |v|^2 + 1 \quad (23в)$$

и в формуле (9) число $n = 0$ при $4 \operatorname{Im} v < |v|^2 + 1$ и $n = 2$ при $4 \operatorname{Im} v > |v|^2 + 1$.

Доказательство достаточно провести для случая (ii) различных корней v_k , в котором можно воспользоваться леммой 1.

(а) Положим $v_k = e^{i\theta_k}$, и пусть для краткости

$$\theta_{(1)} = \frac{\pi}{3} + \frac{\theta_1 - \theta_2}{2}, \quad \theta_{(2)} = \frac{\pi}{3} + \frac{\theta_2 - \theta_1}{2}, \quad \theta_{\pm} = \theta_1 \pm \theta_2.$$

Тогда в обозначениях (6)

$$\delta_k = -4 \sin \theta_{(k)} e^{i\theta_+/2}, \quad 1 - v_1 v_2 = -2i \sin(\theta_+/2) e^{i\theta_+/2},$$

и $\Delta = e^{i\theta_+} [-12 \sin^2(\theta_+/2) + 8 \cos \theta_- + 4]$. Заметим, что условие (23а), которое можно записать в форме первого условия в (5), равносильно $0 < |\theta_{(k)}| < \pi$. Поэтому при выполнении этого условия можем записать

$$a_k^{\pm} = [4 \sin \theta_{(k)}]^{-1} (ib \mp \sqrt{c - b^2}),$$

где положено $b = 2\sqrt{3} \sin(\theta_+/2)$, $c = 4(2 \cos \theta_- + 1)$. Таким образом,

$$[4 \sin \theta_{(k)}] (\operatorname{Im} a_k^{\pm}) = \begin{cases} b, & c \geq b^2, \\ b \mp \sqrt{b^2 - c}, & b^2 > c. \end{cases} \quad (24)$$

Заметим, что число c имеет одинаковый знак с разностью $\pi/3 - |\theta_1 - \theta_2|$. В частности, второе условие в (5) равносильно $c \neq 0$, что, в свою очередь, эквивалентно (23а).

Что касается подсчета n в лемме 1, то при $|\theta_2 - \theta_1| < \pi/2$ числа $\sin \theta_{(k)}$, $k = 1, 2$, и c положительны. Поэтому согласно (24) все точки $a_k^{\pm}$ лежат в верхней полуплоскости и, значит, $n = 0$. При $|\theta_2 - \theta_1| > \pi/2$ числа $\sin \theta_{(k)}$, $k = 1, 2$, имеют противоположные знаки. Поскольку знак левой части (24) не зависит от $k = 1, 2$, четыре точки $a_k^{\pm}$ распределены поровну в верхней и нижней полуплоскостях, т.е. $n = 2$.

(б) В этом случае первое условие в (5) очевидным образом выполнено, и

$$\delta_k a_k^{\pm} = \sqrt{3}(1 + \rho_1\rho_2) \pm \sqrt{\Delta}, \quad k = 1, 2,$$

по отношению к вещественным числам $\delta_k = (-1)^k(\rho_1 - \rho_2) - i\sqrt{3}(\rho_1 + \rho_2)$ и $\Delta = 3(1 + \rho_1\rho_2)^2 - 4(\rho_1^2 + \rho_2^2 + \rho_1\rho_2)$. В частности,

$$\operatorname{Im} a_k^\pm > 0 \quad \text{при} \quad \Delta \geq 0, \quad k = 1, 2. \quad (25)$$

Пусть далее $\Delta < 0$. Тогда

$$|\rho_k^2| \operatorname{Im} a_k^\pm = \pm(-1)^k(\rho_1 - \rho_2)\sqrt{-\Delta} + 3(\rho_1 + \rho_2)(1 + \rho_1\rho_2), \quad k = 1, 2.$$

Поэтому (5) равносильно

$$\beta = 9(\rho_1 + \rho_2)^2(1 + \rho_1\rho_2)^2 + (\rho_1 - \rho_2)^2\Delta \neq 0, \quad (26)$$

причем

$$\operatorname{Im} a_k^\pm > 0 \quad \text{при} \quad \beta > 0, \quad k = 1, 2; \quad \pm(-1)^k(\rho_1 - \rho_2)(\operatorname{Im} a_k^\pm) > 0 \quad \text{при} \quad \beta < 0. \quad (27)$$

Поскольку

$$\beta = 4(\rho_1^2 + \rho_2^2 + \rho_1\rho_2) \left[3(1 + \rho_1\rho_2)^2 - (\rho_1 - \rho_2)^2 \right],$$

условие (26) равносильно (23б), и соотношения (25), (27) приводят к соответствующим значениям числа n леммы 1.

(в) В этом случае первое условие в (5) также выполнено. Что касается второго условия, то по отношению к v оно записывается довольно громоздко. Однако этот случай подробно разобран в [1, теорема 3]. Именно, условие (4) равносильно тому, что точка $z_0 = e^{\pi i/6}$ не принадлежит окружности $\Gamma_0(v)$, определяемой уравнением $(|z|^2 + 1)(\operatorname{Im} v) = (|v|^2 + 1)(\operatorname{Im} z)$, т.е. должно выполняться условие $|v|^2 + 1 \neq 4 \operatorname{Im} v$. При этом имеет место равенство (6) с

$$n = \begin{cases} 0, & |v|^2 + 1 < 4 \operatorname{Im} v, \\ 2, & |v|^2 + 1 > 4 \operatorname{Im} v, \end{cases}$$

что на основании (10) завершает доказательство (в) и теоремы.

Заметим, что для $v = ip$, $p > 1$, условие (23в) теоремы согласуется с (23б). В самом деле, в этом случае уравнение $p^2 + 1 = 4p$ дает корень $p = 2 + \sqrt{3}$. С другой стороны, этот же корень дает и уравнение $p - 1/p = 2\sqrt{3}$.

СПИСОК ЛИТЕРАТУРЫ

1. Солдатов А.П. О фредгольмовости и индексе обобщенной задачи Неймана // Дифференц. ур-ния. 2020. Т. 56. № 2. С. 217–225.
2. Назаров С.А., Пламеневский Б.А. Эллиптические задачи в областях с кусочно гладкой границей. М.: Наука, 1991. 336 с.
3. Лионс Ж.Л., Мадженес Э. Неоднородные граничные задачи и их приложения. М.: Мир, 1971.
4. Кошанов Б.Д., Солдатов А.П. Краевая задача с нормальными производными для эллиптического уравнения на плоскости // Дифференц. ур-ния. 2016. Т. 52. № 12. С. 1666–1681.
5. Солдатов А.П. Об одной краевой задаче для эллиптического уравнения на плоскости в многосвязной области // Владикавказ. матем. журн. 2017. Т. 19. Вып. 3. С. 51–58.
6. Пале Р. Семинар по теореме Атьи-Зингера об индексе. М.: Мир, 1970.
7. Солдатов А.П., Чернова О.В. Задача Римана–Гильберта для эллиптических систем первого порядка на плоскости с постоянными старшими коэффициентами // Итоги науки и техн. Сер. Современ. матем. и ее прилож. Темат. обз. 149. С. 95–102.
8. Солдатов А.П. Сингулярные интегральные операторы и эллиптические краевые задачи. Современная математика // Фундамент. направления. 2016. Т. 63. С. 1–179.

**МАТЕМАТИЧЕСКАЯ
ФИЗИКА**

УДК 519.63

*Светлой памяти Б.В. Рогова посвящается***ВЛИЯНИЕ МОНОТОНИЗАЦИИ НА СПЕКТРАЛЬНОЕ РАЗРЕШЕНИЕ
БИКОМПАКТНЫХ СХЕМ В ЗАДАЧЕ
О НЕВЯЗКОМ ВИХРЕ ТЕЙЛОРА–ГРИНА¹⁾**

© 2022 г. М. Д. Брагин

125047 Москва, Миусская пл., 4, ИПМ РАН, Россия

e-mail: michael@bragin.cc

Поступила в редакцию 28.06.2021 г.
Переработанный вариант 29.08.2021 г.
Принята к публикации 16.12.2021 г.

Рассматривается одна локально-одномерная бикомпактная схема для трехмерных уравнений Эйлера, имеющая четвертый порядок аппроксимации по пространству и второй по времени. На ее примере в задаче о вихре Тейлора–Грина в невязком совершенном газе исследуется, в какой мере метод консервативной монотонизации, применяемый в бикомпактных схемах, практически влияет на их теоретически высокое спектральное разрешение. Предлагаются два алгоритма параллельного счета по локально-одномерным бикомпактным схемам, один из которых используется для проведения вычислений. Показывается, что выбранная бикомпактная схема при наличии монотонизации разрешает 70–85% спектра кинетической энергии жидкости. Эта схема сравнивается с высокоточными схемами серии WENO5 по поведению кинетической энергии и энтропии. Демонстрируется, что бикомпактная схема имеет заметно меньшую диссипацию и слабее подавляет вихри среднего масштаба. Библ. 32. Фиг. 9.

Ключевые слова: уравнения Эйлера, вихрь Тейлора–Грина, высокоточные схемы, неявные схемы, компактные схемы, бикомпактные схемы, параллельные алгоритмы.

DOI: 10.31857/S0044466922040032**ВВЕДЕНИЕ**

Решение задач аэрогидродинамики аппаратов, двигателей, конструкций, задач о движениях атмосферы, природных вод (океанов, морей, рек и т.д.) нередко приводит к необходимости моделирования турбулентных течений вязкой жидкости. Важнейшей особенностью таких течений является широкий диапазон характеризующих их пространственно-временных масштабов (см. [1]). Прямое численное моделирование (DNS) состоит в непосредственном численном решении уравнений Навье–Стокса с полным разрешением на сетке всех масштабов и структур течения. Этот подход лучше всего соответствует исходной физико-математической модели ньютоновской жидкости. Однако с ростом числа Рейнольдса он быстро становится нереализуемым из-за ограниченности имеющихся на данный момент вычислительных ресурсов. Численное решение уравнений Навье–Стокса, осредненных по Рейнольдсу (RANS) (см. [1]), является одним из способов справиться с данной проблемой. Уравнения и замыкающие модели RANS совместно с классом конечно-объемных (либо конечно-разностных) схем второго порядка точности образуют современный промышленный стандарт решения аэрогидродинамических задач. Позволяя достичь успехов в предсказании средних характеристик установившихся безотрывных течений, он оказывается плохо пригодным для изучения, например, отрывных течений, акустических шумов, генерируемых свободными струями либо обтекаемыми телами, смешивающихся потоков жидкости и т.д. (см. [2], [3]). Актуальной альтернативой стандарту RANS является моделирование крупных вихрей (LES) (см. [4]). Суть LES в общих чертах сводится к тому, что динамика структур течения размером крупнее ячеек сетки рассчитывается, исходя из про-

¹⁾ Работа выполнена при финансовой поддержке РФФИ (код проекта № 21-11-00198).

странственно фильтрованных уравнений Навье–Стокса, а динамика оставшихся мелкомасштабных структур, так называемая подсеточная турбулентность, параметризуется с помощью тех или иных априорных или полуэмпирических моделей. Можно выделить три вида LES (см. [5], [6]): собственно, вышеописанное LES с явным выделением и разделением масштабов через фильтр и подсеточную модель; неявное LES (ILES), при котором уравнения Навье–Стокса решаются в исходном виде, а функции фильтра и подсеточной модели выполняют члены из погрешности аппроксимации используемой численной схемы (см. [7]); “LES без модели”, справедливо называемое также “DNS с недостаточным разрешением” (uDNS), которое представляет собой ILES вообще без анализа погрешности аппроксимации на предмет ее связи с какой-либо физической моделью подсеточных процессов.

Чтобы быть применимой в LES или DNS, численная схема должна обладать малыми диссипативными и дисперсионными погрешностями или, другими словами, высоким спектральным разрешением (см. [3], [8], [9]). Это требование вытекает из того, что схема должна как можно точнее воспроизводить динамику вихрей всех разрешаемых масштабов, притом нередко на больших интервалах интегрирования (и по пространству, и по времени) (см. [9]). Наиболее очевидным способом понижения численной диссипации/дисперсии является повышение порядка аппроксимации схемы. Более того, в [10], [11] на примере разрывных схем Галеркина (DG) показывается, что при равном числе степеней свободы схемы более высокого порядка превосходят по точности свои аналоги более низкого порядка. Однако сам по себе высокий порядок аппроксимации еще ничего не говорит о пригодности схемы для LES или DNS. Например, классическая схема WENO5 (см. [12]) (с аппроксимацией пятого порядка по пространству и третьего порядка по времени) из-за довольно высокого уровня численной диссипации мало пригодна для этих приложений (см. [13]), что служит одной из причин для разработки модификаций этой схемы (см. [13], [14]). Напротив, схема “КАБАРЕ” (см. [15]), имеющая формально второй порядок аппроксимации по всем переменным, обладает хорошим спектральным разрешением и потому применима для моделирования турбулентных течений. Если рассматривать схемы высокого порядка (выше второго), то среди них для LES и DNS лучше всего подходят DG-схемы (см. [5], [6], [9]–[11], [16]) и компактные схемы (см. [8], [17], [18]), в том числе мультиоператорные (см. [19]).

В [20]–[28] развивается класс высокоточных бикompактных схем. В их основе лежит симметричная компактная аппроксимация пространственных производных на шаблоне, который целиком помещается в одну ячейку сетки (при любом порядке схемы). Минимизация шаблона осуществляется через введение дополнительных искоемых функций, для отыскания которых привлекаются дифференциальные следствия (производные) решаемых уравнений. Преимущество бикompактных схем заключается в сочетании нескольких положительных свойств: высокой устойчивости, экономичной реализации, равном числе граничных условий в разностной и дифференциальной постановках задачи, высоком спектральном разрешении, лучшем по сравнению с классическими компактными схемами равного порядка аппроксимации (см. [22], [24], [25]). Последнее свойство, по-видимому, делает бикompактные схемы перспективными для применения в LES. Здесь как раз возникает следующая проблема: высокое спектральное разрешение бикompактных схем доказано лишь теоретически в случае модельного линейного уравнения переноса. На практике же, при решении настоящих гидродинамических уравнений, бикompактные схемы (как и схемы других классов) подвергаются монотонизации, которая никак не учитывалась в анализе (см. [22], [24], [25]).

Итак, основная цель настоящей работы – выяснить, как метод консервативной монотонизации из [26], применяемый в бикompактных схемах, влияет на их спектральное разрешение. Это исследование мы проведем для одной локально-одномерной бикompактной схемы четвертого порядка аппроксимации по пространству и второго порядка аппроксимации по времени. В качестве конкретного гидродинамического примера возьмем задачу о вихре Тейлора–Грина (см. [29]), которая является стандартным тестом для схем, разрабатываемых для моделирования турбулентных течений. Данную задачу мы решим в предположении, что жидкость идеальная (т.е. невязкая): отсутствие физической диссипации позволит яснее проанализировать действие численной диссипации от метода из [26]. Кроме того, мы обсудим параллельную реализацию локально-одномерных бикompактных схем и сделаем сравнение по численной диссипации с некоторыми высокоточными схемами серии WENO5.

Работа построена следующим образом. В разд. 1 формулируется постановка задачи о вихре Тейлора–Грина. В разд. 2 кратко описывается локально-одномерная бикompактная схема с консервативной монотонизацией. В разд. 3 предлагаются два алгоритма параллельного счета по локально-одномерным бикompактным схемам. В разд. 4 обсуждаются результаты расчетов.

1. ПОСТАНОВКА ЗАДАЧИ

Сделаем еще одно отступление от оригинальной постановки задачи о вихре Тейлора–Грина из [29]: предположим, что жидкость не только идеальная, но и сжимаемая, причем ее движение происходит при достаточно большом фоновом давлении. Данный прием избавит нас от необходимости решать эллиптическую задачу для давления и сделает систему динамических уравнений гиперболической, но при этом оставит плотность примерно постоянной.

Трехмерные течения идеальной сжимаемой жидкости описываются уравнениями Эйлера

$$\partial_t \mathbf{Q} + \sum_{d=x,y,z} \partial_d \mathbf{F}^d(\mathbf{Q}) = \mathbf{0}, \quad (1)$$

где

$$\mathbf{Q} = \begin{bmatrix} \rho \\ \rho \mathbf{v} \\ E \end{bmatrix}, \quad \mathbf{F}^d(\mathbf{Q}) = \begin{bmatrix} \rho v_d \\ \rho v_d \mathbf{v} + p \mathbf{e}_d \\ v_d(E + p) \end{bmatrix}, \quad E = \rho \epsilon + \frac{\rho v^2}{2}.$$

Поясним обозначения: x, y, z – декартовы пространственные координаты; t – время; ρ , $\mathbf{v} = [v_x v_y v_z]^T$, p , ϵ и E – соответственно плотность, вектор скорости, давление, удельная внутренняя энергия и удельная полная энергия газа; $v = |\mathbf{v}|$ – абсолютная величина скорости; $\mathbf{Q} = [Q_1 \dots Q_5]^T = \mathbf{Q}(x, y, z, t)$ – искомый вектор консервативных переменных; $\mathbf{F}^d(\mathbf{Q})$ – вектор потоков в направлении оси Od ; $\mathbf{e}_d$ – единичный орт оси Od ; ∂_x – символ, обозначающий частную производную $\partial/\partial x$. Поскольку сжимаемость жидкости вводится искусственно, имеется свобода в выборе уравнения состояния. Возьмем наиболее простое – уравнение состояния совершенно-го газа

$$p = p(\rho, \epsilon) = (\gamma - 1)\rho\epsilon, \quad (2)$$

где $\gamma = \text{const}$ – показатель адиабаты.

Система уравнений (1) решается в кубе $x, y, z \in (0, 2\pi L)$ при периодических граничных условиях. Начальным условием является одночастотный вихрь:

$$v_x|_{t=0} = v_0 \sin \frac{x}{L} \cos \frac{y}{L} \cos \frac{z}{L}, \quad v_y|_{t=0} = -v_0 \cos \frac{x}{L} \sin \frac{y}{L} \cos \frac{z}{L}, \quad v_z|_{t=0} \equiv 0, \\ p|_{t=0} = p_0 + \frac{\rho_0 v_0^2}{16} \left(\cos \frac{2x}{L} + \cos \frac{2y}{L} \right) \left(2 + \cos \frac{2z}{L} \right), \quad \rho|_{t=0} \equiv \rho_0.$$

Величины ρ_0 , v_0 и p_0 – заданные постоянные параметры.

В несжимаемой постановке в силу известных свойств соответствующих уравнений движения фоновое давление p_0 не играет никакой роли, а потому может быть задано любым. Иначе обстоит дело в сжимаемом случае: p_0 как минимум должно быть задано так, чтобы $p|_{t=0} > 0$; кроме того, если p_0 недостаточно велико, то начальное возмущение скорости приведет к значительным отклонениям ρ от ρ_0 при $t > 0$, что нежелательно. Чтобы плотность жидкости оставалась примерно постоянной, p_0 должно быть много большим, чем $\rho_0 v_0^2$:

$$\frac{p_0}{\rho_0 v_0^2} = \frac{1}{\gamma} \frac{\gamma p_0 / \rho_0}{v_0^2} = \frac{1}{\gamma} \frac{c_0^2}{v_0^2} = \frac{1}{\gamma M_0^2} \gg 1,$$

где $c_0 = \sqrt{\gamma p_0 / \rho_0}$ – фоновая скорость звука, $M_0 = v_0 / c_0$ – начальное число Маха. Во всех расчетах, результаты которых представлены в разд. 4, полагалось $\gamma = 1.4$, $M_0 = 0.1$. Далее в тексте все числовые значения переменных, функций, параметров приводятся без размерностей в системе единиц, в которой $\rho_0, v_0, L = 1$ (после обезразмеривания по такой системе в уравнениях (1), (2) не появляются дополнительные безразмерные множители или числа подобия).

2. ЛОКАЛЬНО-ОДНОМЕРНАЯ БИКОМПАКТНАЯ СХЕМА

Для численного решения задачи о вихре Тейлора–Грина мы применим одну локально-одномерную бикомпактную схему четвертого порядка аппроксимации по пространству и второго по-

рядка аппроксимации по времени. Выбор схемы с расщеплением продиктован двумя факторами: во-первых, высокое спектральное разрешение бикомпактных схем определяется преимущественно их пространственной, а не временной дискретизацией (если последняя, разумеется, имеет хотя бы второй порядок); во-вторых, расщепленные бикомпактные схемы экономичнее своих нерасщепленных аналогов. Отметим, что локально-одномерные бикомпактные схемы не являются новыми (см., например, [21], [27], [28]), поэтому мы дадим минимальное описание выбранной схемы, которое понадобится далее в разд. 3 при рассмотрении алгоритмов параллельного счета по ней.

Итак, согласно методу локально-одномерного расщепления по пространству и методу расщепления Лакса—Фридрихса по потоку, система многомерных уравнений (1) разбивается на шесть систем одномерных уравнений

$$\partial_t \mathbf{Q} + \partial_d \mathbf{F}^{d,\pm}(\mathbf{Q}) = \mathbf{0}, \quad d = x, y, z, \quad (3)$$

где

$$\mathbf{F}^{d,\pm}(\mathbf{Q}) = \frac{1}{2} \mathbf{F}^d(\mathbf{Q}) \pm C_2^d \mathbf{Q}.$$

Параметры $C_2^d > 0$ подбираются так, чтобы матрицы Якоби $\mathbf{A}^{d,\pm}(\mathbf{Q}) = \partial_{\mathbf{Q}} \mathbf{F}^{d,\pm}(\mathbf{Q})$ были положительно/отрицательно определены на некотором множестве возможных значений $\mathbf{Q}$ (конкретная формула для вычисления C_2^d приводится в конце раздела).

Затем в замкнутом кубе $(x, y, z) \in [0, 2\pi]^3$ вводится декартова сетка $\Omega = \Omega_x \times \Omega_y \times \Omega_z$, где

$$\Omega_d = \{0 = d_0 < d_{1/2} < d_1 < d_{3/2} < d_2 < \dots < d_{N_d} = 2\pi\},$$

$$h_{d,i+1/2} = d_{i+1} - d_i - \text{шаг по координате } d, \quad d_{i+1/2} = \frac{d_i + d_{i+1}}{2}, \quad i = \overline{0, N_d - 1}, \quad d = x, y, z.$$

Для каждой из систем (3) строится одномерная полудискретная бикомпактная схема (см. [26]). Например, для системы с вектором потоков $\mathbf{F}^{x,+}(\mathbf{Q})$ она принимает вид

$$\begin{aligned} \frac{d}{dt} (A_0^x \mathbf{Q}_{j+1/2,k,l}) + \Lambda_1^x \mathbf{F}_{j+1/2,k,l}^{x,+} &= \mathbf{0}, \\ \frac{d}{dt} (\Lambda_1^x \mathbf{Q}_{j+1/2,k,l}) + \Lambda_2^x \mathbf{F}_{j+1/2,k,l}^{x,+} &= \mathbf{0}. \end{aligned} \quad (4)$$

В этих уравнениях $j = 0, 1, \dots, N_x - 1$, $k = 0, 1/2, 1, 3/2, 2, \dots, N_y$, $l = 0, 1/2, 1, 3/2, 2, \dots, N_z$. Векторы $\mathbf{F}_{j,k,l}^{d,\pm} = \mathbf{F}^{d,\pm}(\mathbf{Q}_{j,k,l})$ для любых j, k, l . Разностные операторы $A_0^x, \Lambda_1^x, \Lambda_2^x$ определяются формулами

$$\begin{aligned} A_0^x U_{j+1/2,k,l} &= \frac{U_{j,k,l} + 4U_{j+1/2,k,l} + U_{j+1,k,l}}{6}, \quad \Lambda_1^x U_{j+1/2,k,l} = \frac{U_{j+1,k,l} - U_{j,k,l}}{h_{x,j+1/2}}, \\ \Lambda_2^x U_{j+1/2,k,l} &= \frac{4(U_{j,k,l} - 2U_{j+1/2,k,l} + U_{j+1,k,l})}{h_{x,j+1/2}^2}, \end{aligned}$$

где $U_{j,k,l}$ — произвольная сеточная функция, заданная во всех узлах сетки Ω . Для интегрирования по времени в схемах вида (4) используется L -устойчивый жестко точный однократно диагонально-неявный метод Рунге—Кутты (SDIRK) второго порядка (см. [30]) с таблицей Бутчера

$$a \left| \begin{array}{c} a \\ 1 - a \quad a \\ 1 - a \quad a \end{array} \right., \quad a = 1 - \frac{\sqrt{2}}{2}. \quad (5)$$

Как промежуточный результат для одномерных систем (3) получаются одномерные полностью дискретные бикомпактные схемы. Обозначим их операторы послыного перехода через $S_B^{d,\pm}(\tau)$, где τ — шаг по времени, при котором применяется метод (5). Удобно также ввести обозначение $S_B^d(\tau) = S_B^{d,-}(\tau) S_B^{d,+}(\tau)$.

Локально-одномерная бикомпактная схема (немонотонная) для уравнений Эйлера (1) определяется своим оператором послойного перехода $S_B(\tau)$:

$$S_B(\tau) = S_B^x(\tau/2)S_B^y(\tau/2)S_B^z(\tau)S_B^y(\tau/2)S_B^x(\tau/2). \quad (6)$$

Видно, что для достижения второго порядка аппроксимации по времени использована известная схема расщепления Марчука–Стрэнга (см. [31]).

При решении задач газовой динамики схема с оператором послойного перехода (6) обычно подвергается консервативной монотонизации по методу из [26]. Цель настоящей работы, как уже было сформулировано во Введении, состоит в том, чтобы изучить влияние численной диссипации от данного метода на спектральное разрешение выбранной бикомпактной схемы. Отметим, что задача о вихре Тейлора–Грина может быть посчитана по ней от начала до конца и без привлечения метода из [26], как показывается в разд. 4.

Остановимся кратко на методе консервативной монотонизации, чтобы обсуждение его параллельной реализации в разд. 3 было более понятным.

Рассмотрим две численные схемы для системы (1): монотонную схему A первого порядка и немонотонную бикомпактную схему B высокого порядка. В качестве второй выберем схему с оператором послойного перехода (6), а в качестве первой – аналогично расщепленный “явный уголок” с оператором послойного перехода

$$S_A(\tau) = S_A^z(\tau)S_A^y(\tau)S_A^x(\tau).$$

Оператор $S_A^d(\tau) = S_A^{d,-}(\tau)S_A^{d,+}(\tau)$, а операторы $S_A^{d,\pm}(\tau)$ соответствуют одномерным “явным уголкам” для одномерных систем (3), например,

$$S_A^{x,+}(\tau) : \frac{\mathbf{Q}_{j+1,k,l}^{**} - \mathbf{Q}_{j+1,k,l}^*}{\tau} + \frac{\mathbf{F}^{x,+}(\mathbf{Q}_{j+1,k,l}^*) - \mathbf{F}^{x,+}(\mathbf{Q}_{j,k,l}^*)}{h_{x,j+1/2}} = 0, \quad \mathbf{Q}_{j+1/2,k,l}^{**} = \frac{\mathbf{Q}_{j,k,l}^{**} + \mathbf{Q}_{j+1,k,l}^{**}}{2},$$

где $\mathbf{Q}^*$ – некоторое известное решение, $\mathbf{Q}^{**} = S_A^{x,+}(\tau)\mathbf{Q}^*$ – некоторое новое решение.

Вернемся к методу консервативной монотонизации. Пусть нам дано решение $\mathbf{Q}^n$ на временном слое t^n , $n = 0, 1, \dots$. Используя схемы A и B , посчитаем два новых решения на следующем слое t^{n+1} : $\mathbf{Q}^A = S_A(\tau)\mathbf{Q}^n$ и $\mathbf{Q}^B = S_B(\tau)\mathbf{Q}^n$. Вычислим по ним в каждой ячейке $K_C = [x_j, x_{j+1}] \times [y_k, y_{k+1}] \times [z_l, z_{l+1}]$ сетки Ω ($C = (j + 1/2, k + 1/2, l + 1/2)$ – мультииндекс) монотонизированное решение

$$\tilde{Q}_s(\mathbf{r}; K_C) = \bar{Q}_s^B(K_C) + \beta_s(K_C)[Q_s^B(\mathbf{r}) - \bar{Q}_s^B(K_C)], \quad \mathbf{r} = (x, y, z) \in \Omega \cap K_C, \quad s = \overline{1, 5}, \quad (7)$$

где $\bar{Q}^B(K_C) = A_0^z A_0^y A_0^x \mathbf{Q}_C^B$ – интегральное среднее решения $\mathbf{Q}^B$ в ячейке K_C . Множитель $\beta_s(K_C)$ определяется по формуле

$$\beta_s(K_C) = \frac{1}{1 + w_s^2(K_C)}, \quad w_s(K_C) = \frac{C_1 \max_{\mathbf{r} \in \Omega \cap K_C} |Q_s^A(\mathbf{r}) - Q_s^B(\mathbf{r})|}{\max_{\mathbf{r} \in \Omega \cap K_C} |Q_s^A(\mathbf{r})|}. \quad (8)$$

Число $C_1 \geq 0$ является настраиваемым параметром метода из [26]. Поскольку локальная сеточная функция $\tilde{Q}_s(\mathbf{r}; K_C)$ в каждой ячейке K_C строится независимо от аналогичных функций в соседних ячейках, решение (7) оказывается, вообще говоря, многозначным всюду, кроме центров ячеек. В то же время схемы A и B оперируют лишь сеточными функциями, заданными однозначно во всех узлах сетки Ω . Это противоречие разрешается в следующей формуле для результирующего (окончательного) решения:

$$\mathbf{Q}^{n+1}(\mathbf{r}) = \frac{\sum_{K_C: \mathbf{r} \in K_C} \omega(\mathbf{r}; K_C) \tilde{\mathbf{Q}}(\mathbf{r}; K_C) V(K_C)}{\sum_{K_C: \mathbf{r} \in K_C} \omega(\mathbf{r}; K_C) V(K_C)}, \quad \mathbf{r} \in \Omega, \quad (9)$$

где $\omega(\mathbf{r}; K_C)$ — коэффициент (квадратурный вес), с которым значение $\mathbf{Q}^{n+1}(\mathbf{r})$ входит в интегральное среднее $A_0^z A_0^y A_0^x \mathbf{Q}_C^{n+1}$, $V(K_C)$ — объем ячейки K_C . Чтобы учесть периодические граничные условия в формуле (9), сетку Ω и все связанные с ней величины необходимо формально продолжить на все пространство с периодом 2π по всем направлениям. Вследствие этого, например, узел $(x_0, y_{k+1/2}, z_{l+1/2})$ (индексы k, l принимают целые значения) окажется принадлежащим не только ячейке $K_{1/2, k+1/2, l+1/2}$, но и ячейке $K_{N_x-1/2, k+1/2, l+1/2}$ с противоположной стороны расчетной области.

Поясним смысл параметра C_1 и формул (7), (8). При $C_1 = 0$ множители $\beta_s(K_C) \equiv 1$, и монотонизация не происходит; в пределе при $C_1 \rightarrow \infty$ множители $\beta_s(K_C) \equiv 0$, и монотонизация максимальна, так как решение (7) сводится к набору интегральных средних и становится кусочно-постоянным. Сами же множители $\beta_s(K_C)$ при $\tau \rightarrow 0$, $C_1 = O(1) > 0$ ведут себя так: в областях гладкости точного решения $1 - \beta_s(K_C) = O(\tau^2)$ (в этих точках $Q_s^A - Q_s^B = O(\tau)$), а в окрестностях сильных разрывов $1 - \beta_s(K_C) = O(1)$ (в этих точках $Q_s^A - Q_s^B = O(1)$).

Методу консервативной монотонизации (7)–(9) поставим в соответствие оператор $L(C_1)$. Это оператор того же типа, что и операторы послойного перехода $S_A(\tau)$ и $S_B(\tau)$ у схем A и B . Очевидно, $L(0) = E$ — единичному оператору. Оператор послойного перехода искомой монотонизированной локально-одномерной бикомпактной схемы примет вид

$$S(\tau, C_1) = L(C_1)S_B(\tau).$$

Далее мы называем эту схему аббревиатурой SDIRK2B4.

В завершение раздела приведем формулу для параметров потокового расщепления C_2^d :

$$C_2^d = \frac{1+2\delta}{2} V_{\max}^d, \quad V_{\max}^d = \max_{\mathbf{r} \in \Omega} \left(|v_d''(\mathbf{r})| + c''(\mathbf{r}) \right), \quad d = x, y, z, \quad (10)$$

где $\delta > 0$ — настраиваемый параметр, отвечающий за “запас положительной/отрицательной определенности” матриц $A^{d,\pm}(\mathbf{Q})$, $c = \sqrt{\gamma p / \rho}$ — локальная скорость звука. Формула (10) применяется в начале каждого шага по времени. Аналогично, временной шаг τ пересчитывается перед каждым переходом со слоя на слой:

$$\tau = \kappa \min_{d=x,y,z} \frac{\min_{i=0, N_d-1} h_{d,i+1/2}}{\frac{1}{2} V_{\max}^d + C_2^d}, \quad (11)$$

где $\kappa > 0$ — число Куранта.

3. АЛГОРИТМЫ ПАРАЛЛЕЛЬНОГО СЧЕТА

Мотивация. Замена течения несжимаемой жидкости на течение сжимаемой жидкости при малых числах Маха обладает тем недостатком, что при равном числе Куранта расчет течения сжимаемой жидкости потребует равномерно меньших шагов по времени. Это объясняется тем, что число Куранта в несжимаемом случае прямо пропорционально скорости v , а в сжимаемом случае — сумме скоростей $v + c$, где c существенно превосходит v (локальные числа Маха $M = v/c$ малы). Как следствие, расчет задачи о вихре Тейлора—Грина в сжимаемой постановке по бикомпактной схеме SDIRK2B4 на одном ядре CPU на сетках от 100^3 ячеек оказывается неприемлемо долгим из-за довольно большого числа шагов по времени. Попытка воспользоваться неявностью бикомпактных схем через задание большего числа Куранта (равного 2–10) приводит к неудовлетворительным результатам из-за существенного повышения численной диссипации. Таким образом, возникает необходимость в параллельной реализации бикомпактных схем.

Мы предлагаем два алгоритма параллельного счета по локально-одномерным бикомпактным схемам. Оба алгоритма соответствуют парадигме “параллелизма задач” и предназначены для многопроцессорных вычислительных систем с распределенной памятью. Рассуждения достаточно провести на примере схемы SDIRK2B4, так как локально-одномерные бикомпактные схе-

мы не имеют принципиальных различий в своей реализации. Отметим, что параллельная реализация нерасщепленных бикомпактных схем рассмотрена в [23].

Параллелизм локально-одномерных бикомпактных схем. Прежде чем перейти непосредственно к алгоритмам, сделаем несколько общих замечаний о составных частях схемы SDIRK2B4.

Оператор послыйного перехода $S_B(\tau)$, согласно своему определению (6), является композицией операторов $S_B^{d,\pm}(\tau)$, $d = x, y, z$. Каждый из них реализуется посредством итерационного метода Ньютона (необходимого для решения нелинейных разностных уравнений) и совокупности независимых одномерных двухточечных матричных прогонок (циклических, с матрицами размера 5×5); последние устойчивы в силу знакоопределенности матриц $A^{d,\pm}(\mathbf{Q})$. Ясно, что прогонок на разных линиях сетки (т.е. прямых, проходящих через узлы сетки параллельно осям координат) могут выполняться параллельно. Тем не менее каждая прогонка в отдельности должна выполняться строго последовательно, так как, например, при реализации оператора $S_B^{x,+}(\tau)$ значение $\mathbf{Q}_{j+1,k,l}$ может быть вычислено только после нахождения $\mathbf{Q}_{j,k,l}, \mathbf{Q}_{j-1/2,k,l}, \dots, \mathbf{Q}_{0,k,l}$.

На оператор послыйного перехода $S_A(\tau)$ мы смотрим так же, как на оператор $S_B(\tau)$, хотя реализация $S_A^{d,\pm}(\tau)$ для любого $d = x, y, z$ обладает параллелизмом, в том числе и по оси Od (схема A – явная).

Метод консервативной монотонизации (7)–(9) и соответствующий ему ограничивающий оператор $L(C_1)$ реализуются явно и локально, так как найдены решения $\mathbf{Q}^A = S_A(\tau)\mathbf{Q}^n$ и $\mathbf{Q}^B = S_B(\tau)\mathbf{Q}^n$. Это следует из того, что (а) для применения формул (7), (8) в каждой ячейке достаточно только данных из этой ячейки, (б) для применения формулы (9) в каждом узле достаточно данных из тех ячеек, которым принадлежит этот узел. Значит, метод (7)–(9) обладает параллелизмом по ячейкам.

Алгоритм 1. Введем n_p процессов. Совокупность всех ячеек сетки Ω разделим поровну на n_p горизонтальных (относительно оси Oz) слоев и n_p вертикальных слоев. Горизонтальный слой ячеек номер $r = 0, 1, \dots, n_p - 1$ включает в себя узлы

$$W_r = \left\{ (x_j, y_k, z_l) : j, k - \text{любые}, l = \frac{rN_z}{n_p} + i, i = 0, 1/2, 1, \dots, \frac{N_z}{n_p} \right\} \subset \Omega,$$

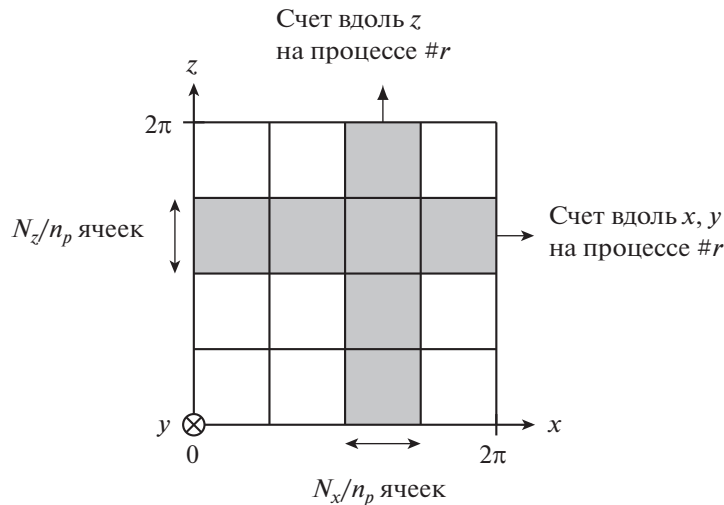
а вертикальный слой номер r – узлы

$$W_r^* = \left\{ (x_j, y_k, z_l) : k, l - \text{любые}, j = \frac{rN_x}{n_p} + i, i = 0, 1/2, 1, \dots, \frac{N_x}{n_p} \right\} \subset \Omega.$$

Ради меньшей громоздкости формул мы предполагаем, что числа N_z и N_x делятся нацело на n_p . Если это не так, то остатки от деления, выраженные в ячейках, равномерно распределяются по слоям. Процесс номер r работает с горизонтальным и вертикальным слоями ячеек с тем же номером, иными словами, с двумя массивами векторов $\mathbf{Q}^n$: один задан на W_r , другой – на W_r^* . Разделение расчетной области по процессам схематически изображено на фиг. 1.

В начале расчета каждый процесс инициализирует свою сеточную функцию $\mathbf{Q}^0$ на W_r с помощью известного начального условия. Переход со слоя t^n на слой t^{n+1} ($n = 0, 1, \dots$) на процессе номер r состоит из следующих действий:

1. На W_r вычисляются местные максимумы $V_{\max}^d$. При $r \neq 0$ числа $V_{\max}^d$ посылаются на процесс номер 0, после чего рассматриваемый процесс (номер r) ожидает от него в ответ параметры C_2^d и шаг τ . При $r = 0$ числа $V_{\max}^d$ сравниваются с аналогичными величинами, принимаемыми от остальных процессов, в результате чего становятся известными глобальные максимумы $V_{\max}^d$ на Ω . Они позволяют посчитать параметры C_2^d и шаг τ по формулам (10), (11). Эти четыре числа рассылаются всем остальным процессам.



Фиг. 1. Схема распараллеливания в алгоритме 1 ($n_p = 4$, $r = 2$).

2. На W_r вычисляются два решения: $\mathbf{Y}^A = S_A^y(\tau)S_A^x(\tau)\mathbf{Q}^n$ и $\mathbf{Y}^B = S_B^y(\tau/2)S_B^x(\tau/2)\mathbf{Q}^n$. Значения $\mathbf{Y}^A$ и $\mathbf{Y}^B$ в плоскости $z = z_{rN_z/n_p}$ не рассчитываются по схемам A и B , а принимаются в готовом виде от процесса номер $r - 1$ (если $r = 0$, то в силу периодических граничных условий принимаются данные с плоскости $z = z_N = 2\pi$ от процесса номер $n_p - 1$). Это делается для того, чтобы избежать удвоения вычислений на границах между горизонтальными слоями ячеек.

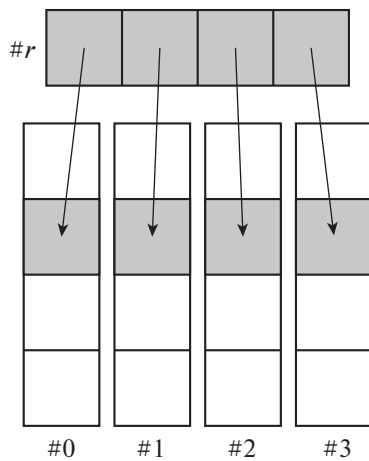
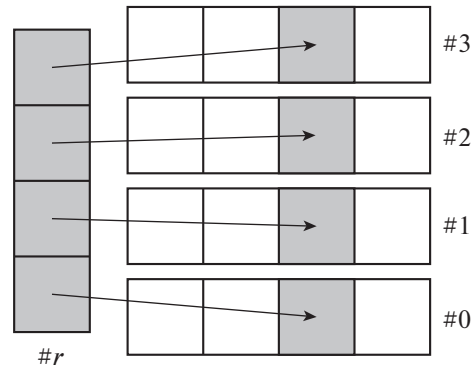
3. Выполняется барьерная синхронизация процессов. Массив, отвечающий решению $\mathbf{Y}^A$ на W_r , разбивается вдоль оси Ox на n_p частей сообразно разделению области на вертикальные слои (фиг. 2). Часть номер $r' \neq r$ этого массива посылается на процесс с номером r' ; часть номер $r' = r$ записывается (на соответствующие ей места) в “сопряженный” массив, отвечающий решению $\mathbf{Y}^A$ на W_r^* . Данные, принимаемые от остальных процессов, заполняют этот “сопряженный” массив до конца. Над массивом, в котором хранится решение $\mathbf{Y}^B$ на W_r , совершаются те же операции.

4. На W_r^* вычисляется два решения: $\mathbf{Q}^A = S_A^z(\tau)\mathbf{Y}^A = S_A(\tau)\mathbf{Q}^n$ и $\tilde{\mathbf{Y}}^B = S_B^z(\tau)\mathbf{Y}^B$. Дублирование вычислений на границах между вертикальными слоями устраняется аналогично п. 2.

5. Выполняется барьерная синхронизация процессов. Над массивами, в которых хранятся решения $\mathbf{Q}^A$ и $\tilde{\mathbf{Y}}^B$ на W_r^* , выполняются операции разбиения и рассылки, симметричные (обратные) тому, что описано в п. 3 (фиг. 2). В итоге процессу номер r становятся известны решения $\mathbf{Q}^A$ и $\tilde{\mathbf{Y}}^B$ на W_r .

6. На W_r , аналогично п. 2, вычисляется решение $\mathbf{Q}^B = S_B^x(\tau/2)S_B^y(\tau/2)\tilde{\mathbf{Y}}^B = S_B(\tau)\mathbf{Q}^n$.

7. На W_r по решениям $\mathbf{Q}^A$ и $\mathbf{Q}^B$, согласно формулам (7)–(9), строится искомое финальное решение $\mathbf{Q}^{n+1} = L(C_1)\mathbf{Q}^B = S(\tau, C_1)\mathbf{Q}^n$. Благодаря тому, что множество W_r включает узлы с границ между горизонтальными слоями, формулы (7), (8) реализуются без обращения к соседним слоям и процессам, связанным с ними. Формулу (9) проще всего реализовывать не через обход по узлам, а через обход по ячейкам: сначала в каждом узле W_r суммы, участвующие в правой части (9), полагаются нулевыми; затем в процессе обхода по ячейкам к ним добавляются составляющие их слагаемые. Однако для применения (9) на плоскостях $z = z_{rN_z/n_p}$ и $z = z_{(r+1)N_z/n_p}$ придется обратиться к процессам с номерами $r - 1$ и $r + 1$, чтобы получить нужные данные от ячеек, примыка-

Перед реализацией $S_A^x(\tau)$, $S_B^x(\tau)$:После реализации $S_A^x(\tau)$, $S_B^x(\tau)$:Фиг. 2. Обмен данными между процессами в алгоритме 1 ($n_p = 4$, $r = 2$).

ющих к этим плоскостям снаружи рассматриваемого r -го слоя. В силу периодических граничных условий слоем с номером (-1) является слой $n_p - 1$, а слоем с номером n_p — слой 0.

Отметим, что все обращения за границы расчетной области вызваны исключительно периодическими граничными условиями, а не особенностями схем A и B (пространственные шаблоны составляющих их одномерных схем и метода (7)–(9) минимальны); любые другие граничные условия не порождают таких обращений.

Дублирование вычислений на границах между слоями ячеек исключается (см. шаги 2, 4, 6 алгоритма) по двум причинам. Во-первых, расчет решения на одной плоскости при $N_d \geq 10$ занимает на порядок больше машинного времени, чем пересылка этих данных с другого процесса. Во-вторых, например, при $n_p = N_z/4$ объем лишних (дублированных) вычислений на шаге 2 составляет $N_z/(4(2N_z + 1)) \approx 13\%$ от общего объема необходимых вычислений, что представляется достаточно неэффективным.

Все расчеты, результаты которых представлены в разд. 4, выполнены по алгоритму 1. Из-за этого далее алгоритм 2 описывается менее подробно; он нужен для того, чтобы показать альтернативный способ параллельной реализации локально-одномерных бикомпактных схем.

Алгоритм 2. Применим стратегию “управляющий–рабочие” (см. [32]). Назначим процесс с номером 0 управляющим процессом; предположим, что он имеет доступ ко всему массиву векторов $\mathbf{Q}^n$, т.е. решению на всей сетке Ω . Процессы с номерами $1, 2, \dots, n_p - 1$ принимают тогда роль рабочих процессов.

Рассмотрим реализацию схемы B на одном временном шаге. Процесс номер 0 берет решение $\mathbf{Q}^n$ с прямых

$$\{(x, y, z) : y = y_0; z = z_0, z_{1/2}, z_1, \dots, z_{n_p/2-1}\} \parallel O_x$$

и рассылает его по процессам номер $1, 2, \dots, n_p - 1$ соответственно; обратно он получает новое решение $S_B^x(\tau/2)\mathbf{Q}^n$ с этих же прямых. Рабочий процесс, который закончил считать первым, получает решение $\mathbf{Q}^n$ с прямой $y = y_0$, $z = z_{n_p/2-1/2}$ и возвращает управляющему процессу решение $S_B^x(\tau/2)\mathbf{Q}^n$ с этой же прямой. Следующий освободившийся рабочий процесс получает решение $\mathbf{Q}^n$ с прямой $y = y_0$, $z = z_{n_p/2}$ и т.д. После обхода по всем прямым $y = y_k \in \Omega_y$, $z = z_l \in \Omega_z$ получается решение $S_B^x(\tau/2)\mathbf{Q}^n$ на всей сетке Ω . Затем оно подвергается действию оператора $S_B^y(\tau/2)$; рассылка на рабочие процессы производится с прямых, проходящих через узлы сетки Ω парал-

тельно оси Oy (а не Ox). Аналогично осуществляется действие операторов $S_B^z(\tau)$, $S_B^y(\tau/2)$, $S_B^x(\tau/2)$, и мы получаем искомое решение $Q^B = S_B(\tau)Q^n$ схемы B . Схема A реализуется тем же способом, что и схема B .

При реализации метода консервативной монотонизации (7)–(9) управляющий процесс рассылает по рабочим процессам решения Q^A и Q^B с подмножеств ячеек, порожденных любым удобным разбиением всего множества ячеек. Например, это могут быть горизонтальные слои ячеек из алгоритма 1. Рабочие процессы получают по формулам (7), (8) данные, необходимые для применения формулы (9), и отправляют их на управляющий процесс, на котором “собирается” результирующее решение Q^{n+1} .

Вычисление максимумов $V_{\max}^d$, необходимых для пересчета параметров C_2^d и шага τ перед каждым переходом на новый слой, выполняется по аналогии с методом (7)–(9) (с той разницей, что каждый из рабочих процессов возвращает управляющему лишь 3 числа, а не решение с некоторого подмножества ячеек).

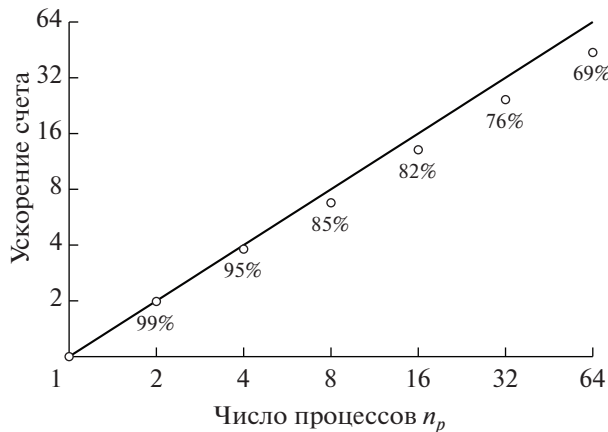
Сравнение алгоритмов. Напомним, что реализация любого из операторов послойного перехода $S_B^{d,\pm}(\tau)$ на любой линии сетки включает в себя итерационный метод Ньютона (для решения нелинейных разностных уравнений), а значит, не является детерминированной по времени счета. Например, если решение тождественно постоянно на какой-то линии сетки, то метод Ньютона сойдется на ней всего за одну итерацию, в несколько раз быстрее предполагаемой длительности итерационного процесса. Это обстоятельство вкупе с фиксированным разбиением расчетной области на слои в алгоритме 1 создает потенциальную проблему неравномерной загрузки процессов. Заметим все же, что в задаче о вихре Тейлора–Грина решение существенно отличается от константы во всей расчетной области во все моменты времени, а потому в этой и подобной ей задачах данный недостаток алгоритма 1 не должен проявляться в заметной степени. Кроме того, эту потенциальную проблему можно решить динамической регулировкой разбиения совокупности ячеек на слои, что, однако, выходит за рамки настоящей работы. Другим недостатком алгоритма 1 является ограничение на число процессов: $n_p \leq \min(N_z, N_x)$ (очевидно, число слоев, на которые разбивается вся совокупность ячеек, не может превосходить числа ячеек по соответствующему направлению). В то же время алгоритм 2 лишен этих недостатков: линии сетки динамически распределяются по процессам, а число процессов ограничено довольно большим числом этих линий, порядка $4N^2$ при $N_x = N_y = N_z = N$. Тем не менее алгоритм 2 проигрывает алгоритму 1 в объеме пересылаемых данных. Пусть размер массива векторов Q на всей сетке Ω равен M . Тогда за один временной шаг в алгоритме 1 суммарно пересылается объем данных $4M$ (решения Y^A, Y^B до реализации операторов $S_A^z(\tau), S_B^z(\tau)$, решения $Q^A, \tilde{Y}^B$ после; обментами на границах между слоями можно пренебречь), а в алгоритме 2 – $19M$ (по $2M$ на каждый из операторов вида $S_A^d(\tau), S_B^d(\tau)$, $2M$ на оператор $L(C_1)$, M на вычисление C_2^d и τ), что почти в пять раз больше. Из сказанного следует вывод, что в задаче о вихре Тейлора–Грина алгоритм 1 является более подходящим для параллельной реализации бикомпактной схемы SDIRK2B4.

4. АНАЛИЗ РЕЗУЛЬТАТОВ

Рассмотрим результаты расчетов задачи о вихре Тейлора–Грина по бикомпактной схеме SDIRK2B4 на сетках $32^3, 64^3, 128^3$ ячеек. Укажем значения параметров схемы: число Куранта $\kappa = 0.8$; параметр метода консервативной монотонизации $C_1 = 0, 0.2, 0.5$ (по три варианта для каждой сетки); параметр потокового расщепления $\delta = 0.2$. Обратим внимание на то, что при расчетах других задач газовой динамики параметр C_1 брался примерно на порядок большим (см. [26]). Его уменьшение обусловлено тем, что характерное число Маха в задаче о вихре Тейлора–Грина примерно на порядок меньше по сравнению с задачами, которые решались ранее.

