ISSN 0044-4669

Том 64, Номер 12

Декабрь 2024

ФИЦ ИУ РАН

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ

НАУКА

1727

СОДЕРЖАНИЕ

Том 64, номер 12, 2024

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

МАТЕМАТИЧЕСКАЯ ФИЗИКА

ИНФОРМАТИКА

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2229–2242

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 517.58

ОЦЕНКА ОСТАТОЧНОГО ЧЛЕНА ПРИ СУММИРОВАНИИ НЕКОТОРЫХ ГИПЕРГЕОМЕТРИЧЕСКИХ РЯДОВ ГОРНА

© 2024 г. С. И. Безродных[1,*], О. В. Дунин-Барковская[1,2,]**

[1] *119333 Москва, ул. Вавилова, 44, ФИЦ ИУ РАН, Россия*

[2] *119234 Москва, Университетский пр-т, 13, Государственный астрономический институт им. П.К. Штернберга, Россия*

**e-mail: sbezrodnykh@mail.ru*

***e-mail: olga.ptitsyna@gmail.com*

Поступила в редакцию 20.04.2024 г.

Переработанный вариант 31.05.2024 г.

Принята к публикации 05.06.2024 г.

Построены интегральные представления и асимптотические оценки остатков суммирования гипергеометрического ряда Аппеля F_1 и родственного ему ряда G_2, указанных в списке Горна гипергеометрических рядов двух переменных. Найденные формулы имеют приложение к разработке алгоритмов вычисления функции F_1 с помощью формул аналитического продолжения во все пространство $\mathbb{C}^2$. Результаты могут найти приложение в задачах математической физики, и вычислительной теории функции, в том числе, при построении конформного отображения сложных многоугольников на основе интеграла Кристоффеля–Шварца. Библ. 24.

Ключевые слова: гипергеометрические функции Аппеля и Горна, формулы аналитического продолжения, эффективное вычисление гипергеометрических функций.

DOI: 10.31857/S0044466924120016, **EDN**: KCZCNY

1. ВВЕДЕНИЕ

1.1. Функция Аппеля F_1 и родственные ряды Горна двух переменных

Гипергеометрическая функция Аппеля F_1 возникает при конструктивном решении ряда важных задач математической физики и теории функций (см., например, некоторые современные работы [1]– [9]). Эта функция в единичном бикруге

$$\mathbb{U}^2 := \left\{(z_1, z_2) \in \mathbb{C}^2 : |z_1| < 1,\ |z_2| < 1\right\} \tag{1.1}$$

определяется с помощью следующего двойного ряда (см. [10]– [12]):

$$F_1(a_1, a_2; b, c; z_1, z_2) := \sum_{k_1,k_2=0}^{\infty} \frac{(b)_{k_1+k_2}(a_1)_{k_1}(a_2)_{k_2}}{(c)_{k_1+k_2}k_1!k_2!} z_1^{k_1} z_2^{k_2}, \tag{1.2}$$

где z_1 и z_2 — комплексные переменные, а величины a_1, a_2, b и c — комплексные параметры, от которых зависит функция F_1. Символ Похгаммера $(a)_k$ выражается через гамма-функцию $\Gamma(s)$ по формуле

$$(a)_k := \frac{\Gamma(a+k)}{\Gamma(a)} \tag{1.3}$$

и представляет собой для целых неотрицательных k произведение вида

$$(a)_0 = 1, \qquad (a)_k = a\,(a+1)\cdots(a+k-1), \qquad k = 1, 2, \ldots . \tag{1.4}$$

Параметр c в формуле (1.2) не принимает целых неположительных значений, $c \notin \mathbb{Z}^-$. В формуле (1.1), как обычно, через $\mathbb{C}^2$ обозначено двумерное комплексное пространство.

Вне бикруга $\mathbb{U}^2$ под функцией Аппеля F_1 понимается аналитическое продолжение ряда (1.2). Прежде чем перейти к обсуждению такого продолжения, отметим, что ряд F_1, рассматриваемый как функция z_1 и z_2, удовлетворяет следующей системе уравнений с частными производными (см. [10]– [12]):

$$\begin{aligned} z_1(1-z_1)\,\partial_{11}u \,+\, z_2(1-z_1)\,\partial_{12}u \,+\, \big[c-(a_1+b+1)z_1\big]\,\partial_1 u - a_1\,\partial_2 u - a_1 b\,u = 0,\\ z_2(1-z_2)\,\partial_{22}u \,+\, z_1(1-z_2)\,\partial_{12}u \,+\, \big[c-(a_2+b+1)z_2\big]\,\partial_2 u - a_2\,\partial_1 u - a_2 b\,u = 0, \end{aligned} \tag{1.5}$$

где $u = u(z_1, z_2)$ — искомая функция, через $\partial_{jk}u(z,\zeta)$ обозначены частные производные, так что, например, выражения $\partial_{12}u$ и $\partial_{11}u$ означают соответственно $\frac{\partial^2 u}{\partial z_1 \partial z_2}$ и $\frac{\partial^2 u}{\partial z_1^2}$. Размерность пространства решений системы (1.5) равна трем (см. [12], [13]). Если $z \notin \mathbb{U}^2$, то F_1 можно представить с помощью *формул аналитического продолжения*, дающих выражение в виде линейной комбинации функций, образующих базис в пространстве решений (1.5). Например, если выполнены соотношения

$$b - a_1 \notin \mathbb{Z}, \qquad b - a_1 - a_2 \notin \mathbb{Z}, \tag{1.6}$$

то формула аналитического продолжения функции F_1, первоначально определенной с помощью ряда (1.2), в область

$$\mathbb{V}^2 := \{(z_1, z_2) \in \mathbb{C}^2 : |z_1| > |z_2| > 1, |\arg(-z_j)| < \pi, j = 1, 2\} \tag{1.7}$$

имеет следующий вид (см. [14], [15]):

$$\begin{aligned} F_1(a_1, a_2; b, c; z_1, z_2) = \frac{\Gamma(c)\,\Gamma(b-a_1-a_2)}{\Gamma(b)\,\Gamma(c-a_1-a_2)}\mathcal{U}_0^{(\infty)}(a_1, a_2; b, c; z_1, z_2) + \\ + \frac{\Gamma(c)\Gamma(a_1-b)}{\Gamma(a_1)\Gamma(c-b)}\mathcal{U}_1^{(\infty)}(a_1, a_2; b, c; z_1, z_2) + \frac{\Gamma(c)\Gamma(b-a_1)\Gamma(a_1+a_2-b)}{\Gamma(a_2)\Gamma(c-b)\Gamma(b)}\mathcal{U}_2^{(\infty)}(a_1, a_2; b, c; z_1, z_2), \end{aligned} \tag{1.8}$$

где функции $\mathcal{U}_j^{(\infty)}$, $j = 0, 1, 2$, определяются равенствами

$$\begin{aligned} &\mathcal{U}_0^{(\infty)}(a_1, a_2; b, c; z_1, z_2) := \\ &\qquad := (-z_1)^{-a_1}(-z_2)^{-a_2} F_1\Big(a_1, a_2; 1-c+a_1+a_2, 1-b+a_1+a_2; \frac{1}{z_1}, \frac{1}{z_2}\Big), \end{aligned} \tag{1.9}$$

$$\begin{aligned} &\mathcal{U}_1^{(\infty)}(a_1, a_2; b, c; z_1, z_2) := \\ &\qquad := (-z_1)^{-b} F_1\Big(1-c+b, a_2; b, 1+b-a_1; \frac{1}{z_1}, \frac{z_2}{z_1}\Big), \end{aligned} \tag{1.10}$$

$$\begin{aligned} &\mathcal{U}_2^{(\infty)}(a_1, a_2; b, c; z_1, z_2) := \\ &\qquad := (-z_1)^{-a_1}(-z_2)^{a_1-b} G\Big(a_1, 1-c+b; b-a_1, 1+b-a_1-a_2; \frac{z_2}{z_1}, \frac{1}{z_2}\Big). \end{aligned} \tag{1.11}$$