Расчетный код реализован на языке C++ с применением библиотеки MPI. Вычисления проводились на гибридном суперкомпьютере К-60, установленном в Центре коллективного пользования ИПМ им. М.В. Келдыша РАН.

Сначала проверим эффективность распараллеливания по алгоритму 1. На фиг. 3 представлена зависимость ускорения счета от числа процессов n_p ; она построена по временам счета первого



Фиг. 3. Ускорение счета по бикомпактной схеме SDIRK2B4 в зависимости от числа процессов n_p . Сплошная прямая соответствует идеальному распараллеливанию, маркеры — фактически наблюдаемому. Подписи под маркерами означают эффективность распараллеливания. График построен по временам счета первого временного шага на сетке 128^3 ячеек при $C_1 = 0.2$.

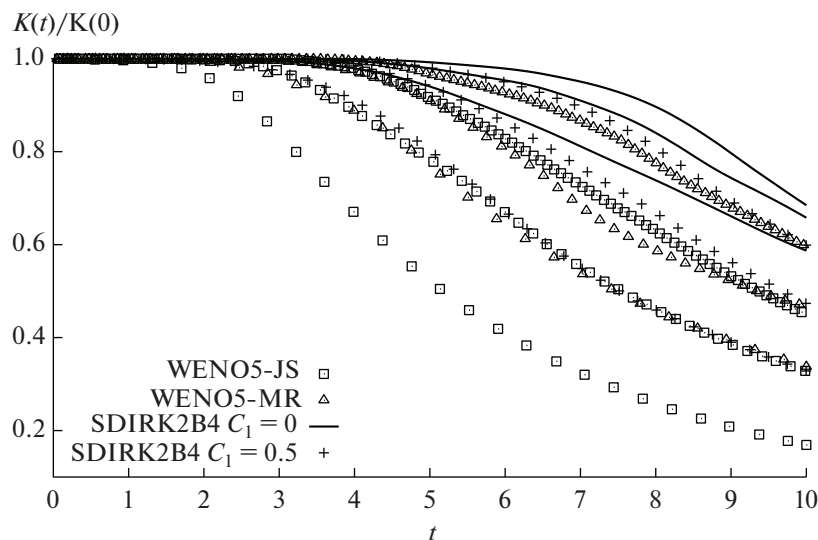
временного шага на сетке 128^3 ячеек при $C_1 = 0.2$. Видно, что при $n_p \leq 32$ эффективность распараллеливания достаточно высокая, не меньше 76%. Чем больше n_p (и чем оно ближе к N_z или N_x), тем тоньше слои ячеек, с которыми работает каждый процесс, тем больше актов обмена между процессами и тем ниже эффективность параллельного счета. Полные расчеты задачи до момента времени $t = 10$ на сетках 32^3 , 64^3 , 128^3 ячеек делались соответственно на $n_p = 8, 16, 32$ процессах.

Сравним бикомпактную схему SDIRK2B4 с другими схемами высокого порядка по уровню численной диссипации. Для сопоставления выберем две схемы серии WENO5: классическую схему WENO5-JS (см. [12]) и одну из ее последних улучшенных версий — схему WENO5-MR (см. [13]) (“nested multi-resolution WENO5”), несколько превосходящую популярную схему WENO5-Z из [14]). Обе схемы имеют аппроксимацию пятого порядка по пространству и третьего порядка по времени. Результаты для схемы WENO5-JS заимствованы из [13]. О численной диссипации мы будем судить по эволюции двух величин — кинетической энергии $K(t)$ и энтропии $\mathcal{E}(t)$ жидкости во всем кубе $\mathbf{r} \in [0, 2\pi]^3$:

$$K(t) = \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \frac{\rho v^2}{2} dx dy dz, \quad \mathcal{E}(t) = \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \frac{\rho |\text{rot } \mathbf{v}|^2}{2} dx dy dz.$$

Если жидкость идеальная и несжимаемая, то кинетическая энергия сохраняется: $K(t) = K(0)$. Следовательно, чем меньше в процессе счета $K(t)$ отклоняется от $K(0)$, тем ниже численная диссипация рассматриваемой схемы, и тем лучше последняя воспроизводит точное решение. Безусловно, интеграл $K(t) = K(0)$ нарушается не только из-за численной диссипации, но и благодаря искусственно введенной сжимаемости, однако ее влиянием на не слишком подробных сетках можно пренебречь. Что же касается энтропии, то она дает следующий критерий для оценки схемы: чем ниже максимум $\mathcal{E}(t)$, тем больше численная диссипация подавляет средние и более мелкие вихри, и тем сильнее она искажает их истинный инерционный каскад.

На фиг. 4 изображены кривые нормированной кинетической энергии $K(t)/K(0)$ для всех схем и всех сеток. Для схемы SDIRK2B4 показаны лишь кривые, полученные при $C_1 = 0$ (без монотонизации) и при $C_1 = 0.5$, как соответственно наилучший и наихудший (в смысле диссипации) варианты. Из данной фигуры очевидно, что схема WENO5-JS имеет самую большую диссипацию; чтобы быть сопоставимой со схемами WENO5-MR и SDIRK2B4, ей требуется сетка с двукратно большим числом ячеек по всем направлениям. Схема SDIRK2B4 при $C_1 = 0.5$ немного превосходит схему WENO5-MR. В отсутствие монотонизации схема SDIRK2B4 демонстрирует значительное преимущество перед схемами WENO5-JS и WENO5-MR. Интересно, что L -устой-



Фиг. 4. Сравнение бикомпактной схемы SDIRK2B4 со схемами WENO5-JS, WENO5-MR по эволюции кинетической энергии жидкости. На графике представлены данные для всех трех сеток 32^3 , 64^3 , 128^3 ячеек: чем ближе та или иная зависимость к единице, тем больше размер соответствующей сетки.

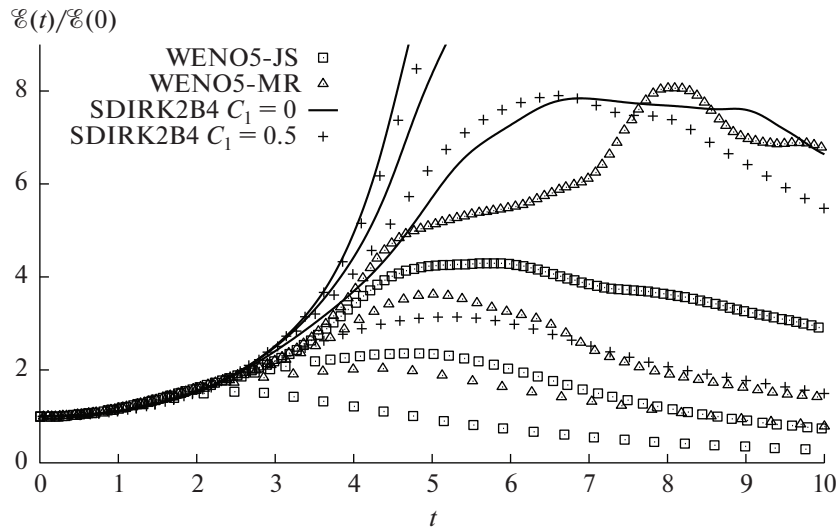
чивость временной дискретизации в бикомпактной схеме проявляет себя как своего рода “встроенный” фильтр нежелательных осцилляций, делающий возможным полный расчет задачи без монотонизации. Свойство L -устойчивости здесь действительно по существу, так как, например, при дискретизации по времени методом трапеций (как известно, не обладающим этим свойством) бикомпактная схема неустойчива при $C_1 = 0$.

Продолжим наше сравнение схем анализом кривых нормированной энтропии $\mathcal{E}(t)/\mathcal{E}(0)$, изображенных на фиг. 5, 6. Из этих фигур следует, что по энтропии схема WENO5-MR соотносится с монотонизированной схемой SDIRK2B4 так же, как по кинетической энергии схема WENO5-JS соотносится со схемой WENO5-MR: чтобы достичь примерно того же максимума энтропии, схеме WENO5-MR нужна сетка с удвоенным числом ячеек по каждому направлению. Схема SDIRK2B4 без монотонизации дает самые высокие максимумы $\mathcal{E}(t)$: уже на сетке 32^3 ячеек получается примерно такое же значение, как у схемы WENO5-MR на сетке 128^3 ячеек и у монотонизированной схемы SDIRK2B4 на сетке 64^3 ячеек; на сетке 128^3 ячеек максимум энтропии составляет около 56, что намного превосходит все остальные результаты, участвующие в сравнении. Добавим, что кривые энтропии, полученные по бикомпактной схеме на сетке 128^3 ячеек при всех трех значениях C_1 , находятся в очень хорошем согласии с полуаналитической зависимостью для времен $t \leq 4$ из [29] (фиг. 6).

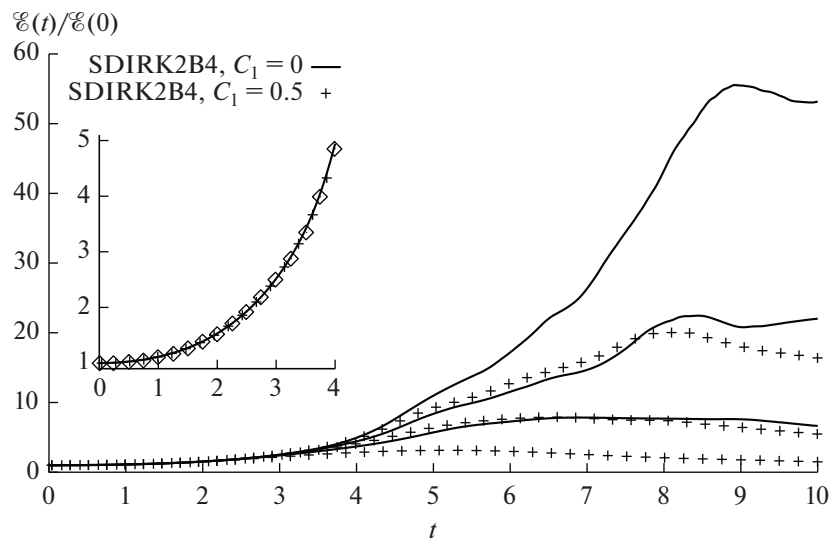
Для полноты изложения поясним, что относительное отклонение плотности жидкости от начального значения составляет не более 2% для всех сеток, для всех выбранных значений C_1 , на всех слоях по времени, во всех узлах. Следовательно, в процессе движения жидкость действительно остается примерно несжимаемой.

Перейдем к основному вопросу настоящей работы — вопросу о влиянии метода консервативной монотонизации из [26] на разрешение бикомпактной схемой SDIRK2B4 спектра $\sigma_K(k, t)$ кинетической энергии жидкости. Уточним, что $\sigma_K(k, t)$ вычисляется по формуле

$$\sigma_K(k, t) = \sum_{k - \frac{1}{2} \leq |\mathbf{k}'| < k + \frac{1}{2}} \frac{|\mathbf{A}(\mathbf{k}', t)|^2}{2},$$



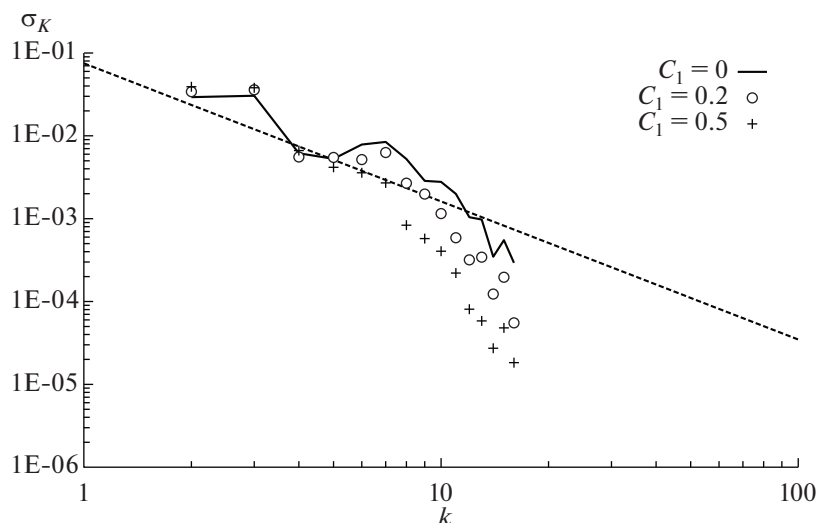
Фиг. 5. Сравнение бикompактной схемы SDIRK2B4 со схемами WENO5-JS, WENO5-MR по эволюции энтропии. На графике представлены данные для всех трех сеток 32^3 , 64^3 , 128^3 ячеек: чем выше максимум той или иной зависимости, тем больше размер соответствующей сетки. Ради лучшей читаемости фигуры значения $\mathcal{E}(t)/\mathcal{E}(0)$ ограничены девятью; недостающие участки кривых изображены на фиг. 6.



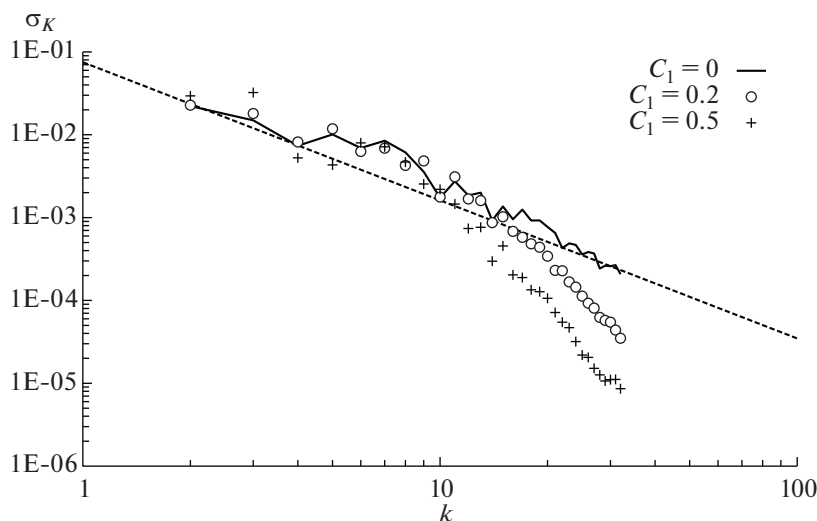
Фиг. 6. Дополнение к фиг. 5. На увеличенном фрагменте показаны данные только для сетки 128^3 ячеек и ромбами — полуаналитическая эталонная зависимость из [29].

где $\mathbf{k} = [k_x \ k_y \ k_z]^T$ — вектор волновых чисел $k_d = 0, \pm 1, \pm 2, \pm \frac{1}{2} N_d$ (пусть все N_d делятся на 2), $k = |\mathbf{k}| = 0, 1, \dots, \frac{1}{2} \min_{d=x,y,z} N_d$ — абсолютное волновое число, $A(\mathbf{k}, t)$ есть $\mathbf{k}$ -й коэффициент ряда Фурье вектора скорости жидкости $\mathbf{v} = \mathbf{v}(\mathbf{r}, t)$:

$$A(\mathbf{k}, t) = \frac{1}{(2\pi)^3} \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \mathbf{v}(\mathbf{r}, t) \exp(-i\mathbf{k} \cdot \mathbf{r}) dx dy dz.$$

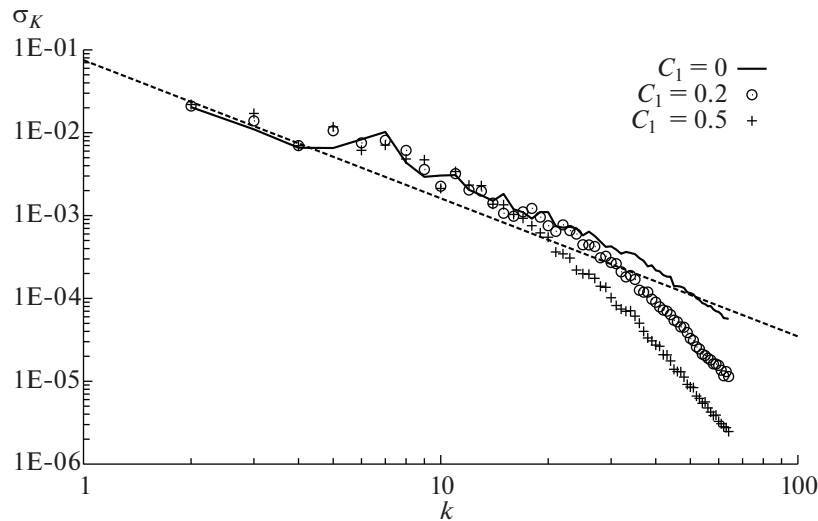


Фиг. 7. Спектры кинетической энергии жидкости, полученные по бикомпактной схеме SDIRK2B4 при разных значениях параметра монотонизации C_1 , сетка 32^3 ячеек. Пунктирной прямой показан наклон $(-5/3)$.



Фиг. 8. Спектры кинетической энергии жидкости, полученные по бикомпактной схеме SDIRK2B4 при разных значениях параметра монотонизации C_1 , сетка 64^3 ячеек. Пунктирной прямой показан наклон $(-5/3)$.

На фиг. 7–9 приведены искомые графики спектров, полученных по бикомпактной схеме SDIRK2B4 на сетках 32^3 , 64^3 , 128^3 ячеек соответственно (при $C_1 = 0, 0.2, 0.5$). Каждый спектр построен в момент времени, отвечающий точке максимума энтропии $\mathcal{E}(t)$. Из-за малого числа точек фиг. 7 не позволяет со всей ясностью оценить численную диссипацию при $C_1 \neq 0$, однако эта фигура все же показывает, что даже на весьма грубой сетке 32^3 ячеек схема SDIRK2B4 без монотонизации более-менее верно (не считая трех точек на правом конце спектра) воспроизводит колмогоровский наклон $(-5/3)$. Фиг. 8, 9 дают более понятную картину: видно, что при $C_1 = 0$ спектры идут вдоль прямой с наклоном $(-5/3)$ без существенных отклонений вплоть до самых больших волновых чисел, а при $C_1 = 0.2$ и $C_1 = 0.5$ бикомпактная схема разрешает соответственно 85 и 70% спектра (по логарифмической шкале), если принимать за эталон спектр при $C_1 = 0$.



Фиг. 9. Спектры кинетической энергии жидкости, полученные по бикомпактной схеме SDIRK2B4 при разных значениях параметра монотонизации C_1 , сетка 128^3 ячеек. Пунктирной прямой показан наклон $(-5/3)$.

и полагать допустимым искажение амплитуды гармоники не более, чем в два раза. Нельзя не заметить, что монотонизированная схема SDIRK2B4 не подавляет полностью вихри наименьших масштабов, поэтому получаемые по ней спектры не имеют резких “уклонов вниз” на правом конце.

Отметим напоследок одну любопытную деталь спектра, полученного по схеме SDIRK2B4 на сетке 128^3 ячеек при $C_1 = 0$ (фиг. 9): несмотря на то, что монотонизация отсутствует (“выключена”), он все же в очень незначительной степени “загибается вниз” после $k \approx 40$. Такое поведение является следствием L -устойчивости временной дискретизации; выше мы уже говорили о том, что данное свойство выступает в роли некоего встроенного минимального фильтра самых высокочастотных составляющих решения.

ЗАКЛЮЧЕНИЕ

В работе предложены два алгоритма параллельной реализации локально-одномерных бикомпактных схем для систем трехмерных уравнений гиперболического типа. Оба алгоритма предназначены для многопроцессорных вычислительных систем с распределенной памятью. В основе алгоритма 1 лежит статическое разбиение совокупности ячеек пространственной сетки на горизонтальные и вертикальные слои, в основе алгоритма 2 — динамическое распределение вычислений с линий сетки между процессами. Алгоритм 1 проигрывает алгоритму 2 в равномерности загрузки процессов (в случае, если рассчитываемое решение имеет вид локализованного возмущения на постоянном фоне), но существенно выигрывает в объеме пересылаемых данных.

По локально-одномерной бикомпактной схеме SDIRK2B4 четвертого порядка аппроксимации по пространству и второго порядка аппроксимации по времени выполнены расчеты задачи о вихре Тейлора—Грина в невязком совершенном газе. На примере данной задачи и выбранной схемы выяснялось то, в какой мере теоретически высокое спектральное разрешение бикомпактных схем ухудшается практически из-за применяемого в них метода консервативной монотонизации из [26] — это важно для понимания перспективы приложения схем данного класса к LES.

Вычисления проводились на сетках 32^3 , 64^3 , 128^3 ячеек, при этом применялась параллельная реализация по алгоритму 1 (постановка задачи делает его более предпочтительным); эффективность распараллеливания составила примерно 75%.

Анализ результатов позволяет сделать несколько выводов. Во-первых, благодаря L -устойчивости временной дискретизации, бикомпактная схема SDIRK2B4 имеет встроенный механизм стабилизации счета, который позволяет решать некоторые задачи газовой динамики вообще в отсутствие какой-либо специальной монотонизации, разрешение спектра кинетической энер-

гии при этом оказывается высоким. Во-вторых, при использовании метода консервативной монотонизации из [26] данное разрешение ухудшается в умеренной, допустимой степени — разрешаются 70–85% спектра. В-третьих, если судить по поведению энтропии, то по сравнению с современными схемами серии WENO5 схема SDIRK2B4 лучше воспроизводит вихревой каскад и не так сильно подавляет вихри среднего масштаба. Таким образом, направлением дальнейших исследований являются обобщение бикомпактных схем на уравнения Навье–Стокса для сжимаемой жидкости и затем применение этих новых схем в LES.

СПИСОК ЛИТЕРАТУРЫ

1. *Андерсон Д., Таннехилл Дж., Плетчер Р.* Вычислительная гидромеханика и теплообмен: В 2-х т. Т. 1: Пер. с англ. М.: Мир, 1990. 384 с.
2. *Kawai S., Lele S.K.* Large-eddy simulation of jet mixing in supersonic crossflows // *AIAA J.* 2010. V. 48. № 9. P. 2063–2083.
3. *Bull J.R., Jameson A.* Simulation of the Taylor–Green vortex using high-order flux reconstruction schemes // *AIAA J.* 2015. V. 53. № 9. P. 2750–2761.
4. *Sagaut P.* Large Eddy simulation for incompressible flows. 3rd ed. Berlin: Springer, 2006.
5. *Moura R.C., Mengaldo G., Peiró J., Sherwin S.J.* On the eddy-resolving capability of high-order discontinuous Galerkin approaches to implicit LES under-resolved DNS of Euler turbulence // *J. Comput. Phys.* 2017. V. 330. P. 615–623.
6. *Flad D., Beck A., Guthke P.* A large Eddy simulation method for DGSEM using non-linearly optimized relaxation filters // *J. Comput. Phys.* 2020. V. 408. 109303.
7. *Drikakis D., Fureby C., Grinstein F., Youngs D.* Simulation of transition and turbulence decay in the Taylor–Green vortex // *J. Turbul.* 2007. V. 8. № 20. P. 1–12.
8. *Liu X., Zhang S., Zhang H., Shu C.-W.* A new class of central compact schemes with spectral-like resolution I: Linear schemes // *J. Comput. Phys.* 2013. V. 248. P. 235–256.
9. *De La Llave Plata M., Couaillier V., Pape M.-C.* On the use of a high-order discontinuous Galerkin method for DNS and LES of wall-bounded turbulence // *Comput. Fluids.* 2018. V. 176. P. 320–337.
10. *Gassner G., Beck A.* On the accuracy of high-order discretizations for underresolved turbulence simulations // *Theor. Comp. Fluid. Dyn.* 2013. V. 27. P. 221–237.
11. *Diosady L.T., Murman S.M.* DNS of flows over periodic hills using a discontinuous-Galerkin spectral-element method // 44th AIAA Fluid Dynam. Conf. 2014.
12. *Jiang G.-S., Shu C.-W.* Efficient implementation of weighted ENO schemes // *J. Comput. Phys.* 1996. V. 126. № 1. P. 202–228.
13. *Wang Z., Zhu J., Tian L., Zhao N.* A low dissipation finite difference nested multi-resolution WENO scheme for Euler/Navier–Stokes equations // *J. Comput. Phys.* 2021. V. 429. 110006.
14. *Borges R., Carmona M., Costa B., Don W.S.* An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws // *J. Comput. Phys.* 2008. V. 227. № 6. P. 3191–3211.
15. *Головизнин В.М., Зайцев М.А., Карабасов С.А., Короткин И.А.* Новые алгоритмы вычислительной гидродинамики для многопроцессорных вычислительных комплексов. М.: Изд-во МГУ, 2013. 472 с.
16. *Тишкин В.Ф., Гасилов В.А., Змитренко Н.В. и др.* Современные методы математического моделирования развития гидродинамических неустойчивостей и турбулентного перемешивания // *Матем. моделирование.* 2020. Т. 32. № 8. С. 57–90.
17. *Grimich K., Cinnella P., Lerat A.* Spectral properties of high-order residual-based compact schemes for unsteady compressible flows // *J. Comput. Phys.* 2013. V. 252. P. 142–162.
18. *Толстых А.И.* Компактные разностные схемы и их применение в задачах аэрогидродинамики. М.: Наука, 1990. 230 с.
19. *Толстых А.И.* Компактные и мультиоператорные аппроксимации высокой точности для уравнений в частных производных. М.: Наука, 2015. 350 с.
20. *Брагин М.Д., Рогов Б.В.* Гибридные бикомпактные схемы с минимальной диссипацией для уравнений гиперболического типа // *Ж. вычисл. матем. и матем. физ.* 2016. Т. 56. № 6. С. 958–972.
21. *Брагин М.Д., Рогов Б.В.* О точном пространственном расщеплении многомерного скалярного квазилинейного гиперболического закона сохранения // *Докл. АН.* 2016. Т. 469. № 2. С. 143–147.
22. *Рогов Б.В., Брагин М.Д.* О свойствах спектрального разрешения симметричных бикомпактных схем четвертого порядка аппроксимации // *Докл. АН.* 2017. Т. 475. № 2. С. 140–144.
23. *Чикиткин А.В., Рогов Б.В.* Два варианта параллельной реализации высокоточных бикомпактных схем для многомерного неоднородного уравнения переноса // *Препринты ИПМ им. М.В. Келдыша.* 2018. № 177. 24 с.

24. *Chikitkin A.V., Rogov B.V.* Family of central bicomcompact schemes with spectral resolution property for hyperbolic equations // *Appl. Numer. Math.* 2019. V. 142. P. 151–170.
25. *Rogov B.V.* Dispersive and dissipative properties of the fully discrete bicomcompact schemes of the fourth order of spatial approximation for hyperbolic equations // *Appl. Numer. Math.* 2019. V. 139. P. 136–155.
26. *Bragin M.D., Rogov B.V.* Conservative limiting method for high-order bicomcompact schemes as applied to systems of hyperbolic equations // *Appl. Numer. Math.* 2020. V. 151. P. 229–245.
27. *Брагин М.Д., Рогов Б.В.* Бикомпактные схемы для задач газовой динамики: обобщение на сложные расчетные области методом свободной границы // *Компьют. исслед. и моделирование.* 2020. Т. 12. № 3. С. 487–504.
28. *Брагин М.Д., Рогов Б.В.* Высокоточные бикомпактные схемы для численного моделирования течений многокомпонентных газов с несколькими химическими реакциями // *Матем. моделирование.* 2020. Т. 32. № 6. С. 21–36.
29. *Brachet M.E., Meiron D.I., Orszag S.A., et al.* Small-scale structure of the Taylor–Green vortex // *J. Fluid Mech.* 1983. V. 130. P. 411–452.
30. *Alexander R.* Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s // *SIAM J. Numer. Anal.* 1977. V. 14. № 6. P. 1006–1021.
31. *Марчук Г.И.* Методы расщепления. М.: Наука, 1988. 263 с.
32. *Лупин С.А., Посыпкин М.А.* Технологии параллельного программирования. М.: ИД “ФОРУМ”: ИНФРА-М, 2011. 208 с.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 532.5:533.6.011

ОБ УПРОЩЕНИИ ЧИСЛЕННЫХ И АНАЛИТИЧЕСКИХ
“ИНСТРУМЕНТОВ” ОПИСАНИЯ “ЗВУКОВОГО УДАРА”¹⁾© 2022 г. Х. Ф. Валиев¹, А. Н. Крайко^{1,*}, Н. И. Тилляева¹¹ 111116 Москва, ул. Авиамоторная, 2, ЦИАМ, Россия

*e-mail: akraiko@ciam.ru

Поступила в редакцию 27.06.2021 г.
Переработанный вариант 28.07.2021 г.
Принята к публикации 17.11.2021 г.

В свете современного состояния и тенденций развития методов математического моделирования звукового удара от сверхзвуковых летательных аппаратов обсуждается место численных и аналитических инструментов его описания. Отмечена возрастающая роль численных расчетов на адаптированных к особенностям течения сетках в рамках стационарных уравнений Эйлера (в координатах летательных аппаратов на “крейсерском” режиме полета) до расстояний от нескольких до пары десятков его длин. Другая важная тенденция — замена развивавшегося с середины XX в. сложного численно-аналитического аппарата описания “среднего” и “дальнего” полей звукового удара более простыми подходами, в том числе, без обращения к функции Уизема. В развитии этих тенденций в рамках уравнений Эйлера продемонстрирована возможность численного расчета типичных для звукового удара волновых структур без ограничений на расстояния и на интенсивности ударных волн, включая крайне малые. Описание эволюции звукового удара с удаления в 15–20 длин летательных аппаратов и до Земли сведено к мгновенному решению следующих из осесимметричных уравнений Эйлера задач Коши для обыкновенных дифференциальных уравнений. Вязкое размазывание слабых ударных волн описывает известное одномерное стационарное решение уравнений Навье–Стокса. Библ. 60. Фиг. 8.

Ключевые слова: численные и аналитические инструменты описания звукового удара, адаптация разностных сеток, приближение короткой волны, вязкое размазывание слабых ударных волн.

DOI: 10.31857/S0044466922040123

ВВЕДЕНИЕ

Достаточно полное представление о развитии и современном состоянии инструментов математического моделирования звукового удара, возникающего при полете сверхзвуковых летательных аппаратов (ЛА), и аналогичных нестационарных ударно-волновых структур дают статьи и монографии [1]–[45] и доклад [46]. Важную роль в этих интенсивно развивавшихся с середины XX в. инструментах играют приемы и формулы линейной теории и нелинейные поправки к ним.

Если возмущения течения и интенсивность ударных волн (УВ) превышают допускаемые линейной теорией, то необходимы расчеты ближнего и, возможно, среднего полей обтекания ЛА в рамках более полных уравнений. Часто при моделировании звукового удара обтекание ЛА описывают уравнениями течения идеального (невязкого и нетеплопроводного) газа — уравнениями Эйлера. В этом приближении на “крейсерском” режиме полета в системе координат ЛА течение стационарное и в основном сверхзвуковое. Его расчет возможен почти всюду — маршем по координате x , направленной по скорости V_0 невозмущенного (“набегающего”) потока. Программы численного решения уравнений Эйлера или RANS-приближения уравнений Навье–Стокса стали еще одним инструментом, правда, как правило, для расчета ближнего поля с размерами от нескольких до 8–10 длин ЛА (см. [23], [25], [28], [33], [38], [41], [44], [46]). Известное авторам исключение — статья [37] с расчетами до $r = 20$. Здесь и далее линейный масштаб L° — длина ЛА или заменяющего его в разд. 2 источника возмущений (величины с верхним “градусом” размерные),

¹⁾ Работа выполнена при финансовой поддержке РФФИ (код проекта 20-01-00100).

r — отнесенная к L° радиальная переменная цилиндрических координат $xr\phi$ с $\phi = 0$ на полуплоскости xu с направленной вниз (к Земле) осью u декартовых координат xuz с началом в передней точке ЛА. В сравнении с расчетами обтекания ЛА в RANS-приближении расчет по уравнениям Эйлера без сгущения сетки у обтекаемых без пограничных слоев поверхностей проще и быстрее.

Возникает вопрос, что препятствует широкому применению чисто численных инструментов описания звукового удара до $r \gg 1$. Существует мнение, что это невозможно из-за погрешностей счета малых возмущений параметров набегающего потока на больших расстояниях от ЛА. В действительности, однако, такое снижение точности присуще разностным сеткам, неадаптированным к особенностям рассчитываемых течений. В [47]–[49] это показали расчеты ударно-волновых структур перед сверхзвуковыми решетками. В связи со звуковым ударом достоинства адаптированных сеток и бесперспективность неадаптированных прямоугольных продемонстрированы в [35]. Еще раньше адаптированные сетки при расчете ближнего поля осесимметричных и простых пространственных тел применили авторы [23]. Кстати, сравнения выполненных в [35] расчетов с экспериментами из [12] по обтеканию тел вращения вопреки утверждениям [35] показали, что при определении ударно-волновой структуры звукового удара уравнения Эйлера не уступают уравнениям Навье–Стокса и их RANS-приближению. Это принципиально, особенно в свете данных [46] о несопоставимых временах RANS-расчетов ближних полей ЛА (при $r \leq 1/6$) с расчетами маршем на адаптированных сетках в рамках уравнений Эйлера их средних полей (при $1/6 \leq r \leq 10$). В [37] неструктурированные адаптированные сетки строились при расчете в рамках уравнений Эйлера сверхзвукового обтекания тел вращения (в том числе, из [12]) и пространственных моделей ЛА.

С учетом опыта расчета плоских (см. [47]–[49]), осесимметричных (см. [23], [35]) и пространственных (см. [23], [35], [37], [46]) ударно-волновых структур для правильного определения параметров течения в конически-кольцевой зоне звукового удара между “головной” и “замыкающей” УВ нужно явно или неявно (сгущением сетки, адаптированной к структуре звукового удара) выделять эти УВ. С удалением от ЛА ширина таких кольцевых зон растет много медленнее чем r .

В приводимых ниже примерах марш в направлении оси x по стационарному сверхзвуковому аналогу схемы Годунова (см. [50]), модернизированной согласно предложениям (см. [51]–[54]), с высокой точностью рассчитывает звуковой удар до $x \gg 1$. Кроме того, при $r \gg 1$ в силу известной (см., например, [25]) осесимметричной асимптотики изучаемые пространственные течения в фиксированных меридиональных плоскостях, отвечая обтеканию разных тел вращения, удовлетворяют осесимметричным уравнениям Эйлера. Благодаря этому с $r = 15$ – 20 их описание при $\phi = \text{const}$ и при однородной, и при стратифицированной атмосфере сведено к практически мгновенному решению задачи Коши для обыкновенных дифференциальных уравнений — следствий осесимметричных уравнений Эйлера для слабо возмущенных коротких волн. Разработанный подход аналогичен, но не тождественен подходу, представленному в [15], [16].

1. УРАВНЕНИЯ ОСЕСИММЕТРИЧНОГО СТАЦИОНАРНОГО СВЕРХЗВУКОВОГО ТЕЧЕНИЯ В СТРАТИФИЦИРОВАННОЙ АТМОСФЕРЕ. ПРИБЛИЖЕНИЕ “КОРОТКОЙ” ВОЛНЫ

При удалении от ЛА интенсивность возникших при его сверхзвуковом обтекании УВ, измеряемая отношением давлений p_+/p_- за и перед ними, почти монотонно приближается к единице. В слабых УВ приращениями энтропии и одного из инвариантов Римана — величинами порядка $(p_+/p_- - 1)^3$ (см. [55]–[58]) — будем пренебрегать. Это, тем не менее, не означает изэнтропичности рассматриваемых далее течений. В общем случае удельная энтропия невозмущенного потока s_0 переменна из-за неоднородности атмосферы по координате $y = r \cos \phi$. Масштаб изменения s_0 много больше длины ЛА.

В предполагаемой далее неподвижной в земной системе цилиндрических координат невозмущенной атмосфере градиент давления уравновешивает соответствующая проекция силы земного притяжения:

$$\frac{dp_0}{dy} = \rho_0 g \Leftrightarrow \frac{\partial p_0}{\partial r} = \rho_0 g \cos \phi. \quad (1.1)$$

Здесь и далее ρ — плотность газа, g — ускорение свободного падения, индекс “нуль”, как и выше, отмечает параметры набегающего потока, а при стратифицированной по $y = r \cos \phi$ атмосфере

$p_0 = p_0(y)$ и $\rho_0 = \rho_0(y)$. В используемых ниже координатах ЛА скорость набегающего потока $\mathbf{V}_0^\circ$ – константа. Взяв V_0° за масштаб скорости, получим $u_0 = V_0 = 1$, $v_0 = \theta_0 = 0$, $a_0 = 1/M_0$, где u и v – x - и r -компоненты вектора скорости $\mathbf{V}$, θ – угол его наклона к оси x , a – скорость звука и $M = V/a$ – число Маха.

Важный результат исследований звукового удара – упомянутая выше осесимметричная асимптотика. Согласно ей, при $r \gg 1$ изучаемое пространственное стационарное течение в фиксированных меридиональных плоскостях описывается осесимметричными уравнениями Эйлера ($h = h(p, s)$ – удельная энтальпия)

$$\begin{aligned} \frac{dp}{dt} + \rho a^2 \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial r} + \frac{v}{r} \right) &= 0, \quad \frac{du}{dt} + \frac{1}{\rho} \frac{\partial p}{\partial x} = 0, \quad \frac{dv}{dt} + \frac{1}{\rho} \frac{\partial p}{\partial r} = g \cos \phi, \\ \frac{ds}{dt} = 0 &\Leftrightarrow \frac{dh}{dt} = \frac{1}{\rho} \frac{dp}{dt}, \quad \frac{d}{dt} = u \frac{\partial}{\partial x} + v \frac{\partial}{\partial r} \end{aligned} \quad (1.2)$$

с нулевыми производными по ϕ и равной нулю ϕ -компонентой скорости.

Следствие второго, третьего и пятого уравнений (1.2) и равенства $dr/dt = v$ – уравнение

$$\frac{d}{dt} \left(h + \frac{V^2}{2} - rg \cos \phi \right) = 0.$$

Введением функции тока ψ оно и четвертое уравнение системы (1.2) интегрируются:

$$h + V^2/2 - rg \cos \phi = H(\psi), \quad s = s_0 = S(\psi), \quad (1.3)$$

с ψ , определяемым дифференциальным равенством (см. [55]–[58])

$$d\psi = krp(udr - vdx) = krpV(\cos \theta dr - \sin \theta dx),$$

в котором k – произвольный нормирующий множитель. Функции $H(\psi)$ и $S(\psi)$ определяются набегающим потоком: первая – всегда, а вторая – при слабых УВ.

Из четырех независимых дифференциальных уравнений (1.2) проинтегрировались два, для которых линии тока $dr/dx = \operatorname{tg} \theta$ – характеристики. Для получения характеристик, отличных от линий тока, в три первых уравнения (1.2) подставим $u = V \cos \theta$ и $v = V \sin \theta$. Производные от V войдут в эти уравнения как производная dV/dt , исключив которую, придем к уравнениям

$$\begin{aligned} (M^2 - 1) \left(\frac{\partial p}{\partial x} + \operatorname{tg} \theta \frac{\partial p}{\partial r} \right) - \rho V^2 \left(\operatorname{tg} \theta \frac{\partial \theta}{\partial x} - \frac{\partial \theta}{\partial r} - \frac{\operatorname{tg} \theta}{r} \right) + \rho g \cos \phi \operatorname{tg} \theta &= 0, \\ \operatorname{tg} \theta \frac{\partial p}{\partial x} - \frac{\partial p}{\partial r} - \rho V^2 \left(\frac{\partial \theta}{\partial x} + \operatorname{tg} \theta \frac{\partial \theta}{\partial r} \right) + \rho g \cos \phi &= 0, \end{aligned}$$

а после добавления к первому из них второго, умноженного на пока неопределенный множитель λ , – к уравнению

$$\begin{aligned} (1 - \lambda \operatorname{tg} \theta) \left(\frac{\partial \theta}{\partial r} + \frac{\operatorname{tg} \theta + \lambda}{\lambda \operatorname{tg} \theta - 1} \frac{\partial \theta}{\partial x} \right) + \frac{(M^2 - 1) \operatorname{tg} \theta - \lambda}{\rho V^2} \left[\frac{\partial p}{\partial r} + \frac{M^2 - 1 + \lambda \operatorname{tg} \theta}{(M^2 - 1) \operatorname{tg} \theta - \lambda} \frac{\partial p}{\partial x} \right] + \\ + \frac{\operatorname{tg} \theta}{r} + \frac{\operatorname{tg} \theta + \lambda}{V^2} g \cos \phi = 0. \end{aligned}$$

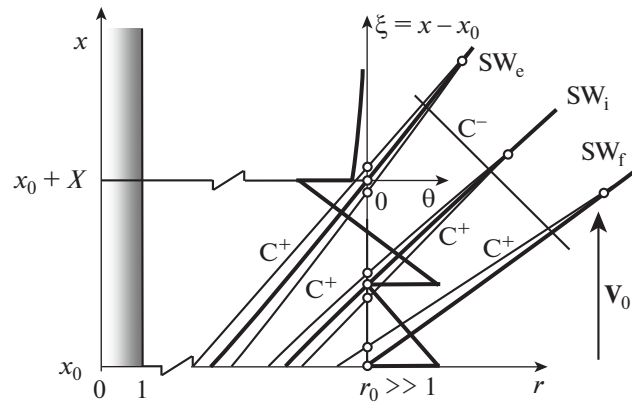
Это уравнение будет характеристическим при одинаковых дифференциальных операторах, действующих на θ и p , т.е. при одинаковых множителях при $\partial \theta / \partial x$ и $\partial p / \partial x$. Приравняв их, получим квадратное уравнение для λ , два его решения

$$\lambda = \pm \sqrt{M^2 - 1} = \pm \operatorname{ctg} \mu$$

и уравнения C^+ - и C^- -характеристик с “условиями совместности”

$$C^\pm : \frac{dx}{dr} = \operatorname{ctg}(\theta \pm \mu), \quad \frac{d\theta}{dr} \pm \frac{\operatorname{ctg} \mu}{\rho V^2} \frac{dp}{dr} + \frac{\sin \theta \sin \mu}{r \sin(\mu \pm \theta)} \mp \frac{\operatorname{ctg}(\mu \pm \theta)}{V^2} g \cos \phi = 0. \quad (1.4)$$

Здесь μ – угол Маха, определенный формулой $\operatorname{ctg} \mu = \sqrt{M^2 - 1}$ или $\sin \mu = 1/M$.



Фиг. 1. Схема фрагмента “короткой” волны звукового удара с тремя УВ и приходящими на них C^+ -характеристиками.

Для слабо возмущенных течений $p = p_0 + \delta p$, $\rho = \rho_0 + \delta \rho$, $V = 1 + \delta V$ и $\mu = \mu_0 + \delta \mu$ со вторыми слагаемыми, много меньшими первых. Кроме того, $|\theta| \ll 1$. В силу этого и равенства (1.1) уравнения (1.4) для таких течений примут вид

$$C^\pm: \frac{dx}{dr} = \pm \operatorname{ctg} \mu_0 \mp \frac{\theta \pm \delta \mu}{\sin^2 \mu_0}, \quad \frac{d\theta}{dr} \pm \frac{\operatorname{ctg} \mu_0}{\rho_0} \frac{d\delta p}{dr} + \frac{\theta}{r} + \left(\theta \mp \frac{\delta \rho}{2\rho_0} \sin 2\mu_0 \right) \frac{g \cos \phi}{\sin^2 \mu_0} = 0. \quad (1.5)$$

Для стратифицированной по $y = r \cos \phi$ атмосферы $\mu_0 = \mu_0(y)$.

Из-за малости θ смещение линий тока по r в примыкающем к слабой головной УВ слабо возмущенном течении так мало, что им можно пренебречь. Как следствие этого, можно пренебречь изменениями ψ в правых частях интегралов (1.3). Благодаря чему справедливы формулы

$$\delta \rho = \frac{\delta p}{a_0^2} = M_0^2 \delta p = \frac{\delta p}{\sin^2 \mu_0}, \quad \delta \mu = \frac{\gamma - 1 + 2 \sin^2 \mu_0}{\rho_0 \sin 2\mu_0} \delta p. \quad (1.6)$$

Вторая из них получена для совершенного газа с постоянными теплоемкостями и их отношением (показателем адиабаты) γ . Подстановка формулы для δp в условия совместности из (1.5) приводит их к виду

$$C^\pm: \frac{d\theta}{dr} \pm \frac{\operatorname{ctg} \mu_0}{\rho_0} \frac{d\delta p}{dr} + \frac{\theta}{r} + \left(\theta \mp \frac{\operatorname{ctg} \mu_0}{\rho_0} \delta p \right) \frac{g \cos \phi}{\sin^2 \mu_0} = 0. \quad (1.7)$$

С удалением от ЛА слабо возмущенное течение в кольцевом слое между головной и замыкающей УВ становится “короткой” волной, ибо приращения r при ее пересечениях C^- -характеристиками (Δr) малы по сравнению с r , т.е. $|\Delta r|/r \ll 1$. Сказанное поясняет фиг. 1. На ней нарисованы фрагмент короткой волны с тремя УВ (головной SW_f , замыкающей SW_e и внутренней SW_i), приходящими на них C^+ -характеристиками возмущенного течения, с отрезком C^- -характеристики и $X = O(1)$, а также частично залитый цилиндр $r = 1$, на котором в примерах разд. 2 ставится условие непротекания: $v = 0$. Из-за того, что ось x направлена вверх, а ось r — вправо, система координат xr левая с положительными углами θ при повороте по часовой стрелке.

Интегрируя от SW_f по отрезку C^- -характеристики условие совместности (1.7), пренебрежем изменениями на этом отрезке величин с индексом “нуль” и кубичным по $\delta p/\rho_0$ изменением на УВ отвечающего C^- -характеристикам инварианта Римана. Если $|\theta|^m$ — максимум $|\theta|$ в примыкающей к SW_f ударно-волновой структуре (короткой волне), то в ней с погрешностью, меньшей $|\theta|^m \Delta r/r \ll 1$, равно нулю возмущение “левого” инварианта Римана

$$\theta - \frac{\operatorname{ctg} \mu_0}{\rho_0} \delta p = 0. \quad (1.8)$$

Насколько погрешность выполнения полученного решения меньше, чем $|\theta|^m \Delta r/r$, зависит от знакопеременного распределения θ поперек короткой волны. Для распределения θ , заданного

при $r = r_0$ на фиг. 1, погрешность, обусловленная предпоследним слагаемым в (1.7), меньше примерно втрое. Вклад последнего слагаемого еще меньше. Как видно из (1.4), безразмерное отношение в (1.7) $g/\sin^2 \mu_0 = g^\circ L^\circ / (V_0^\circ \sin \mu_0)^2 = g^\circ L^\circ / a_0^{\circ 2}$. Здесь ускорение свободного падения g отнесено к $V_0^{\circ 2}/L^\circ$, а “градус”, как и ранее, отмечает размерные величины. Отсюда для $L^\circ = 30$ м и земных условий ($g^\circ = 9.8$ м/с², $a_0^\circ = 330$ м/с) отношение $g/\sin^2 \mu_0 \approx 0.003$.

Чтобы найти, как эволюционирует при $r > r_0 \gg 1$ осесимметричная (своя на каждой меридиональной плоскости) волна, фрагмент которой изображен на фиг. 1, подставим (1.8) в условие совместности для C^+ -характеристики из (1.7). Проинтегрировав полученное уравнение, найдем

$$\theta(\xi, r) = \Theta(\xi)\Omega(r, r_0), \quad \delta p(\xi, r) = \theta(\xi, r)\rho_0 \operatorname{tg} \mu_0, \quad \Omega(r, r_0) = \sqrt{\frac{r_0 \rho_0^0 \operatorname{ctg} \mu_0}{r \rho_0 \operatorname{ctg} \mu_0}}. \quad (1.9)$$

Здесь и далее верхний индекс “нуль” отмечает параметры в общем случае неоднородного (зависящего от $y = r \cos \phi$) невозмущенного набегающего потока при $r = r_0$, ξ — “характеристическая” переменная, постоянная на каждой C^+ -характеристике возмущенного течения, и $\Theta(\xi)$ — известное из расчета среднего поля обтекания ЛА (“начальное”) распределение θ :

$$r = r_0 : \xi = x(r_0, \xi) - x_0, \quad \theta(r_0, \xi) = \Theta(\xi), \quad (1.10)$$

где x_0 — начальная координата SW_F . На фиг. 1 $\Theta(\xi)$ — кусочно-непрерывная кривая.

Подстановка в уравнение C^+ -характеристик (1.5) $\delta \mu$ из (1.6), а затем θ и δp из (1.9) приводит к уравнению

$$\frac{dx}{dr} = \operatorname{ctg}(\theta + \mu) = \operatorname{ctg} \mu_0 - 2\Theta(\xi)\Phi(r, r_0), \quad \Phi(r, r_0) = \frac{\gamma + 1}{\sin^2 2\mu_0} \Omega(r, r_0), \quad (1.11)$$

справедливому на любой C^+ -характеристике от $r = r_0$ до ее прихода на одну из УВ. При рассмотрении внутренних УВ, с обеих сторон от которых, как в случае SW_i , поток возмущен, переменную ξ будем отмечать нижним индексом “минус” (“плюс”) для C^+ -характеристик, приходящих на УВ сверху (снизу) по течению. С учетом такого выбора, проинтегрировав уравнение (1.11) при постоянном ξ_- от сечения $r = r_0$ до УВ, придем к равенству

$$x_{SW} = x_{SW0} + \xi_- + \int_{r_0}^r \operatorname{ctg} \mu_0(\bar{r} \cos \phi) d\bar{r} - 2\Theta(\xi_-) \int_{r_0}^r \Phi(\bar{r}, r_0) d\bar{r},$$

в котором r и x_{SW} — координаты УВ. Продифференцировав его по r , получим

$$\frac{dx_{SW}}{dr} = \operatorname{ctg} \mu_0 + \left[1 - 2\Theta'(\xi_-) \int_{r_0}^r \Phi(\bar{r}, r_0) d\bar{r} \right] \frac{d\xi_-}{dr} - 2\Theta(\xi_-)\Phi(r, r_0). \quad (1.12)$$

Здесь и далее “штрих” означает дифференцирование по ξ_- или по ξ_+ .

Для слабых УВ, догоняющих C^+ -характеристики и догоняемых характеристиками того же семейства (см. [56], [60]),

$$\frac{dx_{SW}}{dr} = \frac{\operatorname{ctg}(\theta_- + \mu_-) + \operatorname{ctg}(\theta_+ + \mu_+)}{2}, \quad (1.13)$$

а в силу равенства (1.11)

$$\frac{dx_{SW}}{dr} = \operatorname{ctg} \mu_0 - [\Theta(\xi_-) + \Theta(\xi_+)] \Phi(r, r_0). \quad (1.14)$$

Разность (1.12) и (1.14) дает уравнение

$$\left[1 - 2\Theta'(\xi_-) \int_{r_0}^r \Phi(\bar{r}, r_0) d\bar{r} \right] \frac{d\xi_-}{dr} = [\Theta(\xi_-) - \Theta(\xi_+)] \Phi(r, r_0).$$

Введя переменную χ с начальным условием $\chi(r_0) = 0$, заменим его на два:

$$\frac{d\chi}{dr} = 2\Phi(r, r_0), \quad \frac{d\xi_-}{dr} = \frac{\Theta(\xi_-) - \Theta(\xi_+)}{1 - \chi\Theta'(\xi_-)} \Phi(r, r_0). \quad (1.15)$$

Аналогичным образом с той же функцией χ получим уравнение для ξ_+

$$\frac{d\xi_+}{dr} = \frac{\Theta(\xi_+) - \Theta(\xi_-)}{1 - \chi\Theta'(\xi_+)} \Phi(r, r_0). \quad (1.16)$$

Итак, каждая внутренняя УВ определяется решением задачи Коши для обыкновенных дифференциальных уравнений первого порядка: (1.15) и (1.16) – для ξ_- и ξ_+ , (1.14) – для x_{SW} и первого уравнения (1.15) – для χ (одного на все УВ) с начальными условиями $\xi_- = \xi_+ = x_{\text{SW}} = x_{\text{SW}0}$ и $\chi = 0$ при $r = r_0$. По найденным $\xi_-(r)$ и $\xi_+(r)$ величины $\theta_{\pm}$ и $\delta p_{\pm}$ с обеих сторон от УВ определяют формулы (1.9), а $\delta p_{\pm}$ и $\delta \mu_{\pm}$ – формулы (1.6). Для головной УВ, поток перед которой не возмущен, остаются три уравнения (1.14) и (1.16) с $\Theta(\xi_-) \equiv 0$. Если не возмущен поток за замыкающей УВ, то также остаются три уравнения, но с $\Theta(\xi_+) \equiv 0$.

При однородном набегающем потоке уравнение для χ интегрируется и дает

$$\chi = \frac{4(\gamma+1)r_0}{\sin^2 2\mu_0} (\sqrt{r/r_0} - 1).$$

Если при $r = r_0$ за головной УВ $\Theta(\xi_+) = \alpha - \beta\xi_+$ с положительными константами α и β , то, проинтегрировав уравнения (1.12) и (1.14) с $\Theta(\xi_-) = 0$, для SW_f придем к решениям и к их следствиям для $r/r_0 \gg 1$:

$$\frac{\delta p(r)}{\delta p(r_0)} = \frac{\theta(r)}{\theta(r_0)} = \frac{\sqrt{r_0/r}}{\sqrt{1+\beta\chi}} \underset{r/r_0 \gg 1}{\approx} \frac{\sin 2\mu_0}{2\sqrt{(\gamma+1)\beta}r_0} \left(\frac{r_0}{r}\right)^{3/4}, \quad (1.17)$$

$$x_{\text{SW}} - x_{\text{SW}0} - (r - r_0) \text{ctg } \mu_0 = -\frac{\alpha}{\beta} (\sqrt{1+\beta\chi} - 1) \underset{r/r_0 \gg 1}{\approx} -\frac{2\alpha\sqrt{(\gamma+1)r_0}}{\sqrt{\beta} \sin 2\mu_0} \left(\frac{r}{r_0}\right)^{1/4}.$$

Первым закон (1.17) затухания головной УВ установил О.С. Рыжов (см. [10]), а справедливую для $r/r_0 \gg 1$ степенную зависимость $\delta p \sim 1/r^{3/4}$ – еще Л.Д. Ландау (см. [1]).

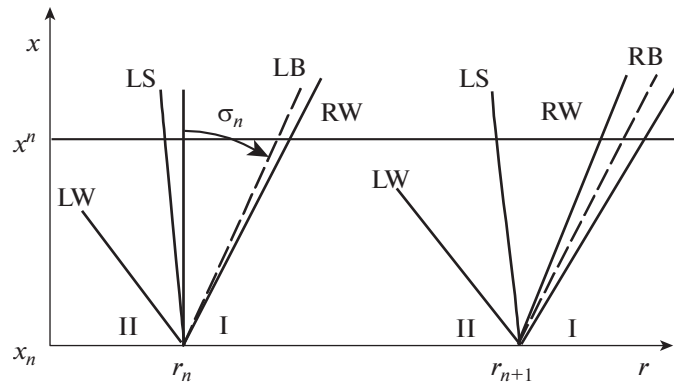
Обычны ситуации, когда одна УВ (“вторая” – SW_{II}) догоняет другую (“первую” – SW_{I}), например, головную. Так как обе УВ слабые, то при их слиянии интенсивность единственной результирующей УВ определяют величины θ_- перед SW_{I} и θ_+ за SW_{II} в момент слияния. При численном решении описанных выше задач Коши момент слияния (значение $r = r_{\text{I+II}}$, при котором оно происходит) определит равенство $x_{\text{SWII}} = x_{\text{SWI}}$.

Интегрирование уравнения (1.11) для одной-двух “промежуточных” характеристик позволяет построить распределения θ по x между УВ при $r = r_{\text{I+II}}$, т.е. новые $\Theta(\xi)$ для продолжения счета при $r > r_{\text{I+II}}$. Это может понадобиться для представления волн сжатия с их “стремлением опрокинуться” и лишь иногда для еще искривленных волн разрежения. Те же волны разрежения, которые были прямолинейными при $r = r_0$, останутся такими и далее. Для них промежуточных характеристик считать не надо. Описанный подход назовем “приближением короткой волны” (ПКВ). В связи с газодинамическими приложениями термин “короткая волна” введен С.А. Христиановичем и соавт. (см. [7]). Первые результаты по затуханию таких волн, не называя их “короткими”, получил Л.Д. Ландау (см. [1]).

Обыкновенные дифференциальные уравнения (1.14)–(1.16) ПКВ – простые следствия уравнений Эйлера (1.2). Этим ПКВ отличается от развитого в [15], [16] подхода, также сводящего расчет дальнего поля звукового удара к решению обыкновенных дифференциальных уравнений.

2. ПРИМЕРЫ РАСЧЕТА УДАРНО-ВОЛНОВЫХ СТРУКТУР В СТАЦИОНАРНЫХ СВЕРХЗВУКОВЫХ ТЕЧЕНИЯХ

Ниже для однородной атмосферы приводятся результаты расчета стационарных ударно-волновых структур в идеальном совершенном газе с $\gamma = 1.4$ и $M_0 = 2$. Для слабо возмущенных осесимметричных сверхзвуковых течений полученные результаты демонстрируют возможность акуратного расчета таких структур на любых расстояниях от источника возмущений со сколь угодно малыми возмущениями параметров. Расчет таких структур велся по явной маршевой в направлении оси x схеме с построением по уравнению (1.13) головной и замыкающей УВ-адаптированных к рассчитываемому течению сеточных линий и в рамках ПКВ, развитого в разд. 1 для $r \gg 1$.



Фиг. 2. Схема взаимодействия сверхзвуковых потоков для определения параметров на левой и правой границах (LB и RB) ячейки (LW – левая волна, RW – правая волна, LS – линия тока).

При маршевом счете на слоях $x = \text{const}$ число $K = 2^{k-1}$ разностных ячеек между двумя УВ было фиксировано, а между замыкающей УВ и цилиндром $r = r_0$ в начале счета увеличивалось от 1 до K , а затем сохранялось неизменным. В итоге общее число ячеек на слоях $x = \text{const}$ становилось равным 2^k при равномерном разбиении по r в каждой из двух областей. Решение на границах всех ячеек задачи взаимодействия равномерных сверхзвуковых потоков (фиг. 2) – аналога нестационарной задачи о распаде разрыва здесь упрощается из-за малого отличия параметров взаимодействующих потоков. Подобно нестационарной задаче [59], благодаря этому, условия, справедливые для слабых УВ и центрированных волн разрежения, записываются единообразно через разности p и θ . В приближении слабых УВ на раздельных линиях тока непрерывны все параметры. В точной постановке эти линии тока – тангенциальные разрывы, на которых непрерывны только p и θ .

Согласно условиям совместности из (1.4), для совершенного газа на LW и RW – “левой” и “правой” слабых ударных или центрированных волнах – с равной погрешностью (кубичной по приращению давления) выполняются равенства

$$\begin{aligned} \text{LW: } \quad \Theta &= \theta_{\text{II}} + \left[\frac{G(M_{\text{II}})}{p_{\text{II}}} + \frac{G(M)}{P} \right] \frac{p_{\text{II}} - P}{2\gamma}, \\ \text{RW: } \quad \Theta &= \theta_{\text{I}} - \left[\frac{G(M_{\text{I}})}{p_{\text{I}}} + \frac{G(M)}{P} \right] \frac{p_{\text{I}} - P}{2\gamma}, \quad G(M) = \frac{\sqrt{M^2 - 1}}{M^2}. \end{aligned} \quad (2.1)$$

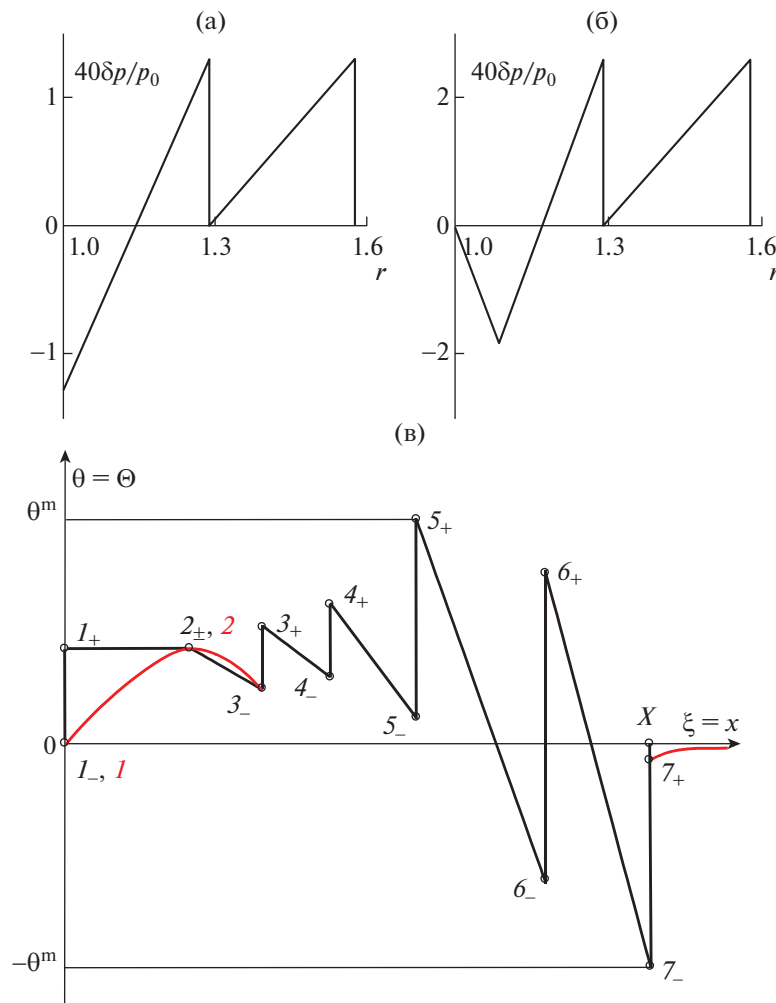
Здесь и ниже индексы “I” и “II” отмечают параметры взаимодействующих потоков, полученные интерполяцией в согласии с [51] на границах, отличных от SW_f и SW_e , экстраполяцией для SW_e и за SW_f и с заменой индекса “I” на “0” перед SW_f , а “большие величины” без индексов P , Θ , R , A и M – значения p , θ , ρ , a и числа Маха между LW и RW. Исключив Θ из уравнений (2.1), придем к квадратному уравнению для P и к его корню

$$\begin{aligned} P &= \left(C + \sqrt{C^2 + BF} \right) / B, \\ B &= \frac{G(M_{\text{I}})}{p_{\text{I}}} + \frac{G(M_{\text{II}})}{p_{\text{II}}}, \quad C = \frac{G(M_{\text{I}}) + G(M_{\text{II}})}{2} - G(M) + \gamma(\theta_{\text{II}} - \theta_{\text{I}}), \quad F = (p_{\text{I}} + p_{\text{II}})G(M). \end{aligned} \quad (2.2)$$

Зависимость от P числа Маха M , входящего в C и F , определяют формулы

$$R = \frac{P^{1/\gamma}}{p_0^{1/\gamma}}, \quad A^2 = \frac{\gamma P}{R}, \quad M^2 = \frac{I_0 - 2A^2}{(\gamma - 1)A^2}, \quad I_0 = \frac{2 + (\gamma - 1)M_0^2}{M_0^2}.$$

Уравнение (2.2) решалось итерациями с $P^{(0)} = p_{\text{II}}$, после сходимости которых R , A и M известны, $V = AM$, а Θ определяет любое из уравнений (2.1). При решении задач взаимодействия для всех ячеек сначала на полуслое $x = x_n + \Delta x/2$ находят координаты r выделяемых УВ. По ним при равномерном разбиении по r отрезков между SW_f и SW_e и между SW_e и цилиндром $r = 1$ находят наклоны всех границ и их координаты на полуслое. По наклонам границ устанавливается, куда

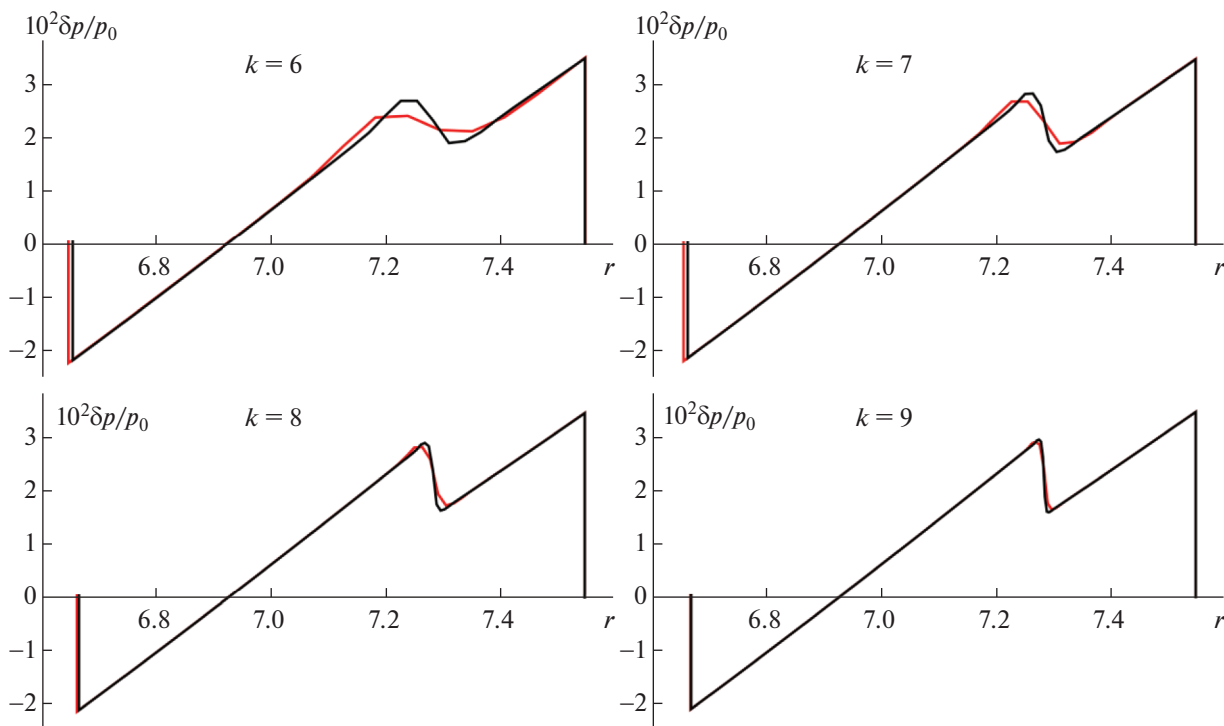


Фиг. 3. Начальные распределения $\delta p/p_0$ и θ (см. текст).

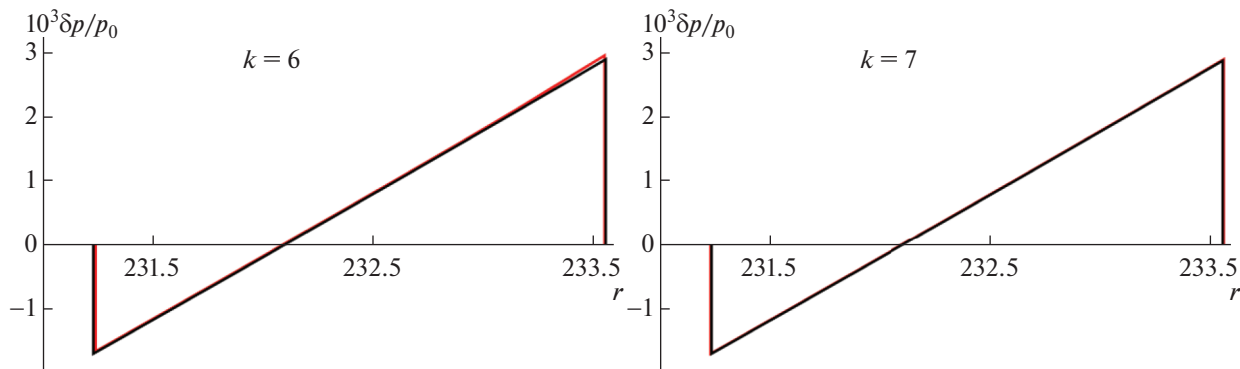
в задачах взаимодействия они попадают, и определяются параметры течения на этих границах (“большие величины” в законах сохранения, записанных для каждой двумерной ячейки между слоями $x = x_n$ и $x = x_n + \Delta x/2$) и делается полушаг, в результате которого находятся параметры во всех ячейках полуслоя. По найденным параметрам на полуслое определяются параметры с индексами “I” и “II”, решаются новые задачи взаимодействия и делается полный шаг до $x^n = x_n + \Delta x$. Не вдаваясь в дальнейшие детали, отметим, что, согласно расчетам, максимальное число Куранта, с которым созданный алгоритм реализует устойчивый счет, равно $0.8 \operatorname{ctg} \mu_0$.

Начальное возмущение (фиг. 3а) однородного потока с $M_0 = 2$, $V_0 = \rho_0 = 1$, $a_0 = 1/M_0 = 0.5$, $p_0 = 1/(\gamma M_0^2) \approx 0.179$ задавалось в кольце $x = x_0 = 1$, $1 \leq r \leq r_0$, что близко к фиг. 1, но с $r_0 = 1 + \operatorname{tg} \mu_0 \approx 1.58$. Максимальные $\theta^m \approx 0.01$ у передней границы кольца (SW_f) и на его среднем радиусе (SW_i), минимальное $\theta = -\theta^m$ на замыкающей границе (SW_e) и линейные по r между ними. Начальное возмущение p принималось равным $\delta p = \theta \operatorname{tg} \mu_0 \approx 0.58\theta$, начальная плотность определялась по p условием $p/\rho^\gamma = p_0$, а модуль скорости газа – условием постоянства полной энтальпии. При $\theta^m \approx 0.01$ максимальные $|\delta p|^m/p_0 = 0.032$. Это настолько мало, что энтропия и “левый” инвариант Римана в УВ практически не изменяются. При $x \geq 1$ на цилиндре $r = 1$ выполнялось условие непротекания.

На фиг. 4 приведены распределения $\delta p/p_0$, в которые начальное возмущение превратилось при $x = 10$. Сравниваются результаты расчетов при разном числе ячеек 2^k в радиальном направлении с выделением SW_f и SW_e – двух границ слегка расширяющейся узкой конически-кольцевой волны между ними. На каждом фрагменте фиг. 4 черные (красные) кривые – результаты рас-



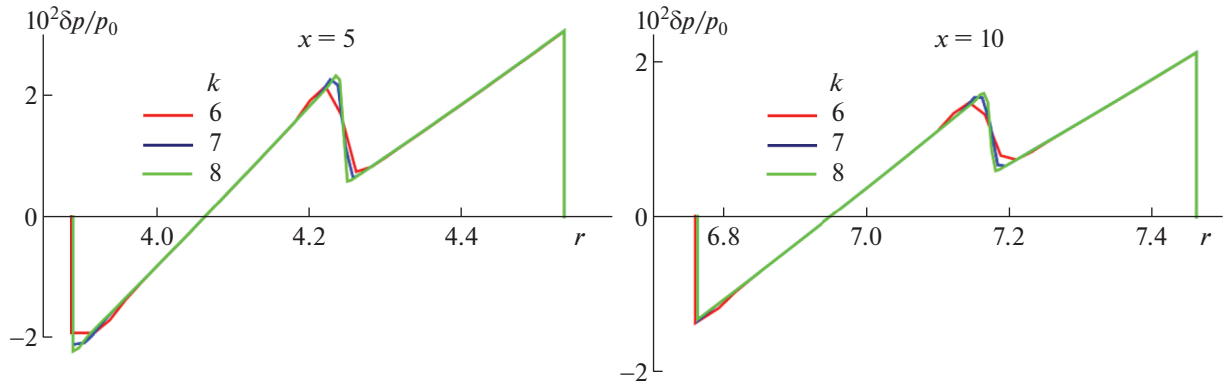
Фиг. 4. Кривые $\delta p/p_0 = f(r)$ при $x = 10$, рассчитанные маршем на сетках с 2^k (черные кривые) и с 2^{k-1} (красные кривые) ячейками в слоях $x = \text{const}$.



Фиг. 5. N -волны при счете маршем сверхзвукового осесимметричного течения до $x = 400$ с 2^k (черные кривые) и 2^{k-1} (красные кривые) ячейками в слоях $x = \text{const}$.

четов с 2^k (2^{k-1}) ячейками. На фиг. 5 таким же способом сравниваются N -волны, в которые при разном числе ячеек это возмущение превратилось при $x = 400$. Расчет уже на 32 ячейках в столь дальнем поле дает близкую к точной N -волну за 50 с работы одноядерного ПК. На N -волну здесь приходится всего 16 ячеек, а их размер по r при $x = 400$ в широкой приосевой зоне больше, чем между SW_f и SW_e в 100 раз.

Обычно замыкающая УВ возникает не сразу, а на некотором удалении от ЛА. При счете маршем ее заменит УВ нулевой интенсивности, т.е. C^+ -характеристика, которая с ростом x станет замыкающей УВ. Как это происходит, показывает пример с начальными параметрами без замыкающей УВ (фиг. 3б). Главные их отличия от рассмотренных — при $1 \leq r < 1 + 1/(2\sqrt{3})$ минимум



Фиг. 6. Кривые $\delta p/p_0 = f(r)$ для $x = 5$ и 10 , рассчитанные на трех сетках при начальных параметрах без замыкающей УВ.

$\theta = -\theta^m/\sqrt{2}$ в вершине $r = 1 + (\sqrt{2} - 1)/(2\sqrt{6})$ равнобедренного треугольника, равного по площади треугольнику, примыкающему к внутренней УВ. Кроме того, в этом примере $\theta^m \approx 0.005$, т.е. вдвое меньше, чем в предыдущем примере.

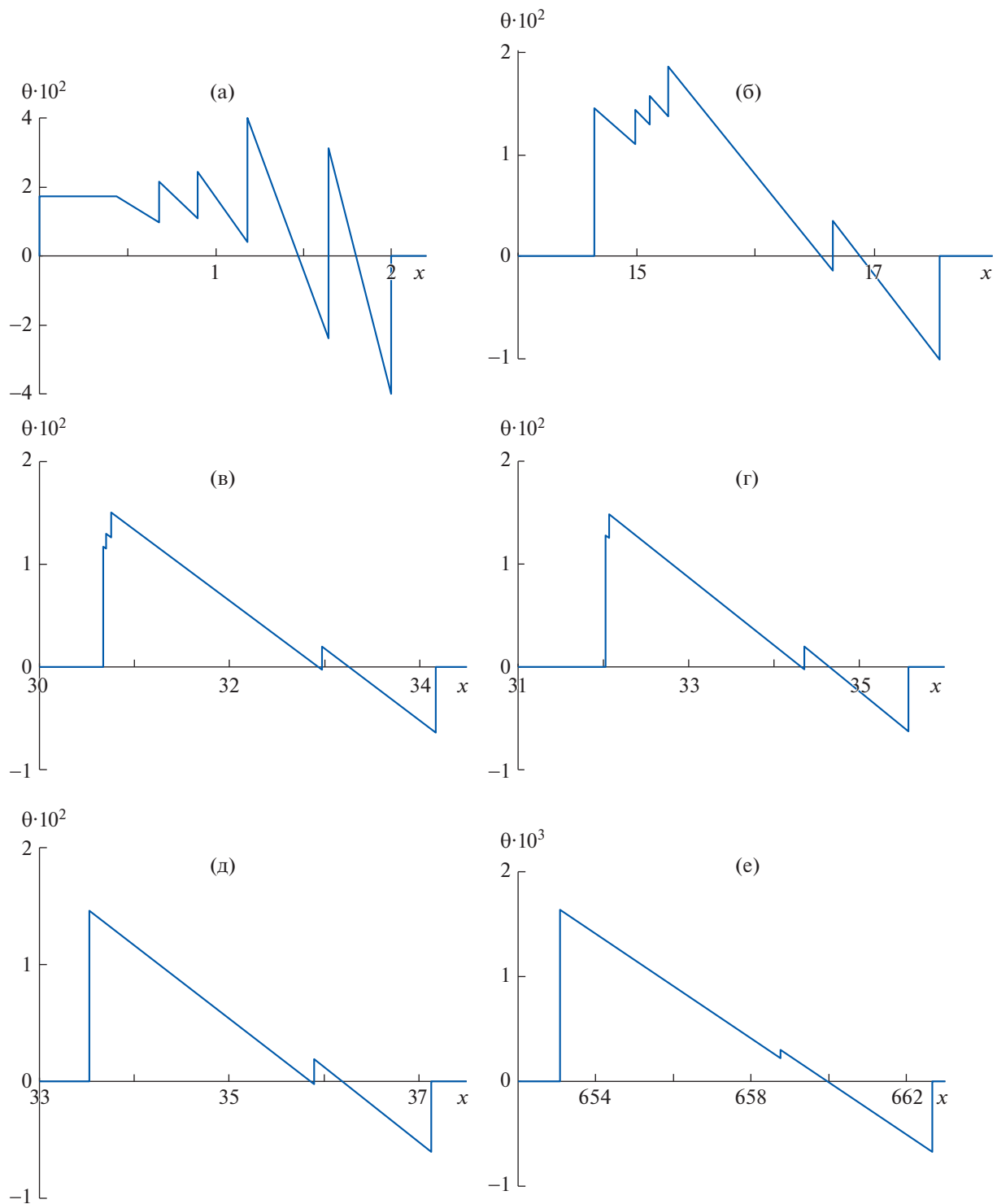
Описание формирования замыкающей УВ требует большего числа ячеек, чем расчет, когда такая УВ есть изначально. Получившиеся в данном случае профили $\delta p/p_0 = f(r)$ для $x = 5$ и 10 при счете на сетках с $k = 6, 7$ и 8 представлены на фиг. 6. При $x = 10$ замыкающая УВ полностью сформировалась, и продолжать счет на столь тонких сетках нет необходимости. Для дальнейшего счета достаточно сетки с $k = 5$.

Последние примеры относятся к описанию звукового удара с удалений в несколько десятков длин ЛА, когда применимо ПКВ. Для смещенных по оси r на $\Delta r \approx 19$ начальных распределений $\delta p/p_0$ (на фиг. 3а это отвечает $r_0 \approx 20.6$) сравнивались результаты ПКВ, счета установлением и маршем. В данном примере при $M_0 = 2$ в силу равенства (1.8) $\delta p/p_0 \approx 3.2\theta$. Сравнение параметров в сечении слияния SW_f и SW_i ($r \approx 103$, $x \approx 141$) и N -волн при $x = 200, 300$ и 400 показало, что для не столь большой величины $r_0 \approx 21$ погрешности ПКВ в определении θ и δp не превышают 1%, а в определении координат N -волн — 0.005%.

После такого подтверждения точности ПКВ покажем его возможности на начальных распределениях, более сложных, чем на фиг. 3а и 3б. Два таких распределения $\theta = \theta(x, r_0) = \Theta(\xi)$ с константами $0 < \theta^m \ll 1$ и $X > 1$ приведены на фиг. 3в. Первое образуют отрезки прямых, из которых вертикальные — УВ, а наклонные — волны разрежения. Горизонтальный участок I_+-2_- введен из-за интереса к таким участкам, например, в [26], [29], [32], [34], [38], [43]. Во второе распределение введена красная волна сжатия при $\xi > X$, а начальная ломаная $I_- - I_+ - 2_+ - 3_-$ заменена гладкой красной волной сжатия-разрежения $I-2-3_-$ без головной УВ (с “головной C^+ -характеристикой” C_1^+). Нечто похожее на такую волну есть в [41], [42].

Первое начальное распределение не содержит непрерывных волн сжатия, а волны разрежения и горизонтальный участок I_+-2_- — отрезки прямых. С ростом r они остаются прямыми. К значениям $\xi_{\pm}$, $\theta_{\pm}$ и координат x всех УВ как функций r , которые определяются уравнениями (1.14)–(1.16), в данном случае добавляется $x_{2\pm}$ — координата C^+ -характеристики C_2^+ , вышедшей из точки $2_{\pm}$. Координаты x любых чем-то интересных C^+ -характеристик можно определять либо интегрированием уравнения (1.11) с постоянным $\xi = \xi_{2\pm}$ (в случае точки $2_{\pm}$), либо для упрощения логики программы — как УВ нулевой интенсивности.

Результаты расчетов для первого распределения $\theta(x, r_0) = \Theta(\xi)$ (фиг. 3в) с $r_0 = 20$, $X = 2$ и $\theta^m = 0.04$ приведены на фиг. 7. Наряду с начальным профили $\theta = \theta(x)$ даны для r , при котором C_2^+ догнала SW_1 , для тех r , при которых одна из УВ догнала другую (по координате x — “предыдущую”) и слилась с ней, и для $r = 400$. За УВ, получившейся при слиянии, сохранялся номер предыдущей. В рассчитанном примере после четырех слияний остались три УВ: SW_1 , SW_6 и SW_7 . Интенсивность SW_6 настолько мала, что она SW_1 уже не догонит, т.е. сохраняется слегка отличная от N -волны структура с третьей очень слабой УВ (фиг. 7е).



Фиг. 7. Профили $\theta = \theta(x)$: начальный $r = r_0 = 20$ (а), при слияниях SW_1 с C_2^+ $r = 28.71$ (б), с SW_3 $r = 38.17$ (в), с SW_4 $r = 38.96$ (г) и с SW_5 $r = 39.85$ (д) и при $r = 400$ (е).

В примере с волной сжатия-разрежения $I-2-3_-$, головной SW_1 нулевой интенсивности (с “головной C^+ -характеристикой” C_1^+) и волной сжатия за SW_7 уточним некоторые моменты. При описании эволюции волны $I-2-3_-$ находятся координата x_1 ее начальной точки, координата x_2 и

значение θ_2 максимума и те же величины в ее концевой точке (x_3 и θ_{3-}). Начальная точка и точка максимума принадлежат фиксированным C^+ -характеристикам, переменные ξ которых изменяются только при пересчетах, вызванных слияниями УВ или образованием “висячей” УВ на C_1^+ . Значение r , при котором это случится, зависит от величины $\Theta'(0)$. Для определения такого r и описания движения SW_3 разобьем волну $I-2-3_-$ на два участка и, согласно фиг. 3в, начальные $\Theta(\xi)$ на них зададим многочленами

$$\begin{aligned} 0 \leq \xi \leq \xi_2 = 0.22X, \quad \Theta(\xi) &= \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3, \quad \Theta'(\xi) = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2; \\ \xi_2 \leq \xi \leq \xi_3 = 0.34X, \quad \Theta(\xi) &= \beta_4 + \beta_5 (\xi - \xi_2)^2, \quad \Theta'(\xi) = 2\beta_5 (\xi - \xi_2), \\ \beta_1 &= 3.3 \frac{\theta^m}{X}, \quad \beta_2 = -3.6 \frac{\theta^m}{X^2}, \quad \beta_3 = -12 \frac{\theta^m}{X^3}, \quad \beta_4 = \beta_1 \xi_2 + \beta_2 \xi_2^2 + \beta_3 \xi_2^3, \\ \beta_5 &= \frac{\beta_1 + \beta_2 (\xi_3 + \xi_2) + \beta_3 (\xi_3^2 + \xi_3 \xi_2 + \xi_2^2)}{\xi_3 - \xi_2}. \end{aligned} \quad (2.3)$$

Поскольку $\Theta(\xi_+ = 0) = 0$, то $\xi_+ = 0$ в силу уравнения (1.16) с $\Theta(\xi_-) \equiv 0$, пока не обратится в нуль знаменатель правой части этого уравнения

$$1 - \Theta'(0)\chi(r_*, r_0) = 1 - \beta_1 \chi(r_*, r_0) = 0. \quad (2.4)$$

При $x = x(\xi_+, r)$ на C_1^+ найдем, как на ней с ростом r изменяется производная $\partial x / \partial \xi_+$. Проинтегрировав уравнение (1.11) по ξ при постоянном r и преобразовав результат с учетом первого уравнения (1.15), придем к задаче

$$\frac{d(\partial x / \partial \xi_+)}{d\chi} = -\Theta'(0), \quad r = r_0, \quad \chi = 0, \quad (\partial x / \partial \xi_+)_{r=r_0} = 1$$

с начальным условием для $\partial x / \partial \xi_+$, вытекающим из определения ξ в (1.10), и с производной по χ при $\xi_+ = \text{const} = 0$. Решение этой задачи – равенство

$$\partial x / \partial \xi_+ = 1 - \Theta'(0)\chi(r, r_0), \quad (2.5)$$

в котором $\Theta'(0) > 0$, а χ – монотонно растущая функция r . Поэтому при $r = r_*$ из (2.4) $\partial x / \partial \xi_+$ становится нулем на C_1^+ (если раньше ее не догнала SW_3), свидетельствуя о зарождении SW_1 . Наконец, из (1.9) и (2.5) следует, что на C_1^+

$$\frac{\partial \Theta(\xi_+, r)}{\partial x} = \frac{\Theta'(0)\Omega(r, r_0)}{1 - \Theta'(0)\chi(r, r_0)}.$$

Таким образом, при том же $r = r_*$ эта производная обращается в бесконечность.

При $r = r_*$ пересчитаем ξ и $\Theta(\xi)$: на отрезках $I-2$ и $2-3_-$ по известным координатам x_1, x_2 и x_3 и значениям θ_2 и θ_{3-} найдем $\xi_{1N} = x_1$, $\xi_{2N} = x_2$ и $\xi_{3N} = x_3$, а “старые” формулы (2.3) заменим на “новые”:

$$\begin{aligned} \xi_{1N} \leq \xi \leq \xi_{2N}: \quad \Theta(\xi) &= \beta_{1N} \sqrt{\xi - \xi_{1N}} + \beta_{2N} (\xi - \xi_{1N}), \quad \Theta'(\xi) = \frac{\beta_{1N}}{2\sqrt{\xi - \xi_{1N}}} + \beta_{2N}; \\ \xi_{2N} \leq \xi \leq \xi_{3N}: \quad \Theta(\xi) &= \theta_2 + \beta_{3N} (\xi - \xi_{2N})^2, \quad \Theta'(\xi) = 2\beta_{3N} (\xi - \xi_{2N}), \\ \beta_{1N} &= \frac{2\theta_2}{\sqrt{\xi_{2N} - \xi_{1N}}}, \quad \beta_{2N} = \frac{-\theta_2}{\xi_{2N} - \xi_{1N}}, \quad \beta_{3N} = \frac{\theta_{3-} - \theta_2}{(\xi_{3N} - \xi_{2N})^2}. \end{aligned}$$

Если же точки 2 на $I-3_-$ уже нет (SW_3 догнала C_2^+), то формулы (2.3) заменим на

$$\xi_{1N} \leq \xi \leq \xi_{3N}: \quad \Theta(\xi) = \beta_{1N} \sqrt{\xi - \xi_{1N}}, \quad \Theta'(\xi) = \frac{\beta_{1N}}{2\sqrt{\xi - \xi_{1N}}}, \quad \beta_{1N} = \frac{\theta_{3-}}{\sqrt{\xi_{3N} - \xi_{1N}}}.$$

Кроме того, здесь “старое” r_0 заменяется на r_* , а при слияниях УВ – на r_{I+II} , и интегрирование уравнений (1.15) возобновляется с начальным условием $\chi(r_0, r_0) = 0$.

При $r = r_*$ в правой части уравнения (1.16):

$$\frac{d\xi_{\pm}}{dr} = \frac{\Theta(\xi_{\pm})\Phi(r, r_0)}{1 - \chi\Theta'(\xi_{\pm})}, \quad \Theta(\xi_{\pm} \rightarrow \xi_{1N}) \rightarrow 0, \quad \chi(r \rightarrow r_0, r_0) \rightarrow 0, \quad \Theta'(\xi_{\pm} \rightarrow \xi_{1N}) \rightarrow \infty,$$

возникли неопределенности. Для их разрешения воспользуемся его следствиями:

$$\frac{d\xi_{\pm}}{d\chi} = \frac{\Theta(\xi_{\pm})}{2 - 2\Theta'(\xi_{\pm})\chi} \Leftrightarrow \frac{d\chi}{d\xi_{\pm}} + 2\frac{\Theta'(\xi_{\pm})}{\Theta(\xi_{\pm})}\chi - \frac{2}{\Theta(\xi_{\pm})} = 0.$$

Решив это линейное уравнение, найдем

$$\chi = \frac{2}{\Theta^2(\xi_{\pm})} \int_{\xi_{1N}}^{\xi_{\pm}} \Theta(\xi_{\pm}) d\xi_{\pm} \approx \frac{2}{\beta_{1N}(\xi_{\pm} - \xi_{1N})} \int_{\xi_{1N}}^{\xi_{\pm}} \sqrt{\xi_{\pm} - \xi_{1N}} d\xi_{\pm} = \frac{4\sqrt{\xi_{\pm} - \xi_{1N}}}{3\beta_{1N}}.$$

Таким образом, для r , близких к новому $r_0 = r_*$, имеем

$$\xi_{\pm} - \xi_{1N} \approx \frac{9}{16} \beta_{1N}^2 \chi^2.$$

С той же точностью $\chi \approx 2\Phi(r_0, r_0)(r - r_0)$, и с учетом формулы для $\Phi(r_0, r_0)$ получим

$$r - r_0 \ll 1: \quad \xi_{\pm} - \xi_{1N} \approx \frac{9}{4} \beta_{1N}^2 \Phi^2(r_0, r_0)(r - r_0)^2, \quad \frac{d\xi_{\pm}}{dr} \approx 9 \frac{\beta_{1N}^2 (\gamma + 1)^2}{2 \sin^4 2\mu_0} (r - r_0).$$

Результаты расчетов для второго распределения $\theta(x, r_0) = \Theta(\xi)$ (фиг. 3в) с $r_0 = 20$, $X = 2$ и $\theta^m = 0.01$ приведены на фиг. 8. Наряду с начальным, профили $\theta = \theta(x)$ даны для значения $r = r_*$ образования висячей SW_1 на C_1^+ для тех r , при которых SW_3 , SW_4 и SW_5 одна за другой догоняли SW_1 , сливаясь с ней, и для $r = 400$. Из-за вчетверо меньшего θ^m слияния УВ происходили при заметно больших r , чем в примере фиг. 7. При интегрировании обыкновенных дифференциальных уравнений методом Рунге–Кутты до $r = 10^3$ времена счета в этих примерах – доли секунды на ПК.

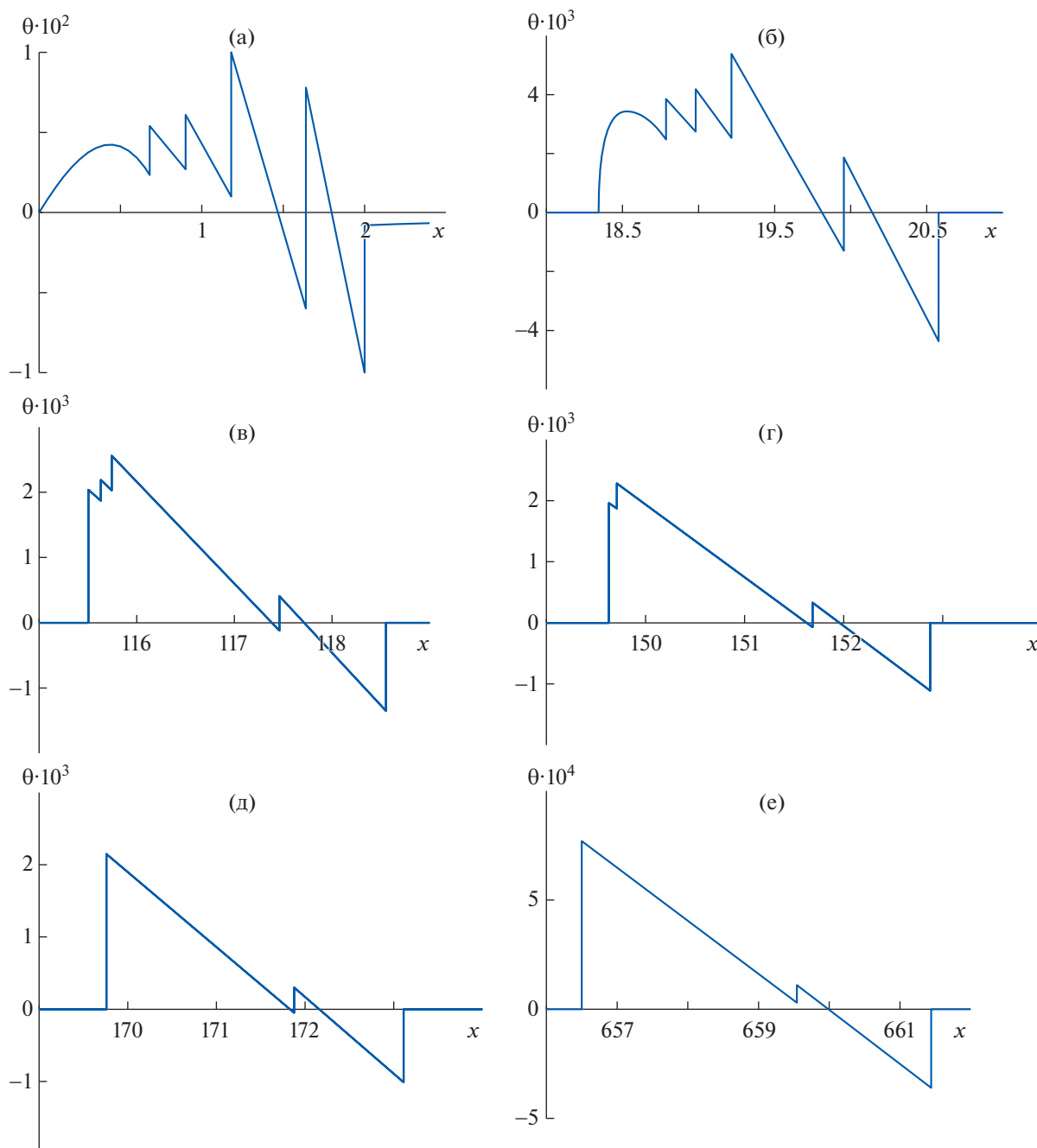
УВ, рассматриваемые как поверхности разрыва, на самом деле имеют конечную ширину Δ° (“градус” по-прежнему отмечает размерные величины). Для УВ умеренной и большой интенсивности она пренебрежимо мала (равна нескольким длинам свободного пробега). Однако размазывание слабых УВ с Δ° , обратно пропорциональным разности $p_+^\circ - p_-^\circ = \delta p_+^\circ - \delta p_-^\circ$, может изменить профиль давления в звуковом ударе и, как следствие, – высокочастотную часть его спектра (см. [60]). Решение из [55], определяющее в рамках одномерных стационарных уравнений Навье–Стокса профиль δp размазанной УВ и ее ширину Δ , можно записать так:

$$\delta p = \frac{\delta p_+ + \delta p_-}{2} + \frac{\delta p_+ - \delta p_-}{2} \tanh \frac{n}{\Delta}, \quad \Delta = \frac{\Delta^\circ}{L} = \frac{4\gamma[4Pr + 3(\gamma - 1)]M_0}{3(\gamma + 1)(\delta p_+ - \delta p_-)PrRe}, \quad (2.6)$$

$$Re = \frac{L^\circ V_0^\circ}{\nu^\circ}, \quad Pr = \frac{\nu^\circ}{\kappa^\circ} \rho_0^\circ c_p^\circ, \quad \nu^\circ = \frac{4\eta^\circ + 3\zeta^\circ}{4\rho_0^\circ}.$$

Здесь координата n растет в направлении потока от “середины” УВ по нормали к ней, δp отнесены к p_0° , η° и ζ° – коэффициенты “первой” и “второй” вязкости, κ° – теплопроводность воздуха, а Re и Pr – числа Рейнольдса и Прандтля.

При полете с $M_0^0 = 2$ на высоте 20 км в рамках “Стандартной атмосферы” у Земли $M_0 \approx 1.73$, и формула для Δ при $Pr = 0.8$ и длине l ЛА в метрах принимает вид $\Delta \approx 1.7/[l(\delta p_+ - \delta p_-)10^6]$. Отсюда для $l = 30$ ширина Δ крайне слабой УВ с приращением давления всего в 1 Па не превысит 6×10^{-3} , что много меньше равной нескольким единицам ширины дошедшей до Земли волны звукового удара. Еще меньшее размазывание УВ при распространении от ЛА до Земли практически не может влиять на их эволюцию. При расчете громкости звукового удара эффекты вязкого размазывания слабых УВ правильно и легко учесть с помощью решения (2.6). Воспользовавшись уравнением Ландау–Теллера, еще проще учесть размазывание слабых УВ из-за неравновесности колебательных степеней свободы молекул воздуха. Это на порядки быстрее и много проще, чем численное решение уравнения Бюргерса в том же, что и ПКВ, осесимметричном приближении (см. [39]).



Фиг. 8. Профили $\theta = \theta(x)$: начальный $r = r_0 = 20$ (а), при образовании SW_1 $r = 30.59$ (б), при слияниях SW_1 с SW_3 $r = 86.92$ (в), с SW_4 $r = 106.69$ (г), с SW_5 $r = 118.36$ (д) и при $r = 400$ (е).

ЗАКЛЮЧЕНИЕ

До настоящего времени инструменты предсказания звукового удара вдали от предполагаемых высоколетящих сверхзвуковых гражданских ЛА во многом опираются на весьма сложные методы и решения линейной и нелинейной акустики движущихся неоднородных сред. Отмеченная сложность проявляется уже в процессе получения применяемых затем уравнений и решений.

Упрощение аналитической части данного исследования обязано интеграции численного и аналитического подходов. “Аналитика” подсказывает правильную адаптацию разностных сеток, а “вычислительный инструмент” ограничивает применение аналитики расстояниями, начиная

с которых исследуемое течение — слабая короткая волна при том, что с удалением от ЛА в каждой меридиональной плоскости она близка к осесимметричной. Учет этих особенностей позволил начинать анализ не с пространственных, а с более простых осесимметричных уравнений и затем работать с уравнениями их характеристик и решениями последних для слабых коротких волн. Существенно, что при таком подходе эффекты осевой симметрии, нелинейности и стратификации невозмущенной атмосферы учитываются не последовательно, а одновременно. Еще одно упрощение связано со скоростью слабых УВ как полусуммы скоростей приходящих к ним характеристик. Вместо этого большинство исследователей, начиная с Л.Д. Ландау (см. [1]), пользуются более сложным “правилом площадей” (исключение — К.Е. Губкин, см. [5]).

В свете выполненного исследования и в развитие полученных ранее результатов можно надеяться на реальность многократного сокращения времен расчета в приближении пространственных уравнений Эйлера ближнего и среднего полей стационарного (на крейсерском режиме полета) обтекания ЛА на адаптированных к их особенностям разностных сетках. Если указанные времена станут соизмеримы с временами счета дальнего поля существующими программными комплексами (типа “sBOOM”), замена их развитым выше также в рамках уравнений Эйлера ПКВ будет обусловлена не только его простотой, но и на порядки меньшими временами счета. Привлечение уравнений Навье—Стокса в их одномерном стационарном варианте при этом свелось к известному локальному решению, описывающему размазывание слабых УВ без влияния на их распространение от ЛА до Земли.

Авторы признательны К.С. Пьянкову и В.С. Горбовскому за полезные обсуждения.

СПИСОК ЛИТЕРАТУРЫ

1. Ландау Л.Д. Об ударных волнах на далеких расстояниях от места их возникновения // Прикл. матем. и механ. 1945. Т. 9. Вып. 4. С. 286—292.
2. Whitham G.B. The Propagation of Spherical Blast // Proceed. of the Royal Soc. Ser. A. 1950. V. 203. № 1075. P. 571—581.
3. Whitham G.B. The Flow Pattern of a Supersonic Projectile // Comm. Pure Appl. Mathem. 1952. V. 5. № 3. P. 301—348.
4. Христианович С.А. Ударная волна на значительном расстоянии от места взрыва // Прикл. матем. и механ. 1956. Т. 20. Вып. 5. С. 599—605.
5. Губкин К.Е. Распространение разрывов в звуковых волнах // Прикл. матем. и механ. 1958. Т. 22. Вып. 4. С. 561—564.
6. Walkden F. The Shock Pattern of a Wing-Body Combination Far from the Flight Path // Aeronautical Quarterly. 1958. V. 9. P. 164—194.
7. Гриб А.А., Рыжов О.С., Христианович С.А. Теория коротких волн // Прикл. механ. и техн. физ. 1960. № 1. С. 63—74.
8. Рыжов О.С. Затухание ударных волн в неоднородных средах // Прикл. механ. и техн. физ. 1961. № 2. С. 15—25.
9. Полянский О.Ю. О затухании ударных волн в движущейся среде с переменными плотностью и температурой // Прикл. матем. и механ. 1960. Т. 24. Вып. 5. С. 912—915.
10. Рыжов О.С. Затухание ударных волн в стационарных течениях // Прикл. механ. и техн. физ. 1961. № 6. С. 36—43.
11. Carlson H.W. Influence of Airplane Configuration on Sonic Boom Characteristics // J. Aircraft. 1964. V. 1. № 2. P. 82—86.
12. Carlson H.W., Mack R.J., Morris O.A. A Wind-Tunnel Investigation of the Effect of Body Shape on Sonic-Boom Pressure Distributions. NASA TN D-3106. 1965. 36 p.
13. Seebass R. Sonic Boom Theory // J. Aircraft. 1969. V. 6. № 3. P. 177—184.
14. Жилин Ю.Л. О звуковом ударе // Уч. зап. ЦАГИ. 1971. Т. 2. № 3. С. 1—11.
15. Thomas Ch.L. Extrapolation of Wind-Tunnel Sonic Boom Signatures without Use of a Whitham F-Function. NASA SP-255. 1970. P. 205—217.
16. Thomas Ch.L. Extrapolation of Sonic Boom Pressure Signatures by the Waveform Parameter Method. NASA TN D-6832. 1972. 35 p.
17. Carlson H.W., Maglieri D.J. Review of Sonic Boom Generation Theory and Prediction Methods // J. Acoustical Soc. Am. 1972. V. 51. P. 675—682.
18. Жилин Ю.Л. Теория звукового удара // Сборник работ по звуковому удару. Тр. ЦАГИ. 1973. Вып. 1489. С. 3—24.
19. Уизем Дж. Линейные и нелинейные волны. М.: Мир, 1977. 622 с.