В (1.11) фигурирует гипергеометрический ряд G, определяемый равенством

$$G(a_1, a_2; b, c; z_1, z_2) = \sum_{k_1, k_2 = 0}^{\infty} \frac{(b)_{k_2-k_1}\,(a_1)_{k_1}\,(a_2)_{k_2}}{(c)_{k_2-k_1}\;k_1!\;k_2!}\, z_1^{k_1} z_2^{k_2}, \tag{1.12}$$

а областью его сходимости является бикруг (1.1); здесь, также как и выше, z_1, z_2 — переменные, a_1, a_2, c — параметры. Фигурирующий в (1.12) символ Похгаммера (1.3) при отрицательных значениях k имеет вид произведения

$$(a)_k = (-1)^k\Big((1-a)(2-a)\cdots\big((1-a)-k+1\big)\Big)^{-1}, \qquad k = -1, -2, \ldots . \tag{1.13}$$

В области (1.7) функции (1.9)–(1.11) образуют базис в пространстве решений системы (1.5). Необходимо подчеркнуть, что формулы продолжения в области

$$\begin{aligned} \mathbb{W}^2 &:= \{(z_1, z_2) \in \mathbb{C}^2 : |1-z_1| < |1-z_2| < 1, |\arg(1-z_j)| < \pi, j = 1, 2\},\\ \widetilde{\mathbb{V}}^2 &:= \{(z_1, z_2) \in \mathbb{C}^2 : (z_2, z_1) \in \mathbb{V}^2\}, \qquad \widetilde{\mathbb{W}}^2 := \{(z_1, z_2) \in \mathbb{C}^2 : (z_2, z_1) \in \mathbb{W}^2\} \end{aligned}$$

также, как и (1.8)–(1.11), содержат ряды (1.2) и (1.12) (см. [14], [15]).

Для построения аналитического продолжения функции Аппеля F_1 могут быть также применены интегральные представления типа Эйлера. Например, если $\operatorname{Re} c > \operatorname{Re} b > 0$, то в области $\{(z_1, z_2) \in \mathbb{C}^2 : |\arg(1-z_j)| < \pi,$ $j = 1, 2\}$ функция Аппеля F_1 представима следующим интегралом (см. [11], [12]):

$$F_1(a_1, a_2; b, c; z_1, z_2) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-z_1 t)^{a_1}(1-z_2 t)^{a_2}}\, dt, \tag{1.14}$$

где $\Gamma(s)$ — гамма-функция. Вместе с тем, подобные интегралы в случае произвольных параметров (a_1, a_2, b, c), вообще говоря, не могут быть вычислены в элементарных функциях. Для адекватного представления функции Аппеля в $\mathbb{C}^2$ естественным аппаратом являются формулы аналитического продолжения (см. [15]–[17]), обеспечивающие ее вычисление с помощью набора экспоненциально сходящихся подходящих подобластях $\mathbb{C}^2$ рядов.

Таким образом, для разработки алгоритма вычисления функции F_1 на основе формул (1.2) и (1.8), а также некоторых других формул аналитического продолжения, необходимо построить эффективный способ оценки и вычисления остатка суммирования рядов (1.2) и (1.12). В разд. 2 найдено интегральное представление и асимптотическая оценка для остатка суммирования ряда (1.2), а в разд. 3 получены аналогичные результаты для ряда (1.12). Основные результаты работы представлены в виде теорем 2 и 3. В следующем п. 1.2 приведены вспомогательные соотношения и теорема, примененные в работе для получения таких интегральных представлений и асимптотик. Отметим, что разработка алгоритмов вычисления гипергеометрических функций двух переменных на основе формул аналитического продолжения привлекает внимание специалистов (см. [18]–[21]). Появление современных работ подчеркивает актуальность построения эффективных алгоритмов для оценки остатка суммирования кратных гипергеометрических рядов.

Нетрудно увидеть, что ряд G, определенный по формуле (1.12), совпадает с точностью до обозначения переменных и параметров с гипергеометрическим рядом G_2, указанным в списке Горна (см. [11, разд. 5.7.1], [12, разд. 1.6]), точнее, справедливо равенство

$$G(a_1, a_2; b, c; z_1, z_2) = G_2(a_1, a_2; b, 1-c; -z_1, -z_2). \tag{1.15}$$

В дальнейшем нам будет удобно использовать обозначение (1.12), применявшееся также в [15]. Напомним, что ряд Аппеля также присутствует в списке гипергеометрических рядов Горна (см. [11, разд. 5.7.1], [12, разд. 1.6]). Отметим еще, что в [15] были сняты указанные выше ограничения (1.6) на параметры функции F_1 и построены аналоги формул (1.8)–(1.11) при $b = a_1 + m$ или $b = a_1 + a_2 + m$, где $m \in \mathbb{Z}$.

В завершение отметим, что функция Аппеля F_1 является частным случаем функции Лауричеллы $F_D^{(N)}$ при $N = 2$ и обратим внимание на то, что проблема построения полного набора формул аналитического продолжения функции $F_D^{(N)}$, области сходимости которых покрывают $\mathbb{C}^N$, при произвольном N решена в работах [22], [23]. В работе [23] также представлено применение алгоритма вычисления $F_D^{(N)}$ на основе формул аналитического продолжения для разработки метода решения проблемы параметров интеграла Кристоффеля–Шварца в ситуации “кроудинга”.

1.2. Некоторые используемые формулы и асимптотические оценки

Далее будут применяться следующие известные тождества, связывающие комбинацию гамма-функций и интегральные представления для бета-функции $\mathrm{B}(x, y)$ (см. [11]):

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_0^1 t^{x-1}(1-t)^{y-1}dt =: \mathrm{B}(x, y), \qquad \operatorname{Re} x > 0, \quad \operatorname{Re} y > 0, \tag{1.16}$$

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = -\frac{1}{2}\frac{1}{\operatorname{sh}(\pi i x)}\int_1^{(0+)} (-t)^{x-1}(1-t)^{y-1}dt, \qquad x \notin \mathbb{Z}, \quad \operatorname{Re} y > 0. \tag{1.17}$$

Контур интегрирования в формуле (1.17) начинается в точке $t = 1$, продолжается в верхней полуплоскости $\operatorname{Im} t > 0$, обходит точку $t = 0$ против часовой стрелки и возвращается в $t = 1$ по точкам нижней полуплоскости $\operatorname{Im} t < 0$. Приведем также формулу Эйлера для гипергеометрической функции Гаусса:

$$F(a, b; c; z) := \sum_{k=0}^{\infty} \frac{(b)_k (a)_k}{(c)_k k!} z^k = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-zt)^a}\, dt, \qquad \operatorname{Re} c > \operatorname{Re} b > 0; \tag{1.18}$$

здесь $|\arg(1-z)| < \pi$, а второе равенство в (1.18) имеет место при $|z| < 1$. Кроме того, нам потребуется следующее тождество:

$$F(a, b; c; z) = (1-z)^{-a} F\Big(a, c-b; c; \frac{z}{z-1}\Big). \tag{1.19}$$

При выводе асимптотик для остатков суммирования двойных гипергеометрических рядов (1.2) и (1.12) будет использована теорема 1 из [24, с. 58], приведенная ниже для удобства изложения материалов разд. 2 и 3.

Теорема 1 (см. [24, с. 58, теорема 1]). *Пусть функции $h(x)$ и $\phi(x)$ имеют следующие асимптотические разложения при $x \to a$:*

$$h(x) \sim h(a) + \sum_{s=0}^{\infty} a_s (x-a)^{s+\mu}, \qquad \phi(x) \sim \sum_{s=0}^{\infty} b_s (x-a)^{s+\alpha-1}, \tag{1.20}$$

где $\mu > 0$ *и* $\operatorname{Re}\alpha > 0$*, причем*

$$\forall x \in (a,b): \quad h(x) > h(a), \qquad \forall \delta > 0: \quad \inf_{x\in[a+\delta,b)} (h(x) - h(a)) > 0. \tag{1.21}$$

Предположим, что следующий интеграл

$$I(\lambda) = \int_a^b \phi(x) e^{-\lambda h(x)} dx \tag{1.22}$$

сходится абсолютно для всех достаточно больших λ*. Тогда имеет место следующее асимптотическое разложение:*

$$I(\lambda) \sim e^{-\lambda h(a)} \sum_{s=0}^{\infty} \Gamma\left(\frac{s+\alpha}{\mu}\right) \frac{c_s}{\lambda^{(s+\alpha)/\mu}}, \qquad \lambda \to \infty, \tag{1.23}$$

где величины c_s *выражаются через коэффициенты рядов* (1.20) *и , в частности, справедливы равенства*

$$c_0 = \frac{b_0}{\mu a_0^{\alpha/\mu}}, \qquad c_1 = \Big(\frac{b_1}{\mu} - \frac{(\alpha+1)a_1 b_0}{\mu^2 a_0}\Big) \frac{1}{a_0^{(\alpha+1)/\mu}}. \tag{1.24}$$

В следующих разделах будут вначале построены интегральные представления для остаточных членов при суммировании гипергеометрических рядов (1.2) и (1.12). Затем при получении асимптотик с помощью теоремы 1 в качестве большого параметра λ будет выбрано максимальное значение индексов суммирования, которое выбирается для вычисления частичных сумм в (1.2) и (1.12).

2. ОЦЕНКА ОСТАТОЧНОГО ЧЛЕНА ПРИ СУММИРОВАНИИ РЯДА АППЕЛЯ F_1

2.1. Суммирование ряда F_1 *и теорема об оценке остатка*

Представим двойной ряд (1.2) в виде

$$F_1\,(a_1, a_2; b,\, c;\, z_1, z_2\,) = \mathcal{F}_1^{(m)}(a_1, a_2; b,\, c;\, z_1, z_2\,) + \mathcal{R}^{(m)}(a_1, a_2; b,\, c;\, z_1, z_2\,), \tag{2.1}$$

где первое слагаемое $\mathcal{F}_1^{(m)}$ является суммой конечного числа членов ряда:

$$\mathcal{F}_1^{(m)}(a_1, a_2; b,\, c;\, z_1, z_2\,) := \sum_{k_1,k_2=0}^{m-1} \frac{(b)_{k_1+k_2}(a_1)_{k_1}(a_2)_{k_2}}{(c)_{k_1+k_2} k_1! k_2!} z_1^{k_1} z_2^{k_2}, \tag{2.2}$$

а второе слагаемое $\mathcal{R}^{(m)}$ — остаток суммирования. Далее для этих величин будем использовать обозначения $\mathcal{F}_1^{(m)}(z_1, z_2)$ и $\mathcal{R}^{(m)}(z_1, z_2)$. Отметим, что для вычисления $\mathcal{F}_1^{(m)}$ удобно воспользоваться следующим рекуррентным соотношением:

$$\begin{aligned} \mathcal{F}_1^{(m)}(z_1, z_2) = \mathcal{F}_1^{(m-1)}(z_1, z_2) &+ \frac{(a_2)_{m-1}}{(m-1)!} z_2^{m-1} \sum_{k_1=0}^{m-2} \frac{(b)_{k_1+m-1}(a_1)_{k_1}}{(c)_{k_1+m-1} k_1!} z_1^{k_1} + \\ &+ \frac{(a_1)_{m-1}}{(m-1)!} z_1^{m-1} \sum_{k_2=0}^{m-2} \frac{(b)_{m-1+k_2}(a_2)_{k_2}}{(c)_{m-1+k_2} k_2!} z_2^{k_2} + \frac{(b)_{2(m-1)}(a_1)_{m-1}(a_2)_{m-1}}{(c)_{2(m-1)}(m-1)!(m-1)!} z_1^{m-1} z_2^{m-1}. \end{aligned} \tag{2.3}$$

Если расположить члены ряда (1.2) в виде бесконечной двумерной таблицы, где строки и столбцы нумеруются индексами k_1 и k_2 соответственно, то в правой части формулы (2.3) к функции $\mathcal{F}_1^{(m-1)}$ добавлены: 1) сумма элементов с индексами $k_1 = m-1, k_2 = 0, \ldots, m-2$, т.е. первых $m-1$ членов m-й строки, 2) сумма элементов с номерами $k_2 = m-1, k_1 = 0, \ldots, m-2$, т.е. первых $m-1$ элементов m-го столбца, а также 3) элемент с индексами $k_1 = k_2 = m-1$.

При вычислении функции Аппеля в точке $(z_1, z_2) \in \mathbb{U}^2$ мы полагаем

$$F_1\,(a_1, a_2; b,\, c;\, z_1, z_2\,) \simeq \mathcal{F}_1^{(m)}(a_1, a_2; b,\, c;\, z_1, z_2), \tag{2.4}$$

при этом число число m требуется выбрать таким, чтобы при некотором заданном числе $\delta > 0$ для остатка $\mathcal{R}^{(m)}$ выполнялось неравенство

$$|\mathcal{R}^{(m)}(z_1, z_2\,)| < \delta. \tag{2.5}$$

При реализации алгоритма вычисления функции F_1 по формулам (2.2) и (2.4) условие (2.5) удобно проверять, приближенно вычисляя $\mathcal{R}^{(m)}(z_1, z_2\,)$ на основе асимптотики или интегрального представления, которые дает приведенная ниже теорема.

Теорема 2. *Предположим, что параметры ряда Аппеля, определенного по формуле* (1.2)*, удовлетворяют условию* $\operatorname{Re} \mathrm{c} > \operatorname{Re} \mathrm{b} > 0$*, а переменные лежат в бикруге* (1.1)*, т.е.* $(z_1, z_2) \in \mathbb{U}^2$*. Тогда имеют место следующие утверждения.*

(i) *Справедлива формула суммирования*

$$F_1\left(a_1, a_2; b,\, c;\, z_1, z_2\right) = \sum_{k_1,k_2=0}^{m-1} \frac{(b)_{k_1+k_2}(a_1)_{k_1}(a_2)_{k_2}}{(c)_{k_1+k_2} k_1! k_2!} z_1^{k_1} z_2^{k_2} + \mathcal{R}^{(m)}(z_1, z_2), \tag{2.6}$$

где для остаточного члена $\mathcal{R}^{(m)}(z_1, z_2)$ *имеет место интегральное представление*

$$\begin{aligned} \mathcal{R}^{(m)}(a_1, a_2; b, c; z_1, z_2) = \\ = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \Bigg(\frac{z_1^m (a_1)_m}{m!} \int_0^1 \frac{t^{b+m-1}(1-t)^{c-b-1}}{(1-z_2 t)^{a_2}}\, \mathcal{I}(z_1 t, a_1, m) dt + \\ + \frac{z_2^m (a_2)_m}{m!} \int_0^1 \frac{t^{b+m-1}(1-t)^{c-b-1}}{(1-z_1 t)^{a_1}}\, \mathcal{I}(z_2 t, a_2, m) dt - \\ - \frac{z_1^m z_2^m (a_1)_m (a_2)_m}{(m!)^2} \int_0^1 t^{b+2m-1}(1-t)^{c-b-1} \mathcal{I}(z_1 t, a_1, m)\, \mathcal{I}(z_2 t, a_2, m)\, dt \Bigg); \end{aligned} \tag{2.7}$$

здесь функция $\mathcal{I}$ *определяется равенством*

$$\mathcal{I}(z, a, m) = m \int_0^1 \frac{(1-\xi)^{m-1}}{(1-\xi z)^{a+m}}\, d\xi, \quad |\arg(1-z)| < \pi. \tag{2.8}$$

(ii) *Для остаточного члена* $\mathcal{R}^{(m)}(z_1, z_2)$ *справедливы следующие асимптотики при* $m \to \infty$.