20. *Darden C.M.* Minimization of Sonic-Boom Parameters in Real and Isothermal Atmospheres. NASA TN D-7842. 1975. 28 p.
21. *Darden C.M.* Sonic Boom Theory: Its Status in Prediction and Minimization // *J. Aircraft.* 1977. V. 14. № 6. P. 569–576.
22. *Mack R.J., Darden C.M.* Wind-Tunnel Investigation of the Validity of a Sonic-Boom-Minimization Concept. NASA Technical Paper 1421. 1979. 44 p.
23. *Siclari M.J., Darden C. M.* An Euler Code Prediction of Near Field to Midfield Sonic Boom Pressure Signatures. AIAA-90-4000-CP. 1990. 14 p.
24. *Darden C.M.* Limitation of Linear Theory for Sonic Boom Calculations // *J. Aircraft.* 1993. V. 30. № 3. P. 309–314.
25. *Жилин Ю.Л., Коваленко В.В.* О связывании ближнего и дальнего полей в задаче о звуковом ударе // Уч. зап. ЦАГИ. 1998. Т. 29. № 3–4. С. 111–122.
26. *Makino Y., Aoyama T., Iwamiya T. et al.* Numerical Optimization of Fuselage Geometry to Modify Sonic-Boom Signature // *J. Aircraft.* 1999. V. 36. № 4. P. 668–674.
27. *Plotkin K.J.* State of the Art of Sonic Boom Modeling // *J. Acoustical Soc. Am.* 2002. V. 111. № 1. Pt. 2. P. 530–536.
28. *Makino Y., Suzuki K., Noguchi M. et al.* Nonaxisymmetrical Fuselage Shape Modification for Drag Reduction of Low-Sonic-Boom Airplane // *AIAA J.* 2003. V. 41. № 8. P. 1413–1420.
29. *Жилин Ю.Л., Ивантеева Л.Г., Чернышев С.Л.* Эквивалентные тела вращения, обладающие минимальным уровнем звукового удара // Тр. ЦАГИ. 2005. Вып. 2670. С. 12–19.
30. *Жилин Ю.Л., Ивантеева Л.Г., Чернышев С.Л.* О телах вращения минимального звукового удара // Уч. зап. ЦАГИ. 2006. Т. 37. № 3. С. 36–45.
31. *Коваленко В.В., Чернышев С.Л.* К вопросу о снижении звукового удара // Уч. зап. ЦАГИ. 2006. Т. 37. № 3. С. 53–62.
32. *Farhat C., Maute K., Argrow B. et al.* Shape Optimization Methodology for Reducing the Sonic Boom Initial Pressure Rise // *AIAA J.* 2007. V. 45. № 5. P. 1007–1018.
33. *Choi S., Alonso J., Kroo I. et al.* Multifidelity Design Optimization of Low-Boom Supersonic Jets // *J. Aircraft.* 2008. V. 45. № 1. P. 106–118.
34. *Li W., Shields E., Le D.* Interactive Inverse Design Optimization of Fuselage Shape for Low-Boom Supersonic Concepts // *J. Aircraft.* 2008. V. 45. № 4. P. 1381–1397.
35. *Bui Trong T.* CFD Analysis of Nozzle Jet Plume Effects on Sonic Boom Signature, NASA dryden flight research center, California. 2009. 28 p.
36. *Кючул Чо.* Анализ звукового удара с использованием полей возмущений, рассчитанных по нелинейной теории // Уч. зап. ЦАГИ. 2009. Т. 40. № 4. С. 45–55.
37. *Alauzet F., Loseille A.* High-Order Sonic Boom Modeling Based on Adaptive Methods // *J. Comput. Phys.* 2010. V. 229. P. 561–593.
38. *Чернышев С.Л.* Звуковой удар. М.: Наука, 2011. 351 с.
39. *Rallabhandi S.* Advanced Sonic Boom Prediction Using Augmented Burgers Equation // *J. Aircraft.* 2011. V. 48. № 4. P. 1245–1253.
40. *Chiba K., Makino Y., Takatoya T.* Design-Informatics Approach for Intimate Configuration of Silent Supersonic Technology Demonstrator // *J. Aircraft.* 2012. V. 49. № 5. P. 1200–1211.
41. *Ordaz I., Li W.* Integration of Off-Track Sonic Boom Analysis in Conceptual Design of Supersonic Aircraft // *J. Aircraft.* 2014. V. 51. № 1. P. 23–28.
42. *Li W., Rallabhandi S.* Inverse Design of Low-Boom Supersonic Concepts Using Reversed Equivalent-Area Targets // *J. Aircraft.* 2014. V. 51. № 1. P. 29–36.
43. *Minelli A., Salah el Din I., Carrier G.* Inverse Design Approach for Low-Boom Supersonic Configurations // *AIAA J.* 2014. V. 52. № 10. P. 2198–2212.
44. *Ordaz I., Geiselhart K., Fenbert J.* Conceptual Design of Low-Boom Aircraft with Flight Trim Requirement // *J. Aircraft.* 2015. V. 52. № 3. P. 932–939.
45. *Киселев А.Ф., Коваленко В.В., Притуло Т.М.* Исследование звукового удара: расчет и эксперимент // Инженерн. журн.: наука и инновации. 2017. № 8. С. 1–11.
46. *Housman J., Kenway G., Jensen J. et al.* Efficient Near-Field to Mid-Field Sonic Boom Propagation Using a High-Order Space Marching Method // AIAA Aviation 2019, Session: APA-25, Commercial Supersonic Technologies III, Dallas, Texas, June 20, 2019. 24 p.
47. *Браилко И.А., Крайко А.Н., Пьянков К.С. и др.* Аэродинамические и акустические характеристики сверхзвуковой вентиляционной решетки с дозвуковой осевой компонентой вектора скорости // Аэромеханика и газовая динамика. 2003. № 4. С. 9–22.

48. *Efremov N.L., Kraiko A.N., Pyankov K.S. et al.* Mathematical Simulation of Shock-Wave Structures Arising ahead of Fan Plane Cascades and Wheels // In "Turbomachines: Aeroelasticity, Aeroacoustics, and Unsteady Aerodynamics". Moscow: Torus Press, 2006. P. 281–294.
49. *Ефремов Н.Л., Крайко А.Н., Пьянков К.С. и др.* Ударно-волновые структуры перед неоднородной вентиляторной решеткой // Изв. РАН. Механика жидкости и газа. 2010. № 2. С. 135–152.
50. *Годунов С.К., Забродин А.В., Иванов М.Я. и др.* Численное решение многомерных задач газовой динамики. М.: Наука, 1976. 400 с.
51. *Колган В.П.* Применение принципа минимальных значений производной к построению конечно-разностных схем для расчета разрывных решений газовой динамики // Уч. зап. ЦАГИ. 1972. Т. 3. № 6. С. 68–77.
52. *Тилляева Н.И.* Обобщение модифицированной схемы С.К. Годунова на произвольные нерегулярные сетки // Уч. зап. ЦАГИ. 1986. Т. 17. № 2. С. 18–26.
53. *Ганжело А.Н., Крайко А.Н., Макаров В.Е. и др.* О повышении точности решения газодинамических задач // Кн. "Современные проблемы аэромеханики". М.: Машиностр., 1987. С. 87–102.
54. *Родионов А.В.* Повышение порядка аппроксимации схемы С.К. Годунова // Ж. вычисл. матем. и матем. физ. 1987. Т. 27. № 12. С. 1853–1860.
55. *Ландау Л.Д., Лифшиц Е.М.* Теоретическая физика. Т. VI. Гидродинамика. Изд. третье, переработанное. М.: Наука, 1986. 733 с.
56. *Черный Г.Г.* Газовая динамика. М.: Наука, 1988. 424 с.
57. *Овсянников Л.В.* Лекции по основам газовой динамики. Москва—Ижевск: Ин-т комп. исслед., 2003. 336 с.
58. *Крайко А.Н.* Теоретическая газовая динамика: классика и современность. М.: Торус пресс, 2010. 440 с.
59. *Валиев Х.Ф., Крайко А.Н., Тилляева Н.И.* Устойчивость одномерных стационарных течений с детонационной волной в канале переменной площади // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. № 4. С. 711–724.
60. *Горбовской В.С., Кажан А.В., Кажан В.Г. и др.* О влиянии формы эпюры избыточного давления на громкость звукового удара // Уч. зап. ЦАГИ. 2020. Т. 51. № 2. С. 3–17.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 519.63

ИССЛЕДОВАНИЕ И ПРИМЕНЕНИЕ МЕТОДА ДЕКОМПОЗИЦИИ
ОБЛАСТИ ДЛЯ МОДЕЛИРОВАНИЯ
ТЕПЛОВЫДЕЛЯЮЩЕГО ЭЛЕМЕНТА© 2022 г. М. П. Галанин^{1,*}, А. С. Родин^{1,**}¹ 125047 Москва, Миусская пл., 4, ИПМ им. М.В. Келдыша РАН, Россия

*e-mail: galan@keldysh.ru

**e-mail: rals@bk.ru

Поступила в редакцию 12.01.2021 г.
Переработанный вариант 15.11.2021 г.
Принята к публикации 16.12.2021 г.

Предложены вычислительные алгоритмы, основанные на применении двухуровневого аддитивного метода Шварца (вариант метода декомпозиции области с перекрытием) и метода Неймана–Дирихле (вариант метода декомпозиции области без перекрытия) для решения задачи контактного взаимодействия системы тел. Представлены результаты применения алгоритмов для численного моделирования напряженно-деформированного состояния участка тепловыделяющего элемента, включающего в себя от 2 до 100 топливных таблеток. Использовано осесимметричное термоупругое приближение. Выполнено исследование зависимости количества итераций алгоритма, требуемого для достижения заданного уровня точности, от количества тел в системе, шага расчетной сетки и граничных условий, поставленных на контактных поверхностях. Библ. 18. Фиг. 8. Табл. 6.

Ключевые слова: контактная задача, метод декомпозиции области, двухуровневый аддитивный метод Шварца, метод Неймана–Дирихле, тепловыделяющий элемент.

DOI: 10.31857/S004446692204007X

ВВЕДЕНИЕ

При расчете напряженно-деформированного состояния (НДС) реалистичной конструкции часто нужно учитывать контактное взаимодействие ее составных частей. Если отдельных деталей достаточно много, а итоговая конфигурация контактных поверхностей является неизвестной, то стандартные алгоритмы, хорошо работающие для случая контакта двух тел, могут оказаться недостаточно точными и быстрыми. Для эффективного решения общей мультиконтактной задачи могут потребоваться новые итерационные алгоритмы, учитывающие особенности конструкции. В данной работе предложены подобные алгоритмы, основанные на использовании вариантов метода декомпозиции области (МДО) и многосеточного метода.

МДО применяется при решении достаточно широкого класса краевых задач для дифференциальных уравнений в частных производных эллиптического и параболического типа, в том числе для статических задач механики деформируемого твердого тела (см. [1], [2]). Суть метода заключается в сведении численного решения задачи в большой области к решению ряда локальных задач в подобластях меньшего размера в рамках дополнительного (внешнего) итерационного процесса. Для учета контактного взаимодействия тел также используют ряд модернизированных вариантов МДО (см. [3]–[7]). Поскольку МДО является итерационным, то эффективность его применения во многом определяется скоростью сходимости алгоритма. В [1] на примере решения уравнения Пуассона показано, что некоторые варианты МДО обладают привлекательным свойством: количество внешних итераций, необходимое для достижения требуемой точности, практически не зависит от размерности глобальной системы уравнений. Для одноуровневых вариантов МДО (с использованием одной сетки) количество итерацийкратно возрастает при увеличении количества подобластей. Для устранения данного недостатка разработаны двухуровневые варианты МДО, в которых кроме основной сетки используется дополнительная грубая сетка (аналог многосеточных методов из [8]–[10]). В [1] показано, что возможно построение двухуровневого метода, для которого количество внешних итераций, необходимое для достиже-

ния требуемой точности, практически не зависит от количества вводимых подобластей. Для ряда вариантов МДО решение локальных задач в каждой подобласти может выполняться параллельно. Подобная возможность является критически важной для задач большой размерности.

Разрабатываемые алгоритмы применены для решения задачи контактного взаимодействия тел в тепловыделяющем элементе (“твэл”). В “твэле” обычно находятся сотни топливных таблеток, каждая из которых вступает в контакт с соседними таблетками и с оболочкой, причем конфигурация и положение контактных поверхностей могут значительно меняться в процессе функционирования устройства. В общем случае для моделирования работы “твэла” в режиме нормальной эксплуатации нужно решать трехмерную квазистационарную задачу с учетом ряда нелинейных эффектов: возникновения пластических и радиационных деформаций, деформаций ползучести, растрескивания таблеток и т.д. В данной работе рассмотрена осесимметричная термоупругая статическая постановка задачи.

Целью работы является построение алгоритма на основе двухуровневого аддитивного метода Шварца, для которого количество внешних итераций, требуемое для достижения заданной точности, меняется незначительным образом при измельчении шага сетки и при увеличении количества топливных таблеток в моделируемой области. Алгоритм должен обладать возможностью распараллеливания (для локальных задач) и должен естественным образом обобщаться на трехмерный нестационарный случай с учетом упомянутых нелинейных эффектов.

1. МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ

Рассмотрим в многомерном пространстве конструкцию из N термоупругих тел, занимающую область $G = \bigcup_{i=1}^N G_i$ с границей ∂G ($\bar{G} = G \cup \partial G$). Математическая модель статической задачи механики деформируемого твердого тела в рассматриваемой постановке состоит из следующих соотношений (см. [1]):

– уравнение равновесия

$$L\mathbf{u} = -\nabla \cdot \boldsymbol{\sigma}(\mathbf{u}) = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in G, \quad (1)$$

– граничные условия (кинематические и силовые)

$$\mathbf{u}(\mathbf{x}) \cdot \mathbf{n} = 0, \quad \mathbf{x} \in \partial G_D, \quad (2)$$

$$\boldsymbol{\sigma}(\mathbf{u}) \cdot \mathbf{n} = \mathbf{p}(\mathbf{x}), \quad \mathbf{x} \in \partial G_N, \quad (3)$$

– соотношение Коши для тензора деформации

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T), \quad (4)$$

– закон Гука для тензора напряжений

$$\boldsymbol{\sigma}(\mathbf{u}) = \mathbf{C} : (\boldsymbol{\varepsilon}(\mathbf{u}) - \boldsymbol{\varepsilon}^{th}(\mathbf{x})). \quad (5)$$

Здесь $\mathbf{u}(\mathbf{x})$ – вектор перемещения точки, определяемой радиус-вектором $\mathbf{x} = x_i \mathbf{e}_i$, $\mathbf{f}(\mathbf{x}) = f_i(\mathbf{x}) \mathbf{e}_i$ – вектор массовых сил, ∂G_D – участок границы, на котором заданы кинематические условия (условия Дирихле) для нормальной компоненты вектора перемещения, ∂G_N – участок границы, на котором заданы силовые условия (условия Неймана), $\mathbf{p}(\mathbf{x}) = p_i(\mathbf{x}) \mathbf{e}_i$ – вектор внешней нагрузки, действующей на поверхности ∂G_N , $\mathbf{C}$ – тензор коэффициентов упругости, $\boldsymbol{\varepsilon}^{th}$ – тензор температурной деформации.

Различные тела контактируют друг с другом, всего в конструкции есть P пар потенциально контактирующих поверхностей $(\Gamma_{c,1}^t, \Gamma_{c,2}^t)$, $t = 1, 2, \dots, P$. Введем обозначения

$$\Gamma_{c,1} = \bigcup_{t=1}^P \Gamma_{c,1}^t, \quad \Gamma_{c,2} = \bigcup_{t=1}^P \Gamma_{c,2}^t, \quad \Gamma_c = \Gamma_{c,1} \cup \Gamma_{c,2}.$$

При использовании на контактных границах условия прилипания дополнительные граничные условия на $\Gamma'_{c,1}$ состоят из следующих соотношений (для границы $\Gamma'_{c,2}$ можно записать аналогичные соотношения) (см. [3], [4]):

$$\begin{aligned}\sigma_\tau(\mathbf{x}) &= \sigma_\tau(\bar{\mathbf{x}}), \\ \sigma_n(\mathbf{x}) &= \sigma_n(\bar{\mathbf{x}}) \leq 0, \\ u_n(\mathbf{x}) + u_n(\bar{\mathbf{x}}) &\leq \delta_{0n}(\mathbf{x}), \\ \sigma_\tau(\mathbf{x})[u_\tau(\mathbf{x}) + u_\tau(\bar{\mathbf{x}}) - \delta_{0\tau}(\mathbf{x})] + \sigma_n(\mathbf{x})[u_n(\mathbf{x}) + u_n(\bar{\mathbf{x}}) - \delta_{0n}(\mathbf{x})] &= 0,\end{aligned}\tag{6}$$

где $\bar{\mathbf{x}} \in \Gamma'_{c,2}$ — сходственная (ближайшая) точка для $\mathbf{x} \in \Gamma'_{c,1}$, $\delta_{0\tau}(\mathbf{x}) \geq 0$, $\delta_{0n}(\mathbf{x}) \geq 0$ — функции, задающие начальный зазор (участки поверхностей в начальный момент могли не соприкасаться друг с другом), $\mathbf{n}$ и $\boldsymbol{\tau}$ — вектор внешней нормали и касательный вектор в рассматриваемой точке, $u_\tau = \mathbf{u} \cdot \boldsymbol{\tau}$, $u_n = \mathbf{u} \cdot \mathbf{n}$, $\sigma_\tau = (\boldsymbol{\sigma} \cdot \mathbf{n}) \cdot \boldsymbol{\tau}$, $\sigma_n = (\boldsymbol{\sigma} \cdot \mathbf{n}) \cdot \mathbf{n}$.

Последнее равенство в (6) гарантирует, что реализуется один из двух следующих вариантов:

1) рассматриваемая точка $\mathbf{x} \in \Gamma'_{c,1}$ находится в контакте, и тогда верны соотношения

$$u_\tau(\mathbf{x}) + u_\tau(\bar{\mathbf{x}}) = \delta_{0\tau}(\mathbf{x}), \quad u_n(\mathbf{x}) + u_n(\bar{\mathbf{x}}) = \delta_{0n}(\mathbf{x}), \quad \sigma_n(\mathbf{x}) < 0;$$

2) рассматриваемая точка не находится в контакте, и тогда верны соотношения

$$u_n(\mathbf{x}) + u_n(\bar{\mathbf{x}}) < \delta_{0n}(\mathbf{x}), \quad \sigma_\tau(\mathbf{x}) = 0, \quad \sigma_n(\mathbf{x}) = 0.$$

Если использовать на контактных границах условия скольжения без трения, то дополнительные условия на $\Gamma'_{c,1}$ принимают следующий вид (аналогично для границы $\Gamma'_{c,2}$) (см. [4]):

$$\begin{aligned}\sigma_\tau(\mathbf{x}) &= 0, \\ \sigma_n(\mathbf{x}) &= \sigma_n(\bar{\mathbf{x}}) \leq 0, \\ u_n(\mathbf{x}) + u_n(\bar{\mathbf{x}}) &\leq \delta_{0n}(\mathbf{x}), \\ \sigma_n(\mathbf{x})[u_n(\mathbf{x}) + u_n(\bar{\mathbf{x}}) - \delta_{0n}(\mathbf{x})] &= 0.\end{aligned}\tag{7}$$

В дальнейшем для обозначения блока граничных условий (6) будем использовать векторное неравенство $\mathbf{g}_1(\mathbf{u}) \leq 0$, а для обозначения блока граничных условий (7) — $\mathbf{g}_2(\mathbf{u}) \leq 0$.

2. ВЫЧИСЛИТЕЛЬНАЯ МОДЕЛЬ

Задача (1)–(6) (или (1)–(5), (7)) является нелинейной, поскольку итоговая конфигурация контактирующих поверхностей заранее неизвестна. Для ее решения обычно применяют различные итерационные численные методы, наиболее популярными из которых являются метод множителей Лагранжа, метод штрафных функций или их комбинация (см. [3], [11]).

2.1. Метод множителей Лагранжа

Применение метода множителей Лагранжа позволяет свести исходную задачу к задаче минимизации функционала (см. [3])

$$\Pi = \frac{1}{2} \int_G \sigma_{ij} \varepsilon_{ij} dG - \int_G u_i f_i dG - \int_{\partial G_N} u_i p_i d\Gamma + \int_{\Gamma'_{c,1}} \lambda_i(\mathbf{x}) (u_i(\mathbf{x}) + u_i(\bar{\mathbf{x}}) - \delta_{0i}(\mathbf{x})) d\Gamma,\tag{8}$$

где $\boldsymbol{\lambda} = (\lambda_\tau \ \lambda_n)^\top$ — множитель Лагранжа.

Для пространственной дискретизации полученных уравнений используем метод конечных элементов (МКЭ) (см. [11]). Построим в расчетной области сетку из четырехугольных конечных элементов, в i -м теле сетка состоит из n_i узлов, общее количество узлов n .

После применения стандартных процедур МКЭ задачу (8) можно свести к решению следующей системы линейных уравнений:

$$\begin{aligned} \begin{bmatrix} K & M^{(m)} \\ M^{(m)T} & 0 \end{bmatrix} \begin{Bmatrix} U^{(m+1)} \\ \lambda^{(m+1)} \end{Bmatrix} &= \begin{Bmatrix} R \\ 0 \end{Bmatrix}, \\ [K] &= \text{diag}[K_{11}, \dots, K_{NN}], \\ [K_{ii}] &= \int_{G_i} [B]^T [D] [B] dG, \\ \{R\} &= \int_{\partial G_N} [L_S]^T p(x) dS + \int_G [L_G]^T f(x) dG + \int_G [B]^T [D] \{\epsilon^{th}\} dG. \end{aligned} \quad (9)$$

Здесь $[K_{ii}]$ — глобальная матрица жесткости размерности $2n_i \times 2n_i$ (для двумерной задачи) для i -го тела, $[B]$ — матрица связи вектора деформаций с вектором перемещений, $[D]$ — матрица коэффициентов упругости, матрица $[M]$ размерности $2n \times l$ возникает в результате аппроксимации интеграла по контактной поверхности, $\{U\}$ — глобальный вектор узловых перемещений размерности $2n$, $\{\lambda\}$ — глобальный вектор множителей Лагранжа размерности l , $\{R\}$ — глобальный вектор узловых сил размерности $2n$, учитывающий вклад поверхностных сил, массовых сил и температурных деформаций.

Матрица системы (9) имеет нулевой диагональный блок, поэтому для решения системы уравнений нельзя использовать большую часть стандартных итерационных методов. Для решения используют либо прямые методы, либо специальные итерационные методы, предназначенные для решения седловых задач (например, метод Узавы, модифицированный метод симметричной последовательной верхней релаксации, см. [14], [15]).

Матрица является плохообусловленной, поэтому, если в последнем интеграле в (8) учитывать вклад всех поверхностных элементов, лежащих на потенциально контактной поверхности $\Gamma_{c,l}$, то численное решение может не соответствовать граничным условиям (6) или (7). Для получения более точного решения будем проводить итерационный процесс и на $(m+1)$ -й итерации учитывать в (8) вклад только тех поверхностных элементов, которые лежат на поверхности $\tilde{\Gamma}_{c,l}^{(m)}$ (объединение участков поверхностей, которые на текущей итерации считаются находящимися в контакте). В конце итерации выполняется вычисление новой контактной конфигурации $\tilde{\Gamma}_{c,l}^{(m+1)}$: поверхностные узлы, в которых возникли растягивающие нормальные напряжения, исключаются из $\tilde{\Gamma}_{c,l}^{(m)}$ (по силовому критерию). Итерационный процесс (в дальнейшем будем называть его внешним процессом) продолжается до тех пор, пока не устанавливается окончательная конфигурация контактной поверхности.

2.2. Метод Неймана–Дирихле

Постановка задачи (1)–(6) во многом аналогична постановке задачи для одного тела, разбитого на подобласти, при использовании варианта МДО без перекрытия. Различия связаны с тем, что при использовании классического МДО условия равенства векторов перемещений и напряжений заданы на всем объединении фиктивных границ между подобластями (которые заранее известны и не меняются в ходе расчета), а в контактной задаче аналогичные равенства будут выполняться только на тех участках поверхности $\tilde{\Gamma}_c$, которые в итоге будут находиться в контакте. Благодаря подобному сходству, для решения контактных задач (с различными условиями на контактирующих поверхностях) применяют модифицированные варианты МДО без перекрытия (см. [1]). Суть этих методов заключается в том, что в рамках одной итерации для каждого тела на контактных поверхностях ставится либо кинематическое условие (условие Дирихле), либо силовое (условие Неймана). Затем полученные на границах значения перемещений или напряжений корректируются с помощью специальных процедур (при необходимости корректируется и конфигурация находящихся в контакте поверхностей). Итерационный процесс продолжается до тех пор, пока оба условия (Дирихле и Неймана) не будут выполнены с достаточной точностью.

В частности, в методе Неймана–Дирихле на одной половине контактных поверхностей $\tilde{\Gamma}_{c,1}^{(m)}$ всегда ставится условие Неймана, а на противоположащих им поверхностях $\tilde{\Gamma}_{c,2}^{(m)}$ всегда ставится условие Дирихле (таким образом, чтобы поле перемещений обеспечивало выполнение кинематической части условий (6) или (7)). Для условий скольжения граничные условия на контактных поверхностях на $(m + 1)$ -й итерации принимают вид

$$\boldsymbol{\sigma}(\mathbf{u}^{(m+1)}) \cdot \mathbf{n} = \mathbf{p}_c^{(m)}(\mathbf{x}), \quad \mathbf{x} \in \Gamma_{c,1}, \quad (10)$$

$$u_n^{(m+1)}(\mathbf{x}) + u_n^{(m+1)}(\bar{\mathbf{x}}) - \delta_{0n}(\mathbf{x}) \leq 0, \quad \mathbf{x} \in \Gamma_{c,2}, \quad (11)$$

где $\mathbf{p}_c^{(m)}(\mathbf{x})$ – вектор распределенных контактных сил (он равен нулю, если $\mathbf{x} \notin \tilde{\Gamma}_{c,1}^{(m)}$), который корректируется таким образом, чтобы в конце итерационного процесса выполнялись силовые условия в (6).

Рассмотрим пространство $H_0^1(G, \partial G_D, \Gamma_{c,2})$ функций $\mathbf{u}$, для которых выполнены кинематические условия (2) и (11). Введем билинейную форму для функций $\mathbf{u}, \mathbf{v} \in H_0^1(G, \partial G_D, \Gamma_{c,2})$:

$$a(\mathbf{u}, \mathbf{v}) = \int_G \mathbf{C} : \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) dG,$$

и линейные функционалы

$$g(\mathbf{v}) = \int_G \mathbf{f} \cdot \mathbf{v} dG + \int_{\partial G_N} \mathbf{p} \cdot \mathbf{v} dS + \int_G \mathbf{C} : \boldsymbol{\varepsilon}^{th}(\mathbf{x}) : \boldsymbol{\varepsilon}(\mathbf{v}) dG,$$

$$g_c^{(m)}(\mathbf{v}) = \int_{\Gamma_{c,1}} \mathbf{p}_c^{(m)} \cdot \mathbf{v} dS.$$

Тогда слабая постановка задачи (1)–(5), (10), (11) на $(m + 1)$ -й итерации принимает следующий вид (см. [1], [16]):

$$a(\mathbf{u}^{(m+1)}, \mathbf{v}) = g(\mathbf{v}) + g_c^{(m)}(\mathbf{v}) \quad \forall \mathbf{v} \in H_0^1(G, \partial G_D, \Gamma_{c,2}). \quad (12)$$

После использования МКЭ для пространственной дискретизации (12) сводится к системе линейных уравнений

$$[K^{(m)}]\{U^{m+1}\} = \{R\} + \{R_c^{(m)}\}, \quad m = 0, 1, \dots \quad (13)$$

Опишем структуру матрицы $[K^{(m)}]$ для условия скольжения без трения. Множество узлов сетки J , которые относятся к области G , можно представить в виде объединения подмножеств $J = J^1 \cup J^2 \cup J^3 \cup J^4$, где J^1 – внутренние узлы и узлы, относящиеся к ∂G_N , J^2 – узлы, относящиеся к $\tilde{\Gamma}_{c,1}^{(m)}$, J^3 – узлы, относящиеся к ∂G_D , J^4 – узлы, относящиеся к $\tilde{\Gamma}_{c,2}^{(m)}$.

Для узлов $j \in J^1 \cup J^2$ строки матрицы $[K^{(m)}]$ соответствуют строкам стандартной матрицы жесткости. Компоненты вектора узловых сил $\{R\}$ формируются из вкладов поверхностных сил, массовых сил и температурной деформации. Компоненты вектора узловых контактных сил $\{R_c^{(m)}\}$ отличны от нуля только для узлов $j \in J^2$. Для узлов $j \in J^3$ компонента вектора $\{R\}$ равна нулю, а строка матрицы $[K^{(m)}]$, соответствующая большей (по модулю) компоненте вектора нормали $\mathbf{n}_j = (n_r, n_z)^T$, имеет вид $[0, \dots, 0, k_{2j-1}, k_{2j}, 0, \dots, 0]$, $k_{2j-1} = n_r$, $k_{2j} = n_z$. Для узлов $j \in J^4$ компонента вектора $\{R\}$ равна δ_{0n}^j , а строка матрицы $[K^{(m)}]$, соответствующая большей компоненте вектора нормали, имеет вид (узел j находится напротив поверхностного элемента, в который входят узлы с номерами p и q) $[0, \dots, 0, k_{2j-1}, k_{2j}, 0, \dots, 0, k_{2p-1}, k_{2p}, 0, \dots, 0, k_{2q-1}, k_{2q}, 0, \dots, 0]$, $k_{2p-1} = -n_r L_{S,p}(\bar{\mathbf{x}}_j)$, $k_{2p} = -n_z L_{S,p}(\bar{\mathbf{x}}_j)$, $k_{2q-1} = -n_r L_{S,q}(\bar{\mathbf{x}}_j)$, $k_{2q} = -n_z L_{S,q}(\bar{\mathbf{x}}_j)$, где $L_{S,p}$ – базисная функция поверхностного элемента для узла p .

В конце итерации выполняется вычисление новой контактной конфигурации $\tilde{\Gamma}_c^{(m+1)}$: поверхностные узлы, в которых возникли растягивающие нормальные контактные силы, исключаются из $\tilde{\Gamma}_c^{(m)}$ (по силовому критерию); узлы, которые в результате сдвига перестали находиться напротив противоположащей поверхности, выходят из контакта по кинематическому критерию; а новые узлы, для которых возник захлест относительно противоположащей поверхности, добавляются в список контактных узлов.

Затем вычисляются новые значения векторов распределенных контактных сил $\{p_{c,t}^{(m+1)}\}$ для каждой поверхности $\Gamma_{c,1}^t$ (для узлов, которые не находятся в контакте, соответствующие компоненты вектора равны нулю):

$$\{p_{c,t}^{(m+1)}\} = (1 - \theta_t) \{p_{c,t}^{(m)}\} - \theta_t \{\bar{p}_{c,t}^{(m+1)}\}, \quad (14)$$

где $\{\bar{p}_{c,t}^{(m+1)}\}$ – вектор значений распределенных контактных сил в сходственных точках, лежащих на поверхности $\Gamma_{c,2}^t$, θ_t – итерационный параметр.

После этого определяется новое приближение для вектора узловых контактных сил:

$$\{R_c^{(m+1)}\} = \int_{\tilde{\Gamma}_{c,1}^{(m+1)}} [L_S]^T p_c^{(m+1)}(x) dS. \quad (15)$$

Применяется следующая аппроксимации функции распределенных контактных сил:

$$p_c^{(m+1)}(x) = \sum_j \chi_j \{p_c^{(m+1)}\}_j,$$

где в качестве базисных функций χ_j применяются кусочно-постоянные в ячейке Дирихле для узла j функции.

В расчетах с несогласованными сетками более точные результаты получаются, если формулу (11) применять не для $\{p_{c,t}^{(m+1)}\}_j$, а для значения контактных сил $\{F_{c,t}^{(m+1)}\}_j = S_{D,j} \{p_{c,t}^{(m+1)}\}_j$ ($S_{D,j}$ – площадь ячейки Дирихле для узла j):

$$\{F_{c,t}^{(m+1)}\} = (1 - \theta_t) \{F_{c,t}^{(m)}\} - \theta_t \{\bar{F}_{c,t}^{(m+1)}\}, \quad (16)$$

где $\{\bar{F}_{c,t}^{(m+1)}\}_j$ – значение распределенных контактных сил, проинтегрированных по участку поверхности $\Gamma_{c,2}^t$, расположенному напротив ячейки Дирихле для узла j .

После вычисления $\{F_c^{(m+1)}\}$ определяется вектор $\{p_c^{(m+1)}\}$ и по формуле (15) находится вектор $\{R_c^{(m+1)}\}$.

Если достигнутая точность не удовлетворяет заданному критерию, то происходит переход к следующей итерации.

2.3. Аддитивный метод Шварца для решения систем линейных уравнений

Для эффективного решения системы линейных уравнений (13) можно использовать аддитивный метод Шварца (вариант МДО с перекрытием подобластей).

Представим область G в виде объединения $G = \bigcup_{k=1}^M \Omega_k$ конечного числа своих подобластей $\Omega_1, \dots, \Omega_M$ с границами $\partial\Omega_1, \dots, \partial\Omega_M$ ($\bar{\Omega}_k = \Omega_k \cup \partial\Omega_k$). В области Ω_k находится m_k узлов сетки.

Для решения (13) применим предобусловленный итерационный метод Ричардсона (его итерации будем называть внутренними) (см. [1])

$$\{U^{m+1}\}^{(l+1)} = \{U^{m+1}\}^{(l)} + \alpha [B_{as}]^{-1} \{r\}^{(l)}, \quad (17)$$

где $\{r\}^{(l)} = \{R\} + \{R_c^{(m)}\} - [K^{(m)}]\{U^{m+1}\}^{(l)}$ — вектор невязки решения на внутренней итерации с номером l , α — некоторый итерационный параметр, существенно влияющий на сходимость данного итерационного процесса, $[B_{as}]^{-1}$ — предобуславливающая матрица для аддитивного метода Шварца. Она задается равенством

$$[B_{as}]^{-1} = \sum_{k=1}^M [T_k][B_k]^{-1}[T_k]^T, \quad (18)$$

где $[T_k]$ — матрица размерности $2n \times 2m_k$, у которой j -й столбец совпадает с j -м столбцом единичной матрицы размерности $2n \times 2n$, $[B_k]$ — матрица размерности $2m_k \times 2m_k$, которая по сути является блоком матрицы $[K^{(m)}]$, который, в свою очередь, соответствует узлам сетки, принадлежащим подобласти Ω_k (вектор $\{r_k\} = [T_k]^T \{r\}^{(l)}$ состоит из компонент вектора $\{r\}^{(l)}$, соответствующих данным узлам сетки).

Как показано в [1], при увеличении количества вводимых подобластей количество внешних итераций в аддитивном методе Шварца, необходимое для достижения заданного уровня точности, может кратно увеличиваться. Подобная зависимость существенно ухудшает эффективность метода. Для преодоления данного недостатка можно использовать аддитивный метод Шварца с грубосеточной коррекцией, предполагающий, что наряду с базовой расчетной сеткой в области задана грубая сетка из N_c узлов, размеры ячеек которой сопоставимы с размерами введенных подобластей. Узлы грубой сетки могут не совпадать с узлами основной сетки, границы расчетной грубой области могут отличаться от границ основной расчетной области. Двухуровневый метод является некоторым аналогом многосеточного метода (V-цикла) (см. [8]–[10]).

В этом случае предобуславливатель задается формулой (см. [1], [17])

$$[B_{as2}]^{-1} = [T_0][B_0]^{-1}[T_0]^T + [B_{as}]^{-1}, \quad (19)$$

где $[T_0]$ — матрица размерности $2n \times 2N_c$, у которой столбцы, соответствующие узлу грубой сетки j , заполнены значениями базисной функции L_{cj} в точках расположения узлов основной сетки, $[B_0] = [T_0]^T [K^{(m)}] [T_0]$ — матрица размерности $2N_c \times 2N_c$.

С учетом (18) и (19) формулу (17) можно переписать следующим образом:

$$\{U^{m+1}\}^{(l+1)} = \{U^{m+1}\}^{(l)} + \alpha \sum_{k=0}^M [T_k]\{\omega_k\}^{(l)}, \quad (20)$$

где для нахождения векторов $\{\omega_k\}^{(l)}$ нужно решить $(M+1)$ систему линейных уравнений

$$[B_k]\{\omega_k\}^{(l)} = \{r_k\}, \quad k = 0, \dots, M. \quad (21)$$

Локальные задачи (21) можно решать независимо друг от друга.

Кроме предобусловленного метода Рундсона для решения (10) можно использовать предобусловленные итерационные метода крыловского подпространства. Матрица системы является несимметричной, поэтому применим предобусловленный метод сопряженных невязок. Данный метод задается соотношениями (см. [17], [18])

$$\begin{aligned} \{r\}^{(0)} &= \{R\} + \{R_c^{(m)}\} - [K^{(m)}]\{U^m\}, \quad \{p\}^{(0)} = \{r\}^{(0)}, \\ v_l &= \frac{\{r\}^{(l)T} [K^{(m)}]\{p\}^{(l)}}{([K^{(m)}]\{p\}^{(l)})^T [K^{(m)}]\{p\}^{(l)}}, \quad l = 0, 1, \dots, \\ \{U^{m+1}\}^{(l+1)} &= \{U^{m+1}\}^{(l)} + v_l \{p\}^{(l)}, \\ \{r\}^{(l+1)} &= \{r\}^{(l)} - v_l [K^{(m)}]\{p\}^{(l)}, \end{aligned} \quad (22)$$

$$\mu_{l,s} = - \frac{\left([K^{(m)}] \{p\}^{(s)} \right)^T [K^{(m)}] [B_{as2}]^{-1} \{r\}^{(l+1)}}{\left([K^{(m)}] \{p\}^{(s)} \right)^T [K^{(m)}] \{p\}^{(s)}}, \quad s = 0, \dots, l,$$

$$\{p\}^{(l+1)} = [B_{as2}]^{-1} \{r\}^{(l+1)} + \sum_{s=0}^l \mu_{l,s} \{p\}^{(s)}.$$

В формулах (22) вычисление вектора $[B_{as2}]^{-1} \{r\}^{(l+1)}$ также сводится к решению локальных задач (аналогичных (21)) в подобластях и на грубой сетке.

2.4. Двухуровневый аддитивный метод Шварца

Кроме использования аддитивного метода Шварца для эффективного решения систем линейных уравнений можно рассмотреть его применение в качестве итерационного процесса в соответствующих функциональных пространствах (см. [2], [16]).

Введем обозначения $\partial\Omega_{N,k} = \partial G_N \cap \partial\Omega_k$, $\partial\Omega_{D,k} = \partial G_D \cap \partial\Omega_k$, $\Gamma_{k,j} = \Gamma_{c,j} \cap \Omega_k$, $\Gamma_c^k = \Gamma_{k,1} \cup \Gamma_{k,2}$.

Выберем некоторую функцию $\mathbf{u}^{(0)}$, удовлетворяющую граничному условию (2). Построим последовательность функций $\{\mathbf{u}^{(m)}\}$. На $(m+1)$ -й итерации сначала нужно решить локальные контактные задачи (с использованием метода Неймана–Дирихле) в подобластях Ω_k ($k = 1, \dots, M$):

$$\begin{aligned} L\mathbf{u}_k^{(m+1)} &= \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \Omega_k \setminus \Gamma_c^k, \\ \boldsymbol{\sigma} \cdot \mathbf{n} &= \mathbf{p}(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega_{N,k}, \\ \boldsymbol{\sigma}(\mathbf{u}_k^{(m+1)}) \cdot \mathbf{n} &= \mathbf{p}_c^{(m+1)}(\mathbf{x}), \quad \mathbf{x} \in \Gamma_{k,1}, \\ \mathbf{u}_{k,n}^{(m+1)}(\mathbf{x}) + \mathbf{u}_{k,n}^{(m+1)}(\bar{\mathbf{x}}) - \delta_{0n}(\mathbf{x}) &\leq 0, \quad \mathbf{x} \in \Gamma_{k,2}, \\ \mathbf{u}_k^{(m+1)}(\mathbf{x}) &= \mathbf{u}^{(m)}(\mathbf{x}), \quad \mathbf{x} \in \bar{G} \setminus (\Omega_k \cup \partial\Omega_{N,k}). \end{aligned} \quad (23)$$

Выделим в пространстве $H_0^1(G, \partial G_D, \Gamma_{c,2})$ замкнутые подпространства

$$H_k = \{ \mathbf{v} \in H_0^1(G, \partial G_D, \Gamma_{c,2}) : \mathbf{v}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in \bar{G} \setminus (\Omega_k \cup \partial\Omega_{N,k}) \}$$

(см. [2]). Функция $\mathbf{w}_k^{(m+1)} = \mathbf{u}_k^{(m+1)} - \mathbf{u}^{(m)}$ принадлежит подпространству H_k . Тогда слабая постановка для локальной задачи (23) принимает вид

$$a(\mathbf{w}_k^{(m+1)}, \mathbf{v}_k) = g(\mathbf{v}_k) + g_c^{(m+1)}(\mathbf{v}_k) - a(\mathbf{u}^{(m)}, \mathbf{v}_k) \quad \forall \mathbf{v}_k \in H_k. \quad (24)$$

В конце каждой итерации итоговое решение вычисляется по формуле

$$\mathbf{u}^{(m+1)} = \mathbf{u}^{(m)} + \alpha \sum_{k=1}^M \mathbf{w}_k^{(m+1)}.$$

В уравнении (24) в правой части стоит слагаемое с интегралом от распределенных контактных сил $\mathbf{p}_c^{(m+1)}$, которые не известны. Поэтому для решения нужно проводить дополнительный внутренний итерационный процесс. После использования МКЭ для пространственной дискретизации (24) сводится к следующей линейной системе уравнений на $(l+1)$ -й внутренней итерации:

$$[K_k^{(m)}] \{w_k\}^{(l+1)} = \{r_k\} + \{\Delta R_{k,c}\}^{(l)}, \quad k = 1, \dots, M. \quad (25)$$

Здесь $[K_k^{(m)}]$ — матрица размерности $2m_k \times 2m_k$, которая является блоком матрицы $[K^{(m)}]$, $\{r_k\} = \{R_k\} + \{R_{k,c}^{(m)}\} - [K_k^{(m)}] \{U_k^{(m)}\}$ — вектор невязки на предыдущей внешней итерации, $\{\Delta R_{k,c}\}^{(l)} = \{R_{k,c}^{(m+1)}\}^{(l)} - \{R_{k,c}^{(m)}\}$, где вектор узловых контактных сил $\{R_{k,c}^{(m+1)}\}^{(l)}$ определяется по формулам, аналогичным (15), (16).

Так же, как делалось выше, можно вместе с базовой расчетной сеткой использовать дополнительную грубую сетку из N_c узлов. В полученном двухуровневом аддитивном методе Шварца глобальный вектор перемещений вычисляется по формуле

$$\{U^{(m+1)}\} = \{U^{(m)}\} + \alpha \sum_{k=0}^M \{w_k^{(m+1)}\}.$$

Вспомогательная задача на грубой сетке имеет вид

$$[K_0]\{\Delta U_0^{(m+1)}\} = \{R_0^{(m)}\}.$$

Здесь $[K_0]$ — матрица размерности $2N_c \times 2N_c$, соответствующая матрице жесткости МКЭ на грубой сетке для эффективных материалов, $\{\Delta U_0^{(m+1)}\}$ — вектор размерности $2N_c$, $\{w_0^{(m+1)}\} = [T_0]\{\Delta U_0^{(m+1)}\}$, компоненты вектора $\{R_0^{(m)}\}$ совпадают с компонентами вектора $[T_0]^T \{r^{(m)}\}$ за исключением узлов, в которых заданы кинематические условия.

3. РЕЗУЛЬТАТЫ РАСЧЕТОВ С ПРИМЕНЕНИЕМ МЕТОДА ДЕКОМПОЗИЦИИ ОБЛАСТИ

В описанных ниже расчетах итерационный процесс продолжается до тех пор, пока не выполняется соотношение

$$E_r = \|\{r^{(m+1)}\}\|_2 / \|\{r^{(0)}\}\|_2 \leq \varepsilon_0.$$

Для расчетов, если не указано иное, использовались конечные элементы первого порядка на четырехугольной сетке (см. [11]), выбирались следующие значения входных параметров: $\{U^0\} = 0$, $\varepsilon_0 = 10^{-6}$, $\alpha = 0.5$.

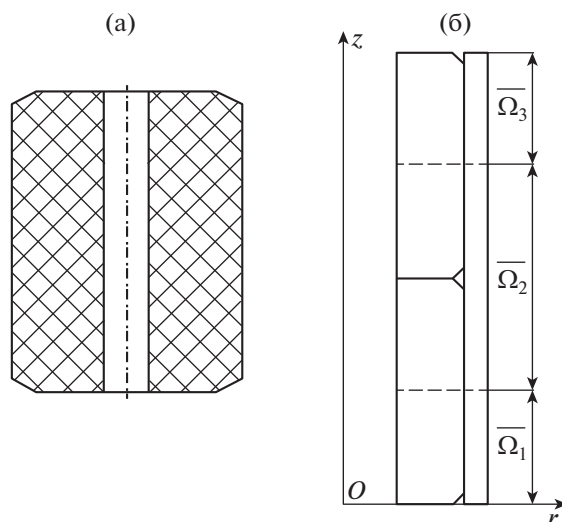
Рассмотренные алгоритмы протестированы на решении осесимметричной задачи о контактно-взаимодействии N однотипных участков упругой трубы ($N = 2, 4, 8$), поставленных в направлении оси Oz друг на друга в виде столба, под действием постоянных давлений: p_1 , приложенного к верхнему торцу верхнего тела, и p_2 , приложенного к внешним боковым поверхностям тел.

В такой постановке контактная задача эквивалентна задаче для единого тела, для которого известны аналитические решения для компонент тензора напряжений (см. [13]). Конфигурация контактных поверхностей не изменяется, задача фактически является линейной, поэтому “tag-метод” (вариант метода множителей Лагранжа) дает корректное решение за одну итерацию. Метод Неймана–Дирихле, если в качестве начального приближения для контактного давления на всех нижних торцах выбрать величину p_1 , тоже сходится за одну итерацию; если выбрать нулевое приближение, то при увеличении количества тел в M раз, количество требуемых итераций будет увеличиваться приблизительно в $M^{1/2}$ раз.

Для двухуровневого аддитивного метода Шварца расчетная область, соответствующая продольному сечению рассматриваемой конструкции, разбивалась на $M = N + 1$ подобластей. Проведенные расчеты показали, что количество итераций, требуемое для достижения заданной точности, незначительно отличалось для различного количества тел и различных значений шага сетки (рассмотрены три последовательно сгущающиеся сетки).

Рассмотрим задачу, позволяющую моделировать ряд термомеханических эффектов, происходящих в “твэле” (см. [7], [15]). В конструкцию “твэла” входят цилиндрическая оболочка и столб из топливных таблеток (их количество достигает нескольких сотен). Оболочка имеет следующие характерные размеры: длина — 3–4 м, диаметр — 1 см, толщина — 1 мм. Продольное сечение топливной таблетки, которая является цилиндром с центральным отверстием и фасками на обоих торцах, приведено на фиг. 1а. Расчетная область для участка “твэла” с двумя таблетками и оболочкой показана на фиг. 1б. В ходе работы “твэла” происходит нагрев таблеток до температуры около 1600 К. Контактное взаимодействие таблеток друг с другом и оболочкой оказывает существенное влияние на НДС “твэла”.

Выполним численное моделирование выхода участка “твэла” на номинальную мощность в осесимметричном термоупругом приближении. Свойства материала таблетки соответствуют ди-



Фиг. 1. (а) — Продольное сечение топливной таблетки с фасками и центральным отверстием; (б) — схема расчетной области с разбиением на подобласти (для случая $N = 2$, $N = 3$).

оксиду урана, свойства материала оболочки соответствуют сплаву циркония, теплофизические характеристики нелинейно зависят от температуры.

Расчетная область включала в себя N таблеток и участок оболочки. Сначала во всей области решалась динамическая температурная задача с заданной постоянной (по пространству) плотностью тепловыделения в таблетках, температура всей оболочки считалась равной 623 К, на боковых поверхностях таблеток и внутренней поверхности оболочки ставилось граничное условие третьего рода.

После выхода температуры на установившийся режим вычисленное температурное поле применялось для решения механической задачи со следующими граничными условиями: нижние торцы первой таблетки и оболочки закреплены в направлении оси Oz , на верхнем торце верхней таблетки задано давление $p_1 = 50$ МПа, на внешней поверхности оболочки задано давление $p_2 = 10$ МПа. На поверхности контактирующих тел поставлено условие скольжения без трения.

Отметим следующие характерные особенности рассмотренной мультиконтактной задачи. Начальный зазор для всех контактных пар равен нулю. Для каждой контактной пары таблетка/таблетка в результате нагрева происходит выход из контакта значительного количества узлов сетки, расположенных ближе к фаске. Из-за большой протяженности конструкции вдоль оси Oz топливные таблетки, особенно в верхней части столба, смещались на значительное расстояние относительно своего начального положения (и относительно оболочки), поэтому поверхностные сетки в контактных парах таблетка/оболочка всегда не согласованы. Например, для 100 таблеток итоговое смещение сопоставимо с высотой одной таблетки.

В табл. 1 приведены количество итераций и время расчетов с использованием “mortar-метода”, в которых количество таблеток менялось от 2 до 50. Для интерполяции перемещений и множителей Лагранжа использовались КЭ второго порядка, в одной таблетке сетка состояла из 10 ячеек в радиальном и 10 ячеек в осевом направлениях ($h_r = 0.308$, $h_z = 1.0$). Для решения системы линейных уравнений (9) применена процедура SparseLU из библиотеки Eigen, предназначенная для работы с разреженными матрицами. Итерационный процесс продолжался до тех пор, пока не устанавливалась окончательная конфигурация контактных поверхностей. Поскольку на каждой итерации матрица менялась, то на каждой итерации LU-разложение делалось заново.

Таблица 1. Количество итераций и время расчета для “mortar-метода”

N	2	5	10	25	50
Количество итераций	12	37	59	140	236
Время, с	2	20	181	3274	36756

Из табл. 1 видно, что при увеличении количества таблеток количество итераций возрастает по зависимости, близкой к линейной ($O(N)$), а время расчета увеличивается по зависимости, близкой к кубической ($O(N^3)$). Поэтому данный алгоритм не является эффективным при моделировании большого количества таблеток.

В [15] для решения уравнения (9) применен модифицированный метод симметричной последовательной верхней релаксации (MSSOR) с предобуславливателем, который позволил существенно сократить время счета. Выполнено сравнение результатов расчета участка “твэла” из 10 таблеток, полученных с помощью “mortar-метода” и двухуровневого аддитивного метода Шварца.

В методе Неймана–Дирихле граничные условия на контактных поверхностях поставлены следующим образом: в контактных парах таблетка/таблетка на торце нижней таблетки задано силовое условие (10), а на торце верхней таблетки – кинематическое (11). В контактных парах таблетка/оболочка на внешней боковой поверхности таблетки задано силовое условие, а на внутренней боковой поверхности оболочки – кинематическое. В качестве начального приближения на первой итерации выбрано нулевое контактное давление. Для всех контактных пар итерационный параметр $\theta_j = 0.5$.

Для решения на каждой итерации системы линейных уравнений (13) могут быть использованы разные методы. Сначала рассмотрим последовательный алгоритм, учитывающий, что при таком выборе граничных условий система уравнений по сути принимает блочный вид: в первую очередь решается блок системы, относящийся к первой таблетке (он не зависит от переменных, соответствующих другим телам). Затем решается блок системы, относящийся ко второй таблетке. Строки матрицы, соответствующие узлам нижнего торца, в которых заданы кинематические условия, переписываются следующим образом: все величины, относящиеся к первой таблетке (они уже известны), переносятся в правую часть локальной системы уравнений. Аналогичным образом происходит последовательное вычисление всех блоков системы, соответствующих таблеткам. В конце выполняется решение блока системы уравнений для оболочки.

В использованном варианте метода Неймана–Дирихле после установления итоговой конфигурации контактных поверхностей для узлов, относящихся к поверхностям $\Gamma_{c,2}^t$, всегда выполняется кинематическая часть граничных условий (6) или (7) (последнее неравенство и последнее равенство). Для корректности решения всей контактной задачи нужно выполнение силовых условий (первые два равенства в (6) или (7)) для узлов, относящихся к поверхностям $\Gamma_{c,1}^t$. Для оценки текущей погрешности для поверхности $\Gamma_{c,1}^t$ будем использовать следующую величину (для условия скольжения):

$$E_{F,t}^{(m+1)} = \sqrt{\frac{\sum_{j=1}^{l_t} \left[\left(\{F_{c,t}^{(m+1)}\}_{2j-1} + \{\bar{F}_{c,t}^{(m+1)}\}_{2j-1} \right) n_{r,j} + \left(\{F_{c,t}^{(m+1)}\}_{2j} + \{\bar{F}_{c,t}^{(m+1)}\}_{2j} \right) n_{z,j} \right]^2}{\sum_{j=1}^{l_t} \left[\{F_{c,t}^{(m+1)}\}_{2j-1} n_{r,j} + \{F_{c,t}^{(m+1)}\}_{2j} n_{z,j} \right]^2}},$$

где $\{F_{c,t}^{(m+1)}\}$ – вектор контактных сил в узлах $\Gamma_{c,1}^t$ (распределенные силы, умноженные на площади ячеек Дирихле) и $\{\bar{F}_{c,t}^{(m+1)}\}$ – вектор контактных сил, вычисленных на участках поверхности $\Gamma_{c,2}^t$, расположенных напротив соответствующих ячеек Дирихле. В дальнейшем будем приводить значение величины $E_F^c = \max_t E_{F,t}^{(m+1)}$ – максимум среди относительных погрешностей вектора контактных сил по всем парам таблетка/оболочка.

В табл. 2 для метода Неймана–Дирихле приведены время расчета iter_out , количество итераций, в течение которого происходит выход из контакта узлов в парах таблетка/таблетка (по силовому критерию), а также количество итераций, необходимое для достижения различных значений погрешности невязки E_r и погрешности E_F^c (приведено в скобках, если заданное значение не достигнуто за проведенное количество итераций, то стоит символ “–”). Расчеты проводились для следующих значений шагов сетки: в таблетке $h_1 = 0.25$, в оболочке $h_2 = 0.165$. В процессе расчета в контактных парах таблетка/таблетка из контакта выходило 5/6 от общего количества элементов на поверхности (10 элементов из 12). Количество таблеток варьировалось от 2 до 100.