1) *Если* $|z_1| < |z_2|$*, то*

$$\mathcal{R}^{(m)}(z_1, z_2) = \frac{\Gamma(c)}{\Gamma(b)} \frac{(a_2)_m}{m^{c-b} m!} \frac{z_2^m}{(1-z_1)^{a_1}(1-z_2)} \Big(1 + C_1 m^{-1} + O\big(m^{-2}\big)\Big), \tag{2.9}$$

где коэффициент $C_1 = C_1(z_1, z_2)$ *определяется по формуле*

$$C_1 = (c-b)\left(\frac{(1-b) - (a_1+1-b)z_1 - (2-b)z_2 + (a_1+2-b)z_1 z_2}{(1-z_1)(1-z_2)} - \frac{c-b+1}{2} \right) + \frac{(a_2-1)z_2}{1-z_2}. \tag{2.10}$$

2) *Если* $|z_2| < |z_1|$*, то*

$$\mathcal{R}^{(m)}(z_1, z_2) = \frac{\Gamma(c)}{\Gamma(b)} \frac{(a_1)_m}{m^{c-b} m!} \frac{z_1^m}{(1-z_1)(1-z_2)^{a_2}} \Big(1 + C_2 m^{-1} + O\big(m^{-2}\big)\Big), \tag{2.11}$$

где коэффициент $C_2 = C_2(z_1, z_2)$ *определяется по формуле*

$$C_2 = (c-b)\left(\frac{(1-b) - (a_2+1-b)z_2 - (2-b)z_1 + (a_2+2-b)z_1 z_2}{(1-z_1)(1-z_2)} - \frac{c-b+1}{2} \right) + \frac{(a_1-1)z_1}{1-z_1}. \tag{2.12}$$

3) *Если* $|z_1| = |z_2|$*, то асимптотика остаточного члена* $\mathcal{R}^{(m)}(z_1, z_2)$ *при* $m \to \infty$ *дается суммой правых частей* (2.9) *и* (2.11).

2.2. Доказательство теоремы 2

2.2.1. Вывод интегрального представления (2.7) для остаточного члена $\mathcal{R}^{(m)}$. Прежде всего, представим остаток суммирования $\mathcal{R}^{(m)}(z_1, z_2)$ в формуле (2.1) в виде

$$\mathcal{R}^{(m)}(z_1, z_2) = \mathcal{R}^{(0,m)}(z_1, z_2) + \mathcal{R}^{(m,0)}(z_1, z_2) - \mathcal{R}^{(m,m)}(z_1, z_2), \tag{2.13}$$

где функции $\mathcal{R}^{(0,m)}$, $\mathcal{R}^{(m,0)}$ и $\mathcal{R}^{(m,m)}$ определяются равенствами

$$\mathcal{R}^{(0,m)}(z_1, z_2) = \sum_{k_1=0}^{\infty} \sum_{k_2=m}^{\infty} \frac{(b)_{k_1+k_2}(a_1)_{k_1}(a_2)_{k_2}}{(c)_{k_1+k_2} k_1! k_2!} z_1^{k_1} z_2^{k_2}, \tag{2.14}$$

$$\mathcal{R}^{(m,0)}(z_1,z_2)=\sum_{k_1=m}^{\infty}\sum_{k_2=0}^{\infty}\frac{(b)_{k_1+k_2}(a_1)_{k_1}(a_2)_{k_2}}{(c)_{k_1+k_2}k_1!k_2!}z_1^{k_1}z_2^{k_2}, \tag{2.15}$$

$$\mathcal{R}^{(m,m)}(z_1,z_2)=\sum_{k_1,k_2=m}^{\infty}\frac{(b)_{k_1+k_2}(a_1)_{k_1}(a_2)_{k_2}}{(c)_{k_1+k_2}k_1!k_2!}z_1^{k_1}z_2^{k_2}. \tag{2.16}$$

Для вывода интегрального представления (2.7) заменим индексы суммирования в (2.14)–(2.16) так, чтобы суммирование начиналось от нуля:

$$\mathcal{R}^{(0,m)}(z_1,z_2)=z_2^m\sum_{k_1,k_2=0}^{\infty}\frac{(b)_{k_1+k_2+m}(a_1)_{k_1}(a_2)_{k_2+m}}{(c)_{k_1+k_2+m}k_1!(k_2+m)!}z_1^{k_1}z_2^{k_2}, \tag{2.17}$$

$$\mathcal{R}^{(m,0)}(z_1,z_2)=z_1^m\sum_{k_1,k_2=0}^{\infty}\frac{(b)_{k_1+k_2+m}(a_1)_{k_1+m}(a_2)_{k_2}}{(c)_{k_1+k_2+m}(k_1+m)!k_2!}z_1^{k_1}z_2^{k_2}. \tag{2.18}$$

$$\mathcal{R}^{(m,m)}(z_1,z_2)=z_1^mz_2^m\sum_{k_1,k_2=0}^{\infty}\frac{(b)_{k_1+k_2+2m}(a_1)_{k_1+m}(a_2)_{k_2+m}}{(c)_{k_1+k_2+2m}(k_1+m)!(k_2+m)!}z_1^{k_1}z_2^{k_2}. \tag{2.19}$$

Для дальнейших вычислений воспользуемся следующими выражениями:
1) тождеством, вытекающим из (1.3) и (1.16):

$$\frac{(b)_{k_1+k_2+\beta m}}{(c)_{k_1+k_2+\beta m}}=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b+k_1+k_2+\beta m-1}(1-t)^{c-b-1}dt, \tag{2.20}$$

2) тождествами, получаемые с помощью определения (1.3):

$$(k_j+m)!=m!(1+m)_{k_j},\quad (a_j)_{k_j+m}=(a_j)_m(a_j+m)_{k_j},\qquad j=1,2, \tag{2.21}$$

3) а также биномиальным рядом

$$\sum_{k_j=0}^{\infty}\frac{(a_j)_{k_j}}{k_j!}(z_j)^{k_j}=(1-z_j)^{-a_j},\qquad |z_j|<1,\quad |\arg(1-z_j)|<\pi,\qquad j=1,2. \tag{2.22}$$

Выражение, указанное в формуле (2.20), при $\beta=1$ возникает в (2.17) и (2.18), а при $\beta=2$ — в формуле (2.19).

Получим интегральное представление, например, для функции $\mathcal{R}^{(m,0)}(z_1,z_2)$, определяемой равенством (2.18). Для этого подставляя в (2.18) соотношения (2.20), (2.21) при $\beta=1$ и $j=1$, меняя порядок суммирования и интегрирования и учитывая (2.22), получаем

$$\mathcal{R}^{(m,0)}(z_1,z_2)=\frac{z_1^m(a_1)_m}{m!}\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1\frac{t^{b+m-1}(1-t)^{c-b-1}}{(1-z_2t)^{a_2}}\Big(\sum_{k_1=0}^{\infty}\frac{(a_1+m)_{k_1}}{(1+m)_{k_1}}(z_1t)^{k_1}\Big)dt. \tag{2.23}$$

Применяя формулу (1.18) для суммирования ряда в круглых скобках, получаем

$$\mathcal{R}^{(m,0)}(z_1,z_2)=\frac{z_1^m(a_1)_m}{m!}\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1\frac{t^{b+m-1}(1-t)^{c-b-1}}{(1-z_2t)^{a_2}}\mathcal{I}(z_1t,a_1,m)dt, \tag{2.24}$$

где функция $\mathcal{I}$ определяется равенством (2.8).

Выполняя аналогичные преобразования для функций $\mathcal{R}^{(0,m)}(z_1,z_2)$ и $\mathcal{R}^{(m,m)}(z_1,z_2)$, приходим к следующим формулам:

$$\mathcal{R}^{(0,m)}(z_1,z_2)=\frac{z_2^m(a_2)_m}{m!}\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1\frac{t^{b+m-1}(1-t)^{c-b-1}}{(1-z_1t)^{a_1}}\mathcal{I}(z_2t,a_2,m)dt, \tag{2.25}$$

$$\mathcal{R}^{(m,m)}(z_1,z_2)=\frac{z_1^mz_2^m(a_1)_m(a_2)_m}{(m!)^2}\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b+2m-1}(1-t)^{c-b-1}\mathcal{I}(z_1t,a_1,m)\,\mathcal{I}(z_2t,a_2,m)dt, \tag{2.26}$$

где $\mathcal{I}$ имеет вид (2.8). Подставляя (2.24)–(2.26) в (2.13), приходим к требуемому представлению (2.7) для остатка суммирования ряда Аппеля F_1.

2.2.2. Асиптотика остаточного члена $\mathcal{R}^{(m)}$, $m \to \infty$. Найдем асимптотики функций $\mathcal{R}^{(0,m)}$, $\mathcal{R}^{(m,0)}$ и $\mathcal{R}^{(m,m)}$, фигурирующих в (2.13). Для этого, прежде всего, оценим интеграл $\mathcal{I}(z,a,m)$, $m \to \infty$, определяемый по формуле (2.8), выполняя следующие преобразования:

$$\begin{aligned}\mathcal{I}(zt,a,m) = F(a+m,1,1+m,zt) &\overset{(1.19)}{=} (1-zt)^{-1}F\Big(1,1-a,1+m,\frac{zt}{zt-1}\Big) = \\ = (1-zt)^{-1}\sum_{k=0}^{\infty}\frac{(1-a)_k}{(1+m)_k k!}\frac{(zt)^k}{(zt-1)^k} &= \frac{1}{1-zt} + \frac{a-1}{1+m}\frac{zt}{(1-zt)^2} + O\big(m^{-2}\big), \qquad m \to \infty.\end{aligned} \tag{2.27}$$

Подставляя (2.27) в представления (2.24)–(2.26) для функций $\mathcal{R}^{(m,m)}$, $\mathcal{R}^{(m,0)}$, $\mathcal{R}^{(0,m)}$, раскрывая все скобки и применяя к каждой из этих функций теорему 1, приходим к следующему утверждению.

Утверждение 1. *Для функций $\mathcal{R}^{(m,m)}$, $\mathcal{R}^{(m,0)}$, $\mathcal{R}^{(0,m)}$, определенных по формулам* (2.14)–(2.16) *для* $\operatorname{Re} c > \operatorname{Re} b > 0$ *при $m \to \infty$ справедливы асимптотические равенства*

$$\mathcal{R}^{(0,m)}(z_1,z_2) = \frac{\Gamma(c)}{\Gamma(b)}\frac{(a_2)_m}{m^{c-b}m!}\frac{z_2^m}{(1-z_1)^{a_1}(1-z_2)}\Big(1 + C_1 m^{-1} + O\big(m^{-2}\big)\Big), \tag{2.28}$$

$$\mathcal{R}^{(m,0)}(z_1,z_2) = \frac{\Gamma(c)}{\Gamma(b)}\frac{(a_1)_m}{m^{c-b}m!}\frac{z_1^m}{(1-z_1)(1-z_2)^{a_2}}\Big(1 + C_2 m^{-1} + O\big(m^{-2}\big)\Big), \tag{2.29}$$

$$\mathcal{R}^{(m,m)}(z_1,z_2) = \frac{\Gamma(c)}{\Gamma(b)}\frac{(a_1)_m(a_2)_m}{(2m)^{c-b}(m!)^2}\frac{(z_1z_2)^m}{(1-z_1)(1-z_2)}\Big(1 + C_3 m^{-1} + O\big(m^{-2}\big)\Big), \tag{2.30}$$

где коэффициенты C_1 и C_2 определены по формулам (2.10) *и* (2.12)*, а C_3 имеет вид*

$$C_3 = \frac{c-b}{2}\left(\frac{(1-b)-(2-b)(z_1+z_2)+(3-b)z_1z_2}{(1-z_1)(1-z_2)} - \frac{c-b+1}{2}\right) + \frac{(a_1-1)z_1}{1-z_1} + \frac{(a_2-1)z_2}{1-z_2}. \tag{2.31}$$

Отметим, что согласно формуле Стирлинга [11] величины $(a_j)_m/m!$ имеют степенные по m асимптотики при $m \to \infty$. Таким образом, если $\{|z_1| < |z_2| < 1\}$, то

$$\mathcal{R}^{(m)}(z_1,z_2) = \mathcal{R}^{(0,m)}(z_1,z_2) + O(|z_2|^m), \qquad m \to \infty; \tag{2.32}$$

а если $\{|z_2| < |z_1| < 1\}$, то

$$\mathcal{R}^{(m)}(z_1,z_2) = \mathcal{R}^{(m,0)}(z_1,z_2) + O(|z_1|^m), \qquad m \to \infty. \tag{2.33}$$

Если же $|z_1| = |z_2| < 1$, то

$$\mathcal{R}^{(m)}(z_1,z_2) = \mathcal{R}^{(m,0)}(z_1,z_2) + \mathcal{R}^{(0,m)}(z_1,z_2) + O(|z_1|^{2m}), \qquad m \to \infty. \tag{2.34}$$

Учитывая формулы (2.32)–(2.34), а также (2.30)–(2.29), убеждаемся в справедливости утверждений (ii) теоремы 2, в том числе, асимптотик (2.9) и (2.11).

Теорема 2 доказана.

3. ОЦЕНКА ОСТАТОЧНОГО ЧЛЕНА ПРИ СУММИРОВАНИИ РЯДА ГОРНА G

3.1. Суммирование ряда G и теорема об оценке остатка

Проводя рассуждения, схема которых аналогична п. 2.1, представим двойной ряд (1.12) в виде

$$G\,(a_1,a_2;b,\,c;\,z_1,z_2\,) = \mathcal{G}^{(m)}(a_1,a_2;b,\,c;\,z_1,z_2\,) + \mathcal{R}_G^{(m)}(a_1,a_2;b,\,c;\,z_1,z_2\,), \tag{3.1}$$

где первое слагаемое $\mathcal{G}^{(m)}$ является суммой конечного числа членов ряда

$$\mathcal{G}^{(m)}(a_1,a_2;b,\,c;\,z_1,z_2\,) := \sum_{k_1,k_2=0}^{m-1}\frac{(b)_{k_2-k_1}\,(a_1)_{k_1}\,(a_2)_{k_2}}{(c)_{k_2-k_1}\;k_1!\;k_2!}\,z_1^{k_1}z_2^{k_2}, \tag{3.2}$$

а второе слагаемое $\mathcal{R}_G^{(m)}$ — остаток суммирования, который в бикруге $\mathbb{U}^2$ стремится к нулю при $m \to \infty$. Далее для этих величин будем использовать обозначения $\mathcal{G}^{(m)}(z_1,z_2)$ и $\mathcal{R}_G^{(m)}(z_1,z_2)$. Для вычисления $\mathcal{G}^{(m)}$ удобно