Таблица 2. Количество итераций для метода Неймана–Дирихле

N	Время, с	iter_out	iter $E_r = E_F^c = 10^{-3}$	iter $E_r = E_F^c = 10^{-4}$	iter $E_r = E_F^c = 10^{-5}$	iter $E_r = E_F^c = 10^{-6}$
2	8	15	20 (29)	25 (35)	31 (40)	36 (—)
4	18	21	28 (39)	34 (44)	39 (50)	45 (—)
10	55	40	46 (58)	53 (63)	57 (69)	63 (—)
25	235	80	87 (95)	92 (101)	96 (106)	101 (—)
50	799	138	142 (154)	147 (159)	154 (165)	161 (—)
100	3372	248	261 (272)	265 (277)	270 (283)	278 (—)

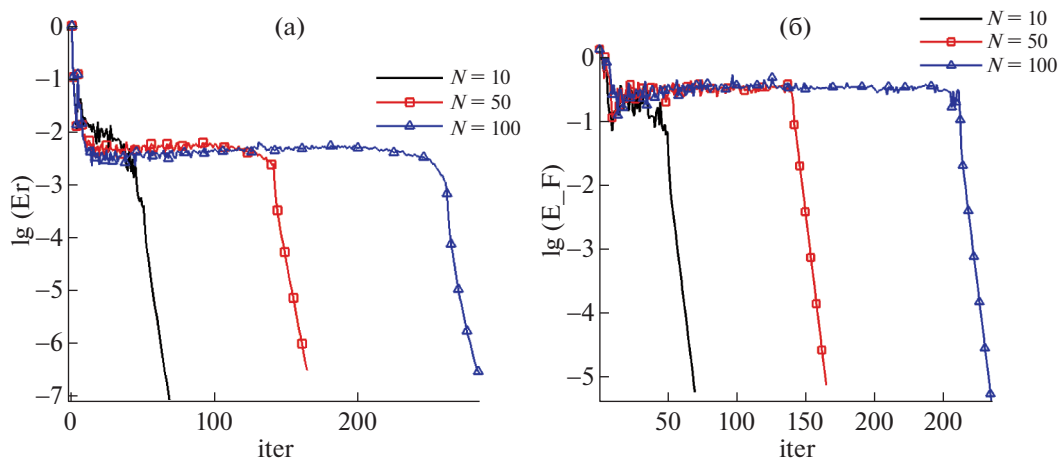
Как видно из табл. 2, значение iter_out значительно возрастает при увеличении количества таблеток. При этом после того, как в контактных парах таблетка/таблетка устанавливается окончательная конфигурация, последующее количество итераций, необходимое для достижения фиксированных значений E_r и E_F^c , для разных расчетов меняется незначительным образом. Для большого количества таблеток время расчета увеличивается по зависимости, близкой к квадратичной ($O(N^2)$).

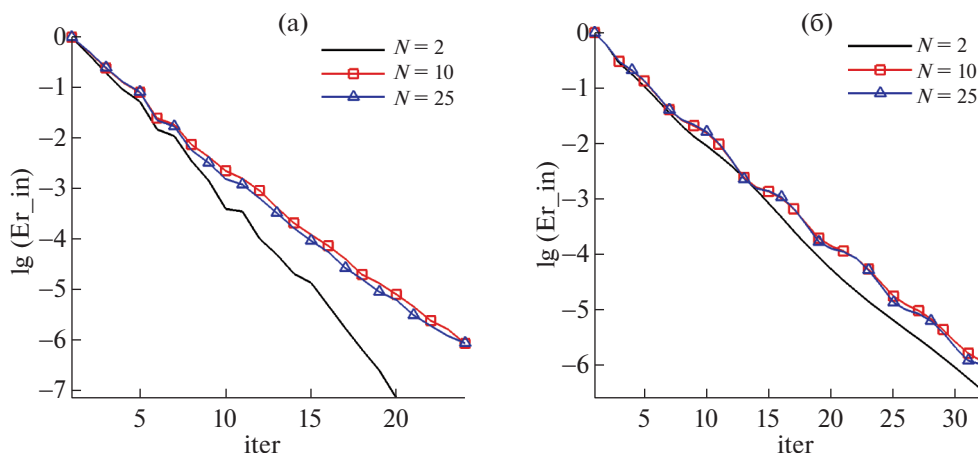
Поскольку для расчетов с большим количеством таблеток происходит значительное смещение узлов сетки на поверхности оболочки относительно поверхностных элементов таблеток, а в местах контакта углов фасок с оболочкой возникают концентраторы напряжений, то для стабилизации сходимости итерационного процесса в определенный момент расчета выполнялась фиксация положения сходственных точек для узлов, находящихся в контакте, и соответствующих им ячеек Дирихле.

Для большей наглядности на фиг. 2 приведены графики зависимости $\lg(E_r)$ (фиг. 2а) и $\lg(E_F^c)$ (фиг. 2б) от номера итерации iter для расчетов с $N = 10, 50, 100$.

Как уже упоминалось, описанный выше алгоритм решения системы уравнений (13) является строго последовательным. Проведены дополнительные расчеты, в которых для решения (13) применялись два варианта алгоритма с предобуславливателем $[B_{as2}]$, основанном на использовании аддитивного метода Шварца с грубосеточной коррекцией: метод Ричардсона (20) для $\alpha = 0.4$ и метод сопряженных невязок (22).

В аддитивном методе Шварца расчетная область разбита на подобласти без перекрытия $\bar{\Omega}_1, \dots, \bar{\Omega}_M$, $M = N + 1$. На фиг. 1б показана расчетная область с разбиением на подобласти для случая $N = 2$, $M = 3$. Подобласть $\bar{\Omega}_1$ является объединением нижней половины нижней таблетки и соответ-

Фиг. 2. Графики зависимости от номера итерации $\lg(E_r)$ (а) и $\lg(E_F^c)$ (б).



Фиг. 3. Графики зависимости $\lg(E_r^{\text{in}})$ от номера итерации: (а) — метод сопряженных невязок, (б) — метод Ричардсона.

ствующего ей участка оболочки, $\bar{\Omega}_M$ является объединением верхней половины верхней таблетки и соответствующего ей участка оболочки, для других случаев $\bar{\Omega}_k$, $k = 2, \dots, M-1$, является объединением верхней половины k -й таблетки, нижней половины $(k+1)$ -й таблетки и соответствующего им участка оболочки. Подобласть $\bar{\Omega}_k$ получается объединением подобласти $\bar{\Omega}_k$ и дополнительных участков соседних подобластей $\bar{\Omega}_{k-1}$ и $\bar{\Omega}_{k+1}$ с размерами, задаваемыми выбранным коэффициентом относительного перекрытия (обычно 0.2). Грубая сетка строилась для двух тел: одно тело соответствовало объединению всех таблеток без учета их фасок, второе тело совпадало с оболочкой. В радиальном направлении помещалась одна ячейка, в осевом направлении размер ячеек брался равным половине высоты таблеток.

На фиг. 3 показаны графики изменения десятичного логарифма величины $E_r^{\text{in}} = \|\{r\}^{(\text{iter})}\|_2 / \|\{r\}^{(0)}\|_2$ в зависимости от номера внутренней итерации iter для первой внешней итерации ($m = 0$) в расчетах с количеством таблеток 2, 10, 25.

Как видно из графиков, благодаря грубосеточной коррекции, количество внутренних итераций, необходимых для решения системы уравнений, полученной для различного количества таблеток, отличается незначительно. При этом локальные задачи для каждой подобласти (и для грубой сетки) можно решать независимо друг от друга. Поэтому данный подход может быть более эффективным при использовании параллельных вычислений.

Рассмотрим применение двухуровневого аддитивного метода Шварца. В этом случае в каждой из введенных подобластей нужно решать нелинейную задачу. Для подобластей $\bar{\Omega}_1$, $\bar{\Omega}_M$ решаются задачи контакта двух термоупругих тел (контактная пара таблетка/оболочка), для остальных подобластей — задачи контакта трех термоупругих тел (одна контактная пара таблетка/таблетка и две контактные пары таблетка/оболочка). Все локальные задачи решаются независимо друг от друга. Для всех контактных пар итерационный параметр $\theta_j = 0.5$. При решении локальных задач число внутренних итераций всегда равнялось двум.

Корректировка конфигурации контактных поверхностей проводилась только в конце внешних итераций, на внутренних итерациях конфигурация считалась неизменной. Также в конце внешней итерации учитывалось изменение участков оболочки, которые входят в подобласти $\bar{\Omega}_k$ (они всегда должны быть расположены напротив соответствующих участков топливных таблеток).

На грубой сетке для каждого из двух тел отдельно решалась линейная задача упругости, значения модуля Юнга и коэффициента Пуассона для эффективного материала определялись путем усреднения значений соответствующих характеристик в элементах базовой сетки, которые попадали в рассматриваемую ячейку грубой сетки. Нижние торцы тел считались закрепленными по оси. На внешней поверхности первого тела и на внутренней поверхности второго тела стави-

Таблица 3. Количество итераций для двухуровневого метода Шварца

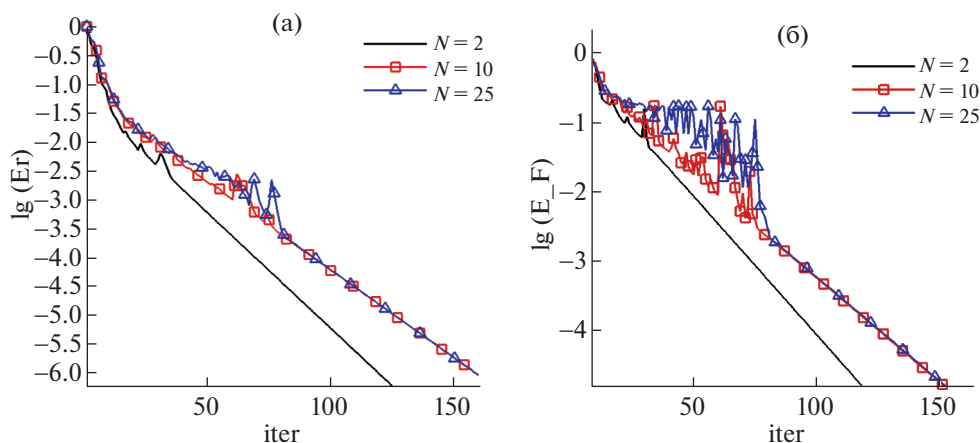
N	Время, с	iter_out	iter $E_r = E_F^c = 10^{-3}$	iter $E_r = E_F^c = 10^{-4}$	iter $E_r = E_F^c = 10^{-5}$	iter $E_r = E_F^c = 10^{-6}$
2	65	22	45 (74)	70 (99)	95 (124)	120 (—)
4	141	26	58 (89)	86 (119)	117 (150)	148 (—)
10	351	28	68 (92)	93 (125)	127 (158)	159 (—)
25	910	28	71 (93)	93 (126)	126 (159)	159 (—)
50	1865	28	78 (91)	91 (124)	123 (158)	155 (—)
100	4127	28	77 (93)	87 (126)	119 (160)	152 (—)

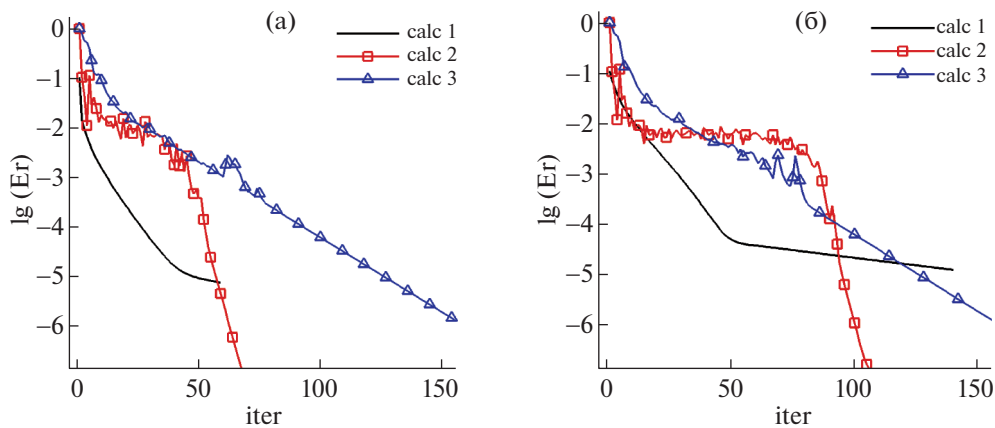
лось условие Дирихле для радиального перемещения. Значения приращений в узлах грубой сетки получались путем усреднения приращений радиальных перемещений с предыдущей итерации в узлах основной сетки, расположенных на контактной границе таблетка/оболочка и относящихся к соответствующим ячейкам грубой сетки.

В табл. 3 приведены время расчетов и количество внешних итераций (аналогично табл. 2) для различного количества таблеток в расчетной области. Из табл. 3 видно, что при увеличении количества таблеток от 2 до 10 наблюдается увеличение количества итераций, но в дальнейшем процесс стабилизируется: увеличение количества таблеток вплоть до 100 не приводит к существенному изменению количества итераций. Время расчета увеличивается по зависимости, близкой к линейной ($O(N)$). Поэтому, хотя для моделирования нескольких таблеток данный алгоритм требует существенно больших затрат, чем «mortar-метод» и метод Неймана–Дирихле, но для расчета большого количества таблеток он может оказаться существенно более эффективным.

Для большей наглядности на фиг. 4 приведены графики зависимости $\lg(E_r)$ (фиг. 2а) и $\lg(E_F^c)$ (фиг. 2б) от номера итерации iter для расчетов с $N = 2, 10, 25$.

На фиг. 5 приведены графики зависимости $\lg(E_r)$ от номера итерации iter в расчетах тремя методами: «mortar-методом» (calc 1), методом Неймана–Дирихле (calc 2) и двухуровневым аддитивным методом Шварца (calc 3). На фиг. 5а представлены графики для 10 таблеток, а фиг. 5б — для 25 таблеток. Для первого метода на каждой итерации решается система уравнений (9). Несмотря на то что для решения используется прямой метод, в силу плохой обусловленности матрицы системы без процедуры корректировки конфигурации контактных поверхностей полученное численное решение имеет весьма низкую точность. Для второго метода на каждой итерации нужно решать систему уравнений (13), это можно сделать, в том числе, с помощью предобусловливателя, использующего аддитивный метод Шварца. Для третьего метода на каждой итерации

Фиг. 4. Графики зависимости от номера итерации $\lg(E_r)$ (а) и $\lg(E_F^c)$ (б).



Фиг. 5. Графики зависимости $\lg(E_r)$ от номера итерации: (а) – $N = 10$, (б) – $N = 25$.

выполняются по два решения локальной задачи (25) в каждой подобласти и решение задачи на грубой сетке; все эти задачи могут быть выполнены независимо друг от друга. В расчетах для 10 и 25 таблеток для первых двух методов количество итераций меньше, чем для третьего. Но для большего количества таблеток для первых двух методов требуемое количество итераций продолжит значительно увеличиваться, а для третьего метода останется практически неизменным.

В табл. 4 приведено количество внешних итераций двухуровневого аддитивного метода Шварца для случая, когда в расчетной области находятся 10 таблеток и используются сетки с разными шагами.

Из табл. 4 видно, что уменьшение шагов сетки не приводит к существенному изменению количества внешних итераций.

В табл. 5 приведено количество внешних итераций для 10 таблеток в расчетах, для которых в контактных парах таблетка/таблетка запрещен выход из контакта вне определенной зоны даже при возникновении растягивающих напряжений (по силовому критерию), но конфигурация контактных пар таблетка/оболочка по-прежнему менялась. Таблица еще раз наглядно демонстрирует, что чем больше меняется конфигурация контактных поверхностей, тем больше требуется итераций для достижения требуемой точности.

Проведена также серия расчетов, в которых на всех контактных поверхностях использованы условия прилипания. В этом случае грубая сетка строилась в прямоугольной области, включающей в себя и таблетки, и оболочку. В радиальном направлении помещалась одна ячейка, в осе-

Таблица 4. Количество итераций аддитивного метода Шварца для разных шагов сеток

Шаги сетки	iter_out	$E_r = E_F^c = 10^{-3}$	$E_r = E_F^c = 10^{-4}$	$E_r = E_F^c = 10^{-5}$	$E_r = E_F^c = 10^{-6}$
$h_1 = 0.25, h_2 = 0.165$	28	68 (92)	93 (125)	127 (158)	159 (—)
$h_1 = 0.125, h_2 = 0.0825$	30	75 (96)	96 (131)	131 (166)	166 (—)
$h_1 = 0.06125, h_2 = 0.04125$	32	79 (99)	100 (136)	137 (173)	173 (—)

Таблица 5. Количество итераций аддитивного метода Шварца для различных конфигураций контакта таблетка/таблетка (10 таблеток)

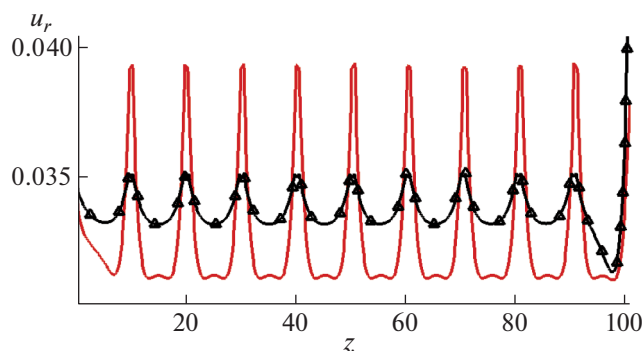
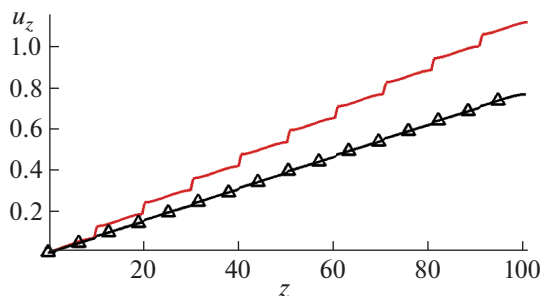
Доля поверхностных элементов, вышедших из контакта	iter_out	$E_r = E_F^c = 10^{-3}$	$E_r = E_F^c = 10^{-4}$	$E_r = E_F^c = 10^{-5}$	$E_r = E_F^c = 10^{-6}$
0	0	16 (20)	21 (25)	25 (29)	30 (36)
1/2	10	36 (44)	46 (59)	59 (70)	71 (—)
5/6	28	68 (92)	93 (125)	127 (158)	159 (—)

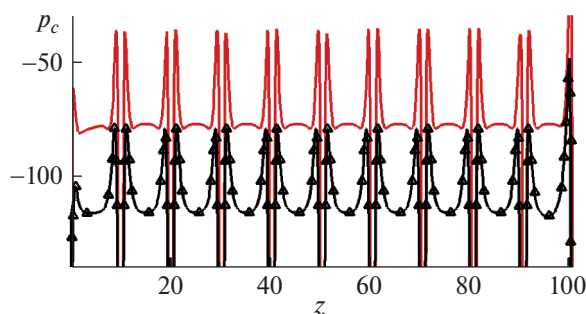
Таблица 6. Количество итераций аддитивного метода Шварца для различного количества таблеток, условия прилипания

N	iter_out	$E_r = E_F^c = 10^{-3}$	$E_r = E_F^c = 10^{-4}$	$E_r = E_F^c = 10^{-5}$	$E_r = E_F^c = 10^{-6}$
2	8	15 (21)	21 (26)	26 (32)	32 (38)
10	9	15 (21)	21 (27)	27 (33)	33 (39)
25	9	17 (24)	23 (30)	29 (36)	35 (42)
100	9	20 (30)	29 (34)	34 (37)	40 (—)

вом направлении размер ячеек брался равным половине высоты таблетки. Разделение расчетной области на подобласти проводилось идентично предыдущим случаям, количество внутренних итераций в локальных задачах не превышало пяти. Значения итерационных параметров в методе Неймана–Дирихле $\theta_j = 0.25$. В контактных парах таблетка/таблетка происходил выход из контакта примерно половины элементов поверхностной сетки, но расхождение поверхностей было существенно меньше, чем в расчетах с условием скольжения. Конфигурация контактных пар таблетка/оболочка в процессе расчета не меняется. В табл. 6 приведено количество внешних итераций для различного количества таблеток в расчетной области. Расчеты проводились для фиксированных шагов сетки (в таблетке – 0.25, в оболочке – 0.125) и фиксированного коэффициента относительного перекрытия (0.1).

В качестве иллюстрации полученных результатов приведем некоторые графики для расчетов с 10 таблетками ($h_1 = 0.125$, $h_2 = 0.08$) для условия скольжения (графики без маркера) и условия прилипания (графики с маркерами). Координаты и перемещения на графиках нормированы на 1 мм, а напряжения – на 1 МПа. На фиг. 6 приведены графики распределения радиального перемещения, на фиг. 7 – осевого перемещения, а на фиг. 8 – контактного давления в узлах, расположенных на внешней поверхности таблеток. Аналогичные графики для узлов, расположен-

**Фиг. 6.** Радиальные перемещения в узлах на внешней поверхности таблеток.**Фиг. 7.** Осевые перемещения в узлах на внешней поверхности таблеток.



Фиг. 8. Контактные давления в узлах на внешней поверхности таблеток.

ных на внутренней поверхности оболочки, визуально неотличимы от показанных (за исключением участков между фасками, где контакта не происходит).

Из графиков следует, что в расчете с условием прилипания в среднем радиальные перемещения на 10% больше, максимальное осевое перемещение на 30% меньше, а контактное давление на 40% больше, чем соответствующие величины в расчете с условием скольжения без трения. Вблизи углов фасок находятся концентраторы напряжений. Значения давлений в этих зонах могут на порядок превосходить значения давлений в центральной части контактной поверхности, они увеличиваются при измельчении сетки, но интегральная контактная сила, действующая на поверхность, остается практически неизменной.

ЗАКЛЮЧЕНИЕ

Представлены результаты применения различных алгоритмов для численного моделирования термомеханического состояния участка “твэла” с учетом контактного взаимодействия топливных таблеток друг с другом и с оболочкой. Задача решена в термоупругом осесимметричном приближении, количество таблеток N в расчетной области изменялось от 2 до 100, на контактных поверхностях использованы либо условия прилипания, либо условия скольжения без трения. В результате нагрева топливного столба происходит существенное изменение конфигурации контактных поверхностей. Анализ полученных результатов позволил сделать вывод, что для большого количества таблеток при использовании “mortar-метода” в сочетании с прямым методом решения системы линейных уравнений время расчета увеличивается по зависимости, близкой к $O(N^3)$, в методе Неймана–Дирихле — по зависимости, близкой к $O(N^2)$, а для двухуровневого аддитивного метода Шварца — по зависимости, близкой к $O(N)$. В то же время для расчетов с небольшим количеством таблеток первые два алгоритма являются гораздо более быстрыми, чем третий. Для рассмотренной задачи при определенном выборе входных параметров (шаг грубой сетки, величина захлеста, значение итерационных параметров) для двухуровневого аддитивного метода Шварца количество внешних итераций, необходимое для достижения заданной точности, незначительно меняется при использовании различных шагов расчетной сетки и различного количества топливных таблеток. Решение локальных задач в каждой подобласти может быть выполнено независимым (в том числе и параллельным) способом. Указанные свойства делают данный алгоритм эффективным инструментом для решения аналогичных мультиконтактных задач с подробными расчетными сетками.

Авторы выражают глубокую благодарность П.С. Аронову за проведенные расчеты контактных задач с использованием mortar-метода.

СПИСОК ЛИТЕРАТУРЫ

1. Toselli A., Widlund O. Domain decomposition methods — algorithms and theory. Berlin-Heidelberg: Springer-Verlag, 2005. 450 p.
2. Марчук Г.И. Методы вычислительной математики. М.: Наука, 1989. 608 с.
3. Wriggers P. Computational contact mechanics. Berlin-Heidelberg: Springer-Verlag, 2006. 520 p.
4. Eck C., Wohlmuth B. Convergence of a contact — Neumann iteration for the solution of two-body contact problems // Math. Model. and Meth. Appl. Sci. 2003. V. 13. № 8. P. 1103—1118.

5. *Bayada G., Sabil J., Sassi T.* Neumann-Neumann domain decomposition algorithm for the Signorini problem // *Appl. Math. Lett.* 2004. V. 17. P. 1153–1159.
6. *Цвик Л.Б.* Принцип поочередности в задачах о сопряжении и контакте твердых деформируемых тел // *Прикл. механ.* 1980. Т. 16. № 1. С. 13–18.
7. *Галанин М.П., Крупкин А.В., Кузнецов В.И., Лукин В.В., Новиков В.В., Родин А.С., Станкевич И.В.* Математическое моделирование контактного взаимодействия системы термоупругих тел методом Шварца для многомерного случая // *Изв. вузов. Машиностр.* 2016. № 12 (681). С. 9–20.
8. *Федоренко Р.П.* Введение в вычислительную физику. Долгопрудный: Изд-кий Дом “Интеллект”, 2008. 504 с.
9. *Мартыненко С.И.* Универсальная многосеточная технология. М.: ИПМ им. М.В. Келдыша, 2013. 243 с.
10. *Галанин М.П., Савенков Е.Б.* Методы численного анализа математических моделей. М.: Изд-во МГТУ им. Н.Э. Баумана, 2010. 591 с.
11. *Бате К.-Ю.* Методы конечных элементов. М.: Физматлит, 2010. 1024 с.
12. *Bayada G., Sabil J., Sassi T.* Convergence of a Neumann-Dirichlet algorithm for two-body contact problems with non local Coulomb’s friction law // *Math. Model. and Numer. Analys.* 2008. V. 42. P. 243–262.
13. *Тимошенко С. П., Гудьер Дж.* Теория упругости. М.: Наука, 1975. 576 с.
14. *Быченков Ю.В., Чижонков Е.В.* Итерационные методы решения седловых задач. М.: БИНОМ, 2010. 349 с.
15. *Аронов П.С., Галанин М.П., Родин А.С.* Численное решение задачи контактного взаимодействия элементов твэла с помощью mortar-метода и метода декомпозиции области // *Вестн. МГТУ им. Н.Э. Баумана. Сер. Естествен. науки.* 2021. № 3. С. 4–22.
16. *Korneev V.G., Langer U.* Dirichlet–Dirichlet domain decomposition methods for elliptic problems. *Word Sci. Publ.*, 2015. 463 p.
17. *Ильин В.П., Гурьева Я.Л.* О некоторых параллельных методах и технологии декомпозиции областей // *Зап. научн. семинаров ПОМИ.* 2014. Т. 428. С. 89–106.
18. *Ильин В.П., Ицкович Е.А.* О методах полусопряженных направлений с динамическим предобуславливателем // *Сиб. журнал индустр. матем.* 2007. Т. 10. № 4. С. 41–54.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 519.634

МОДЕЛИРОВАНИЕ ЛАМИНАРНО-ТУРБУЛЕНТНОГО ПЕРЕХОДА
С ПРИМЕНЕНИЕМ ГИБРИДНЫХ РАЗНОСТНЫХ СХЕМ¹⁾© 2022 г. И. В. Егоров^{1,2}, Н. К. Нгуен^{1,*}¹ 141701 Долгопрудный, М.о., Институтский пер., 9, МФТИ, Россия² 140180 Жуковский, М.о., Жуковского ул., 1, ЦАГИ, Россия*e-mail: nguyennhucan528@gmail.com

Поступила в редакцию 30.06.2021 г.

Переработанный вариант 14.09.2021 г.

Принята к публикации 16.12.2021 г.

Предложена простая гибридная разностная схема, которая применима в течениях с ударными волнами. Схема остается монотонной вблизи ударных волн и переключается на низкодиссипативную немонотонную разностную схему в гладких областях течения, регулируя количество численной диссипации. Это достигается путем плавного уменьшения монотонизирующей поправки до заданного порогового уровня в зависимости от индикатора гладкости решения. Для примера рассмотрена задача моделирования ламинарно-турбулентного перехода в сверхзвуковом пограничном слое над плоской пластиной при числе Маха 3. Результаты расчетов сопоставлены с результатами других работ, в которых применялись как диссипативные, так и низкодиссипативные схемы. Сопоставляются спектральные характеристики возмущений в области их линейного и нелинейного развития, а также структура переходного течения и характеристики осредненного пограничного слоя. Библ. 7. Фиг. 17.

Ключевые слова: гибридная схема, монотонная схема, низкодиссипативная схема, численное моделирование, возмущение, ламинарно-турбулентный переход, сверхзвуковой пограничный слой, нелинейный распад.

DOI: 10.31857/S0044466922040056

ВВЕДЕНИЕ

Разработка перспективных летательных аппаратов включает детальное исследование обтекаемых аэродинамических форм. Экспериментальная часть, как правило, сопряжена не только с финансовыми затратами на постановку и проведение эксперимента в аэродинамических трубах, но и с ограниченностью получаемой в эксперименте информации. Численное моделирование нестационарного движения сжимаемого газа позволяет описывать течения около тел произвольной конфигурации, выявлять тонкую структуру рассматриваемых явлений и получать результаты, которые затруднительно получить экспериментальным путем. По результатам моделирования рассчитываются аэротермодинамические коэффициенты — коэффициенты давления, трения и теплообмена. Последние становятся критически важными для случая больших сверхзвуковых и гиперзвуковых скоростей потока, в особенности, когда происходит турбулизация течения и вместе с ней в разы усиливаются коэффициенты трения и теплопередачи к поверхности.

Данные трубных испытаний по проблеме ламинарно-турбулентного перехода (ЛТП) непостоянны, так как зависят от фона возмущений в потоке конкретной аэродинамической установки. Ввиду этого особую ценность представляют результаты прямого численного моделирования течений на режиме ламинарно-турбулентного перехода, когда такой фон можно строго контролировать. К сожалению, пространственно-временные затраты для такого моделирования велики для режимов, близких к практически важным. Использование высокопроизводительных многопроцессорных вычислительных кластеров (супер ЭВМ) позволяет проводить лишь отдельные

¹⁾ Работа выполнена при финансовой поддержке РНФ (код проекта 21-19-00307) в МФТИ с использованием оборудования центра коллективного пользования “Комплекс моделирования и обработки данных исследовательских установок мега-класса” НИЦ “Курчатовский институт”, <http://ckp.nrcki.ru/>.

расчеты, целью которых является исследование линейных и нелинейных механизмов, лежащих в основе ЛТП.

Повышение точности используемых вычислительных схем помогает снизить требования к пространственному разрешению течения, когда напрямую моделируется именно стадия ЛТП и начального участка турбулентного движения (“молодая” турбулентность). Однако повышение точности дестабилизирует численные алгоритмы, например, в присутствии сильных ударных волн. Поэтому применение монотонных схем типа Годунова может оказаться оправданным. Монотонизация решения достигается путем рассмотрения разрывных решений на базе задачи Римана о распаде произвольного разрыва. При этом в аппроксимацию конвективных потоковых членов добавляется монотонизирующая поправка, которая косвенно вводит дополнительную диссипацию и тем самым стабилизирует численное решение в областях сильных разрывов и подавляет малые возмущения [1]. Последние могут играть существенную роль в процессе ЛТП.

Данная работа является развитием работы [1]: изучается возможность уменьшения диссипации такого рода в областях, где она не является необходимой (например, внутри безградиентного пограничного слоя). Результаты моделирования сопоставляются с результатами работы [1], полученными на диссипативной схеме (описание приведено ниже), и с результатами работы [2], полученными с применением низкодиссипативного метода. В работе [2] использована схема четвертого порядка точности по продольному и нормальному к поверхности направлениям, а также спектральный метод в боковом направлении. Интегрирование по времени в [2] проводится по методу Рунге—Кутты четвертого порядка аппроксимации.

1. ПОСТАНОВКА ЗАДАЧИ

1.1. Численный метод

В настоящей работе использована методика численного интегрирования полных уравнений Навье—Стокса, описанная в [3]. Уравнения решаются в криволинейной системе координат (ξ, η, ζ) в безразмерном дивергентном виде:

$$\frac{\partial \mathbf{Q}}{\partial t} + \frac{\partial \mathbf{E}}{\partial \xi} + \frac{\partial \mathbf{G}}{\partial \eta} + \frac{\partial \mathbf{F}}{\partial \zeta} = 0.$$

Декартовы координаты $x^* = xL$, $y^* = yL$, $z^* = zL$ отнесены к характерному масштабу длины L , время $t^* = tL/V_\infty$ — к характерному масштабу времени L/V_∞ , компоненты вектора скорости $u^* = uV_\infty$, $v^* = vV_\infty$, $w^* = wV_\infty$ — к модулю вектора скорости набегающего потока V_∞ , давление $p^* = p(\rho_\infty V_\infty^2)$ — к удвоенному скоростному напору в набегающем потоке, остальные газодинамические переменные — к их значениям в набегающем потоке. Звездочка в верхнем индексе обозначает размерную величину; если звездочка отсутствует, переменная полагается безразмерной. Символ “ ∞ ” обозначает привязку к набегающему потоку. Для удобства сравнения с результатами других работ будут использоваться размерные переменные.

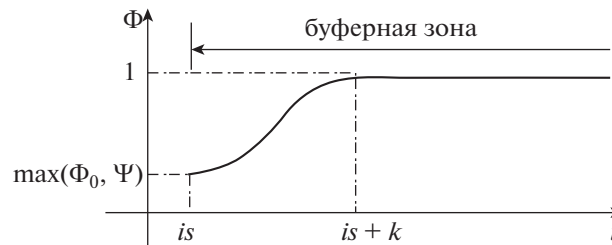
Для аппроксимации дифференциальных уравнений используются полностью неявный метод конечного объема и схема второго порядка аппроксимации по времени:

$$\frac{3\mathbf{Q}_{i,j,k}^{n+1} - 4\mathbf{Q}_{i,j,k}^n + \mathbf{Q}_{i,j,k}^{n-1}}{\Delta t} + \frac{\mathbf{E}_{i+\frac{1}{2},j,k}^{n+1} - \mathbf{E}_{i-\frac{1}{2},j,k}^{n+1}}{h_\xi} + \frac{\mathbf{G}_{i,j+\frac{1}{2},k}^{n+1} - \mathbf{G}_{i,j-\frac{1}{2},k}^{n+1}}{h_\eta} + \frac{\mathbf{F}_{i,j,k+\frac{1}{2}}^{n+1} - \mathbf{F}_{i,j,k-\frac{1}{2}}^{n+1}}{h_\zeta} = 0.$$

При определении конвективных потоковых величин на гранях ячейки сетки, например, $\mathbf{E}_{i+\frac{1}{2},j,k}^{n+1}$, потоки расщепляются по направлениям. Для каждого направления определяется Матрица Якоби ($A = \partial \mathbf{E} / \partial \mathbf{Q}$ для направления ξ), которая диагонализуется в виде $A = B \Lambda B^{-1}$, где B — матрица, составленная из правых собственных векторов, а Λ — диагональная матрица собственных значений матрицы A . Конвективная составляющая потоковых величин $\mathbf{E}$, $\mathbf{G}$, $\mathbf{F}$ на грани ячейки аппроксимируется с использованием монотонной схемы типа Годунова (индексы i, j, k, n опущены для читаемости):

$$\mathbf{E}_{i+\frac{1}{2}} = \frac{1}{2} \left[\mathbf{E}(\mathbf{Q}_L) + \mathbf{E}(\mathbf{Q}_R) - \Phi \times B_{LR} \Lambda(\Phi(\lambda_{LR})) B_{LR}^{-1} (\mathbf{Q}_R - \mathbf{Q}_L) \right],$$

где нижними индексами L и R отмечены величины, рассчитанные слева и справа на рассматриваемой грани с использованием восстановленных с помощью процедуры реконструкции значе-



Фиг. 1. График параметра включения диссипативной компоненты в буферной зоне для случая гибридных схем.

ний газодинамических переменных. Например, для грани $i + 1/2$ индекс L соответствует ячейке i , а индекс R — ячейке $i + 1$. Нижним индексом LR отмечены величины, вычисляемые с помощью приближенного метода Роу решения задачи Римана о распаде произвольного разрыва. Модификация собственных значений $\phi(\lambda)$ обеспечивает физически корректное изменение энтропии на разрывах. В работе использована процедура реконструкции по методу WENO-3.

Функция Φ определяет степень немонотонности схемы и принимает значения от 0 до 1. Для исходной монотонной схемы $\Phi = 1$. Для гибридной схемы в расчетной области $\Phi = \max(\Phi_0, \Psi)$, где Φ_0 — некоторая постоянная величина, Ψ — функция-индикатор Джеймсона [4]:

$$\Psi = \frac{(\operatorname{div} \mathbf{V})^2}{(\operatorname{div} \mathbf{V})^2 + (\operatorname{rot} \mathbf{V})^2 + \delta}, \quad \delta = 10^{-20}, \quad (1)$$

где $\mathbf{V}$ — локальный вектор скорости, $(\operatorname{rot} \mathbf{V})^2$ — квадрат модуля вектора $\operatorname{rot} \mathbf{V}$.

Величина ϵ выбирается малым положительным числом и позволяет устранить численную особенность деления на ноль в областях, где $(\operatorname{div} \mathbf{V})^2$ и $(\operatorname{rot} \mathbf{V})^2$ оба равны нулю. Величина Ψ меняется от 0 при $(\operatorname{div} \mathbf{V})^2 \ll (\operatorname{rot} \mathbf{V})^2$ до 1 в случае $(\operatorname{div} \mathbf{V})^2 \gg (\operatorname{rot} \mathbf{V})^2$. В частности, в пограничном слое $\Psi \rightarrow 0$, а в окрестности ударных волн $\Psi \rightarrow 1$.

Перед выходной границей для подавления возмущений организуется буферная область (см. фиг. 1) с сеточным разрежением. В этой области вниз по потоку плавно восстанавливается монотонная схема (Φ стремится к единице) по формуле:

$$\Phi = \begin{cases} 1 - (1 - \max(\Phi_0, \Psi)) \left(1 - \frac{i - is}{k}\right), & \text{если } is \leq i \leq is + k, \\ 1, & \text{если } i > is + k, \end{cases}$$

где i — продольный сеточный индекс, is — начальный индекс буферной зоны, k — константа. В настоящей работе значение $k = 3$.

В криволинейной системе координат (ξ, η, ζ) расчетная сетка является равномерной по каждому из направлений с шагами h_ξ , h_η , h_ζ . Для аппроксимации диффузионной составляющей векторов $\mathbf{E}$, $\mathbf{G}$, $\mathbf{F}$ на грани элементарной ячейки применена разностная схема типа центральных разностей второго порядка точности:

$$\begin{aligned} \left. \frac{\partial q}{\partial \xi} \right|_{i+\frac{1}{2}, j, k} &= \frac{1}{h_\xi} (q_{i+1, j, k} - q_{i, j, k}), \\ \left. \frac{\partial q}{\partial \eta} \right|_{i+\frac{1}{2}, j, k} &= \frac{1}{4h_\eta} (q_{i+1, j+1, k} + q_{i, j+1, k} - q_{i+1, j-1, k} - q_{i, j-1, k}), \\ \left. \frac{\partial q}{\partial \zeta} \right|_{i+\frac{1}{2}, j, k} &= \frac{1}{4h_\zeta} (q_{i+1, j, k+1} + q_{i, j, k+1} - q_{i+1, j, k-1} - q_{i, j, k-1}), \end{aligned}$$

где q — любая из неконсервативных (“примитивных”) зависимых переменных задачи (u, v, w, p, T) , декартовы компоненты скорости, давление, температура.

После аппроксимаций уравнений Навье—Стокса и граничных условий интегрирование исходных уравнений в частных производных сводится к решению системы нелинейных алгебраических уравнений $\mathbf{R}(\mathbf{U}) = 0$, где $\mathbf{R}$ — оператор дискретизации, вычисляющий вектор невязки согласно аппроксимации уравнений, $\mathbf{U}$ — вектор искомых неконсервативных переменных (u, v, w, p, T) (компоненты скорости, давление, температура) во всех узлах расчетной сетки. Длина вектора $\mathbf{U}$ составляет $n_q N$, где N — полное число узлов расчетной сетки, включая граничные, n_q — число неизвестных в каждом узле ($n_q = 5(u, v, w, p, T)$ в трехмерной постановке и $n_q = 4(u, v, p, T)$ в двухмерной). Система сеточных уравнений $\mathbf{R}(\mathbf{U}) = 0$ решается с помощью модифицированного метода Ньютона—Рафсона

$$\mathbf{U}^{[k+1]} = \mathbf{U}^{[k]} - \tau^{[k+1]} (\mathbf{J}^{[k_0]})^{-1} \mathbf{R}(\mathbf{U}^{[k]}),$$

где k, k_0 — номера итераций по нелинейности, $k_0 \leq k$, $\mathbf{J}^{[k_0]} = (\partial \mathbf{R} / \partial \mathbf{U})^{[k_0]}$ — матрица Якоби системы нелинейных уравнений (матрица Якоби отображения), $\mathbf{R}(\mathbf{U}^{[k]})$ — вектор невязки, τ — параметр регуляризации. Здесь выражение $(\mathbf{J}^{[k_0]})^{-1} \mathbf{R}(\mathbf{U}^{[k]}) \equiv \mathbf{Y}^{[k]}$ является решением линейной системы уравнений

$$(\mathbf{J}^{[k_0]}) \mathbf{Y}^{[k]} = \mathbf{R}(\mathbf{U}^{[k]}).$$

Параметр регуляризации метода Ньютона относительно начального приближения $\tau^{[k]}$ определяется по формуле

$$\tau^{[k+1]} = \frac{(\mathbf{Y}^{[k]} - \mathbf{Y}^{[k-1]}) \mathbf{Y}^{[k]}}{(\mathbf{Y}^{[k]} - \mathbf{Y}^{[k-1]})^2}.$$

По мере сходимости итерационного [5] процесса $\tau^{[k]} \rightarrow 1$, а скорость сходимости теоретически стремится к квадратичной.

Получение аналитического вида матрицы Якоби $\mathbf{J}$ для рассматриваемой численной схемы, включающей решение задачи Римана о распаде разрыва, представляется весьма трудоемким. В данной работе применяется универсальный метод формирования матрицы $\mathbf{J}^{[k_0]} = (\partial \mathbf{R} / \partial \mathbf{U})^{[k_0]}$ на итерации по нелинейности k_0 с помощью процедуры конечных приращений вектора невязки R по вектору искомых переменных $\mathbf{U}$. При этом m -й столбец матрицы $\mathbf{J}^{[k_0]}$ вычисляется в виде

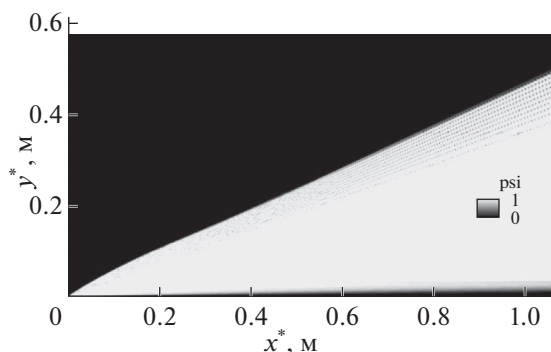
$$\mathbf{J}_m^{[k_0]} = \frac{\mathbf{R}(\mathbf{U}^{[k_0]} + \alpha \mathbf{e}_m) - \mathbf{R}(\mathbf{U}^{[k_0]})}{\alpha}, \quad \alpha = 10^{-8}, \quad m = 1, \dots, n_q N,$$

где $\mathbf{e}_m$ — единичный вектор длины $n_q N$, полностью состоящий из 0, кроме единственной 1 на позиции m . Такая методика вычисления Якобиана применима к произвольной системе сеточных уравнений.

1.2. Параметры потока и условия расчетов

Рассматривается номинально двухмерное течение над заостренной плоской пластиной при числе Маха набегающего потока $M_\infty = 3$ и температуре набегающего потока $T_\infty = 103.6$ К. Расчет эволюции возмущений проводится в подобласти; процедура расчета аналогична процедуре, описанной в [6]. Число Рейнольдса составляет $R_{l,\infty} = 2.181 \times 10^6 \text{ м}^{-1}$. Число Прандтля принимается постоянным: $\text{Pr} = \mu c_p / \lambda = 0.71$. Уравнения Навье—Стокса замыкаются уравнением состояния $\gamma M_\infty^2 p = \rho T$, где $\gamma = 1.4$ — показатель адиабаты. Динамический коэффициент молекулярной вязкости рассчитывается по формуле Сазерленда: $\mu = (1 + T_\mu) / (T + T_\mu) T^{3/2}$, где $T_\mu = T_\mu^* / T_\infty^* = 110.4 \text{ К} / 103.6 \text{ К} \approx 1.07$.

Численное интегрирование проводится в прямоугольной области, показанной ниже. На входной и верхней границах фиксируются безразмерные параметры набегающего потока:



Фиг. 2. Поле функции Джеймсона Ψ .

$(u, v, w, p, T) = (1, 0, 0, 1/\gamma M_\infty^2, 1)$. Для стационарных расчетов стенка принимается теплоизолированной, на стенке ставятся условия прилипания. Выходной границе предшествует буферная зона с укрупненными ячейками по продольной координате x и нормальной к стенке координате y для демпфирования выходящих через границу возмущений. На выходной границе накладываются мягкие условия как линейная экстраполяция примитивных переменных из расчетной области. На боковых границах $z_{\min}$ и $z_{\max}$ ставятся условия симметрии.

Расчет проводится следующим образом. Во-первых, двухмерное стационарное невозмущенное течение над плоской пластиной рассчитывается до достижения невязки величины 10^{-8} . Во-вторых, из полученного решения вырезается подобласть, в которой далее будет моделироваться развитие возмущений; на новых входных границах подобласти фиксируются газодинамические величины из расчета на первом шаге; стационарное поле устанавливается дополнительно до полной сходимости (величина невязки не превосходит 10^{-8}). В-третьих, полученное в подобласти стационарное поле дублируется в третьем поперечном направлении z ; распределение температуры по поверхности фиксируется; в пограничный слой вводятся возмущения типа “вдвух—отсос” по процедуре, описанной ниже. Нестационарный расчет проводится до установления квазистационарного режима течения.

1.3. Функция-индикатор Джеймсона

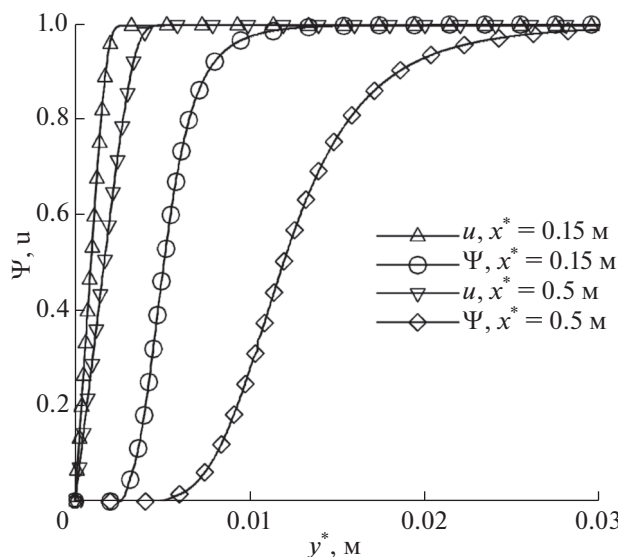
Поле функции-индикатора Джеймсона [4] рассчитано по двухмерному невозмущенному течению и представлено на фиг. 2. “Пятнистость” значения этой функции за слабой ударной волной связана с численными особенностями ее расчета по формуле (1). В этой области значения слагаемых $(\text{div } \mathbf{V})^2$ и $(\text{rot } \mathbf{V})^2$ являются малыми величинами и могут быть одного порядка.

Для расчетов настоящей работы величина Ψ принимается равной единице всюду за исключением окрестности пограничного слоя, где Ψ плавно меняется от 0 на поверхности до 1 по мере удаления от пограничного слоя.

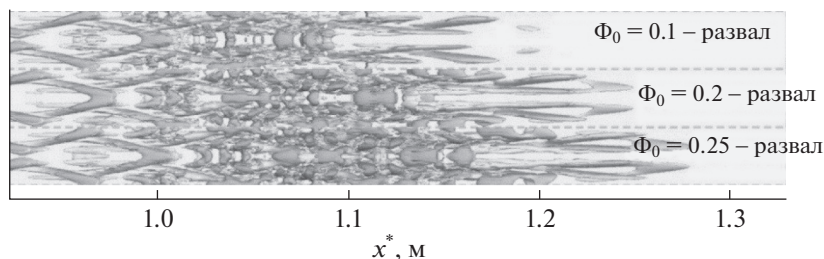
На фиг. 3 приведены профили $u(y)$ и $\Psi(y)$. Следует заметить, что область $\Psi \ll 1$ занимает порядка двух-трех толщин невозмущенного пограничного слоя. Поэтому на начальной стадии турбулизации течения пограничный слой продолжает находиться в области, где для аппроксимации конвективных составляющих потоковых величин используется низкодиссипативная схема типа центральных разностей. При проведении нестационарных расчетов форма профилей $\Psi(y)$ зафиксирована во всех сечениях $x = \text{const}$ и не зависит от текущего возмущенного поля течения.

1.4. Определение значения Φ_0

При описанном подходе не удается полностью перейти на схему типа центральных разностей ($\Phi_0 = 0$), так как при турбулизации течения проявляется численная неустойчивость и процесс расчета перестает сходиться. Эмпирически удается определить наименьшее возможное значение Φ_0 , при котором численный процесс интегрирования дифференциальных уравнений



Фиг. 3. Профили скорости и функции Джеймсона.



Фиг. 4. Развалы решения, Q-критерий=15, вид сверху, сетка 5 миллионов узлов.

остаётся устойчивым. Это значение зависит от качества сетки и от возмущений. На сетке 20 миллионов узлов $\Phi_{0,\min} = 0.35$, а на сетке 80 миллионов узлов — $\Phi_{0,\min} = 0.21$.

На фиг. 4 приведены изоповерхности Q-критерия, полученные на грубой сетке в 5 млн узлов в моменты времени, когда решение перестает сходиться, где Q — второй скалярный инвариант тензора скоростей деформаций

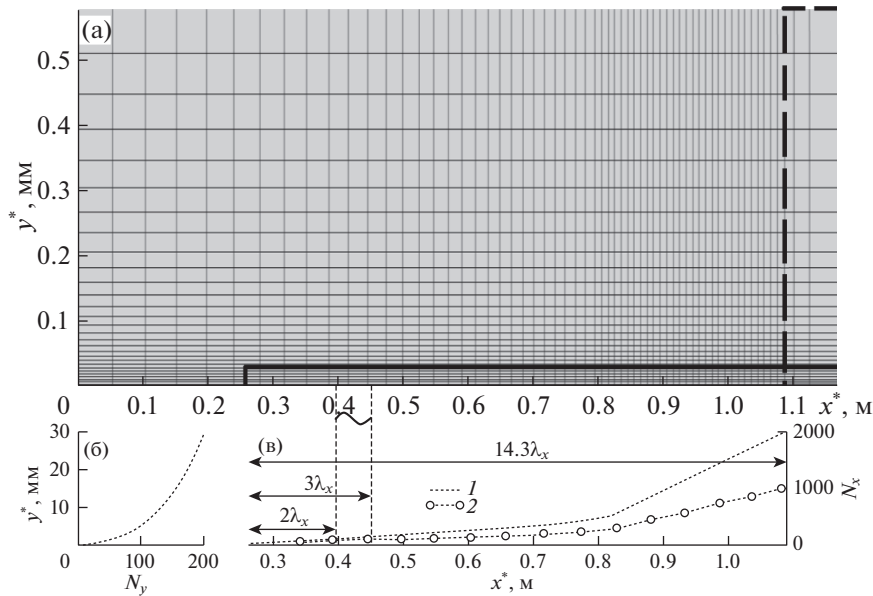
$$Q = \frac{1}{2} [|\Omega|^2 - |S|^2], \quad S = \frac{1}{2} [\nabla V + (\nabla V)^T], \quad \Omega = \frac{1}{2} [\nabla V - (\nabla V)^T].$$

Как видно, во всех случаях возмущения не успевают дойти до конца расчетной области, причем, чем менее диссипативной является схема (чем меньше величина Φ_0), тем раньше проявляются вычислительные проблемы. На рассматриваемой грубой сетке $\Phi_{0,\min} = 0.3$, однако, в этом случае амплитуда генерируемых возмущений $\epsilon = 0.003$. Ниже в настоящей работе, как и в [1] рассматривается случай $\epsilon = 0.00573$.