использовать следующее рекуррентное соотношение:

$$\mathcal{G}^{(m)}(z_1,z_2)=\mathcal{G}^{(m-1)}(z_1,z_2)+\frac{(a_1)_{m-1}}{(m-1)!}z_1^{m-1}\sum_{k_2=0}^{m-2}\frac{(b)_{k_2-m+1}\ (a_2)_{k_2}}{(c)_{k_2-m+1}\ \ k_2!}\,z_2^{k_2}+$$

$$+\frac{(a_2)_{m-1}}{(m-1)!}z_2^{m-1}\sum_{k_1=0}^{m-2}\frac{(b)_{m-1-k_1}\,(a_1)_{k_1}}{(c)_{m-1-k_1}\ k_1!}\,z_1^{k_1}+\frac{(a_1)_{m-1}\ (a_2)_{m-1}}{((m-1)!)^2}\,z_1^{m-1}z_2^{m-1}.$$

Для вычисления ряда Горна G в точке $(z_1,z_2)\in\mathbb{U}^2$ мы полагаем

$$G\,(a_1,a_2;b,\,c;\,z_1,z_2\,)\simeq\mathcal{G}^{(m)}(a_1,a_2;b,\,c;\,z_1,z_2),\tag{3.3}$$

при этом число m требуется выбрать таким, чтобы при некотором заданном числе $\delta>0$ для остатка $\mathcal{R}_G^{(m)}$ выполнялось неравенство

$$|\mathcal{R}_G^{(m)}(z_1,z_2\,)|<\delta.\tag{3.4}$$

При реализации алгоритма вычисления функции G по формулам (3.2) и (3.3) условие (3.4) удобно проверять, приближенно вычисляя $\mathcal{R}_G^{(m)}(z_1,z_2\,)$ на основе асимптотики или интегрального представления, которые дает приведенная ниже теорема.

Теорема 3. *Предположим, что параметры ряда Горна G, определенного по формуле* (1.12)*, удовлетворяют условию* $\operatorname{Re}(c-b)>0, b\notin\mathbb{Z}$*, а переменные лежат в бикруге* (1.1)*, т.е.* $(z_1,z_2)\in\mathbb{U}^2$.

(i) *Справедлива формула суммирования*

$$G(a_1,a_2;b,\,c;\,z_1,z_2\,)\,=\sum_{k_1,k_2=0}^{m-1}\frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}}\frac{(a_1)_{k_1}(a_2)_{k_2}}{k_1!k_2!}z_1^{k_1}z_2^{k_2}+\mathcal{R}_G^{(m)},\tag{3.5}$$

где для остаточного члена $\mathcal{R}_G^{(m)}(z_1,z_2)$ имеет место интегральное представление

$$\begin{aligned}&\mathcal{R}_G^{(m)}(a_1,a_2;\,b,c;z_1,z_2)=\\&=-\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{1}{2i\sin(\pi b)}\biggl(z_1^m\frac{(a_1)_m(-1)^m}{m!}\oint\frac{(-t)^{m-c}(1-t)^{c-b-1}}{(1-z_2/t)^{a_2}}\,\mathcal{I}(z_1t,a_1,m)dt+\\&\quad+z_2^m\frac{(a_2)_m(-1)^m}{m!}\oint\frac{(-t)^{b+m-1}(1-t)^{c-b-1}}{(1-z_1/t)^{a_1}}\,\mathcal{I}(z_2t,a_2,m)dt-\\&\quad-(z_1z_2)^m\frac{(a_1)_m(a_2)_m}{(m!)^2}\oint(-t)^{b-1}(1-t)^{c-b-1}\mathcal{I}\Bigl(\frac{z_1}{t},a_1,m\Bigr)\,\mathcal{I}(z_2t,a_2,m)\,dt\biggr);\end{aligned}\tag{3.6}$$

здесь функция $\mathcal{I}$ определяется равенством (2.8)*, а интегралы берутся по контуру* $\{t\in\mathbb{C}^2:|t|=1,|\arg(-t)|<\pi\}$.

(ii) *Для остаточного члена $\mathcal{R}_G^{(m)}(z_1,z_2)$ справедливы следующие асимптотики при $m\to\infty$.*

1) *Если $|z_1|<|z_2|$, то*

$$\mathcal{R}_G^{(m)}(z_1,z_2)=\frac{\Gamma(c)}{\Gamma(b)}\frac{(a_2)_m}{m^{c-b}m!}\frac{z_2^m}{(1-z_1)^{a_1}(1-z_2)}\Bigl(1+D_1m^{-1}+O\bigl(m^{-2}\bigr)\Bigr),\tag{3.7}$$

где коэффициент D_1 определяется по формуле

$$D_1=(c-b)\biggl(\frac{(1-b)+(a_1-1+b)z_1+(b-2)z_2+(-a_1+2-b)z_1z_2}{(1-z_1)(1-z_2)}-\frac{c-b+1}{2}\biggr)+\frac{(a_2-1)z_2}{1-z_2}.\tag{3.8}$$

2) *Если $|z_2|<|z_1|$, то*

$$\mathcal{R}_G^{(m)}(z_1,z_2)=\frac{\Gamma(1-b)}{\Gamma(1-c)}\frac{(a_1)_m}{m^{c-b}m!}\frac{z_1^m}{(1-z_1)(1-z_2)^{a_2}}\Bigl(1+D_2m^{-1}+O\bigl(m^{-2}\bigr)\Bigr),\tag{3.9}$$

где коэффициент D_2 имеет вид

$$D_2=(c-b)\biggl(\frac{c+(a_2-c)z_2-(c+1)z_1+(-a_2+1+c)z_1z_2}{(1-z_1)(1-z_2)}-\frac{c-b+1}{2}\biggr)+\frac{(a_1-1)z_1}{1-z_1}.\tag{3.10}$$

3) *Если $|z_1|=|z_2|$, то остаточный член $\mathcal{R}_G^{(m)}(z_1,z_2)$ является суммой правых частей равенств* (3.7) *и* (3.9).

Отметим, что учитывая равенство (1.15), утверждение теоремы очевидным образом может быть переформулировано для ряда Горна G_2.

3.2. Доказательство теоремы 3

3.2.1. Интегральное представление для остатка суммирования $\mathcal{R}_G^{(m)}$. Нетрудно убедиться в том, что остаток суммирования $\mathcal{R}_G^{(m)}(z_1, z_2)$ в формуле (3.1) можно записать в виде

$$\mathcal{R}_G^{(m)}(z_1, z_2) = \mathcal{R}_G^{(m,0)}(z_1, z_2) + \mathcal{R}_G^{(0,m)}(z_1, z_2) - \mathcal{R}_G^{(m,m)}(z_1, z_2), \tag{3.11}$$

где введены обозначения

$$\mathcal{R}_G^{(0,m)}(z_1, z_2) = \sum_{k_1=0}^{\infty} \sum_{k_2=m}^{\infty} \frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}} \frac{(a_1)_{k_1}(a_2)_{k_2}}{k_1!k_2!} z_1^{k_1} z_2^{k_2}, \tag{3.12}$$

$$\mathcal{R}_G^{(m,0)}(z_1, z_2) = \sum_{k_1=m}^{\infty} \sum_{k_2=0}^{\infty} \frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}} \frac{(a_1)_{k_1}(a_2)_{k_2}}{k_1!k_2!} z_1^{k_1} z_2^{k_2}, \tag{3.13}$$

$$\mathcal{R}_G^{(m,m)}(z_1, z_2) = \sum_{k_1=m}^{\infty} \sum_{k_2=m}^{\infty} \frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}} \frac{(a_1)_{k_1}(a_2)_{k_2}}{k_1!k_2!} z_1^{k_1} z_2^{k_2}. \tag{3.14}$$

Здесь, напомним, символ Похгаммера $(b)_n$ при отрицательных n определяем по формуле (1.3) или (1.13). Используя те же приемы, что и в п. 2.2, получаем следующие интегральные выражения для $\mathcal{R}_G^{(0,m)}$, $\mathcal{R}_G^{(m,0)}$ и $\mathcal{R}_G^{(m,m)}$:

$$\mathcal{R}_G^{(0,m)}(z_1, z_2) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \frac{(-1)^{m+1}}{2i\sin(\pi b)} \frac{z_2^m (a_2)_m}{m!} \oint \frac{(-t)^{b+m-1}(1-t)^{c-b-1}}{(1-z_1/t)^{a_1}} \mathcal{I}(z_2 t, a_2, m) dt, \tag{3.15}$$

$$\mathcal{R}_G^{(m,0)}(z_1, z_2) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \frac{(-1)^{m+1}}{2i\sin(\pi b)} \frac{z_1^m (a_1)_m}{m!} \oint \frac{(-t)^{m-c}(1-t)^{c-b-1}}{(1-z_2/t)^{a_2}} \mathcal{I}(z_1 t, a_1, m) dt, \tag{3.16}$$

$$\begin{aligned} \mathcal{R}_G^{(m,m)}(z_1, z_2) = -\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \frac{1}{2i\sin(\pi b)} \frac{(z_1 z_2)^m (a_1)_m (a_2)_m}{(m!)^2} \times \\ \times \oint (-t)^{b-1}(1-t)^{c-b-1} \mathcal{I}\Big(\frac{z_1}{t}, a_1, m\Big) \mathcal{I}(z_2 t, a_2, m)\, dt, \end{aligned} \tag{3.17}$$

где интегралы берутся по контуру $\{t \in \mathbb{C}^2 : |t| = 1, |\arg(-t)| < \pi\}$. Складывая (3.15)–(3.17), получаем интегральное представление (3.6).

3.2.2. Асимптотика функций $\mathcal{R}_G^{(0,m)}$ и $\mathcal{R}_G^{(m,0)}$. Заменяя индекс $k_2 \to k_2 + m$ и используя тождества (2.21), получаем для функции $\mathcal{R}_G^{(0,m)}(z_1, z_2)$ следующее выражение:

$$\mathcal{R}_G^{(0,m)}(z_1, z_2) = \frac{(a_2)_m}{m!} z_2^m \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \frac{(b)_{k_2-k_1+m}}{(c)_{k_2-k_1+m}} \frac{(a_1)_{k_1}}{k_1!} z_1^{k_1} \frac{(a_2+m)_{k_2}}{(1+m)_{k_2}} z_2^{k_2}. \tag{3.18}$$

Записывая символы Похгаммера, которые фигурируют в коэффициентах (3.18), через гамма-функции и используя контурный интеграл (1.17) для бета-функции, находим

$$\begin{aligned} \frac{(b)_{k_2-k_1+m}}{(c)_{k_2-k_1+m}} &= \frac{\Gamma(c)}{\Gamma(b)} \frac{\Gamma(b+k_2-k_1+m)}{\Gamma(c+k_2-k_1+m)} = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \mathrm{B}(b+k_2-k_1+m, c-b) = \\ &= \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \frac{(-1)^{k_2-k_1+m+1}}{2i\sin(\pi b)} \int_1^{(0+)} (-t)^{b+k_2-k_1+m-1}(1-t)^{c-b-1} dt, \end{aligned} \tag{3.19}$$

где $|\arg(-t)| \le \pi$, и $\mathrm{Re}(c-b) > 0$, $b \notin \mathbb{Z}$.

Получим асимптотику интеграла в (3.19) при $m \to \infty$. Для этого запишем его в виде

$$\int_1^{(0+)} (-t)^{b+k_2-k_1+m-1}(1-t)^{c-b-1} dt = (-1)^m \int_1^{(0+)} e^{m \ln t} (-t)^{b+k_2-k_1-1}(1-t)^{c-b-1} dt. \tag{3.20}$$

Так как $\ln t < 0$, $t \in (0,1)$, и этот логарифм достигает максимума в точке $t = 1$, то при $m \to \infty$ вклад точек интегрирования, лежащих вне некоторой окрестности $t = 1$, в интеграл является экспоненциально малым.

Точнее говоря, выполняя оценку интеграла и подставляя ее в (3.20) и (3.19), а затем результат — в (3.18), можно показать, что справедливо равенство

$$\mathcal{R}_G^{(0,m)}(z_1,z_2)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(a_2)_m z_2^m}{m!}\int_{1-\epsilon}^{1} r^{b+m-1}(1-r)^{c-b-1}\times \\ \times\sum_{k_1=0}^{\infty}\frac{(a_1)_{k_1}(z_1/r)^{k_1}}{k_1!}\Big(\sum_{k_2=0}^{\infty}\frac{(a_2+m)_{k_2}(z_2r)^{k_2}}{(1+m)_{k_2}}\Big)dr\Big(1+O\big(m^{-k}\big)\Big)\qquad \forall k,\quad m\to\infty, \tag{3.21}$$

где число ϵ выбрано таким, что $|z_j|<|1-\epsilon|$, $j=1,2$. Используя формулу Эйлера (1.18) для преобразования ряда в интеграл и учитывая (2.22), получаем следующее асимптотическое интегральное представление для $\mathcal{R}_G^{(0,m)}(z_1,z_2)$:

$$\mathcal{R}_G^{(0,m)}(z_1,z_2)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(a_2)_m z_2^m}{m!}\int_{1-\epsilon}^{1} r^{b+m-1}(1-r)^{c-b-1}\Big(1-\frac{z_1}{r}\Big)^{-a_1}\times \\ \times\mathcal{I}(z_2r,a_2,m)dr\Big(1+O\big(m^{-k}\big)\Big)\qquad \forall k,\quad m\to\infty, \tag{3.22}$$

где $|\arg(1-z_1/r)|<\pi$, а функция $\mathcal{I}$ определена равенством (2.8).

Подставляя полученную в п. 2.2 асимптотику (2.27) интеграла $\mathcal{I}$ в (3.22), получаем следующее выражение для $\mathcal{R}_G^{(0,m)}$:

$$\mathcal{R}_G^{(0,m)}(z_1,z_2)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(a_2)_m z_2^m}{m!}\times \\ \times\int_{1-\epsilon}^{1} r^{b+m-1}(1-r)^{c-b-1}\frac{1}{(1-z_1/r)^{a_1}}\frac{1}{1-z_2r}\Big(1+\frac{(a_2-1)z_2r}{1-z_2r}m^{-1}+O\big(m^{-2}\big)\Big)dr. \tag{3.23}$$

Раскроем квадратные скобки в (3.23) и очевидным образом введем обозначения I_j, $j=1,2,3$, для получившихся интегралов:

$$\mathcal{R}_G^{(0,m)}(z_1,z_2)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(a_2)_m z_2^m}{m!}\big[I_1(z_1,z_2)+I_2(z_1,z_2)+I_3(z_1,z_2)\big]. \tag{3.24}$$

Выполняя в $I_1(z_1,z_2)$ замену переменного $r=1-x$ и применяя к полученному интегралу теорему 1, находим

$$I_1(z_1,z_2)=\frac{\Gamma(c-b)}{m^{c-b}}\frac{1}{(1-z_1)^{a_1}(1-z_2)}\Big(1+E_1m^{-1}+O\big(m^{-2}\big)\Big),\qquad m\to\infty, \tag{3.25}$$

где коэффициент E_1 имеет вид

$$E_1:=(c-b)\bigg(\frac{(1-b)+(a_1-1+b)z_1+(b-2)z_2+(-a_1+2-b)z_1z_2}{(1-z_1)(1-z_2)}-\frac{c-b+1}{2}\bigg).$$

Аналогично получаем асимптотику интеграла $I_2(z_1,z_2)$:

$$I_2(z_1,z_2)=\frac{\Gamma(c-b)}{m^{c-b+1}}\frac{(a_2-1)z_2}{(1-z_1)^{a_1}(1-z_2)^2}\Big(1+O\big(m^{-1}\big)\Big),\qquad m\to\infty. \tag{3.26}$$

Третий интеграл имеет порядок $I_3(z_1,z_2)=O(m^{-2})$. Складывая полученные асимптотики для I_j и подставляя в (3.24), получаем следующее утверждение.