1.5. Генератор возмущений

Возмущения вводятся в пограничный слой при помощи генератора типа “вдув—отсос” с поверхности пластины. Генератор моделируется при $x^* \in [x_1^*, x_2^*] = [0.394, 0.452]$ м в соответствии с работами [1], [2]. В этом диапазоне нормальная составляющая вектора скорости имеет вид:

$$v(x, y = 0, z, t) = A(t)v_p(x_p)\cos(\beta_0 z)\cos(-\omega_0 t),$$



Фиг. 5. Параметры расчетной сетки: (а) — общий вид (показана каждая 10-я сеточная линия); (б) — сгущение сетки по нормали к стенке; (в) — сгущение сетки в продольном направлении: 1 — сетка 80 миллионов узлов, 2 — сетка 20 миллионов узлов.

где

$$v_p = \begin{cases} 1.5^4(1+x_p)^3(3(1+x_p)^2 - 7(1+x_p) + 4), & -1 \leq x_p \leq 0, \\ -1.5^4(1-x_p)^3(3(1-x_p)^2 - 7(1-x_p) + 4), & 0 \leq x_p \leq 1, \end{cases} \quad x_p = \frac{2x - (x_2 + x_1)}{x_2 - x_1},$$

$$A(t) = \varepsilon \begin{cases} 0, & t < 0, \\ 0.1^{((T_0-t)/(0.9T_0))^2}, & 0 \leq t \leq T_0, \\ 1, & t > T_0, \end{cases}$$

где $T_0 = 2\pi/\omega_0$, $A(t)$ — амплитуда, $\varepsilon = 0.00573$. Остальные параметры потока в области генератора вычисляются как для случая стенки без генератора. Возмущение с частотой $\omega_0^*/2\pi = f_0^* = 6.36$ кГц и волновым числом $\beta_0^* = 211.52$ м⁻¹ будем называть *фундаментальным*.

Как будет показано далее, результаты настоящих расчетов хорошо согласуются с результатами работ [1], [2] как качественно, так и количественно. Однако амплитуда возмущений ε в настоящей работе и в работе [1] отличается от амплитуды из работы [2] практически вдвое (в работе [2] $\varepsilon = 0.003$) и подобрана для совпадения положения начала ЛТП (в [1] показано, что численная диссипация не влияет на положение ЛТП, и допускается, что в работе [2] амплитуда возмущений указана ошибочно).

1.6. Расчетная сетка

Характерный масштаб длины выбран как $L = 0.7239$ м. Продольный размер буферной зоны, которая ограничена пунктирным прямоугольником на фиг. 5, составляет полторы длины волны фундаментального возмущения, или $1.5\lambda_x$ где $\lambda_x = x_2^* - x_1^*$. В настоящей работе используются расчетные сетки из работы [1].

На фиг. 5 показаны расчетная область и расчетная сетка (вид сбоку). Подобласть для проведения основных нестационарных расчетов ограничена сплошным прямоугольником и начинается на расстоянии $x_0^* = 0.258$ м от передней кромки пластины. Длина подобласти в 14.3 раза больше продольной длины волны фундаментального возмущения. Высота подобласти выбрана $y_H^* = 0.03$ м, что составляет не менее пяти местных толщин пограничного слоя на выходной гра-

нице. Размер подобласти в боковом направлении составляет одну длину волны λ_z в поперечном направлении, где $\lambda_z = 2\pi/\beta_0^* \approx 0.0297$ м.

Расчетная сетка представлена на фиг. 5а, соответствующие сеточные сгущения — на фиг. 5б и 5в. Основные расчеты настоящей работы проведены на сетке 80 миллионов узлов (подробная сетка). Эта сетка соответствует сетке из работы [2] в плоскости xOy . Сеточные линии распределены равномерно в поперечном направлении. Количество точек вдоль оси z составляет 201 на подробной сетке. Грубая сетка имеет вдвое меньше узлов по x и по z , чем подробная сетка. В вертикальном направлении количество точек для обеих сеток одинаково; поперек пограничного слоя приходится не менее 100 точек. На подробной сетке фундаментальное возмущение разрешено в боковом направлении (по z) 201 точками на длину волны по z , а в продольном направлении (по x) — 320 точками. Стоит отметить, что для распространения монохроматической акустической волны в равномерном потоке требуется около 40 точек на длину волны, чтобы добиться близкого к естественному уровню вязкого затухания волны для используемого диссипативного численного метода. Поэтому на построенных сетках численная диссипация фундаментального возмущения незначительна. Можно принять, что возмущения с длиной волны вчетверо короче длины фундаментальной волны моделируются достаточно достоверно, что позволяет “ухватить” процесс ламинарно-турбулентного перехода даже на грубых расчетных сетках.

1.7. Анализируемые величины

Данные для обработки и сравнения собираются, начиная с момента безразмерного времени $t = 2.261$ после ввода возмущений в пограничный слой, когда установился квазипериодический режим течения. В следующем разделе анализируются свойства переходного течения.

Анализ проводится для спектрального состава возмущений, где сопоставляются амплитуды отдельных гармоник Фурье или максимумы этих амплитуд по нормали к поверхности в рассматриваемом сечении $x = \text{const}$. Фурье-анализ и обработка нестационарных результатов выполнены с помощью возможностей языка программирования Python (библиотека `numpy`). Результат работы процедур быстрого преобразования Фурье по времени и координате z нормирован на $N_t \times N_z/4$, где N_t и N_z — количество точек анализируемого сигнала по времени и по координате z соответственно. В настоящей работе исследованы амплитуды гармоник пульсаций продольной компоненты скорости, давления, температуры, а также максимумы этих величины по нормали к поверхности.

Вихревая структура полей течения визуализируется с помощью Q -критерия: $Q = 0.5(\Omega_{ij}\Omega_{ij} - S_{ij}S_{ij})$, $S_{ij} = 0.5(\partial_j u_i + \partial_i u_j)$, $\Omega_{ij} = 0.5(\partial_j u_i - \partial_i u_j)$, u_i — компоненты вектора скорости (для записи использованы тензорные обозначения; предполагается соглашение о суммировании по повторяющемуся в произведении индексу).

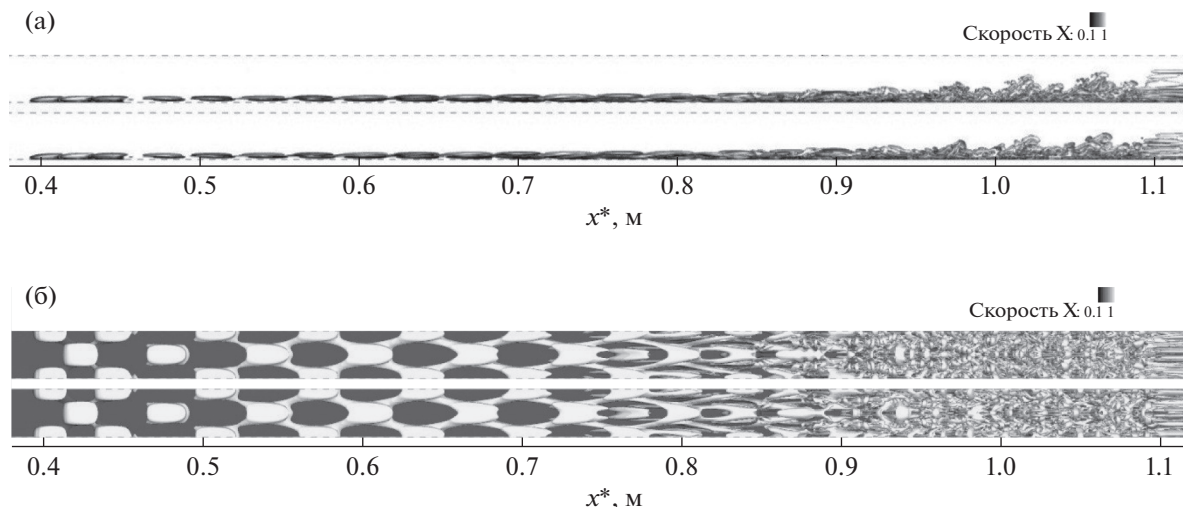
Также сопоставляются поля мгновенных значений продольной и поперечной компонент вектора завихренности и средние параметры течения (коэффициент трения, профили продольной компоненты скорости, осредненные по Фавру и по Рейнольдсу).

2. АНАЛИЗ РЕЗУЛЬТАТОВ

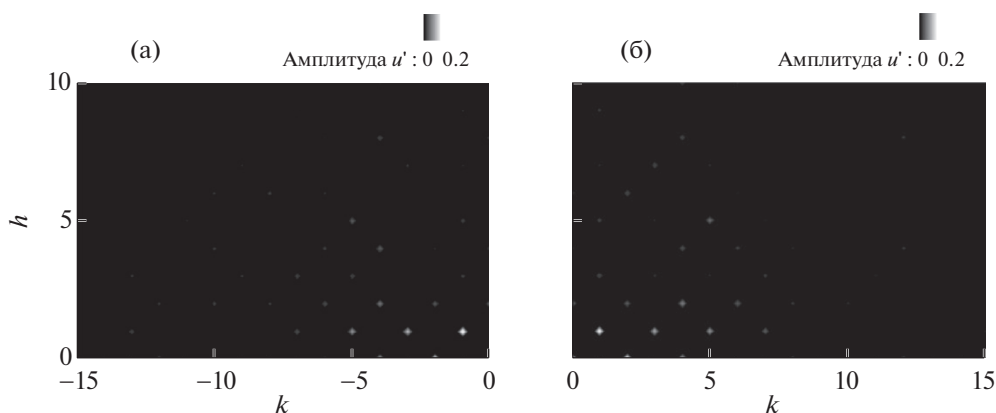
2.1. Структуры течения

Мгновенная структура возмущенного течения представлена на фиг. 6 в виде изоповерхностей Q -критерия. Сразу за источником возмущений расположена область линейного развития возмущений, где они формируют X-образные структуры, усиливающиеся вниз по потоку. Вблизи $x^* \approx 0.6$ мм появляются первые признаки нелинейного взаимодействия — возмущения начинают искажаться. При $x^* \in (0.75, 0.85)$ м наблюдается интенсивный нелинейный распад возмущений. За этой областью формируется зона молодой турбулентности с ростом мелкомасштабных вихрей. Эта зона развивается вниз по потоку. Можно видеть, что для гибридной схемы (настоящая работа) размер разрешаемых мелкомасштабных вихрей меньше, чем в случае монотонной схемы (см. [1]). Это различие наблюдается в области нелинейного режима при $x^* > 0.85$ м.

Так как возбуждающие возмущения носят периодический характер с выделенной частотой, а нелинейные взаимодействия приводят к порождению кратных гармоник, то отклик пограничного слоя на такие возмущения остается периодическим (квазистационарный режим течения). Для изучения спектральных свойств процесса ЛТП нестационарное течение сначала устанавли-



Фиг. 6. Визуализация вихревых структур пограничного слоя с помощью изоповерхностей Q-критерия, $Q = 5$, сетки 20 миллионов узлов, вверху настоящая работа, внизу работа [1]: (а) — вид сбоку со стороны $+z$; (б) — вид сверху со стороны $+y$. Окраска соответствует величине продольной компоненты вектора скорости. Буферная зона начинается при $x^* \approx 1.09$ м.

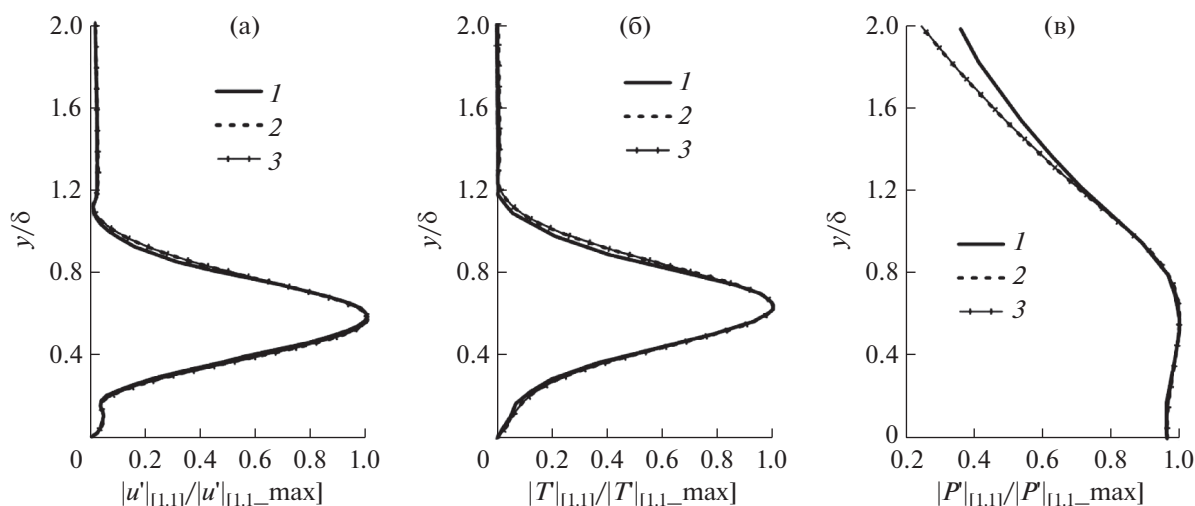


Фиг. 7. Двухмерное преобразование Фурье поля $u'(t, x_0^*, y_0^*, z)$ на линии $(x_0^*, y_0^*) = (0.9201, 0.0035)$ м, сетки 80 миллионов узлов: (а) — работа [1], (б) — настоящая работа.

вается до квазистационарного режима, а затем набирается статистика в течение пяти временных периодов фундаментального возмущения.

В каждом поперечном сечении $x^* = \text{const}$ возмущение можно представить через сумму гармонических колебаний путем преобразования Фурье. Для рассматриваемого течения вдоль плоской пластины разумно делать двухмерное преобразование Фурье по времени и поперечной координате для каждой линии $x^* = \text{const}$, $y^* = \text{const}$. Результат такого двухмерного преобразования можно представить в виде амплитуды гармоники $(f^*, \beta^*) = (hf_0^*, k\beta_0^*)$. Таким образом, результат двухмерного преобразования Фурье можно представить в виде амплитуд двухмерных гармоник пульсации продольной скорости $\hat{u}'_{hk}$. Пример представлен на фиг. 7 для начала области молодой турбулентности.

В описанной постановке Фурье-спектр является симметричным, поэтому интерес представляет только часть спектра при $h \geq 0$, $k \geq 0$. Следует отметить, что пики спектра расположены в шахматном порядке, что объясняется квадратичным (нелинейным) взаимодействием возмущений друг с другом. Например, для стационарного возмущения и других четных частот $h = 0, 2, 4, \dots$ наблюдаются только максимумы на четных волновых числах $k = 0, 2, 4, 6, \dots$, а для не-



Фиг. 8. Амплитуды гармоники $[1, 1]$ на сетке 80 миллионов узлов в сечении $x^* = 0.5$ м: (а) — $|u'_{[1,1]}|$, (б) — $|T'_{[1,1]}|$, (в) — $|P'_{[1,1]}|$: 1 — работа [2], 2 — работа [1], 3 — настоящая работа.

четных частот $h = 1, 3, 5, \dots$ — на нечетных волновых числах $h = 1, 3, 5, \dots$. Такая картина свойственна механизму наклонного распада, когда нелинейно (квадратично) взаимодействуют две гармоники с одинаковыми частотами, но противоположными по знаку волновыми числами. При этом частота удваивается, а волновое число обнуляется $[1, 1] + [1, -1] \rightarrow [2, 0]$. Чем ближе гармоника к фундаментальной гармонике, тем выше ее амплитуда. Это связано как с тем, что нелинейный распад продвигается в область высоких частот постепенно, так и с тем, что имеется численная пространственно-временная диссипация используемого численного метода. Расчеты настоящей работы выполнены на двух разных сетках, одна из которых вдвое мельче в продольном и боковом направлении. Из фиг. 7а и 7б следует, что спектр возмущений, полученных с применением гибридной схемы, оказывается несколько шире по сравнению со случаем монотонной схемы.

Ниже полученные результаты сопоставляются с результатами работ [1] и [2].

2.2. Линейный режим

Рассмотрим линейный режим развития возмущений, который наблюдается примерно от $x^* = 0.4$ м до $x^* = 0.6$ м. На фиг. 8 амплитуды фундаментальной моды $[1, 1]$ пульсаций продольной компоненты вектора скорости u' в сечении $x^* = 0.5$ м, полученные в настоящей работе, сопоставляются с результатами работы [1] и [2]. Для данного случая и для других линий $x^* = \text{const}$, $y^* = \text{const}$, наблюдается хорошее согласование (фиг. 8).

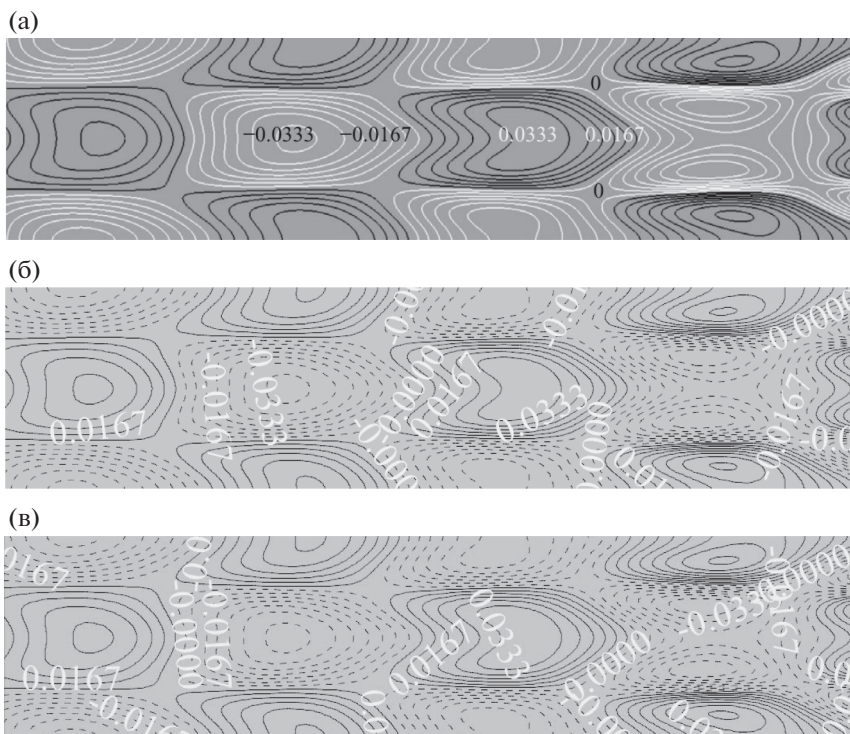
Среди всех возможных линий $y^* = \text{const}$ для данного сечения $y^* = \text{const}$ можно выделить линию y_0^* , на которой амплитуда рассматриваемой гармоники максимальна. Для пульсаций u' или T' эта линия будет находиться в критическом слое пограничного слоя, $y_0/\delta \approx 0.65$.

По сравнению с монотонной схемой работы [1] гибридная схема настоящей работы дает почти одинаковые решения. Это значит, что в линейном режиме диссипативный компонент очень слабо влияет на решение на подробной сетке.

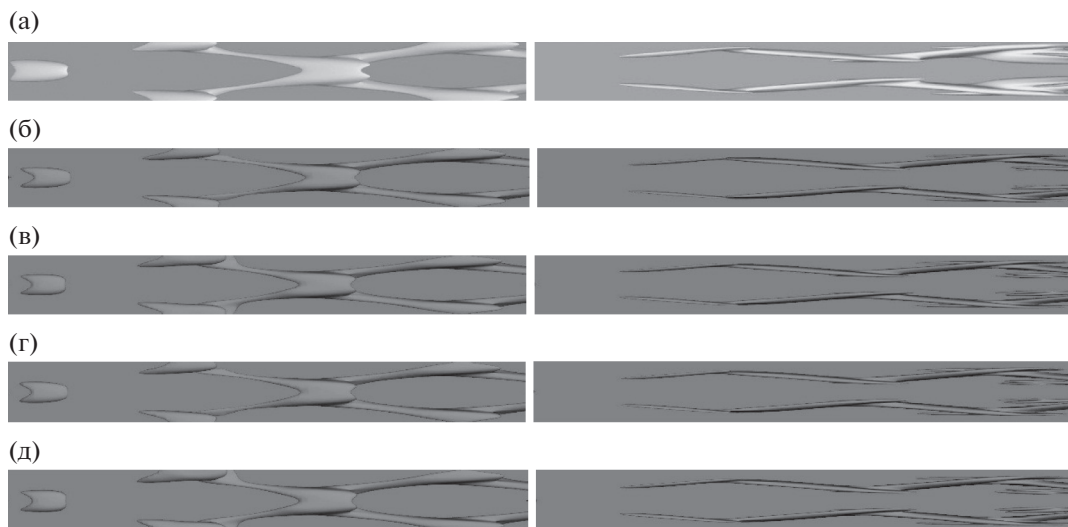
На фиг. 9 представлены структуры возмущений внутри пограничного слоя в сечении $y^* = \text{const}$. Все поля хорошо согласуются друг с другом. Гибридная схема дает чуть более наполненные контуры по сравнению с монотонной схемой и лучше согласуется с работой [2], но эта разница незначительна.

2.3. Нелинейный режим

Рассмотрим нелинейную стадию развития возмущений. Момент проявления нелинейного взаимодействия можно отметить по картинам Q -критерия, на которых заметно усиление “кана-



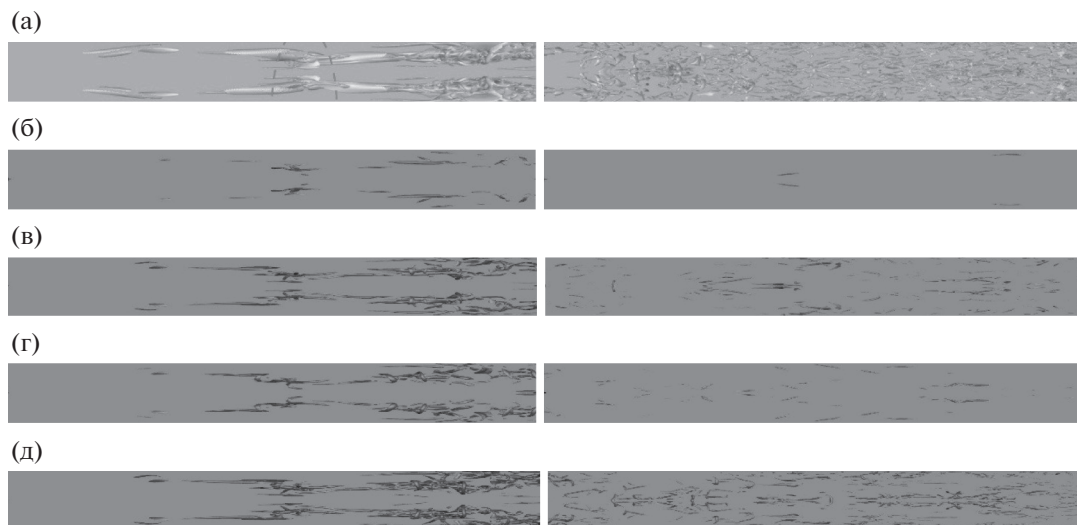
Фиг. 9. Мгновенный контур пульсации u' для сечения $x^* = 0.546\text{--}0.67$ м на сетке 80 миллионов узлов, $y^* = 2.3$ мм: (а) – работа [2], (б) – работа [1], (в) – настоящая работа.



Фиг. 10. Мгновенные изоповерхности Q -критерия, вид сверху; слева – $Q = 15$, $x^* = 0.546\text{--}0.670$ м, справа – $Q = 100$, $x^* = 0.670\text{--}0.798$ м: (а) – работа [2] на сетке 80 миллионов узлов, (б) – работа [1] на сетке 20 миллионов узлов, (в) – работа [1] на сетке 80 миллионов узлов, (г) – настоящая работа на сетке 20 миллионов узлов, (д) – настоящая работа на сетке 80 миллионов.

тообразных” структур (см. фиг. 10 и 11). Для малых значений Q (15 и 100) (фиг. 10) на сетке 20 и 80 миллионов узлов гибридная и монотонная схемы дают почти одинаковые структуры значений Q . Они хорошо согласуются результатами работы [2].

Однако в области молодой турбулентности, где появляются мелкомасштабные структуры и максимальная величина Q растет (10000 и 40000) (фиг. 11), диссипативная схема не позволяет



Фиг. 11. Мгновенные изоповерхности Q -критерия, вид сверху; слева — $Q = 10000$, $x^* = 0.798-0.924$ м, справа — $Q = 40000$, $x^* = 0.924-1.051$ м: (а) — работа [2] на сетке 80 миллионов узлов, (б) — работа [1] на сетке 20 миллионов узлов, (в) — работа [1] на сетке 80 миллионов узлов, (г) — настоящая работа на сетке 20 миллионов узлов, (д) — настоящая работа на сетке 80 миллионов.

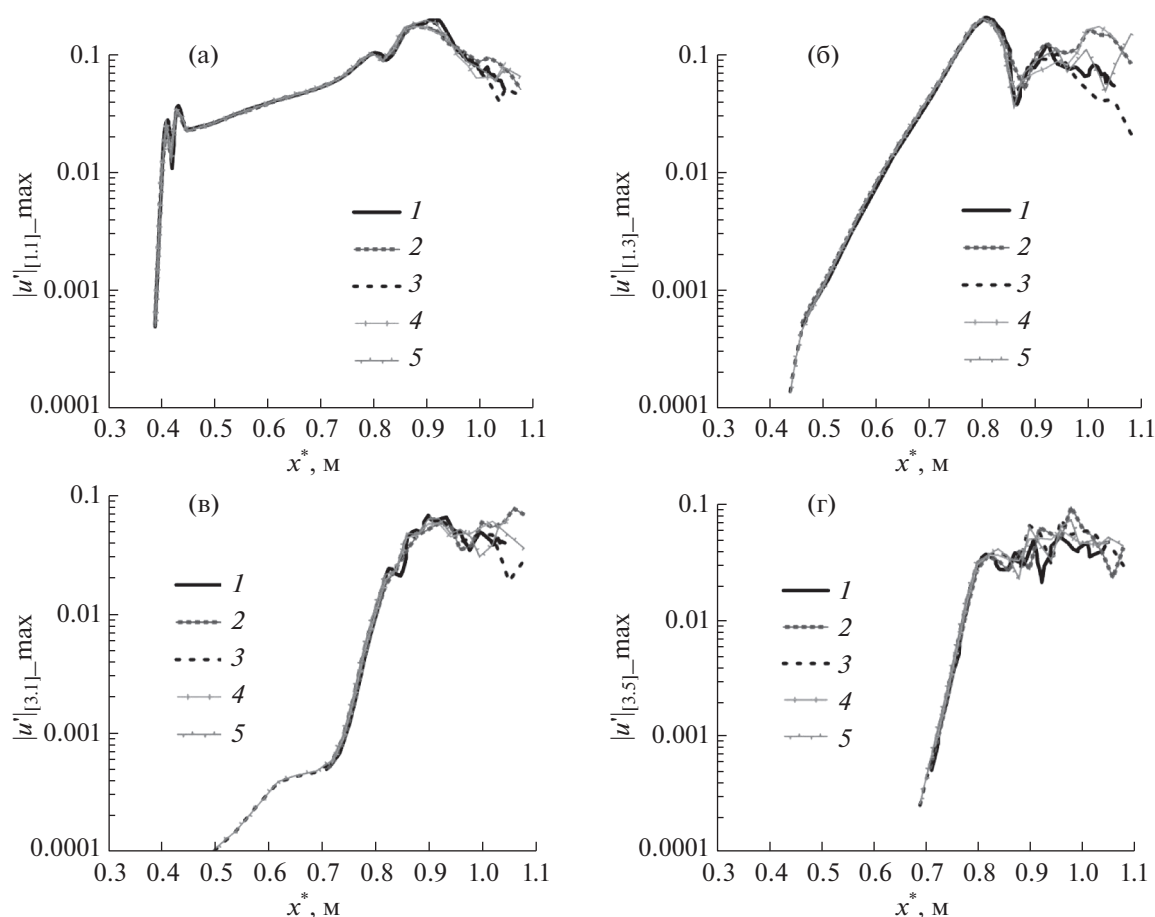
идеально воспроизвести результаты низкодиссипативной схемы [2]. Наиболее вероятными причинами этого расхождения являются применение в [2] спектрального метода в боковом направлении и использование высокочастотных гармоник, которые располагаются вблизи частоты (волнового числа) Найквиста для подробной расчетной сетки настоящей работы и поэтому плохо разрешаются на ней.

Наполненность вихревых структур, полученных с применением гибридной схемы на грубой сетке (см. фиг. 11в), хорошо соответствует случаю подробной сетки (см. фиг. 11в, г) как для монотонной, так и для гибридной схем. Однако соответствующие структуры, полученные на исходной монотонной схеме на грубой сетке, оказываются значительно слабее (см. фиг. 11б). Поэтому можно заключить, что на грубой сетке гибридная схема дает более правдоподобные результаты по сравнению с монотонной схемой.

Сделанный выше вывод подтверждается количественно с помощью эволюции амплитуд отдельных гармоник. Рассмотрим, в частности, эволюцию максимальных по y амплитуд. Механизм наклонного резонанса проявляется последовательно. Изначально растет наиболее неустойчивая фундаментальная наклонная волна $[1, \pm 1]$, что обусловлено чисто неустойчивостью пограничного слоя. При достижении некоторой критической амплитуды она начинает нелинейно взаимодействовать сама с собой, порождая кратные гармоники: $h = 0$ и $h = 2$, $k = 0$, $k = 2$, которые начинают расти благодаря нелинейному взаимодействию фундаментальных гармоник. Когда кратные гармоники достигают достаточных амплитуд, они начинают нелинейно взаимодействовать друг с другом и с фундаментальными гармониками, порождая все больше кратных гармоник. Такой процесс и его временная последовательность прослеживаются на фиг. 12 и фиг. 13, на которых изображена эволюция амплитуд гармоник вниз по потоку. Описанный механизм объясняет шахматную структуру спектра, приведенного на фиг. 7.

На фиг. 12а представлена эволюция фундаментальной моды продольной компоненты скорости, полученная на сетке 80 млн узлов с использованием гибридной схемы (линия 5). Решение оказывается ближе к решению из работы [2], чем в случае монотонной схемы [1] (линия 3). На сетке 20 млн узлов решение с использованием гибридной схемы (линия 4) также аналогично ближе к [2] и почти достигает до решения с использованием монотонной схемы и сетки 80 млн узлов (линия 3). На фиг. 12б—г различие в линиях не такое четкое.

На фиг. 13а продемонстрировано, что на сетке 20 млн узлов решение с использованием монотонной схемы (линия 2) сильно отличается от решения [2]. Однако решение, полученное с использованием гибридной схемы (линия 4), значительно лучше совпадает с решением на базе монотонной схемы на сетке 80 млн узлов (линия 3).

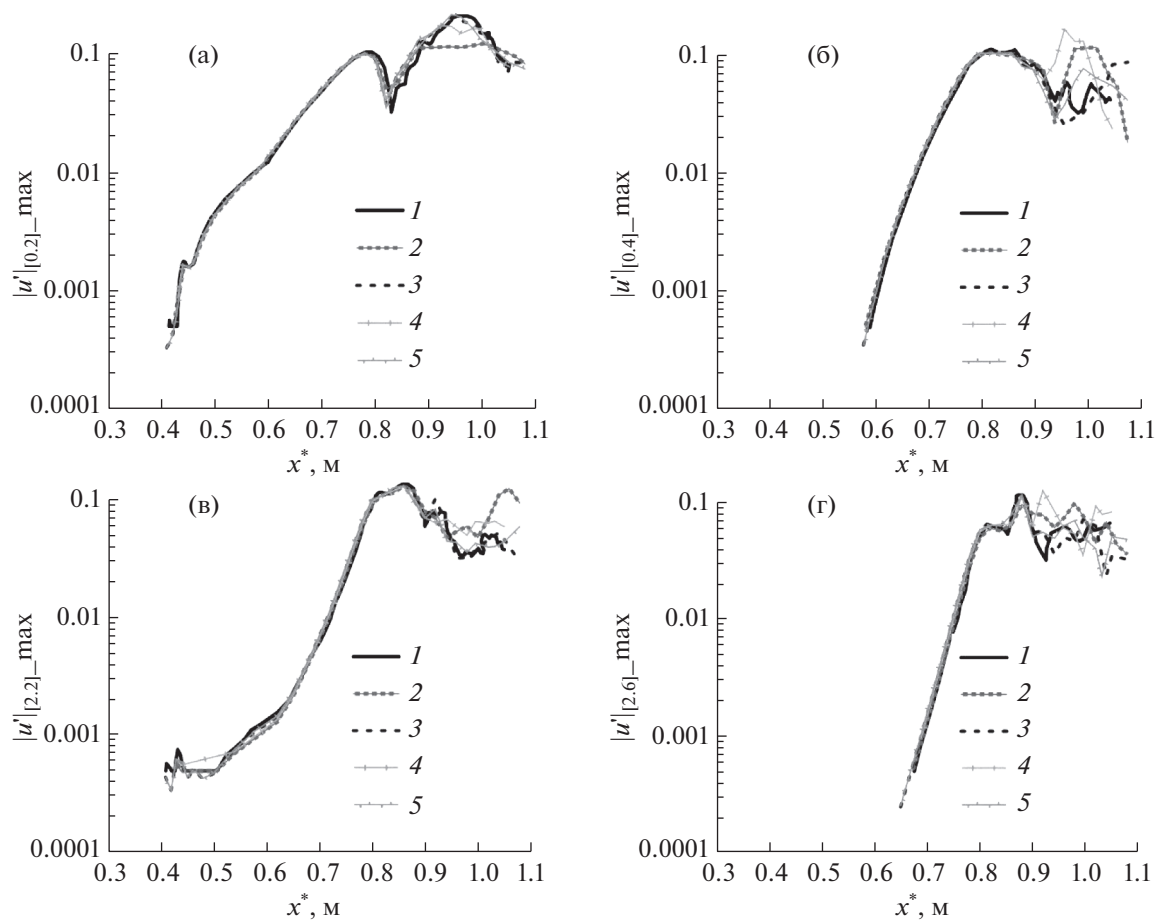


Фиг. 12. Эволюция максимальной по y амплитуды Фурье-гармоники для нечетных частот, 1 — работа [2] на сетке 80 миллионов узлов, 2 — работа [1] на сетке 20 миллионов узлов, 3 — работа [1] на сетке 80 миллионов узлов, 4 — настоящая работа на сетке 20 миллионов узлов, 5 — настоящая работа на сетке 80 миллионов узлов: (а) $[h, k] = [1, 1]$, (б) $[h, k] = [1, 3]$, (в) $[h, k] = [3, 1]$, (г) $[h, k] = [3, 5]$.

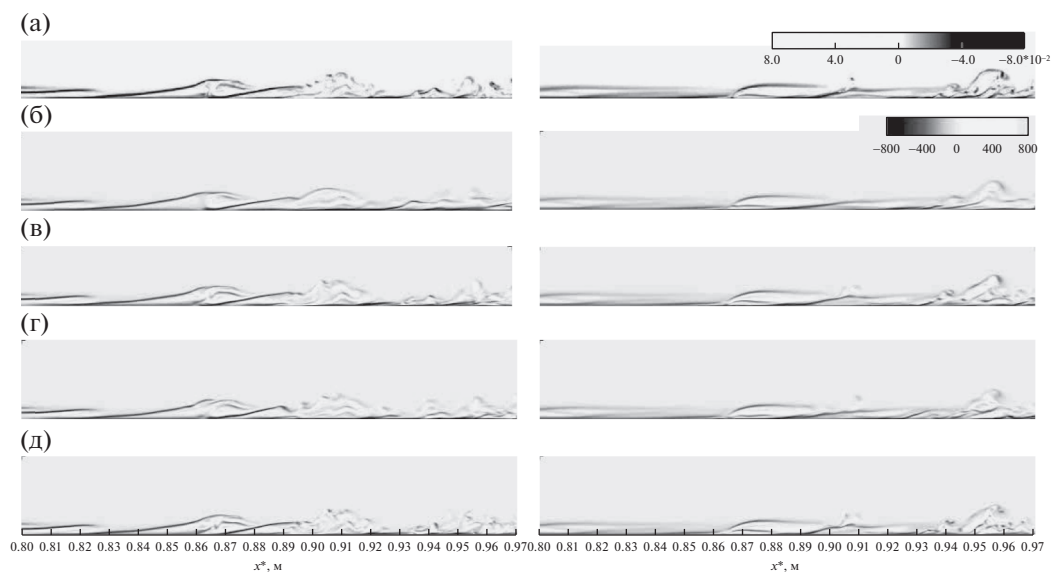
Таким образом, гибридная схема дает решения, которые лучше согласуются с решением [2] по сравнению со случаем монотонной схемы [1]. На грубой сетке решение гибридной схемы может достигать точности решения монотонной схемы на мелкой сетке. Разницы в решениях с использованием монотонной и гибридной схем проявляются сильнее в низках модах.

На фиг. 14 приведены мгновенные структуры поперечного вектора завихренности в различных сечениях $z^* = \text{const}$ и сравнение полученных результатов с результатами работы [1] и [2]. Видно, что вблизи $x^* = 0.865$ м появляются мелкомасштабные структуры, свойственные развитой турбулентности. Основные крупномасштабные структуры настоящей работы и работы [1] хорошо совпадают с [2]. Детальное сравнение показывает, что вихри, полученные в работе [1], менее интенсивны по сравнению с вихрями из работы [2], и содержат меньше мелкомасштабных структур, как обсуждалось выше. При этом гибридная схема настоящей работы разрешает вихревые структуры лучше (фиг. 14г, д), чем монотонная схема [1] (фиг. 14б, в). Визуально гибридная схема позволяет достигать точности подробной сетки с использованием грубой сетки [2].

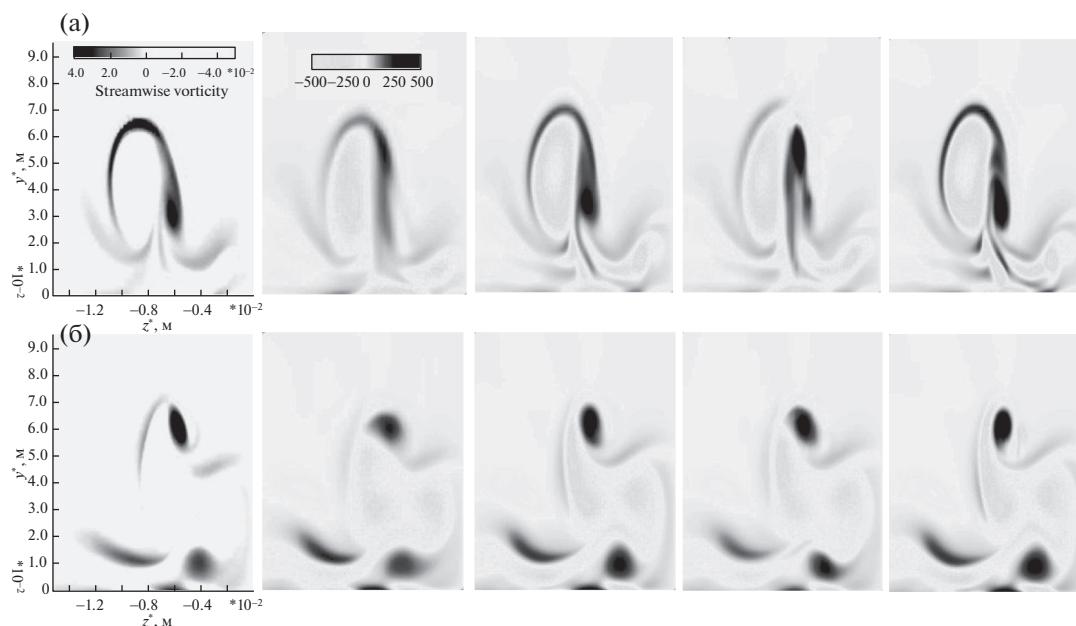
На фиг. 15 показаны мгновенные сечения продольной компоненты вектора завихренности, которые сопоставлены с результатами работы [1] и [2]. Структура течения симметрична относительно плоскости $z^* = 0$; поэтому приводится только половина области по z^* . В сечении $x^* = 0.862$ м (фиг. 15а) расположен большой вихрь. С ростом x он растет и, начиная с $x^* = 0.872$ м (фиг. 15б), распадается с образованием более мелких вихрей. Результаты, полученные на гибридной схеме (настоящая работа), лучше согласуются с результатами работы [2] в случае монотонной схемы (работа [1]).



Фиг. 13. То же, что на фиг. 12, но для четных частот: (а) — $[h, k] = [0, 2]$, (б) — $[h, k] = [0, 4]$, (в) — $[h, k] = [2, 2]$, (г) — $[h, k] = [2, 6]$.



Фиг. 14. Мгновенное поле поперечного вектора завихренности в момент времени $t = 2.82891 + kT$, слева — $z^* = -0.0062$ м, справа — $z^* = -0.0032$ м: (а) — работа [2] на сетке 80 миллионов узлов, (б) — работа [1] на сетке 20 миллионов узлов, (в) — работа [1] на сетке 80 миллионов узлов, (г) — настоящая работа на сетке 20 миллионов узлов, (д) — настоящая работа на сетке 80 миллионов.



Фиг. 15. Мгновенный продольный вектор завихренности в момент $t = 2.82891 + kT$, слева направо работа [2] на сетке 80 миллионов узлов, работа [1] на сетке 20 миллионов узлов, работа [1] на сетке 80 миллионов узлов, настоящая работа на сетке 20 миллионов узлов, настоящая работа на сетке 80 миллионов узлов соответственно: (a) – $x^* = 0.862$ м, (б) – $x^* = 0.870$ м.

Рассмотрим зависимость осредненной величины коэффициента трения c_f от продольной координаты x . Локальный коэффициент трения вычисляется по формуле

$$c_f = \frac{2\mu}{\text{Re}_\infty} \left. \frac{\partial u}{\partial y} \right|_{y=0}.$$

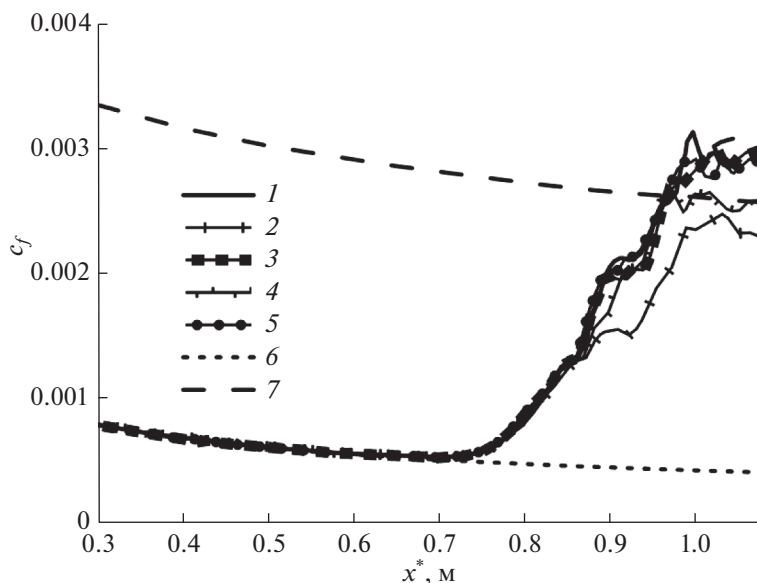
Как и в работе [1], в настоящей работе рассматривается усреднение по времени в течение пяти периодов фундаментальной гармоники $\Delta t = 10\pi/\omega_0$ и по размаху z всей расчетной области:

$$\bar{\varphi} = \frac{1}{\lambda_z} \frac{1}{\Delta t} \int_0^{\lambda_z} \int_{t_0}^{t_0+\Delta t} \varphi(t, z) dt dz.$$

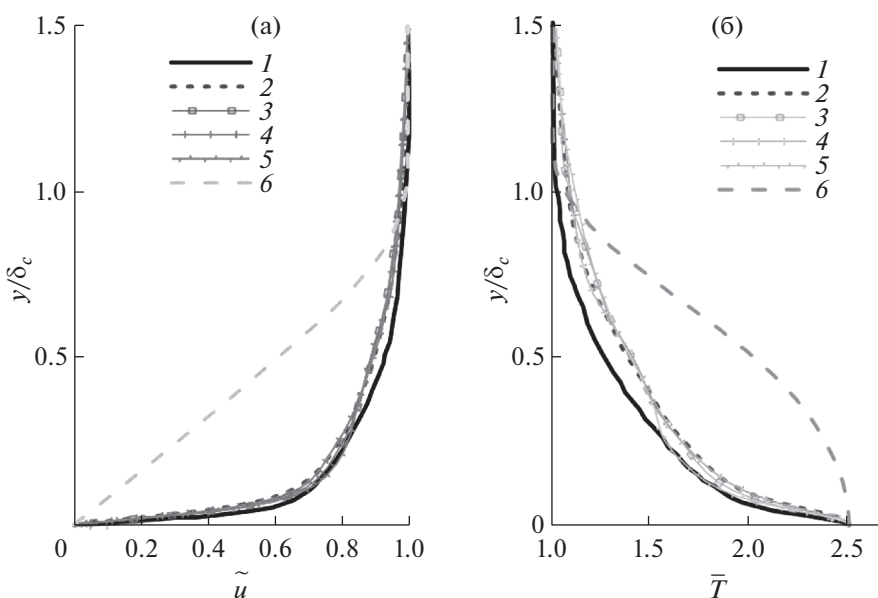
Начиная с $x^* = 0.72$ м, величина c_f резко увеличивается в окрестности точки $x^* = 0.86$ м, что соответствует началу ламинарно-турбулентного. В этом диапазоне результаты для диссипативной и гибридной схем совпадают с результатами работы [2] даже на грубой сетке. Ниже по потоку результаты на грубой сетке оказываются ниже по сравнению со случаем подробной сетки. В особенности это проявляется для монотонной схемы, а в случае гибридной схемы соответствующая кривая мало отклоняется от случая подробной расчетной сетки ($x^* \geq 0.9$ м).

На фиг. 17 показаны средние профили газодинамических переменных в сечении $x^* = 0.996$ м. Чем сильнее эффект нелинейных взаимодействий в рассматриваемом сечении, тем более наполнены средние профили.

Фигура 17 подтверждает, что несмотря на недостаточно детальную картину мелкомасштабных вихрей, полученную с помощью диссипативного метода [1], интегральные характеристики течения (средние профили газодинамических переменных, средний коэффициент трения) оказываются достаточно близки к результатам, полученным на низкодиссипативных схемах [2]. При этом гибридная схема настоящей работы дает лучшие результаты даже на грубой сетке. Этот вывод важен для прикладных задач, где нет необходимости подробно разрешать все структуры развитого турбулентного движения, но требуется получить надежные интегральные характеристики течения. Для таких задач применение гибридной схемы на грубой сетке позволяет получить удовлетворительные результаты с ограниченными вычислительными ресурсами.



Фиг. 16. Осредненный по пространству и по времени коэффициент трения: 1 — работа [2] на сетке 80 миллионов узлов, 2 — работа [1] на сетке 20 миллионов узлов, 3 — работа [1] на сетке 80 миллионов узлов, 4 — настоящая работа на сетке 20 миллионов узлов, 5 — настоящая работа на сетке 80 миллионов узлов, 6 — ламинарная ветвь, 7 — теоретическая турбулентная ветвь [7].



Фиг. 17. Осредненная по Фавру продольная компонента вектора скорости (а) и осредненная по Рейнольдсу температура (б) в сечении $x^* = 0.996$ м: 1 — работа [2] на сетке 80 миллионов узлов, 2 — работа [1] на сетке 20 миллионов узлов, 3 — работа [1] на сетке 80 миллионов узлов, 4 — настоящая работа на сетке 20 миллионов узлов, 5 — настоящая работа на сетке 80 миллионов узлов, 6 — ламинарная ветвь.

ВЫВОДЫ

Предложена гибридная разностная схема для моделирования развития возмущений в сверхзвуковых пограничных слоях. Схема основана на индикаторе гладкого решения и позволяет плавно уменьшать диссипативность исходной монотонной схемы по отношению к конвективным потоковым величинам посредством уменьшения монотонизирующей поправки, которая возникает при решении задачи Римана о распаде произвольного разрыва. Уменьшение возмож-

но до некоторой пороговой величины. При дальнейшем уменьшении схема, по-видимому, теряет устойчивость. Эмпирически установлено, что пороговый уровень монотонизирующей поправки зависит от качества сетки и интенсивности внешних возмущений.

Эволюция возмущений, рассчитанная с использованием гибридной схемы на грубой расчетной сетке, хорошо согласуется с результатами других работ, в том числе с результатами, полученными на подробной сетке с использованием монотонной схемы. При этом монотонная схема на грубой сетке занижает уровень возмущений и значительно хуже разрешает мелкомасштабные вихревые структуры в области молодой турбулентности по сравнению со случаем гибридной схемы.

Предложенная гибридная разностная схема пригодна для моделирования развития нестационарных возмущений в сверхзвуковых пограничных слоях. Она позволяет получать удовлетворительные результаты на ограниченных вычислительных ресурсах.

СПИСОК ЛИТЕРАТУРЫ

1. *Егоров И.В., Нгуен Н.К., Нгуен Т.Ш., Чувахов П.В.* Моделирование ламинарно-турбулентного перехода с применением диссипативных численных схем // Ж. вычисл. матем. и матем. физ. 2021. Т. 61. № 2. С. 268–280.
2. *Mayer C.S.J., Terzi D.A.V., Fasel H.F.* DNS of Complete Transition to Turbulence Via Oblique Breakdown at Mach 3 // AIAA 2008-4398, 2008.
3. *Егоров И.В., Новиков А.В.* Прямое численное моделирование ламинарно-турбулентного обтекания плоской пластины при гиперзвуковых скоростях потока // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 6. С. 145–162.
4. *Ducros F. et.al.* Large-Eddy Simulation of the Shock / Turbulence Interaction // J. of Comput. Phys. 1999. V. 152. Issue 2. 1 July 1999. P. 517–549.
5. *Каримов Т.Х.* О некоторых итерационных методах решения нелинейных уравнений в гильбертовом пространстве // Докл. АН СССР. 1983. Т. 269. № 5. С. 1038–1046.
6. *Chuvakhov P.V., Fedorov A.V., Obraz A.O.* Numerical simulation of turbulent spots generated by unstable wave packets in a hypersonic boundary layer. Computers & Fluids. 2018. V. 162. P. 26–38. Available at: <https://doi.org/10.1016/j.compfluid.2017.12.001>
7. *White F.M.* Viscous Fluid Flow. New York: McGraw-Hill, 1991.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 519.63

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ГРАДИЕНТНЫХ
МЕТОДОВ ОПРЕДЕЛЕНИЯ ИСТОЧНИКА
ДИФфуЗИОННО-ЛОГИСТИЧЕСКОЙ МОДЕЛИ¹⁾© 2022 г. Т. А. Звонарева^{1,*}, О. И. Криворотко^{1,2,**}¹ 630090 Новосибирск, ул. Пирогова, 2, Новосибирский государственный университет, Россия² 630090 Новосибирск, пр-т Акад. Лаврентьева, 6, Институт вычислительной математики
и математической геофизики СО РАН, Россия

*e-mail: t.zvonareva@g.nsu.ru

**e-mail: krivorotko.olya@mail.ru

Поступила в редакцию 10.08.2021 г.

Переработанный вариант 10.08.2021 г.

Принята к публикации 16.12.2021 г.

Проведен сравнительный анализ численного решения задачи определения источника диффузионно-логистической модели по данным о диффузионном процессе в фиксированные моменты времени и точках пространства градиентными методами для случаев непрерывной и дискретной постановок. Получены выражения вычисления градиента целевого функционала для двух постановок, связанных с решением соответствующих сопряженных задач. Показано, что в случае аппроксимации дискретных функций модели кубическими сплайнами точность восстановленного источника совпадает с решением в непрерывной постановке. Численные эксперименты решения задачи об источнике для дискретной модели распространения информации в онлайн социальных сетях показали, что применение дискретного подхода в разы увеличивает вычислительное время по сравнению с использованием непрерывного подхода. Библ. 23. Фиг. 1. Табл. 5.

Ключевые слова: диффузионно-логистическая модель, задача об источнике, обратная задача, градиентные методы, сопряженная задача, сравнительный анализ, регуляризация, оптимизация, социальные процессы.

DOI: 10.31857/S0044466922040147

ВВЕДЕНИЕ

Градиентные методы решения задач об источнике для дифференциальных уравнений применялись в [1]–[7]. Основная их идея состоит в последовательном уменьшении значения целевого функционала $J(q) = \|A(q) - f\|^2$ в виде

$$q^{m+1} = q^m - \alpha_m J'(q^m), \quad q^0 \in Q, \quad \alpha_m > 0, \quad (1)$$

где α_m – параметр спуска, характеризующий тот или иной градиентный метод, $J'(q^m)$ – градиент целевого функционала $J(q)$, $A: Q \rightarrow F$ – оператор обратной задачи.

В зависимости от типа пространства Q (гильбертово, соболевское, евклидово и т.п.) области определения функционала $J(q)$ и его градиента $J'(q)$ могут быть интегрируемые, непрерывно-дифференцируемые или дискретные. В связи с этим возможно использование двух подходов: 1) непрерывный, который состоит в формулировке используемых функций в непрерывных пространствах с последующим представлением задачи в конечно-разностном виде и 2) дискретный, в котором сначала переходят к конечно-разностному аналогу, для которого формулируют дискретные функции $J(q)$ и $J'(q)$ (см. [8]).

¹⁾ Работа выполнена при финансовой поддержке РНФ (проект № 18-71-10044-П).

В данной работе для задачи об источнике для диффузионно-логистической модели, возникающей при описании распространения информации в онлайн социальных сетях и сформулированной в непрерывном и дискретном видах, применены оба подхода. Выявлены области их применимости. Получены градиенты целевых функционалов в непрерывной и дискретной постановках, связанных с решениями соответствующих сопряженных задач. Отметим, что дискретный вариант градиента не является следствием дискретизации непрерывного своего аналога, а получен путем минимизации функции Лагранжа (см. [9]–[12]).

Статья организована следующим образом. Формулировки прямой и обратной задач приведены в разд. 1. Вывод градиента целевого функционала в дискретной форме приведен в п. 2.1. Выражение для градиента целевого функционала в непрерывной форме и соответствующая постановка сопряженной задачи приведены в п. 2.2. Алгоритмы градиентных методов для разных параметров спуска, а также сравнительный анализ результатов численных расчетов для решения задачи об источнике для диффузионно-логистической модели приведены в разд. 3.

1. ПОСТАНОВКИ ПРЯМОЙ И ОБРАТНОЙ ЗАДАЧ

1.1. Постановка прямой задачи

Под прямой задачей в данной работе понимается задача моделирования процесса распространения информации в онлайн социальных сетях, в которой требуется найти функцию плотности активных пользователей $u(x_j, t_n)$ в каждой точке пространства и в момент времени. За расстояние x принимается целочисленная величина, описывающая минимальное количество дружеских связей между пользователем и источником информации и измеряемая в единицах. Время t измеряется в часах, а плотность активных пользователей $u(x_j, t_n)$ — в количестве человек на единицу расстояния.

Данные, представленные в [13], иллюстрируют, что интерес к информации проявляется у пользователей с расстоянием x в пределах от 1 до 6. Именно в этих границах происходит значительный вклад в изменение плотности активных пользователей $u(x_j, t_n)$. Также для социальных сетей справедлива теория шести рукопожатий, согласно которой большинство агентов находится на расстоянии $x \leq 6$. Поэтому в модели рассматриваются граничные условия Неймана, которые описывают отсутствие потока информации через границы при $x = 1, 6$.

Задача моделирования процесса распространения информации основана на законе сохранения информационного потока

$$\int_a^b \left(\frac{\partial u}{\partial t} + \frac{\partial H}{\partial x} - g(u, x, t) \right) dx = 0.$$

А именно, для фиксированного участка $[a, b]$ скорость изменения общего количества вовлеченных пользователей $\frac{\partial}{\partial t} \int_a^b u(x, t) dx$ на этом участке должна равняться сумме потока $H(a, t) - H(b, t)$,

с которой они поступают, и скорости $\int_a^b g(u, x, t) dx$, с которой в пределах $a < x < b$ появляются новые вовлеченные пользователи. Информация в социальной сети распространяется от высокой плотности вовлеченных пользователей к низкой, поэтому поток может быть представлен в виде $H = -d \frac{\partial u}{\partial x}$. В [13] функция $g(u, x, t)$ выбирается в виде $(1 - u/K)r(t)u$ и описывает динамику изменения численности активных пользователей. Неопределенность построенной модели заключается в том, что распределение плотности активных пользователей в начальный момент времени неизвестно и зависит от структуры социальной сети. Поэтому актуальной становится задача идентификации начального распределения пользователей с целью корректного описания распространения информации в конкретной сети и ее дальнейшего контроля/управления.

Обозначим начальные значения плотности вовлеченных пользователей $u(x_i, 1) = q_i$, $i = 0, \dots, 5$. Интерполируя вектор значений $q = (q_0, \dots, q_5)$ кубическими сплайнами, переходим к

Таблица 1. Параметры модели (2), описывающие распространение информации в онлайн социальной сети

Параметр	Описание	Среднее значение	Единицы измерения
d	популярность информации, которая способствует распространению информации через неструктурные действия, такие как целенаправленный поиск пользователем рассматриваемой информации	0.01	расст. ² /ч
K	пропускная способность, которая является максимально возможным числом активных пользователей	25	чел./расст.
$r(t)$	$= \frac{\beta_2}{\beta_1} - e^{-\beta_1(t-1)} \left(\frac{\beta_2}{\beta_1} - \beta_2 \right)$ скорость роста числа вовлеченных пользователей	—	1/ч
β_1	скорость снижения интереса к информации с течением времени	1.5	—
β_2	остаточная скорость	0.375	—
β_3	начальная скорость роста числа активных пользователей	1.65	—

непрерывной постановке начально-краевой задачи для диффузионно-логистической модели, описываемой уравнением в частных производных параболического типа (см. [13]):

$$\begin{aligned}
 u_t &= du_{xx} + \left(1 - \frac{u}{K}\right) r(t)u, \quad l \leq x \leq L, \quad t \geq 1, \\
 u(x, 1) &= Q(x), \quad l \leq x \leq L, \\
 u_x(l, t) &= u_x(L, t) = 0, \quad t \geq 1.
 \end{aligned} \tag{2}$$

Здесь $Q(x) \geq 0$ — начальная функция плотности. Согласно [13], если $Q(x) \not\equiv 0$ и удовлетворяет следующим условиям корректности: 1) $Q(x) \in C^2(l, L)$, 2) $Q'(l) = Q'(L) = 0$, 3) $dQ'' + (1 - Q/K)rQ \geq 0$, то по принципу максимума существует единственное положительное решение $u(x, t) \in C^{2,1}((l, L) \times (1, +\infty)) \cap C^{1,0}([l, L] \times [1, +\infty))$ прямой задачи (2).

Средние значения и описания параметров модели представлены в табл. 1. Параметры β_1, β_2 и β_3 в рассматриваемой модели соответствуют представленным в [13].

1.2. Постановка обратной задачи

Предположим, что известна дополнительная информация следующего вида:

$$u(x_i, t_k; q) = f_{ik}, \quad i = 1, \dots, N_1, \quad k = 1, \dots, N_2, \tag{3}$$

где $u(x, t; q)$ — решение прямой задачи для начальной функции плотности $Q(x)$, определяемой из набора параметров q , $N_1 \leq 6$. Обратная задача (2)–(3) состоит в определении набора параметров q по данным f_{ik} вида (3). Обратная задача может быть записана в виде $Aq = f$, где $f = (f_{11}, \dots, f_{1N_2}, f_{21}, \dots, f_{2N_2}, f_{N_11}, \dots, f_{N_1N_2}) \in \mathbb{R}^{N_1N_2}$, A — оператор обратной задачи.

Отметим, что в линейном приближении при $N_1N_2 = 6$ и $\det A \neq 0$ существует единственное решение системы $Aq = f$. В случае $N_1N_2 > 6$ вводится понятие нормального псевдорешения, т.е. решения, реализующего минимум нормы невязки

$$J(q) = \|Aq - f\|^2.$$

В [14] в линейном приближении решение обратной задачи (2)–(3) исследовано на устойчивость, и показано, что число обусловленности матрицы A имеет порядок 10^{16} , что свидетельствует о неустойчивости решения обратной задачи.

В данной работе невязка имеет вид

$$J(q) = w \sum_{i=1}^{N_1} \sum_{k=1}^{N_2} \int_1^L \int_1^T |u(x, t; q) - f(x, t)|^2 \delta(x - x_i) \delta(t - t_k) dt dx, \quad (4)$$

здесь $w = (L - l)(T - 1)/(N_1 N_2)$.

По аналогии сформулируем дискретную постановку целевого функционала

$$I(q) = w \sum_{i=1}^{N_1} \sum_{k=1}^{N_2} |v_i^k - f_{ik}|^2, \quad (5)$$

где $v_i^k := u(x_i, t_k; q)$.

2. НЕПРЕРЫВНАЯ И ДИСКРЕТНАЯ ПОСТАНОВКИ ГРАДИЕНТОВ ФУНКЦИОНАЛА

Исследование непрерывной и дискретной постановок прямой и обратной задач для диффузионно-логистической модели было навеяно работой А.Л. Карчевского [8], в которой похожий анализ был проделан для решения обратной задачи для гиперболического уравнения.

Существуют два подхода к численному решению поставленной обратной задачи (2)–(3). Первый состоит из следующих шагов:

- перейти от непрерывной постановки прямой задачи $L_q u = 0$ (L_q – оператор прямой задачи) к дискретной $\Lambda_p v = 0$ (Λ_p – конечно-разностный аналог оператора прямой задачи, а функция p является некоторым приближением функции Q и $p_j = q_{s-1}$ при $x_j = s$, $s = 1, \dots, 6$);

- выписать целевой функционал в дискретной постановке $I(q)$;

- получить постановку сопряженной задачи $\Lambda_p^* \phi = 0$ и градиент целевого функционала в дискретной форме $I'(q)$;

- решить задачу минимизации функционала $I(q)$.

Второй подход подразумевает следующую схему действий:

- выписать целевой функционал в непрерывной постановке $J(q)$;

- получить постановку сопряженной задачи $L_q^* \psi = 0$ и выражение градиента целевого функционала в непрерывной форме $J'(q)$;

- перейти к задаче $\Lambda_p v = 0$;

- выписать целевой функционал $I(q)$, аппроксимирующий $J(q)$;

- от постановки сопряженной задачи $L_q^* \psi = 0$ перейти к задаче $\Lambda_p^* \phi = 0$;

- получить соотношение $I'(q)$, которое аппроксимирует градиент целевого функционала $J'(q)$;

- решить задачу минимизации функционала $J(q)$.

Особенность постановки обратной задачи (2)–(3) состоит в дискретном задании дополнительных измерений, а также в определении набора параметров $q = (q_0, \dots, q_5)$ вместо функции $Q(x)$.

2.1. Градиент функционала в дискретной постановке

Дискретный градиент, полученный с помощью подхода из работ Ю.Г. Евтушенко и Ф.Л. Черноусько [9]–[12], формулируется для дискретной постановки рассматриваемой модели, т.е. для задачи

$$\begin{aligned} \frac{v_j^{n+1} - v_j^n}{\tau} &= d \frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{h^2} + \left(1 - \frac{v_j^n}{K}\right) r^n v_j^n, \quad j = 1, \dots, N_x - 1, \quad n = 0, \dots, N_t - 1, \\ v_j^0 &= p_j, \quad j = 0, \dots, N_x, \\ v_0^{n+1} &= \frac{4v_1^{n+1} - v_2^{n+1}}{3}, \quad n = 0, \dots, N_t - 1, \\ v_{N_x}^{n+1} &= \frac{4v_{N_x-1}^{n+1} - v_{N_x-2}^{n+1}}{3}, \quad n = 0, \dots, N_t - 1. \end{aligned} \quad (6)$$

Преобразуем ее в условие связи

$$\Phi(n, v^n; q) = 0, \quad n = 0, \dots, N_t, \quad (7)$$

где v^n — набор векторов v_j , $j = 0, \dots, N_x$.

Каждому условию в (7) ставится в соответствие вектор $\phi^n \in \mathbf{R}^{N_x+1}$, их объединение есть вектор $\phi^T = (\phi_0^T, \dots, \phi_{N_t}^T)$, $\phi \in \mathbf{R}^{(N_t+1) \times (N_x+1)}$. И рассматривается аналог функции Лагранжа для многошагового процесса (7):

$$E(v, \phi; q) = I(q) - \sum_{n=0}^{N_t} \Phi^T(n, v^n; q) \phi^n. \quad (8)$$

Тогда (см. [9])

$$\phi_{i_n}^n = E_{v_{i_n}}(v, \phi; q) = I_{v_{i_n}}(q) - \sum_{n=0}^{N_t} \Phi_{v_{i_n}}^T(n, v^n; q) \phi^n, \quad i_n = 0, \dots, N_t, \quad (9)$$

$$\frac{\partial I(q)}{\partial q_{i_s}} = E_{q_{i_s}}(v, \phi; q) = I_{q_{i_s}}(q) - \sum_{n=0}^{N_t} \Phi_{q_{i_s}}^T(n, v^n; q) \phi^n, \quad i_s = 0, \dots, 5. \quad (10)$$

И справедлива следующая лемма.

Лемма. Градиент функционала (5) имеет вид

$$\begin{aligned} I'(q)[0] &= -\frac{d}{h^2} \phi_1^0, \\ I'(q)[s] &= -\frac{\phi_{j_s}^0}{\tau} - d \frac{\phi_{j(s-1)}^0 - 2\phi_{j_s}^0 + \phi_{j(s+1)}^0}{h^2} - r^0 \phi_{j_s}^0 + \frac{2r^0 q_s}{K} \phi_{j_s}^0, \quad s = 1, 2, 3, 4, \\ I'(q)[5] &= -\frac{d}{h^2} \phi_{N_x-1}^0, \end{aligned} \quad (11)$$

где $\phi_{j_s}^0$ соответствует значению ϕ в точке $(x_j = s+1, t_0)$ и функция ϕ_j^n удовлетворяет сопряженной задаче

$$\begin{aligned} \frac{\phi_j^{n-1} - \phi_j^n}{\tau} &= d \frac{\phi_{j-1}^n - 2\phi_j^n + \phi_{j+1}^n}{h^2} + r^n \phi_j^n - \frac{2r^n v_j^n}{K} \phi_j^n - \frac{[\phi]}{\tau h}, \quad j = 1, \dots, N_x - 1, \quad n = 1, \dots, N_t, \\ \phi_j^{N_t} &= 0, \quad j = 0, \dots, N_x, \\ \phi_0^n &= \frac{d}{h^2} \phi_1^n, \quad n = 0, \dots, N_t - 1, \\ \phi_{N_x}^n &= \frac{d}{h^2} \phi_{N_x-1}^n, \quad n = 0, \dots, N_t - 1, \end{aligned} \quad (12)$$

где $[\phi] = 2w(v_i^k - f_{ik})$ при $(x_j, t_n) = (x_i, t_k)$ и $[\phi] = 0$ при $(x_j, t_n) \neq (x_i, t_k)$.

Доказательство. Функция E , определенная по формуле (8), имеет вид

$$\begin{aligned} \frac{E(v, q, \phi)}{\tau h} &= \frac{I(q)}{\tau h} + \sum_{j=1}^{N_x-1} \sum_{n=1}^{N_t-1} \left[\frac{v_j^{n+1} - v_j^n}{\tau} - d \frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{h^2} - r^n v_j^n + \frac{r^n (v_j^n)^2}{K} \right] \phi_j^n + \\ &+ \sum_{j=1}^{N_x-1} \left[\frac{v_j^1 - p_j}{\tau} - d \frac{p_{j+1} - 2p_j + p_{j-1}}{h^2} - r^0 p_j + \frac{r^0 (p_j)^2}{K} \right] \phi_j^0 + \\ &+ \sum_{n=0}^{N_t-1} \left[v_0^{n+1} - \frac{1}{3} (4v_1^{n+1} - v_2^{n+1}) \right] \phi_0^n + \sum_{n=0}^{N_t-1} \left[v_{N_x}^{n+1} - \frac{1}{3} (4v_{N_x-1}^{n+1} - v_{N_x-2}^{n+1}) \right] \phi_{N_x}^n = 0. \end{aligned}$$

Находим производную функции E по v_j^n , что из (9) есть формула для определения функции ϕ_j^n :

$$\frac{[\phi]}{\tau h} + \frac{\phi_j^{n-1} - \phi_j^n}{\tau} - d \frac{\phi_{j-1}^n - 2\phi_j^n + \phi_{j+1}^n}{h^2} - r^n \phi_j^n + \frac{2r^n u_j^n}{K} \phi_j^n = 0,$$

при $j = 1, \dots, N_x - 1, n = 1, \dots, N_t$. В случаях $j = 0$ и $j = N_x$ для всех $0 \leq n \leq N_t$

$$\phi_0^{n-1} = \frac{d}{h^2} \phi_1^n, \quad \phi_{N_x}^{n-1} = \frac{d}{h^2} \phi_{N_x-1}^n.$$

При всех $0 \leq j \leq N_x$ составляющая v_j^0 не входит в выражение для функции E , поэтому можно положить

$$\phi_j^{N_t} = 0, \quad j = 0, \dots, N_x.$$

Таким образом, получили постановку сопряженной задачи (12).

Так как $p_j = q_{s-1}$ при $x_j = s$, находим производные функции E по q_s :

$$-\frac{\phi_{j_s}^0}{\tau} - d \frac{\phi_{j(s-1)}^0 - 2\phi_{j_s}^0 + \phi_{j(s+1)}^0}{h^2} - r^0 \phi_{j_s}^0 + \frac{2r^0 q_s}{K} \phi_{j_s}^0,$$

при $s = 2, 3, 4, 5$, где $\phi_{j_s}^0$ соответствует значению ϕ в точке $(x_j = s, t_0)$. При $s = 1$ и $s = 6$ производные соответственно

$$E_{q_0} = -\frac{d}{h^2} \phi_1^0, \quad E_{q_5} = -\frac{d}{h^2} \phi_{N_x-1}^0.$$

Тогда из (10) следует (11), что завершает доказательство леммы.

2.2. Градиент функционала в непрерывной постановке

Градиент функционала (4) имеет вид

$$J'(q) = -\psi(x, 1), \quad (13)$$

где функция $\psi(x, t)$ удовлетворяет решению сопряженной задачи

$$\begin{aligned} \psi_t &= -d\psi_{xx} - r(t)\psi + \frac{2r(t)u}{K}\psi + \xi, \quad 1 \leq t \leq T, \quad l \leq x \leq L, \\ \psi(x, T) &= 0, \quad l \leq x \leq L, \\ \psi_x(l, t) &= \psi_x(L, t) = 0, \quad 1 \leq t \leq T, \end{aligned}$$

где $\xi = 2w \sum_{i=1}^{N_1} \sum_{k=1}^{N_2} \int_l^L \int_1^T |u(x, t; q) - f(x, t)|^2 \delta(x - x_i) \delta(t - t_k) dt dx$.

В [15] приведен вывод формулы (13) для более общей постановки задачи (2) с произвольной правой частью.

Отметим, что в силу свойств дельта-функции сопряженная задача может быть записана в виде (см. [16])

$$\begin{aligned} \psi_t &= -d\psi_{xx} - r(t)\psi + \frac{2r(t)u}{K}\psi, \quad 1 \leq t \leq T, \quad l \leq x \leq L, \\ \psi(x, T) &= 0, \quad l \leq x \leq L, \\ \psi_x(l, t) &= \psi_x(L, t) = 0, \quad 1 \leq t \leq T, \\ [\psi]_{x=x_i} &= 2w(u(x_i, t_k; q) - f_{ik}), \quad i = 1, \dots, N_1, \quad k = 1, \dots, N_2. \end{aligned}$$

3. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Для решения обратной задачи (2)–(3) в качестве синтетических данных f_{ik} были взяты значения решения прямой задачи при значениях параметров, представленных в табл. 1 и 2, в каждой

Таблица 2. Дискретизованная функция начальной плотности $Q(x)$, используемая как точное решение q_{ex} обратной задачи

Параметр	q_0	q_1	q_2	q_3	q_4	q_5
Значение	5.8	1.7	1.9	1	0.95	0.7

десятой точке по $x < N_x$ и в каждой двадцать пятой точке по $t \geq 50$, т.е. $N_1 = 5$ и $N_2 = 22$. Значения параметров в табл. 2 эмитируют реальные данные новостного сайта Digg.com, представленные в [13].

Кроме значения функционала также измерялась относительная ошибка

$$\delta = \frac{\|q_{\text{ex}} - \tilde{q}\|^2}{\|q_{\text{ex}}\|^2}.$$

Здесь q_{ex} — точные значения дискретизованной функции $Q(x)$, а $\tilde{q}$ — найденное решение обратной задачи, которое соответствует минимуму функционала $J(q)$ (5).

3.1. Градиентные методы

Решение задачи минимизации целевых функционалов $J(q)$ и $I(q)$ было получено с использованием двух градиентных методов:

1. Метод градиентного спуска (МГС) (см. [17], [18]) — классический одношаговый градиентный метод (1) с постоянным параметром спуска $\alpha_k = \alpha$. Скорость сходимости по функционалу (см. [18])

$$J(q^m) \leq \frac{\|q^0 - q_{\text{ex}}\|^2}{m\alpha(1 - \alpha\|A\|^2)}.$$

2. Многоуровневый градиентный метод (МГМ) (см. [19], [20]) — это метод локальной оптимизации функции от нескольких переменных. Он является модификацией метода градиентного спуска, что позволяет значительно увеличить скорость сходимости. Алгоритм метода имеет следующий вид:

$$\begin{aligned} q^{m+1} &= \theta_{m+1} z^m + (1 - \theta_{m+1}) y^m, \quad \theta_{m+1} \in \arg \min_{\theta \in [0;1]} J(\theta z^m + (1 - \theta) y^m), \\ y^{m+1} &= q^{m+1} - \zeta_{m+1} J'(q^{m+1}), \quad \zeta_{m+1} \in \arg \min_{\zeta \geq 0} J(q^{m+1} - \zeta J'(q^{m+1})), \\ z^{m+1} &= z^m - \eta_{m+1} J'(q^{m+1}), \quad \eta_{m+1} = \frac{1}{2L_{m+1}} + \sqrt{\frac{1}{4L_{m+1}^2} + \eta_m^2}, \quad \eta_0 = 0. \end{aligned}$$

В случае МГМ скорость сходимости по функционалу оценивается следующим образом (см. [21]):

$$J(q^m) \leq \frac{4L\|q^0 - q_{\text{ex}}\|^2}{(m+1)^2},$$

где L — константа Липшица для градиента функционала

$$\|J'(q^{m+1}) - J'(q^m)\|_2 \leq L\|q^{m+1} - q^m\|_2.$$

Были применены и проанализированы МГС и МГМ с непрерывным и дискретным видами градиентов. Причем в случае дискретного градиента МГС имеет параметр спуска $\alpha^{(1)} = 10^{-7}$ и минимум по ζ в МГМ находится на отрезке $[0, 10^{-6}]$ с шагом 10^{-7} . В случае непрерывного градиента $\alpha^{(2)} = 10^{-3}$, а минимум по $\zeta \in [0, 10^{-5}]$ определяется с шагом 10^{-6} .

Таблица 3. Решение обратной задачи, полученное методом роя частиц

Параметр	$\hat{q}_0$	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{q}_4$	$\hat{q}_5$
Значение	5.809	1.697	1.901	0.999	0.949	0.701

Таблица 4. Результаты, полученные методами градиентного спуска и его модификаций

Метод	q^0	$J(q), \times 10^{-4}$	$\delta, \times 10^{-4}$	N_{iter}	$t_{\text{ЭВМ}}$
МГС I	$\hat{q}$	9.604	6.274	5380	3
МГМ I	$\hat{q}$	9.544	6.273	397	3
МГМ II	0	9.029	6.262	245	2
МГС II	0	9.032	6.262	16374	9
МГС III	$\hat{q}$	9.847	6.276	166862	93
МГМ III	$\hat{q}$	9.793	6.276	1372	11
МГМ IV	0	9.047	6.261	1630	12
МГС IV	0	9.101	6.266	1005999	569
МГМ V	0	11.26	6.254	1586	25

Примечание. В столбце начального приближения q^0 под $\hat{q}$ понимается вектор решения обратной задачи, полученный с помощью метода роя частиц (представленный в табл. 3), а 0 — нулевой вектор N_{iter} — число итераций, $t_{\text{ЭВМ}}$ — время работы программы.

3.2. Конечно-разностные схемы решения прямой и сопряженной задач

Для построения разностных схем введем равномерную сетку в замкнутой области $\bar{D} = \{(x, t) | l \leq x \leq L, 1 \leq t \leq T\}$:

$$\bar{\omega} = \{(x_j, t_n) | x_j = l + jh, t_n = 1 + n\tau, j = 0, \dots, N_x, n = 0, \dots, N_t\},$$

где $h = (L - l)/N_x$ и $\tau = (T - 1)/N_t$.

Для применения классического непрерывного подхода функция начального приближения $Q(x)$ определяется из вектора q (см. табл. 2) интерполяцией кубическими сплайнами.

Прямая задача (2) решается с помощью явной конечно-разностной схемы с порядком аппроксимации $O(\tau + h^2)$ в виде (6).

Положим в численных расчетах $l = 1, L = 6, T = 24, N_x = 50$ и $N_t = 575$. Такие значения l, L были выбраны в соответствии с данными, представленными в [13], которые иллюстрируют, что интерес к информации проявляется у пользователей с расстоянием x в пределах от 1 до 6. Именно в этих границах происходит значительный вклад в изменение плотности активных пользователей $u(x, t)$. Значения разбиения сетки N_x и N_t выбраны таким образом, чтобы выполнялось условие Куранта—Фридрихса—Леви

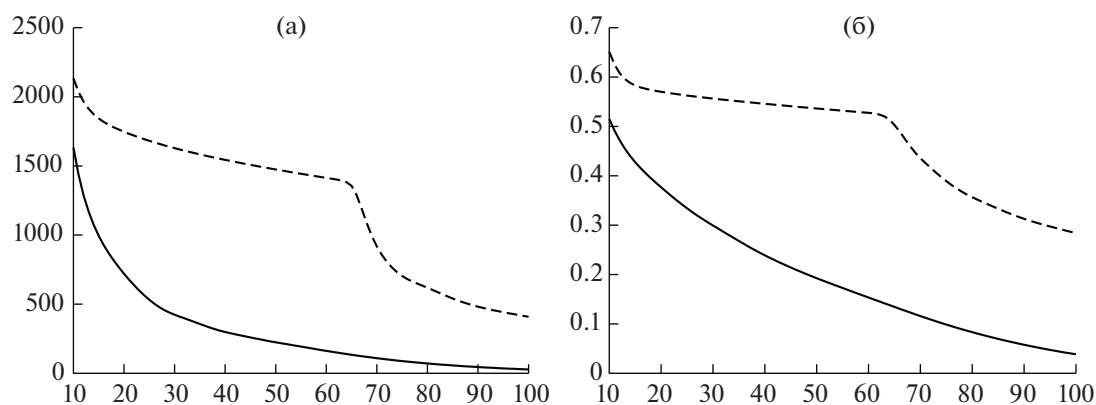
$$\tau \leq \frac{2h^2}{4d + r_c h^2}$$

с учетом того, что $r_c = \max_t r(t) = 0.44, h = 0.1$ и $\tau = 0.04$.

Для анализа устойчивости решения прямой задачи была реализована схема Кранка—Николсон с порядком аппроксимации $O(\tau^2 + h^2)$.

3.3. Сравнительный анализ непрерывной и дискретной постановок

Проведем сравнительный анализ подходов решения обратной задачи: с точным подсчетом градиента для дискретной задачи (6)—(3) и с непрерывной постановкой и дискретизацией при вычислении градиента (2)—(3). Этапы схождения и различия алгоритмов решения указанных обратных задач приведены в таблице 3. Отметим, что в обоих случаях идентично решена прямая задача, а сопряженные задачи решены одной конечно-разностной схемой, которые входят в соответствующие выражения для вычисления градиентов целевого функционала. К решению



Фиг. 1. Графики убывания функционалов $J(q)$ и $I(q)$ (а) и относительной погрешности δ (б) на итерациях с 10-й по 100-ю. Сплошная линия характеризует использование непрерывного градиента, а пунктирная — дискретного градиента целевого функционала.

соответствующих задач минимизации применены одинаковые алгоритмы. В качестве сравнения двух подходов анализируется значение целевого функционала и относительная погрешность восстановленных решений. Решение, полученное с помощью градиентных методов, зависит от начального приближения. Было рассмотрено два варианта выбора начального приближения (см. [19], [22], [23]): решение обратной задачи, полученное с помощью метода глобальной оптимизации роя частиц (МРЧ), представленное в табл. 3, и нулевое начальное приближение. Нулевое начальное приближение выбрано из физической постановки, так как q^0 описывает первую реакцию пользователей на новость в отсутствие дополнительной информации об этой реакции. В приведенной реализации МРЧ функционал $J(\hat{q}) = 1.114 \times 10^{-3}$ и относительная погрешность $\delta = 1.533 \times 10^{-3}$.

В табл. 4 представлены результаты, полученные следующими методами:

- МГС I и МГМ I — методы с непрерывным градиентом и начальным приближением в виде решения, полученного МРЧ;
- МГС II и МГМ II — методы с непрерывным градиентом и нулевым начальным приближением;
- МГС III и МГМ III — методы с дискретным градиентом и начальным приближением в виде решения, полученного МРЧ;
- МГС IV и МГМ IV — методы с дискретным градиентом и нулевым начальным приближением;
- МГМ V — метод с нулевым начальным приближением и дискретным градиентом в случае, когда прямая и сопряженная задачи численно решались конечно-разностной схемой Кранка—Николсон.

Из табл. 4 видно, что МГМ находит решение с заданной точностью, производя более чем в 10 раз меньше итераций, чем МГС при одинаковых начальных условиях и формах нахождения градиента целевого функционала. В случае начального приближения $\hat{q}$ и непрерывного градиента (т.е. МГС I и МГМ I) времена работ программ не отличаются, однако при других вариантах реализации градиентных методов МГМ находит решение с заданной точностью в несколько раз быстрее МГС. При нулевом начальном приближении методы работают в несколько раз дольше, но достигают большей точности только на шестом знаке после запятой. В случае дискретного градиента МГМ достигает решения того же порядка в несколько раз дольше, а МГС — более чем в 30 раз дольше по времени. При этом МГМ V имеет наименьшее среди всех рассмотренных методов значение погрешности, однако при этом метод работает в несколько раз дольше других реализаций МГМ. На фиг. 1 непрерывными линиями изображены кривые, полученные МГМ II в случае непрерывного вычисления градиента функционала, а пунктирными линиями — кривые, полученные МГМ IV в случае дискретной реализации градиента. Из фиг. 1 следует, что значение функционала и погрешность в случае МГМ II убывают быстрее.

Таблица 5. Результаты решения обратной задачи с уровнем зашумленности в данных $\varepsilon = 1, 5, 10\%$, полученные многоуровневым градиентным методом с градиентом в дискретной и непрерывной постановках

Начальное приближение	$\varepsilon, \%$	Дискретный градиент (11)		Непрерывный градиент (13)	
		$I(q)$	δ	$J(q)$	δ
Нулевое	1	1.133×10^{-2}	6.706×10^{-4}	9.628×10^{-2}	7.356×10^{-4}
	5	1.475×10^{-1}	8.447×10^{-4}	3.272×10^{-3}	6.399×10^{-4}
	10	3.461×10^{-2}	7.305×10^{-4}	3.286×10^{-3}	6.407×10^{-4}
Решение МРЧ	1	1.242×10^{-3}	6.271×10^{-4}	1.403×10^{-3}	6.262×10^{-4}
	5	3.165×10^{-2}	6.431×10^{-4}	1.865×10^{-3}	6.261×10^{-4}
	10	16.795	1.707×10^{-2}	11.289	1.388×10^{-2}

Был проанализирован случай зашумленных данных:

$$f_{ik}^{\varepsilon} = f_{ik} + \varepsilon \gamma \max |f_{ik}|, \quad i = 1, \dots, N_1, \quad k = 1, \dots, N_2,$$

где $\varepsilon \in (0, 1)$ — уровень шума, а γ — случайная величина, имеющая стандартное нормальное распределение.

Для зашумленных данных обратной задачи был реализован МГМ с различными начальными приближениями, а прямая задача решалась методом конечных разностей схемой Кранка—Николсон. Результаты численных экспериментов для такой постановки обратной задачи в случае дискретного и непрерывного градиентов приведены в табл. 5. Можно видеть, что в случае нулевого начального приближения метод с дискретным градиентом имеет наименьшее значение относительной ошибки при шуме в 1%, в то время как метод с непрерывным градиентом — при шуме в 5 и 10%. В случае начального приближения в виде решения, полученного с помощью МРЧ, точности восстановленных решений обратной задачи δ различаются незначительно.

В работе проведено исследование двух подходов к решению задачи восстановления источника, в рамках которых были реализованы две конечно-разностные схемы, в которых порядок аппроксимации по пространственной переменной совпадает и равен 2, а по временной имеют 1-й и 2-й (Кранка—Николсон) порядки. В таблицах 5 и 6 экспериментально показано, что результат сходимости не чувствителен к порядку аппроксимации по временной переменной. Исходя из оценок сходимости и итерационных выражений определения решения обратной задачи, порядок аппроксимации по пространственной переменной при решении прямой и сопряженной задач имеет большое влияние на точность полученного решения и, скорее всего, на скорость сходимости градиентных методов. В случае определения плотности начального распределения пользователей относительно популярной новости в онлайн социальных сетях с большой пропускной способностью (Twitter, Facebook, Reddit и др.) требуется качественно определить тип вовлеченности пользователей и структуру распределения. Для получения более точных оценок взаимосвязи скорости сходимости градиентных методов с точностью аппроксимации необходимы отдельные исследования.

ЗАКЛЮЧЕНИЕ

В работе проведен сравнительный анализ численного решения задачи об источнике для диффузионно-логистической модели по данным о диффузионном процессе в фиксированные моменты времени и точках пространства градиентными методами для случаев непрерывной (классической) и дискретной постановок. Получены выражения для вычисления градиента целевого функционала в случае двух постановок, связанных с решением соответствующих сопряженных задач.

Прямая и сопряженные задачи были решены с помощью явной схемы с порядком аппроксимации $O(\tau + h^2)$ и соблюдением условия Куранта—Фридрихса—Леви и полунейвной схемы Кранка—Николсон, имеющей порядок аппроксимации $O(\tau^2 + h^2)$. Задача минимизации целевого функционала была решена методами градиентного спуска с постоянным параметром спуска и многоуровневым градиентным методом. Были проанализированы два варианта начального приближения решения обратной задачи: нулевое и полученное методом роя частиц. Показано, что в случае ненулевого начального приближения с использованием выражения для градиента целевого

функционала в непрерывной постановке времени работы программ не отличаются, однако при других вариантах реализации градиентных методов многоуровневый градиентный метод находит решение с заданной точностью в несколько раз быстрее метода градиентного спуска. При нулевом начальном приближении методы работают в несколько раз дольше, но достигают большей точности. В случае дискретного градиента многоуровневый градиентный метод достигает решения того же порядка в несколько раз дольше, а метод градиентного спуска — более чем в 30 раз дольше по времени.

Показано, что в случае аппроксимации кубическими сплайнами дискретных функций в прямой задаче различия в точности полученных решений обратных задач для непрерывной и дискретной постановок наблюдаются только в шестом знаке после запятой в случае незашумленных синтетических данных, описывающих динамику вовлеченных пользователей в онлайн социальной сети. Однако машинное время решения обратной задачи в дискретной форме больше, чем непрерывный аналог.

СПИСОК ЛИТЕРАТУРЫ

1. *Hestenes M.R., Stiefel E.* Methods of conjugate gradients for solving linear systems // J. Res. Natl. Inst. Stand. Technol. 1952. V. 49. № 6. P. 409–436.
2. *Fletcher R., Reeves C.M.* Function minimization by conjugate gradients // Comput. J. 1964. V. 7. № 2. P. 149–154.
3. *Byrd R.H., Nocedal J., Yuan Y.* Global convergence of a class of quasi-Newton methods on convex problems // SIAM J. Numer. Anal. 1987. V. 24. № 5. P. 1171–1190.
4. *Hasanov A.* Simultaneous determination of source terms in a linear parabolic problem from the final overdetermination: Weak solution approach // J. Math. Anal. Appl. 2007. V. 330. № 2. P. 766–779.
5. *Prilepko A.I., Kamynin V.L., Kostin A.B.* Inverse source problem for parabolic equation with the condition of integral observation in time // J. Inverse Ill-Posed Probl. 2018. V. 26. № 4. P. 523–539.
6. *Aida-zade K.A., Rahimov A.B.* Numerical solution to inverse source problems for linear parabolic equation // IFAC-PapersOnLine. 2018. V. 51. № 30. P. 231–236.
7. *Cheng J., Liu J.* An inverse source problem for parabolic equations with local measurements // Appl. Math. Lett. 2020. V. 103. P. 106213.
8. *Карчевский А.Л.* Корректная схема действий при численном решении обратной задачи оптимизационным методом // Сиб. журн. вычисл. матем. 2008. Т. 11. № 2. С. 139–149.
9. *Евтушенко Ю.Г.* Оптимизация и быстрое автоматическое дифференцирование. М.: ВЦ РАН, 2013.
10. *Евтушенко Ю.Г., Засухина Е.С., Зубов В.И.* О численном подходе к оптимизации решения задачи Бюргерса с помощью граничных условий // Ж. вычисл. матем. и матем. физ. 1997. Т. 37. № 12. С. 1449–1458.
11. *Евтушенко Ю.Г.* Приближенный расчет задач оптимального управления // Приклад. матем. и механика. 1970. Т. 34. № 1. С. 95–104.
12. *Черноусько Ф.Л., Колмановский В.Б.* Вычислительные и приближенные методы оптимального управления // Итоги науки и техн. Сер. Мат. анализ. 1977. Т. 14. С. 101–166.
13. *Wang H., Wang F., Xu K.* Diffusive logistic model towards predicting information diffusion in online social networks // Proceed. 32nd Inter. Conf. on Distribut. Comput. Syst. Workshop. 2012. P. 133–139.
14. *Krivorotko O., Zvonareva T.* Numerical solution of the inverse problem for diffusion-logistic model arising in online social networks // Commun. Comput. Info. Sci. 2021. V. 1476. P. 444–459.
15. *Krivorotko O., Kabanikhin S., Zhang S., Kashtanova V.* Global and local optimization in identification of parabolic systems // J. Inverse Ill-Posed Probl. 2020. V. 28. № 6. P. 899–913.
16. *Владимиров В.С.* Уравнения математической физики. М.: Наука, 1981.
17. *Алифанов О.М., Артюхин Е.А., Румянцев С.В.* Экстремальные методы решения некорректных задач. М.: Наука, 1988.
18. *Kabanikhin S., Penenko A.* Gradient-type methods in inverse parabolic problems // J. Phys. Conf. Ser. 2008. V. 135. P. 012054.
19. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. М.: МФТИ, 2018.
20. *Nesterov Y., Gasnikov A., Guminov S., Dvurechensky P.* Primal–dual accelerated gradient methods with small-dimensional relaxation oracle // Optim. Methods Softw. 2020. P. 1–38.
21. *Gasnikov A.V., Nesterov Y.E.* Universal method for stochastic composite optimization problems // Comput. Math. Math. Phys. 2018. V. 58. № 1. P. 48–64.
22. *Kaltenbacher B., Neubauer A., Scherzer O.* Iterative regularization methods for nonlinear ill-posed problems. New York: De Gruyter, 2008.
23. *Kaltenbacher B.* All-at-once versus reduced iterative methods for time dependent inverse problems // Inverse Probl. 2017. V. 33. P. 064002.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 517.958

ДИСКРЕТНАЯ МОДЕЛЬ УРАВНЕНИЯ БОЛЬЦМАНА ДЛЯ ДЕВЯТИ
СКОРОСТЕЙ: РЕШЕНИЕ В ВИДЕ СУММЫ ВАЙЛДА
И ПРИЛОЖЕНИЯ К МОДЕЛИРОВАНИЮ
НЕСЖИМАЕМЫХ ТЕЧЕНИЙ

© 2022 г. О. В. Ильин

119133 Москва, ул. Вавилова, 40, ВЦ ФИЦ ИУ РАН, Россия

e-mail: oilyin@gmail.com

Поступила в редакцию 27.07.2021 г.
Переработанный вариант 05.09.2021 г.
Принята к публикации 16.12.2021 г.

Рассмотрена дискретная кинетическая модель уравнения Больцмана на плоскости с девятью скоростями. В пределе малых длин свободного пробега и малых потоковых скоростей модель описывает течения вязкой несжимаемой жидкости или газа. Полная дискретизация модели по пространственным и временным переменным, которая в частности необходима для численного решения, осуществляется с помощью усеченной суммы Вайлда. Показано, что построенная схема имеет второй порядок точности. В качестве приложения предлагаемого метода построены численные решения эталонных задач: вихри Тэйлора–Грина, течение в полости с подвижной границей. Проведено сравнение результатов моделирования с решениями на основе классической решеточной модели уравнения Больцмана с девятью скоростями. Библ. 43. Фиг. 6.

Ключевые слова: решеточные уравнения Больцмана, уравнение Бхатнагара–Гросса–Крука (БГК), уравнения вязкой жидкости.

DOI: 10.31857/S0044466922040081

1. ВВЕДЕНИЕ

Метод решеточных уравнений Больцмана (РУБ) в настоящее время популярен для моделирования как академических, так и прикладных задач в области механики сплошной среды и разреженных течений [1]. Данный подход основан на кинетической модели БГК [2], которая дискретизирована специальным образом. В общем виде решеточная модель уравнения Больцмана записывается следующим образом:

$$D_i f_i = \frac{1}{\tau} (f_i^{eq} - f_i), \quad i = 1, \dots, N,$$

где $f_i(t, \mathbf{r})$ — функция распределения (ФР), отвечающая частицам, движущимся со скоростью $\mathbf{c}_i$, $t, \mathbf{r}$ — временные и пространственные переменные, τ — характерное время релаксации системы, f_i^{eq} — локальное распределение Максвелла, N есть число дискретных скоростей, $D_i = \frac{\partial}{\partial t} + \mathbf{c}_i \cdot \frac{\partial}{\partial \mathbf{r}}$.

Для наиболее популярных РУБ локальное состояние равновесия имеет вид полиномиальной функции от потоковой скорости, получаемой из распределения Максвелла разложением в ряд Тейлора по потоковой скорости с конечным числом членов. Отметим, что для данной модели взаимодействия частиц, т.е. столкновения, описываются феноменологическим образом, правая часть (интеграл столкновений) отражает стремление ФР к локальному равновесию со скоростью, пропорциональной отклонению от этого локального равновесия.

Другим способом дискретизации в пространстве скоростей является метод дискретных скоростей для уравнения Больцмана [3]–[5]. В данном подходе столкновения частиц рассматриваются явным образом, в общем виде модель записывается в виде:

$$D_i f_i = \sum_{jkl=1}^N A_{kl}^{ij} (f_k f_l - f_i f_j) \equiv I_i[f_1, \dots, f_N], \quad i = 1, \dots, N, \quad (1)$$

где $A_{kl}^{ij} = A_{ij}^{kl} \geq 0$ есть величины, пропорциональные вероятности того, что частицы со скоростями $\mathbf{c}_i, \mathbf{c}_j$ после столкновения имеют скорости $\mathbf{c}_k, \mathbf{c}_l$. Важным свойством дискретных кинетических моделей уравнения Больцмана является то, что для всех уравнений вида (1) выполняется H -теорема, т.е.

$$\frac{dH}{dt} \leq 0, \quad H(t) = \int d\mathbf{r} \sum_i f_i \lg(f_i),$$

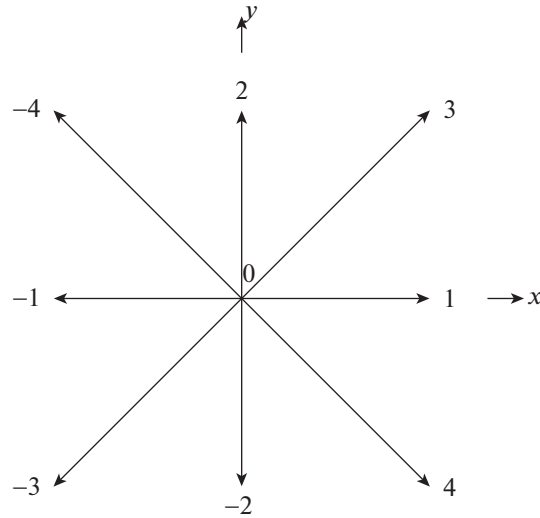
здесь следует отметить, что для стандартных решеточных уравнений Больцмана H -теорема, вообще говоря, неверна [6]–[8], чтобы обеспечить рост энтропии (убывание H функции) в системе, описываемой РУБ, требуется выбирать состояние локального равновесия таким образом, чтобы максвелловское распределение зависело от потоковой скорости существенно неполиномиальным образом [9]–[12].

На первый взгляд метод дискретных скоростей для уравнения Больцмана более перспективен, чем метод РУБ. Действительно, столкновения задаются явным образом аналогично уравнению Больцмана, H -теорема верна по построению. Однако в настоящее время метод дискретных скоростей для уравнения Больцмана менее популярен. Одна из главных проблем распространенных дискретных моделей уравнения Больцмана заключается в том, что они не обладают достаточно изотропным набором дискретных скоростей, а данное свойство необходимо для того, чтобы в пределе малых длин свободного пробега правильно описывалось поведение вязкой жидкости или газа [13]. Например, уравнение Броудэлла на плоскости, которое подробно исследовалось во многих работах [3]–[19], имеет всего четыре возможные скорости движения частиц в сплошнородном пределе не может правильно воспроизвести даже уравнения Эйлера. Также следует отметить еще одну особенность дискретных моделей уравнения Больцмана: гипотеза молекулярного хаоса не выполняется для таких моделей, например, для уравнения Броудэлла на плоскости две сталкивающиеся частицы имеют существенно ненулевую вероятность столкнуться еще раз после одного дополнительного (промежуточного) столкновения, т.е. данные частицы оказываются скоррелированными перед столкновением [20]. С теоретической точки зрения данный эффект нежелателен, но его влияние в пределе сплошной среды на прикладные задачи, рассматриваемые в работе, незначительно. Дискретные модели с достаточным числом скоростей для описания уравнений гидродинамики были рассмотрены относительно недавно [21], [22]. Кроме того, была построена модель для девяти скоростей, правильно описывающая в пределе малых длин свободного пробега уравнения Навье–Стокса для случая медленных несжимаемых течений [23]. Приближенные решения данной модели рассмотрены в настоящей работе.

Для уравнения Больцмана в случае максвелловских псевдомолекул и для дискретных моделей уравнения Больцмана возможно построить решение в виде рядов Вайлда [24], [25], представляющие собой бесконечные суммы со свертками интеграла столкновений разных порядков. Данное свойство позволяет построить приближенные решения этих уравнений в виде конечных сумм Вайлда [26], [27]. В настоящей работе рассмотрена аппроксимация дискретной модели уравнения Больцмана с девятью скоростями с помощью варианта конечной суммы Вайлда. Для данной аппроксимации получены выражения вязкости в сплошнородном пределе и численно решены модельные задачи: двумерные вихри Тэйлора–Грина, течение в квадратной полости с твердыми границами, причем верхняя граница является подвижной. Проведено сравнение результатов моделирования с результатами расчетов на основе стандартной РУБ с девятью скоростями.

2. ДИСКРЕТНАЯ КИНЕТИЧЕСКАЯ МОДЕЛЬ НА ПЛОСКОСТИ С ДЕВЯТЬЮ СКОРОСТЯМИ

В настоящей работе удобно ввести следующие обозначения. Скалярное произведение двух векторов $\mathbf{X} \in \mathbb{R}^D$ и $\mathbf{Y} \in \mathbb{R}^D$ (здесь D – размерность пространства) с координатами X_α, Y_β ;



Фиг. 1. Дискретные скорости для рассматриваемой модели.

$\alpha, \beta = 1, \dots, D$ обозначается как $\mathbf{X} \cdot \mathbf{Y} = \sum_{\alpha} X_{\alpha} Y_{\alpha}$, кроме того, $\mathbf{XY}$ есть тензор второго ранга с компонентами $X_{\alpha} Y_{\beta}$. Если $\mathbf{T}, \mathbf{S}$ – тензоры второго ранга, то $:$ обозначает оператор свертки, т.е. $\mathbf{T} : \mathbf{S} = \sum_{\alpha\beta} T_{\alpha\beta} S_{\alpha\beta}$. Также, если $\mathbf{X}$ вектор, а $\mathbf{T}$ – тензор второго ранга, то произведение $\mathbf{X} \cdot \mathbf{T}$ есть вектор $\mathbf{Z}$ с координатами $Z_{\alpha} = \sum_{\beta} X_{\beta} S_{\alpha\beta}$.

Состояние локального равновесия для дискретных кинетических моделей (1) имеет вид (см. [29]):

$$f_i^{eq} = \exp(a + \mathbf{b} \cdot \mathbf{c}_i + d c_i^2), \quad i = 1, \dots, N, \quad (2)$$

где $a, \mathbf{b}, d$ – параметры, вычисляемые из законов сохранения

$$\sum_i f_i^{eq} = \rho, \quad \sum_i f_i^{eq} \mathbf{c}_i = \rho \mathbf{u}, \quad \sum_i f_i^{eq} c_i^2 = \rho(u^2 + D\theta),$$

здесь $\rho, \mathbf{u}, \theta$ – плотность, потоковая скорость, температура, N – число дискретных скоростей. В случае $\mathbf{u} = 0$, а также постоянных плотности и температуры имеем состояние абсолютного равновесия, которое обозначим через w_i , т.е.

$$w_i = \exp(a_0 + c_0 c_i^2),$$

где a_0, d_0 – числовые параметры, по аналогии с теорией РУБ будем называть величины w_i , соответствующие дискретной скорости $\mathbf{c}_i$, *весами*. Из законов сохранения получаем следующие соотношения для весов:

$$\sum_i w_i = 1, \quad \sum_i w_i \mathbf{c}_i = 0, \quad \sum_i w_i \mathbf{c}_i \mathbf{c}_i = \theta_0 \delta, \quad (3)$$

где $\mathbf{c}_i \mathbf{c}_i$ есть тензор с элементами $c_{i,\alpha} c_{i,\beta}$, $\alpha, \beta = 1, \dots, D$, где D – число пространственных переменных.

Рассматриваемая дискретная модель уравнения Больцмана для двух пространственных переменных состоит из скоростей трех типов [23] (см. фиг. 1): нулевая скорость $\mathbf{c}_0 = (0, 0)$ с весом $w_0 = 16/36$; четыре скорости, параллельные осям x, y , т.е. $\mathbf{c}_{\pm 1} = (\pm 1, 0)c$, $\mathbf{c}_{\pm 2} = (0, \pm 1)c$ с весом $w_0 = 4/36$; четыре скорости, направленные по диагоналям $\mathbf{c}_{\pm 3} = (\pm 1, \pm 1)c$, $\mathbf{c}_{\pm 4} = (\pm 1, \mp 1)c$ с весами $w_0 = 1/36$, где c – положительная величина. Также данная модель описывает среду с темпера-

турой $\theta_0 = \sum_i w_i c_i^2 = c^2/3$. Очевидно, что набор дискретных скоростей для данной модели совпадает с дискретными скоростями решеточной модели Больцмана *D2Q9* (см. [1]).

Девятискоростная дискретная модель уравнения Больцмана модель записывается следующим образом [23]:

$$\frac{\partial f_1}{\partial t} + c \frac{\partial f_1}{\partial x} = \alpha J_0 + \beta(J_1 + J_2) + \lambda(J_6 + J_7), \quad (4)$$

$$\frac{\partial f_{-1}}{\partial t} - c \frac{\partial f_{-1}}{\partial x} = \alpha J_0 + \beta(J_3 + J_4) - \lambda(J_6 + J_7), \quad (5)$$

$$\frac{\partial f_2}{\partial t} + c \frac{\partial f_2}{\partial y} = -\alpha J_0 + \beta(J_1 + J_3) + \lambda(J_8 + J_9), \quad (6)$$

$$\frac{\partial f_{-2}}{\partial t} - c \frac{\partial f_{-2}}{\partial y} = -\alpha J_0 + \beta(J_2 + J_4) - \lambda(J_8 + J_9), \quad (7)$$

$$\frac{\partial f_3}{\partial t} + c \frac{\partial f_3}{\partial x} + c \frac{\partial f_3}{\partial y} = \gamma J_5 - \beta J_1 - \lambda(J_6 + J_9), \quad (8)$$

$$\frac{\partial f_{-3}}{\partial t} - c \frac{\partial f_{-3}}{\partial x} - c \frac{\partial f_{-3}}{\partial y} = \gamma J_5 - \beta J_4 + \lambda(J_7 + J_8), \quad (9)$$

$$\frac{\partial f_4}{\partial t} + c \frac{\partial f_4}{\partial x} - c \frac{\partial f_4}{\partial y} = -\gamma J_5 - \beta J_2 + \lambda(-J_7 + J_9), \quad (10)$$

$$\frac{\partial f_{-4}}{\partial t} - c \frac{\partial f_{-4}}{\partial x} + c \frac{\partial f_{-4}}{\partial y} = -\gamma J_5 - \beta J_3 + \lambda(J_6 - J_8), \quad (11)$$

$$\frac{\partial f_0}{\partial t} = -\beta(J_1 + J_2 + J_3 + J_4), \quad (12)$$

где $\alpha, \beta, \gamma, \lambda$ неотрицательны. Кроме того,

$$J_0 = f_{-2}f_2 - f_{-1}f_1, \quad (13)$$

отвечает за взаимодействия частиц 1 и -1 , которые после столкновения переходят в состояния 2 и -2 , схематично это взаимодействие обозначим как $(1, -1) \rightarrow (2, -2)$.

Также имеются столкновения, которые связывают частицы всех трех типов (т.е. с разными величинами кинетической энергии), имеется четыре реакции: $(1, 2) \rightarrow (0, 3)$, $(1, -2) \rightarrow (0, 4)$, $(-1, -2) \rightarrow (0, -3)$, $(-1, 2) \rightarrow (0, -4)$, т.е.

$$\begin{aligned} J_1 &= f_0f_3 - f_1f_2, & J_2 &= f_0f_4 - f_1f_{-2}, \\ J_3 &= f_0f_{-4} - f_{-1}f_2, & J_4 &= f_0f_{-3} - f_{-1}f_{-2}. \end{aligned} \quad (14)$$

Столкновения между частицами, направленными по диагоналям, т.е. $(3, -3) \rightarrow (4, -4)$, они задаются слагаемым вида

$$J_5 = f_{-4}f_4 - f_{-3}f_3. \quad (15)$$

Столкновения между частицами, направленными по диагоналям, и частицами, движущимися параллельно осям, т.е. $(-4, 1) \rightarrow (-1, 3)$, $(-3, 1) \rightarrow (-1, 4)$, $(-3, 2) \rightarrow (-4, -2)$, $(4, 2) \rightarrow (-2, 3)$, получаем слагаемые вида

$$\begin{aligned} J_6 &= f_3f_{-1} - f_{-4}f_1, & J_7 &= f_{-1}f_4 - f_{-3}f_1, \\ J_8 &= f_{-4}f_{-2} - f_{-3}f_2, & J_9 &= f_3f_{-2} - f_4f_2. \end{aligned} \quad (16)$$

Модель (4)–(16) обладает следующим важным свойством. Разложение Чепмена–Энскога [1] для данной модели дает уравнения вязкой жидкости, которые отличаются от уравнений Навье–

Стокса для несжимаемой среды только членами вида $O(M^3)$, где M есть число Маха, определяемое как $u/\sqrt{\theta_0}$, если выполняется равенство [23]

$$4\alpha = \gamma + 4\beta + 4\lambda, \quad (17)$$

причем вязкость ν равна

$$\nu = \frac{3}{4\alpha}. \quad (18)$$

Очевидно, что в пределе малых чисел Маха уравнение (4)–(16) может применяться для моделирования сплошносредных несжимаемых течений. Отдельно возникает задача численной реализации построенной модели, этот вопрос рассмотрен в следующем разделе.

3. АППРОКСИМАЦИЯ РЕШЕНИЙ УСЕЧЕННОЙ СУММОЙ ВАЙЛДА

Следуя предыдущим результатам [26], [27], рассмотрим обыкновенное дифференциальное уравнение вида

$$\frac{df}{dt} = \frac{1}{\epsilon}(P(f, f) - \lambda f), \quad f|_{t=0} = f_0, \quad (19)$$

где P — билинейный оператор, ϵ, λ — положительные константы. Уравнение (19) имеет решение, которое можно записать в следующем виде (см. [26], [27]):

$$f = \exp\left(-\frac{\lambda t}{\epsilon}\right) \sum_{n=0}^{\infty} (1 - \exp(-\lambda t/\epsilon))^n f^{(n)}, \quad f^{(n+1)} \equiv \frac{1}{(n+1)\lambda} \sum_{m=0}^n P(f^{(m)}, f^{(n-m)}), \quad (20)$$

решение в форме (20) называется *суммой Вайлда* (см. [24]).

Теперь снова рассмотрим дискретную кинетическую модель (1), но в пространственно однородном случае перепишем ее следующим образом:

$$\frac{df_i}{dt} = R_i(f, f) - A \rho f_i, \quad (21)$$

где

$$R_i(f, f) = \sum_{jkl=1}^N A_{kl}^{ij} f_k f_l + \sum_{j=1}^N \left(A - \sum_{kl=1}^N A_{kl}^{ij} \right) f_i f_j, \quad (22)$$

также константа A выбрана так

$$A = \max_{ij} \sum_{kl} A_{kl}^{ij}, \quad (23)$$

отметим, что выбор A в виде (23) гарантирует, что оператор R_i в (21), (22) неотрицателен. Допустим, что нам известно решение системы (21), (22) в момент времени t . Решение в виде суммы Вайлда для (21), (22) в момент $t + \Delta t$ (где Δt мало) по аналогии с [26], [27] аппроксимируем конечной суммой Вайлда

$$f_i(t + \Delta t) = \exp(-A \rho(t) \Delta t) \sum_{n=0}^M (1 - \exp(-A \rho(t) \Delta t))^n f_i^{(n)}(t) + (1 - \exp(-A \rho(t) \Delta t))^{M+1} f_i^{(\infty)}, \quad i = 1, \dots, N, \quad (24)$$

$$f_i^{(n)}(t) = \frac{1}{A \rho(t)(n+1)} \sum_{m=0}^n R_i(f^{(m)}(t), f^{(n-m)}(t)), \quad i = 1, \dots, N, \quad (25)$$

где $M \geq 1$, а также $f_i^{(\infty)} = \lim_{n \rightarrow \infty} f_i^{(n)}$. Отметим, что (24) выводится из бесконечной суммы Вайлда, если $f_i^{(M+k)} = f_i^{(\infty)}$, $k = 1, 2, \dots$. Возникает вопрос, как вычислить величину $f_i^{(\infty)}$, с физической

точки зрения данная ФР соответствует частице, испытавшей бесконечное число столкновений, следовательно, $f_i^{(\infty)}$ близко к локальному максвелловскому распределению, т.е. можно положить

$$f_i^{(\infty)} = f_i^{(eq)}, \quad i = 1, \dots, N.$$

Можно показать, что схема (24), (25) в пространственно однородном случае сходится к точному решению дискретной кинетической модели, при этом ошибка монотонно уменьшается с ростом M [26], [27]. Для доказательства используется свойство неотрицательности оператора R_i , которое гарантируется равенством (23).

В текущей работе основной интерес представляет вопрос описания гидродинамики с помощью дискретной кинетической модели, а не рост порядка аппроксимации суммой Вайлда кинетической модели, таким образом, величину A будем выбирать не с помощью условия (23), а из физических соображений. В дальнейшем с помощью разложения Чепмена—Энскога будет показано, что данный подход позволяет получить уравнения вязкой несжимаемой жидкости. С физической точки зрения $\rho(t)A$ характеризует величину, обратно пропорциональную времени релаксации системы τ , которое, в свою очередь, пропорционально времени свободного пробега, последнее может быть выражено из вязкости рассматриваемой жидкости или газа. Имеем

$$\frac{1}{A\rho(t)} = \tau = \frac{\nu}{\theta_0}, \quad (26)$$

так как моделируется несжимаемый газ, то $\rho(t) = \rho_0$, т.е. постоянная величина, соответственно в левой и правой части (26) стоят выражения, не зависящие от времени.

Тогда, учитывая все изложенное выше, решение системы (1) аппроксимируем суммой Вайлда для $M = 1$, при этом адвекцию учтем аналогично РУБ (см. [1]), в итоге в пространственно неоднородном случае имеем следующую схему:

$$f_i(t + \Delta t, \mathbf{r} + \mathbf{c}_i \Delta t) = \exp\left(-\frac{\Delta t}{\tau}\right) f_i(t, \mathbf{r}) + \\ + \exp\left(-\frac{\Delta t}{\tau}\right) \left(1 - \exp\left(-\frac{\Delta t}{\tau}\right)\right) f_i^{(1)}(t, \mathbf{r}) + \left(1 - \exp\left(-\frac{\Delta t}{\tau}\right)\right)^2 f_i^{eq}(t, \mathbf{r}), \quad (27)$$

$$f_i^{(1)}(t, \mathbf{r}) = \tau R_i(f, f)(t, \mathbf{r}) = \tau \sum_{jkl=1}^N A_{kl}^{ij} (f_k f_l - f_i f_j)(t, \mathbf{r}) + f_i(t, \mathbf{r}), \quad (28)$$

где время релаксации τ выражается через вязкость ν и температуру θ_0

$$\tau = \frac{\nu}{\theta_0}. \quad (29)$$

4. ГИДРОДИНАМИЧЕСКИЙ ПРЕДЕЛ ДЛЯ СЛУЧАЯ МЕДЛЕННЫХ ТЕЧЕНИЙ И ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ

Уравнения гидродинамики, которые получаются из системы (27)–(29), в случае, когда столкновения и дискретные скорости задаются девятискоростной моделью (4)–(12), могут быть получены с помощью разложения Чепмена—Энскога. Детали разложения представлены в Приложении ниже. Схема (27)–(29) аппроксимирует уравнения Навье—Стокса (для несжимаемых течений) со вторым порядком точности по пространственным и временным переменным, причем вязкость выражается формулой

$$\nu = \frac{\theta_0 \Delta t}{\left(1 - e^{-\frac{\Delta t}{\tau}}\right)^2 + \frac{4}{9} \tau \rho_0 e^{-\frac{\Delta t}{\tau}} \left(1 - e^{-\frac{\Delta t}{\tau}}\right) \alpha} - \frac{\theta_0 \Delta t}{2}, \quad \tau = \nu / \theta_0, \quad (30)$$

также дополнительно требуется выполнение условия

$$4\alpha = \gamma + 4\beta + 4\lambda.$$

Для решения тестовых задач используется схема (27)–(29), при этом необходимо определить параметры α , β , γ , λ . При решении модельных задач обычно известна вязкость или число Рейнольдса (из которого можно выразить вязкость). Тогда, используя равенство (30), находим α :

$$\alpha = \frac{1}{\frac{4}{9}\tau\rho_0 e^{-\frac{\Delta t}{\tau}}(1 - e^{-\frac{\Delta t}{\tau}})} \left(\frac{\theta_0 \Delta t}{v + \frac{\theta_0 \Delta t}{2}} - (1 - e^{-\frac{\Delta t}{\tau}})^2 \right), \quad \tau = \nu/\theta_0,$$

здесь следует отметить, что для малых вязкостей экспоненциальные члены, стоящие в знаменателе, приводят к очень большим величинам α , поэтому на практике удобно использовать аппроксимацию $e^{-\frac{\Delta t}{\tau}} \approx 1/\left(1 + \frac{\Delta t}{\tau}\right)$, в этом случае

$$\alpha = \frac{\left(1 + \frac{\Delta t}{\tau}\right)^2}{\frac{4}{9}\rho_0 \Delta t} \left(\frac{\theta_0 \Delta t}{v + \frac{\theta_0 \Delta t}{2}} - \left(1 - \frac{1}{1 + \frac{\Delta t}{\tau}}\right)^2 \right), \quad \tau = \nu/\theta_0, \quad (31)$$

параметры α , β , γ , λ определяются из (17), в текущей работе положим $\lambda = 0$, т.е. столкновения между частицами, направленными по диагоналям, и частицами, движущимися параллельно осям, исключаются, это несколько упрощает систему. Кроме того, положим $\beta = 0.1\alpha$ и тогда $\gamma = 4\alpha - 4\beta = (18/5)\alpha$.

Схема решается в два шага. На первом шаге вычисляется столкновительная часть (правая часть) в (27)–(29), затем проводится шаг адвекции – ФР, полученные на первом шаге, разносятся по узлам $\mathbf{x} + \mathbf{c}_i \Delta t$ в соответствии со скоростями $\mathbf{c}_i$. Данный подход совершенно аналогичен применяемому в теории РУБ [1].

Также из (27)–(29) видно, что для решения необходимо знать локально равновесную функцию распределения (2), коэффициенты которой выражаются из законов сохранения. Уравнения для коэффициентов существенно нелинейные, но в текущей работе рассматриваются медленные изотермические сплошнородные течения, поэтому локально равновесное распределение можно разложить в ряд Тейлора, считая, что отклонения коэффициентов a , b , d от их значений в случае абсолютно максвелловского распределения малы [23]. В случае медленных изотермических течений локальная максвелловская ФР имеет тот же вид, как и для стандартной РУБ D2Q9 (см. [1], [23])

$$f_i^{eq} = \rho w_i \left(1 + \frac{\mathbf{c}_i \mathbf{u}}{\theta_0} + \frac{\mathbf{u} \mathbf{u}}{2\theta_0^2} : (\mathbf{c}_i \mathbf{c}_i - \mathbf{I}\theta_0) \right),$$

очевидно, что локальная максвелловская ФР в данной форме может быть легко вычислена на каждом временном шаге в любой пространственной точке, так как она выражается через потоковую скорость и плотность.

Отметим, что для всех тестовых задач брались следующие значения температуры и скорости частиц: $\theta_0 = 1/3$, $c = 1$, кроме того, временной шаг $\Delta t = 1$.

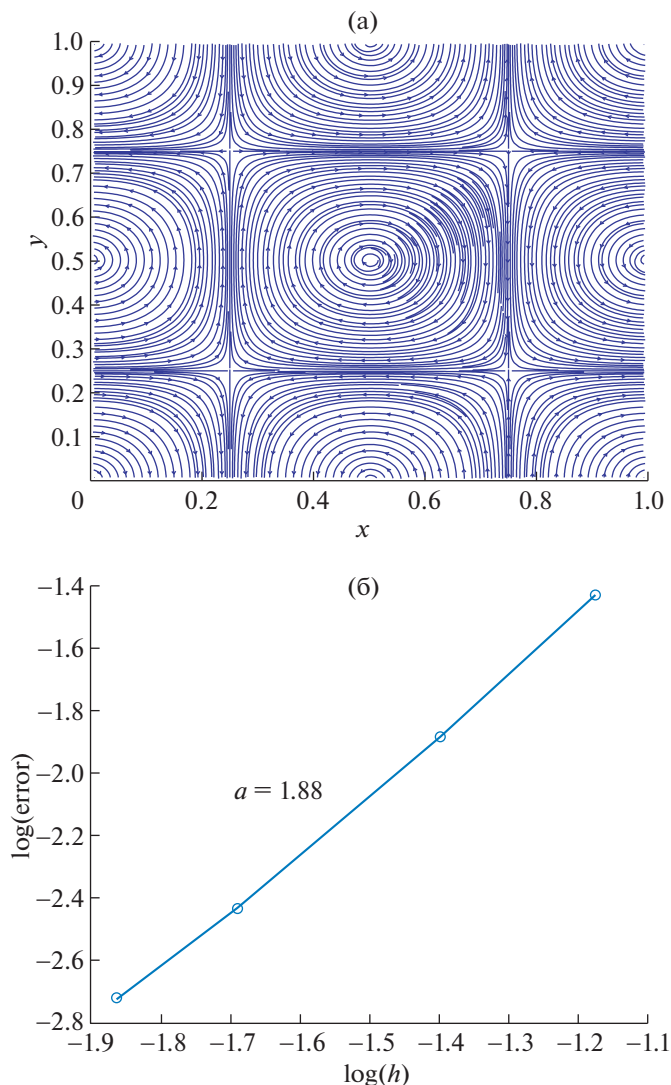
4.1. Вихри Тэйлора–Грина

Рассмотрим область в виде квадрата размером $L \times L$, заполненную несжимаемой жидкостью или газом, в которой начальные скорости заданы следующим образом:

$$u_x(x, y, t = 0) = -U_0 \cos(kx) \sin(ky), \quad u_y(x, y, t = 0) = U_0 \sin(kx) \cos(ky),$$

где $U_0 = 0.01$ есть максимальная амплитуда скорости, вязкость $\nu = 0.01$, временной шаг $\Delta t = 1$, также $k = \frac{2\pi}{L}$. На границах ставятся периодические граничные условия. Аналитическое решение задачи следующее:

$$u_x(x, y, t) = -U_0 \cos(kx) \sin(ky) e^{-2\nu k^2 t}, \quad u_y(x, y, t) = U_0 \sin(kx) \cos(ky) e^{-2\nu k^2 t},$$



Фиг. 2. Вихри Тэйлора–Грина. Линии тока для разрешения 25×25 в момент времени $1/(2k^2\nu)$, см. (а). Пространственные переменные x, y нормализованы на длину L . Логарифмы ошибок схемы относительно логарифма шага пространственной решетки $\lg(h) = \lg(L/K)$, $K = 15, 25, 49, 73$, см. (б), наклон линии линейной регрессии a , где $\lg(\text{error}) = a \lg(h) + b$, построенной на данных логарифма ошибок, равен $a = 1.88$.

из данного решения видно, что вихревая структура сохраняется во все моменты времени, но уменьшается амплитуда скоростей. Для численного решения будем измерять ошибку еггог относительно аналитического решения по формуле

$$\text{error} = \frac{\sqrt{\sum_{x,y} \left(u_x^{(n)}(T, x, y) - u_x(T, x, y) \right)^2 + \left(u_y^{(n)}(T, x, y) - u_y(T, x, y) \right)^2}}{\sqrt{\sum_{x,y} u_x(T, x, y)^2 + u_y(T, x, y)^2}}, \quad (32)$$

где суммирование ведется по всем пространственным узлам схемы x, y , также $u_x^{(n)}, u_y^{(n)}$ есть численные решения, $T = 1/(2k^2\nu)$. Число пространственных узлов N в каждом направлении x или y выбиралось равным $K = 15, 25, 49, 73$. Линии тока для разрешения 25×25 в момент времени

$1/(2k^2\nu)$ представлены на фиг. 2а. Пространственные переменные x, y нормализованы на длину L . Логарифмы ошибок схемы (32) относительно логарифма шага пространственной решетки $\lg(h) = \lg(L/K)$, $K = 15, 25, 49, 73$ представлены на фиг. 2б, наклон линии линейной регрессии a , где $\lg(\text{error}) = a \lg(h) + b$, построенной на логарифмах ошибок, равен $a = 1.88$, т.е. близок 2, что подтверждает второй порядок сходимости схемы.

4.2. Течение в полости с подвижной границей

В данной задаче снова рассматривается квадратная область, но границы области представляют собой твердые непроницаемые стенки. При этом верхняя стенка является подвижной, ее скорость, направленная параллельно стенке, равна $U_0 = 0.03$. Число Рейнольдса определяется следующим образом: $\text{Re} = U_0 H/\nu$, где $L = K \times c$ — длина или ширина области, K есть число узлов в направлении оси x или y (всего в рассматриваемом квадрате $K \times K$ узлов), ν — вязкость. На границах ставятся стандартные условия отскока, применяемые в теории РУБ [1], для данных условий считается, что твердая граница находится на расстоянии половины длины шага решетки от носителя узлов, находящихся рядом с границей. Для таких приграничных узлов имеем

$$f_i = f_i^+ - \frac{2w_i\rho_w}{\theta_0} \mathbf{c}_i U_w,$$

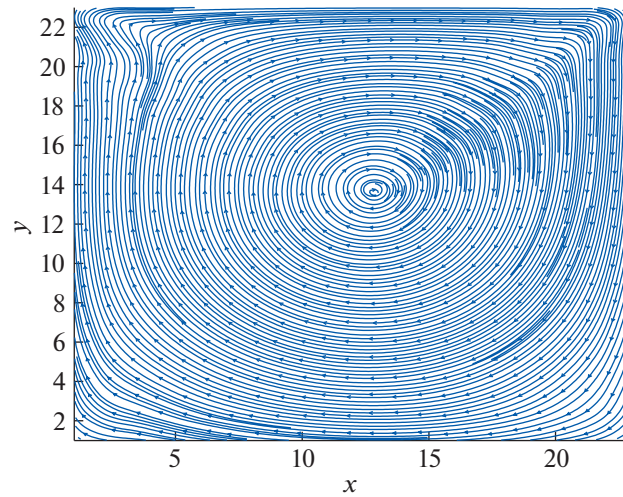
где $\bar{i}$ — индексы, соответствующие скоростям $\mathbf{c}_{\bar{i}}$, которые направлены от стенки, ФР для этих скоростей не могут быть найдены из дискретной кинетической модели. Также индексы i соответствуют скоростям $\mathbf{c}_i = -\mathbf{c}_{\bar{i}}$, а ФР f_i^+ в приграничных узлах, соответствующие скоростям $\mathbf{c}_i$, получаются после применения шага столкновений в этих узлах (но без адвекции); кроме того, U_w — скорость вещества на стенке, она равна 0 или U_0 ; ρ_w — плотность вещества на стенке, в данной задаче считаем $\rho_w = 1$.

Для данной задачи неизвестны аналитические решения, вопрос скорости сходимости здесь не исследуется, но имеются результаты численного моделирования [30], которые часто используют для оценки точности и устойчивости РУБ для разных чисел Рейнольдса и разрешения сеток [28], [31], [32]. Начальное распределение задается в форме абсолютного максвелловского распределения, в ходе решения схемы (27)–(29) профили скорости могут сходиться, не сходясь к какому-то пределу, или вообще разрушаться. Будем считать, что численное решение сходится к некоторому пределу, если существует такой момент времени t , что имеем

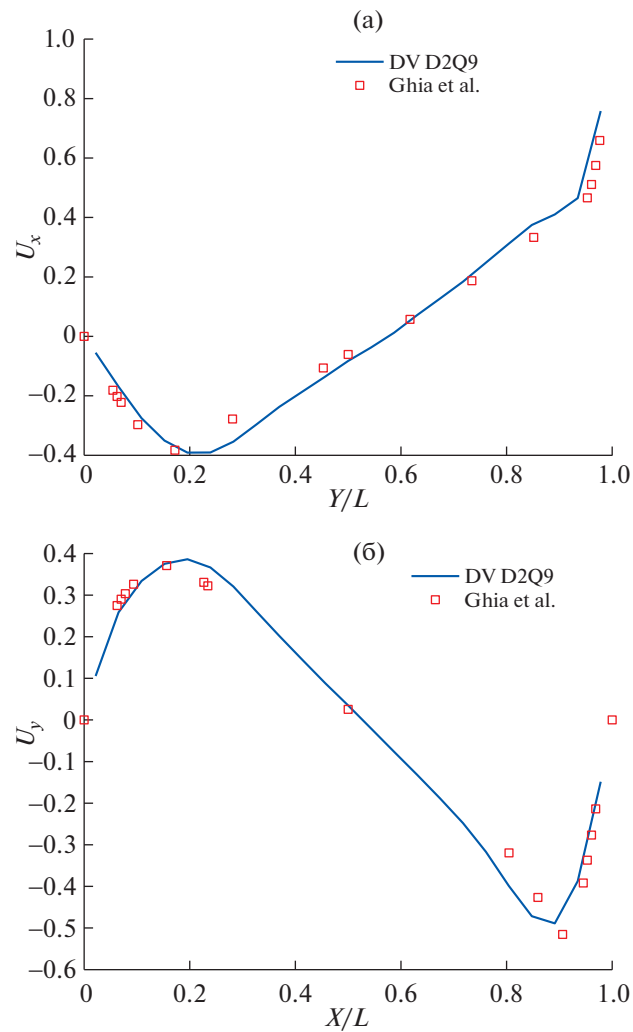
$$\frac{\sqrt{\sum_{x,y} \left(u_x^{(n)}(t+1000, x, y) - u_x^{(n)}(t, x, y) \right)^2 + \left(u_y^{(n)}(t+1000, x, y) - u_y^{(n)}(t, x, y) \right)^2}}{\sqrt{\sum_{x,y} u_x^{(n)}(t, x, y)^2 + u_y^{(n)}(t, x, y)^2}} < 10^{-6},$$

где суммирование идет по всем узлам x, y , также $u_x^{(n)}, u_y^{(n)}$ суть численные решения. Сходимость зависит от числа Рейнольдса, разрешения сетки. В настоящей работе мы рассмотрим случай числа Рейнольдса (Re), равного 1000. Особенностью настоящей схемы является то, что для данной задачи она позволяет получать устойчивые решения при грубом разрешении сетки. Для $\text{Re} = 1000$ рассматриваемая схема позволяет получить сходящиеся устойчивые решения при очень маленьком разрешении, например $K \times K$, равном 23×23 . Результаты моделирования для данного разрешения представлены на фиг. 3 и фиг. 4. На фиг. 3 представлены линии тока, решение устойчиво, качественно получается верная картина, следует отметить, что дополнительные вихри в углах не воспроизводятся, это связано с тем, что разрешение сетки очень грубое. На фиг. 4 представлено сравнение скоростей $u_x(L/2, y)$ и $u_y(x, L/2)$, обозначенные как U_x, U_y , с численным эталонным решением [30].

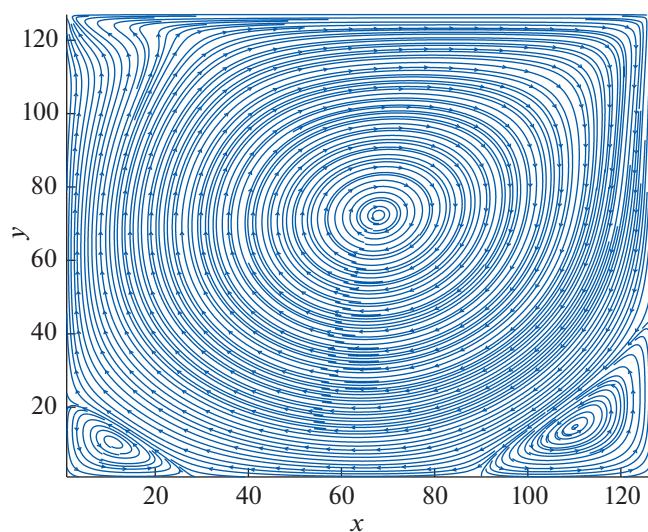
Отдельно возникает вопрос минимального разрешения сетки, при котором сходятся решения стандартной решеточной девятискоростной модели Больцмана $D2Q9$. Для повышения устойчивости модели $D2Q9$ дополнительно проведем стандартную процедуру регуляризации [33]. Для $\text{Re} = 1000$ регуляризованная модель $D2Q9$ при разрешении $K \times K$, где $K < 100$ неустойчива. Решения становятся устойчивыми примерно при $K > 140$. Этот результат согласуется с предыду-



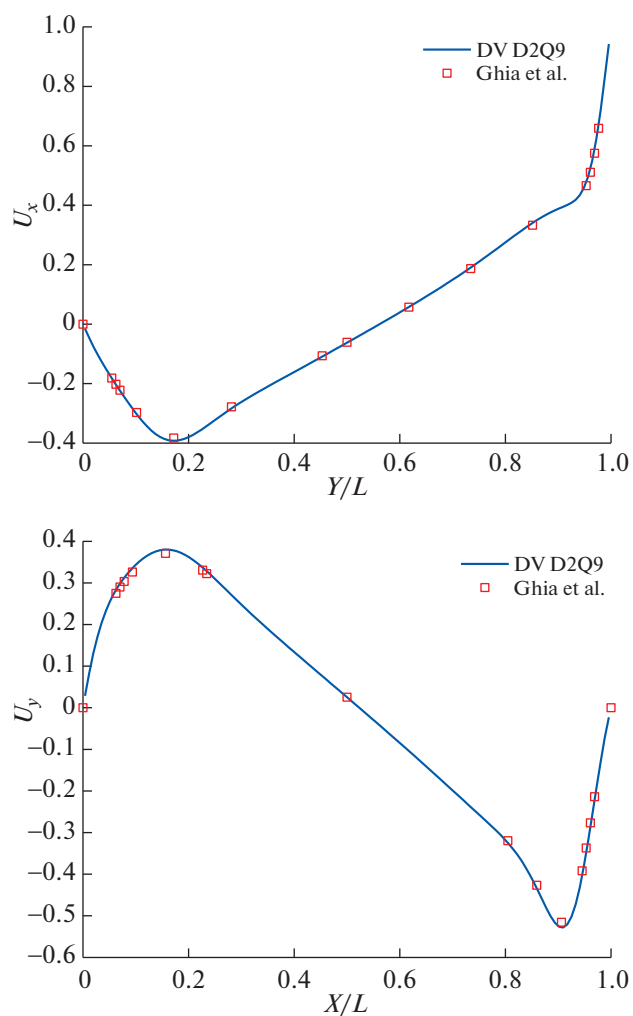
Фиг. 3. Течение в полости с подвижной верхней границей. Показаны линии тока для разрешения 23×23 и числа Рейнольдса $Re = 1000$.



Фиг. 4. Течение в полости с подвижной верхней границей. Показаны скорости $u_x(L/2, y)$ и $u_y(x, L/2)$ для разрешения 23×23 , $Re = 1000$. Квадратами отмечено численное решение Ghia, Ghia, Shin [30].



Фиг. 5. Течение в полости с подвижной верхней границей. Показаны линии тока для разрешения 127×127 и числа Рейнольдса $Re = 1000$.



Фиг. 6. Течение в полости с подвижной верхней границей. Показаны скорости $u_x(L/2, y)$ и $u_y(x, L/2)$ для разрешения 127×127 , $Re = 1000$. Квадратами отмечено численное решение Ghia, Ghia, Shin [30].

щими исследованиями [28]. Можно заключить, что для данной задачи схема (27)–(29) для дискретной кинетической модели (4)–(12) позволяет получать сходящиеся решения для значительно меньших разрешений решетки, чем в случае стандартных РУБ.

При увеличении разрешения численные решения для схемы (27)–(29) приближаются к эталонному решению [30]. Результаты моделирования при $K = 127$ показаны на фиг. 5 и фиг. 6. Очевидно, дополнительные вихри в нижних углах воспроизводятся схемой (фиг. 5), как и должно быть, профили скоростей $u_x(L/2, y)$ и $u_y(x, L/2)$, обозначенные как U_x, U_y на фиг. 6, практически совпадают с эталонным решением [30] (фиг. 6).

5. ЗАКЛЮЧЕНИЕ

В настоящей работе построена аппроксимация для дискретной кинетической модели уравнения Больцмана для девяти скоростей на основе суммы Вайлда, которая в сплошносредном пределе описывает вязкую несжимаемую жидкость. Данная аппроксимация представляет собой разностную схему второго порядка точности по пространственным координатам и времени. Рассмотрены две тестовые задачи: вихри Тэйлора–Грина и течение в полости с подвижной верхней границей. Показано, что схема на основе суммы Вайлда может описывать течение в полости с подвижной верхней границей при очень грубом разрешении сетки и не малых числах Рейнольдса. При этом регуляризованное решеточное уравнение Больцмана $D2Q9$ для тех же условий неустойчиво.

Следует отметить, что для дискретной кинетической модели (4)–(12) выполняется H -теорема, но дальнейшая дискретизация этой системы по пространству и времени может нарушать это свойство. Выполнение H -теоремы, конечно, очень желательно, так как оно означает не только линейную, но и нелинейную устойчивость, что важно для приложений. Для четырехскоростной модели Брудэлла схема, описывающая столкновения в виде сумм Вайлда, обладает H -теоремой [27]. Таким образом, исследование схем на основе сумм Вайлда перспективно, так как здесь имеется большая вариативность — можно построить очень большое количество разностных схем с желаемыми свойствами. Для схемы в виде суммы Вайлда (27)–(29) для модели (4)–(12) доказать H -теорему не получится, хотя моделирование течения в полости с подвижной верхней границей показывает лучшую устойчивость, чем моделирование с использованием регуляризованной модели $D2Q9$.

Это не означает, что сумма Вайлда (27)–(29) устойчивее РУБ $D2Q9$ в любых задачах. Действительно, линейная устойчивость РУБ исследовалась многими авторами [34]–[43]. Важно, что устойчивость РУБ существенно зависит от исследуемой задачи. Например, для возмущений в виде бегущих волн неустойчивость нерегуляризованных РУБ может наблюдаться в узком интервале волновых векторов, причем возмущение должно быть направлено под определенными углами относительно скорости потока. Кроме того, ряд регуляризованных РУБ, которые в модельных задачах показывают хорошую устойчивость, неожиданно оказываются неустойчивы для длинноволновых возмущений [42]. Предварительный анализ показывает, что у схемы (27)–(29) для уравнений (4)–(12) также существуют возмущения, ведущие к неустойчивости, что, впрочем, не означает, что неустойчивости будут наблюдаться для основных модельных течений. Можно заключить, что анализ линейной устойчивости и поиск схемы с выполняющейся H -теоремой представляют интересную задачу для будущих исследований.

ПРИЛОЖЕНИЕ

РАЗЛОЖЕНИЕ ЧЕПМЕНА–ЭНСКОГА

Разложение Чепмена–Энскога для схемы (27)–(29) проведем аналогично тому, как это делается в теории РУБ (см. [1], [28]). Вначале разложим левую часть (27) в ряд Тэйлора

$$f_i(t + \Delta t, \mathbf{r} + \mathbf{c}_i \Delta t) = f_i(t, \mathbf{r}) + \Delta t D_i f_i(t, \mathbf{r}) + \frac{1}{2} \Delta t^2 D_i^2 f_i(t, \mathbf{r}) + o(\Delta t^2),$$

где

$$D_i = \frac{\partial}{\partial t} + \mathbf{c}_i \cdot \frac{\partial}{\partial \mathbf{r}}.$$

Аналогично стандартному подходу в теории РУБ (см. [1], [28]) функцию распределения f_i разложим в ряд по малому параметру ϵ (числу Кнудсена) в окрестности локального равновесия

$$f_i = f_i^{eq} + \epsilon f_{1,i} + \epsilon^2 f_{2,i} + \dots,$$

а также считаем, что имеется несколько временных масштабов (см. [1], [28]) и производные по пространственным переменным и времени записываются следующим образом (см. [1], [28]):

$$\frac{\partial}{\partial t} = \epsilon \frac{\partial}{\partial t_1} + \epsilon^2 \frac{\partial}{\partial t_2} + \dots, \quad \frac{\partial}{\partial \mathbf{r}} = \epsilon \frac{\partial}{\partial \mathbf{r}_1}.$$

Подставляем разложения производных и ФР по параметру ϵ в уравнение (27) и соберем члены одинаковых порядков по ϵ . Члены порядка $O(1)$ равны нулю тождественно. Далее имеем уравнения

$$O(\epsilon): \quad \Delta t D_{1,i} f_i^{eq} = - \left(1 - e^{-\frac{\Delta t}{\tau}} \right)^2 f_{1,i} + \tau e^{-\frac{\Delta t}{\tau}} \left(1 - e^{-\frac{\Delta t}{\tau}} \right) L_i(\mathbf{f}_1), \quad (33)$$

где

$$D_{1,i} = \left(\frac{\partial}{\partial t_1} + \mathbf{c}_i \cdot \frac{\partial}{\partial \mathbf{r}_1} \right),$$

а также L_i — линеаризованный относительно локального максвеллиана оператор столкновений для уравнения (4)–(12). Так как рассматриваются течения только с малыми числами Маха, то вместо локально-максвелловского распределения можно брать абсолютно-максвелловское распределение, которое задается весами $\rho_0 w_i$. В этом случае оператор $L_i(\mathbf{f}_1)$ имеет вид

$$L_i(\mathbf{f}_1) = \rho_0 \sum_{jkl=1}^N A_{kl}^{ij} (w_k f_{1,l} - w_i f_{1,j} + f_{1,k} w_l - f_{1,i} w_j).$$

Умножая уравнения (33) на $\mathbf{c}_i$, суммируя по i , получаем уравнения изменения импульса для невязкой среды (уравнение Эйлера)

$$\frac{\partial}{\partial t_1} \rho_0 \mathbf{u} + \frac{\partial}{\partial \mathbf{r}_1} \cdot (\rho_0 \mathbf{u} \mathbf{u} + \rho_0 \theta_0) = 0. \quad (34)$$

Для членов порядка $O(\epsilon^2)$ получаем:

$$O(\epsilon^2): \quad \Delta t \frac{\partial}{\partial t_2} f_i^{eq} + \Delta t D_{1,i} f_{1,i} + \frac{\Delta t^2}{2} D_{1,i}^2 f_i^{eq} = \dots, \quad (35)$$

где точками обозначены члены второго порядка по ϵ , входящие в правую часть, их явная форма не понадобится. Умножаем уравнения (35) на $\mathbf{c}_i$ и после суммирования по i имеем

$$\frac{\partial}{\partial t_2} \rho_0 \mathbf{u} + \frac{\partial}{\partial \mathbf{r}_1} \cdot \mathbf{P}_1 + \frac{\Delta t}{2} \frac{\partial}{\partial \mathbf{r}_1} \cdot \sum_i \mathbf{c}_i \mathbf{c}_i D_{1,i} f_i^{eq} = 0, \quad (36)$$

где $\mathbf{P}_1 = \sum_i \mathbf{c}_i \mathbf{c}_i f_{1,i}$. Складываем уравнения (34) и (36) и получаем

$$\left(\epsilon \frac{\partial}{\partial t_1} + \epsilon^2 \frac{\partial}{\partial t_2} \right) \rho_0 \mathbf{u} + \frac{\partial}{\partial \mathbf{r}_1} \cdot \left(\epsilon (\rho_0 \mathbf{u} \mathbf{u} + \rho_0 \theta_0) + \epsilon^2 \mathbf{P}_1 + \epsilon^2 \frac{\Delta t}{2} \sum_i \mathbf{c}_i \mathbf{c}_i D_{1,i} f_i^{eq} \right) = 0,$$

возвращаясь к исходным переменным t , $\mathbf{r}$, получаем с точностью до малых членов $O(\epsilon^3)$ уравнение

$$\frac{\partial}{\partial t} (\rho_0 \mathbf{u}) + \frac{\partial}{\partial \mathbf{r}} \cdot \left(\rho_0 \mathbf{u} \mathbf{u} + \rho_0 \theta_0 + \epsilon \mathbf{P}_1 + \epsilon \frac{\Delta t}{2} \sum_i \mathbf{c}_i \mathbf{c}_i D_{1,i} f_i^{eq} \right) = 0, \quad (37)$$

третий и четвертый члены во втором слагаемом в (37) отвечают за вязкие поправки к уравнению Эйлера, которые требуется найти. Для нахождения $\mathbf{P}_1$ нам понадобятся выражения для $f_{1,i}$. Пол-

ная производная от локального максвелловского распределения в (33) в случае малых чисел Маха с точностью до членов порядка $O(M^3)$ и несжимаемых течений равна (см. [23])

$$\rho_0 w_i \frac{\mathbf{c}_i \mathbf{c}_i}{\theta_0} : \frac{\partial}{\partial \mathbf{r}_1} \mathbf{u},$$

напомним, что “:” есть оператор свертки тензоров. Неравновесные ФР $f_{1,i}$ следует искать в виде (см. [23], [29])

$$f_{1,i} = a_i \frac{\mathbf{c}_i \mathbf{c}_i}{\theta_0} : \frac{\partial}{\partial \mathbf{r}_1} \mathbf{u},$$

где коэффициенты a_i одинаковы для индексов i , соответствующих скоростям $\mathbf{c}_i$, для которых кинетическая энергия частиц одинакова, очевидно, имеются три разных коэффициента a_i . Из уравнения (33) для коэффициента a_1 , который пусть отвечает скоростям $\pm 1, \pm 2$, получаем

$$a_1 = - \frac{\frac{1}{9} \rho_0 \Delta t}{\left(1 - e^{-\frac{\Delta t}{\tau}}\right)^2 + \frac{4}{9} \tau \rho_0 e^{-\frac{\Delta t}{\tau}} \left(1 - e^{-\frac{\Delta t}{\tau}}\right) \alpha},$$

аналогично для коэффициента a_3 , который отвечает скоростям $\pm 3, \pm 4$, имеем

$$a_3 = - \frac{\frac{1}{12} \rho_0 \Delta t}{\left(1 - e^{-\frac{\Delta t}{\tau}}\right)^2 + \frac{1}{9} \tau \rho_0 e^{-\frac{\Delta t}{\tau}} \left(1 - e^{-\frac{\Delta t}{\tau}}\right) (\gamma + 4\beta + 4\lambda)},$$

также коэффициент a_0 , который отвечает нулевой скорости, равен нулю. Для удобства далее будем считать, что $\rho_0 = 1$.

Как было показано выше, вязкие поправки тензора напряжений $\mathbf{P}_1$ зависят от неравновесных ФР $f_{1,i}$, т.е. $\mathbf{P}_1 = \sum_i f_{1,i} \mathbf{c}_i \mathbf{c}_i$. Рассмотрим подробнее вязкие слагаемые

$$\epsilon \frac{\partial}{\partial \mathbf{r}} \cdot \mathbf{P}_1 = \epsilon \sum_{\sigma} \frac{\partial}{\partial r_{\sigma}} P_{1,\eta\sigma} = \epsilon \sum_{\sigma} \frac{\partial}{\partial r_{\sigma}} \sum_i f_{1,i} c_{i\eta} c_{i\sigma} = \theta_0^{-1} \sum_{\sigma\kappa w} \sum_i a_i c_{i\eta} c_{i\sigma} c_{i\kappa} c_{iw} \frac{\partial^2}{\partial r_{\kappa} \partial r_{\sigma}} u_w,$$

где σ, η, κ, w равны x или y , также обозначим $r_x = x, r_y = y$. Вспоминая, что векторы $\mathbf{c}_i$ имеют вид $(e_{ix}, e_{iy})c = (e_{ix}, e_{iy})\sqrt{3\theta_0}$, где e_{ix}, e_{iy} принимают значения $0, \pm 1$, то, например, имеем

$$\begin{aligned} \epsilon \frac{\partial}{\partial x} P_{1,xx} + \epsilon \frac{\partial}{\partial y} P_{1,xy} &= 2(9\theta_0)a_1 \frac{\partial^2 u_x}{\partial x^2} + 4(9\theta_0)a_3 \left(\frac{\partial^2}{\partial x \partial y} u_y + \frac{\partial^2}{\partial y^2} u_x \right) = \\ &= 2(9\theta_0)a_1 \frac{\partial^2 u_x}{\partial x^2} + 4(9\theta_0)a_3 \left(\frac{\partial^2}{\partial x \partial y} u_y + \frac{\partial^2}{\partial y^2} u_x \right) - 4(9\theta_0)a_3 \frac{\partial}{\partial x} \left(\frac{\partial}{\partial x} u_x + \frac{\partial}{\partial y} u_y \right) = \\ &= 9\theta_0(2a_1 - 4a_3) \frac{\partial^2}{\partial x^2} u_x + 4(9\theta_0)a_3 \frac{\partial^2}{\partial y^2} u_x, \end{aligned}$$

из последнего выражения следует, что вязкие члены имеют такую же форму, как и в уравнениях Навье–Стокса, если

$$a_1 = 4a_3,$$

что эквивалентно соотношению (17), но вязкость схемы (27)–(29) уже не выражается формулой (18).

Члены вида $\epsilon \frac{\Delta t}{2} \frac{\partial}{\partial \mathbf{r}} \sum_i \mathbf{c}_i \mathbf{c}_i D_{i,1} f_i^{eq}$ рассматривались в предыдущей работе [23], они равны $\rho_0 \theta_0 \frac{\Delta t}{2} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \mathbf{u}$.

Таким образом, вязкость определяется формулой

$$\nu = \frac{\theta_0 \Delta t}{\left(1 - e^{-\frac{\Delta t}{\tau}} \right)^2 + \frac{4}{9} \tau \rho_0 e^{-\frac{\Delta t}{\tau}} \left(1 - e^{-\frac{\Delta t}{\tau}} \right) \alpha} - \frac{\theta_0 \Delta t}{2}, \quad \tau = \nu / \theta_0,$$

также дополнительно требуется выполнение условия

$$4\alpha = \gamma + 4\beta + 4\lambda.$$

СПИСОК ЛИТЕРАТУРЫ

1. Krüger T., Kusumaatmaja H., Kuzmin A., Shardt O., Silva G., Viggen E. The Lattice Boltzmann Method. Principles and Practice. Switzerland: Springer, 2017.
2. Kogan M. Rarefied Gas Dynamics. New York: Plenum Press, 1969.
3. Broadwell J. Shock structure in a simple discrete velocity gas // Phys. Fluids. 1964. V. 7. P. 1243–1247.
4. Godunov S., Sultangazin U. On discrete models of the kinetic Boltzmann equation // Russian Math. Surveys. 1971. V. 26. P. 1–56.
5. Gatignol R. The hydrodynamical description for a discrete velocity model of gas // Complex Systems. 1987. V. 1. P. 709–725.
6. Wagner A. An H-theorem for the lattice Boltzmann approach to hydrodynamics // Europhys. Lett. 1998. V. 44. P. 144–149.
7. Yong W.-A., Luo L.-S. Nonexistence of H theorems for the athermal lattice Boltzmann models with polynomial equilibria // Phys. Rev. E. 2003. 051105.
8. Yong W.-A., Luo L.-S. Nonexistence of H Theorem for some Lattice Boltzmann models // J. Stat. Phys. 2005. V. 121. P. 91–103.
9. Karlin I., Ferrante A., Öttinger H. Perfect entropy functions of the Lattice Boltzmann method // Europhys. Lett. 1999. V. 47. P. 182–188.
10. Ansumali S., Karlin I., Öttinger H. Minimal entropic kinetic models for hydrodynamics // Europhys. Lett. 2003. V. 63. P. 798–804.
11. Karlin I., Ansumali S., Frouzakis C., Chikatamarla S. Elements of the Lattice Boltzmann Method I: Linear Advection Equation // Commun. Comput. Phys. 2006. V. 1. P. 616–655.
12. Karlin I., Chikatamarla S., Ansumali S. Elements of the lattice Boltzmann method II: kinetics and hydrodynamics in one dimension // Commun. Comput. Phys. 2007. V. 2. P. 196–238.
13. Chen H., Goldhirsch I., Orszag S. Discrete rotational symmetry, moment isotropy, and higher order lattice Boltzmann models // J. Sci. Comput. 2008. V. 34. P. 87–112.
14. Bobylev A., Spiga G. On a class of exact two-dimensional stationary solutions for the Broadwell model of the Boltzmann equation // J. Phys. A: Math. Gen. 1994. V. 27. P. 7451–7459.
15. Bobylev A. Exact solutions of discrete kinetic models and stationary problems for the plane Broadwell model // Math. Methods Appl. Sci. 1996. V. 19. P. 825–845.
16. Bobylev A., Toscani G. Two dimensional half-space problems for the Broadwell discrete velocity model // Contin. Mech. Thermodyn. 1996. V. 8. P. 257–274.
17. Bobylev A., Caraffini G., Spiga G. Non-stationary two-dimensional potential flows by the Broadwell model equations // Eur. J. Mech. B, Fluids. 2000. V. 19. P. 303–315.
18. Ilyin O. The analytical solutions of 2D stationary Broadwell kinetic model // J. Stat. Phys. 2012. V. 146. P. 67–72.
19. Ильин О. Симметрии, функция тока и точные решения для двумерной стационарной кинетической модели Бродуэлла // Теор. Мат. Физ. 2014. Т. 179. С. 350–359.
20. Uchiyama K. On the Boltzmann-Grad limit for the Broadwell model of the Boltzmann equation // J. Stat. Phys. 1988. V. 52. P. 331–355.
21. Babovsky H. “Small” kinetic models for transitional flow simulations // AIP Conf. Proc. 2012. V. 1501. P. 272–278.
22. Babovsky H. Discrete kinetic models in the fluid dynamic limit // Comput. and Math. with Appl. 2014. V. 67. P. 256–271.

23. *Ilyin O.* Discrete Velocity Boltzmann Model for Quasi- Incompressible Hydrodynamics // *Math.* 2021. V. 9. P. 993.
24. *Wild E.* On Boltzmann's Equation in the Kinetic Theory of Gases // *Proc. Camb. Philos. Soc.* 1951. V. 47. P. 602–609.
25. *McKean H.* An exponential formula for solving Boltzmann's equation for a Maxwellian gas // *J. of Combinat. Theory.* 1967. V. 2. P. 358–382.
26. *Gabetta E., Pareschi L., Toscani G.* Wild's sums and numerical approximation of nonlinear kinetic equations // *Transp. Theory and Stat. Phys.* 1996. V. 25. P. 515–530.
27. *Gabetta E., Pareschi L., Toscani G.* Relaxation Schemes for Nonlinear Kinetic Equations // *SIAM J. Numer. Anal.* 1997. V. 34. P. 2168–2194.
28. *Latt J.* Hydrodynamic Limit of Lattice Boltzmann Equations. Dissertation. Geneva: University of Geneva, 2007.
29. *Chauvat P., Gatignol R.* Euler and Navier-Stokes description for a class of discrete models of gases with different moduli // *Transp. Theory Stat. Phys.* 1992. V. 21. P. 417–435.
30. *Ghia U., Ghia K., Shin C.* High-Re solutions for incompressible flow using the Navier–Stokes equations and a multigrid method // *J. Comp. Phys.* 1982. V. 48. P. 387–411.
31. *Montessori A., Falcucci G., Prestininzi P., La Rocca M., Succi S.* Regularized lattice Bhatnagar–Gross–Krook model for two- and three-dimensional cavity flow simulations // *Phys. Rev. E.* 2014. V. 89. P. 053317.
32. *Wu J.-S., Shao Y.-L.* Simulation of lid-driven cavity flows by parallel lattice Boltzmann method using multi-relaxation-time scheme // *Int. J. Numer. Meth. Fluids.* 2004. V. 46. P. 921–937.
33. *Latt J., Chopard B.* Lattice Boltzmann method with regularized pre-collision distribution functions // *Math. Comput. in Simul.* 2006. V. 72. P. 165–168.
34. *Sterling J., Chen S.* Stability Analysis of Lattice Boltzmann Methods // *J. Comp. Phys.* 1996. V. 123. P. 196–206.
35. *Lallemand P., Luo L.-S.* Theory of the lattice Boltzmann method: Dispersion, dissipation, isotropy, Galilean invariance, and stability // *Phys. Rev. E.* V. 61. P. 6546–6562.
36. *Siebert D., Hegele Jr. L., Philippi P.* Lattice Boltzmann equation linear stability analysis: Thermal and athermal models // *Phys. Rev. E.* 2008. V. 77. P. 026707.
37. *Servan-Camas B., Tsai F.* Non-negativity and stability analyses of lattice Boltzmann method for advection–diffusion equation // *J. Comp. Phys.* 2009. V. 228. P. 236–256.
38. *Marié S., Ricot D., Sagaut P.* Comparison between lattice Boltzmann method and Navier–Stokes high order schemes for computational aeroacoustics // *J. Comp. Phys.* 2009. V. 228. P. 1056–1070.
39. *Hosseini S.A., Darabiha N., Thévenin D., Eshghinejadfard A.* Stability limits of the single relaxation-time advection–diffusion lattice Boltzmann scheme // *Int. J. Mod. Phys. C.* 2017. V. 28. 1750141.
40. *Wissocq G., Sagaut P., Boussuge J.-F.* An extended spectral analysis of the lattice Boltzmann method: modal interactions and stability issues // *J. Comp. Phys.* 2019. V. 380. P. 311–333.
41. *Masset P.-A., Wissocq G.* Linear hydrodynamics and stability of the discrete velocity Boltzmann equations // *J. Fluid Mech. A.* 2020. V. 897. P. 29.
42. *Wissocq G., Coreixas C., Boussuge J.-F.* Linear stability and isotropy properties of athermal regularized lattice Boltzmann methods // *Phys. Rev. E.* 2020. V. 102. 053305.
43. *Wissocq G., Sagaut P.* Hydrodynamic limits and numerical errors of isothermal lattice Boltzmann schemes // *arXiv:2104.14217v1* 2021.