Утверждение 2. *Пусть параметры ряда Горна* (1.12) *таковы, что* $\operatorname{Re}(c-b)>0$ *и* $b\notin\mathbb{Z}$. *Тогда для функции* $\mathcal{R}_G^{(0,m)}$, *определенной по формуле* (3.12), *справедливо следующее асимптотическое соотношение:*

$$\mathcal{R}_G^{(0,m)}=\frac{\Gamma(c)}{\Gamma(b)}\frac{(a_2)_m}{m^{c-b}m!}\frac{z_2^m}{(1-z_1)^{a_1}(1-z_2)}\Big(1+D_1m^{-1}+O\big(m^{-2}\big)\Big),\qquad m\to\infty, \tag{3.27}$$

где коэффициент D_1 *определяется по формуле* (3.8).

Нетрудно увидеть, что функция $\mathcal{R}_G^{(m,0)}$ получается из $\mathcal{R}_G^{(0,m)}$ с помощью следующих замен переменных и параметров:

$$b\to(1-c),\qquad c\to(1-b),\qquad a_1\leftrightarrow a_2,\qquad k_1\leftrightarrow k_2,\qquad z_1\leftrightarrow z_2.$$

Тогда, используя (3.27), приходим к утверждению, позволяющему оценить $\mathcal{R}_G^{(m,0)}$.

Утверждение 3. *Пусть параметры ряда Горна* (1.12) *таковы, что* $\operatorname{Re}(c-b)>0$ *и* $b\notin\mathbb{Z}$. *Тогда для функции* $\mathcal{R}_G^{(m,0)}$, *определенной по формуле* (3.13), *справедливо следующее асимптотическое равенство:*

$$\mathcal{R}_G^{(m,0)}=\frac{\Gamma(1-b)}{\Gamma(1-c)}\frac{(a_1)_m}{m^{c-b}m!}\frac{z_1^m}{(1-z_1)(1-z_2)^{a_2}}\Big(1+D_2m^{-1}+O\big(m^{-2}\big)\Big),\qquad m\to\infty, \tag{3.28}$$

где коэффициент D_2 *определяется по формуле* (3.10).

3.2.3. Асимптотика $\mathcal{R}_G^{(m,m)}$, при $m\to\infty$. Перепишем функцию $\mathcal{R}_G^{(m,m)}(z_1,z_2)$, определяемую равенством (3.14), в виде

$$\mathcal{R}_G^{(m,m)}(z_1,z_2)=\frac{(a_1)_m}{m!}z_1^m\frac{(a_2)_m}{m!}z_2^m\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}}\frac{(a_1+m)_{k_1}}{(1+m)_{k_1}}z_1^{k_1}\frac{(a_2+m)_{k_2}}{(1+m)_{k_2}}z_2^{k_2}, \tag{3.29}$$

и представим отношение $(b)_{k_2-k_1}/(c)_{k_2-k_1}$ в виде интеграла следующим образом:

$$\begin{aligned}\frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}}&=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(-1)^{k_1-k_2+1}}{2i\sin(\pi b)}\int_1^{(0+)}(-t)^{b+k_2-k_1-1}(1-t)^{c-b-1}dt=\\&=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(-1)^{k_1-k_2+1}}{2i\sin(\pi b)}\oint(-t)^{b+k_2-k_1-1}(1-t)^{c-b-1}dt,\end{aligned} \tag{3.30}$$

где $\oint$ — интеграл, взятый по контуру $\{t\in\mathbb{C}^2:|t|=1,|\arg(-t)|<\pi\}$, а параметры удовлетворяют ограничениям $\operatorname{Re}(c-b)>0$ и $b\notin\mathbb{Z}$. Выполняя замену переменного $-t=e^{i\varphi}$ во втором интеграле (3.30), перепишем отношение $(b)_{k_2-k_1}/(c)_{k_2-k_1}$ в виде

$$\frac{(b)_{k_2-k_1}}{(c)_{k_2-k_1}}=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(-1)^{k_1-k_2+1}}{2\sin(\pi b)}\int_{-\pi}^{\pi}e^{i\varphi b}e^{i\varphi(k_2-k_1)}(1+e^{i\varphi})^{c-b-1}d\varphi. \tag{3.31}$$

Подставляя (3.31) в (3.29), получаем

$$\begin{aligned}\mathcal{R}_G^{(m,m)}(z_1,z_2)=\sum_{k_1,k_2=0}^{\infty}&\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{(-1)^{k_1-k_2+1}}{2\sin(\pi b)}\frac{(a_1)_m z_1^m}{m!}\frac{(a_2)_m z_2^m}{m!}\times\\&\times\int_{-\pi}^{\pi}e^{i\varphi b}e^{i\varphi(k_2-k_1)}(1+e^{i\varphi})^{c-b-1}\frac{(a_1+m)_{k_1}z_1^{k_1}}{(1+m)_{k_1}}\frac{(a_2+m)_{k_2}z_2^{k_2}}{(1+m)_{k_2}}d\varphi,\end{aligned}$$

а меняя здесь порядок суммирования и интегрирования, находим

$$\begin{aligned}\mathcal{R}_G^{(m,m)}(z_1,z_2)=-\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\frac{1}{2\sin(\pi b)}\frac{(a_1)_m z_1^m}{m!}\frac{(a_2)_m z_2^m}{m!}\times\\\times\int_{-\pi}^{\pi}e^{i\varphi b}(1+e^{i\varphi})^{c-b-1}\sum_{k_1=0}^{\infty}\frac{(a_1+m)_{k_1}}{(1+m)_{k_1}}\Big(-\frac{z_1}{e^{i\varphi}}\Big)^{k_1}\sum_{k_2=0}^{\infty}\frac{(a_2+m)_{k_2}}{(1+m)_{k_2}}(-z_2e^{i\varphi})^{k_2}d\varphi.\end{aligned} \tag{3.32}$$

Учитывая выражения (2.22) и (2.27), с помощью (3.32) получаем требуемую асимптотику для $\mathcal{R}_G^{(m,m)}$, которую устанавливает следующее утверждение.

Утверждение 4. *Пусть параметры ряда Горна* (1.12) *таковы, что* $\operatorname{Re}(c-b)>0$ *и* $b\notin\mathbb{Z}$. *Тогда для функции* $\mathcal{R}_G^{(m,m)}$, *определенной по формуле* (3.14), *справедливо асимптотическое равенство*

$$\begin{aligned}\mathcal{R}_G^{(m,m)}(z_1,z_2)=-\frac{\Gamma(c)}{2\sin(\pi b)\Gamma(b)\Gamma(c-b)}\frac{(a_1)_m(a_2)_m}{(m!)^2}(z_1z_2)^m\times\\\times\int_{-\pi}^{\pi}\frac{e^{i\varphi b}(1+e^{i\varphi})^{c-b-1}}{(1+z_1e^{-i\varphi})(1+z_2e^{i\varphi})}\left[1+\left(\frac{(1-a_1)z_1e^{-i\varphi}}{1+z_1e^{-i\varphi}}+\frac{(1-a_2)z_2e^{i\varphi}}{1+z_2e^{i\varphi}}\right)m^{-1}+O\big(m^{-2}\big)\right]d\varphi.\end{aligned} \tag{3.33}$$

Можно получить несколько менее точную, но проще реализуемую оценку для $\mathcal{R}_G^{(m,m)}$ следующим способом. Оценим модуль интеграла, входящего в (3.33), с помощью соотношений

$$\begin{aligned}|I|&=\left|\int_{-\pi}^{\pi}\frac{e^{i\varphi b}(1+e^{i\varphi})^{c-b-1}}{(1+z_1e^{-i\varphi})(1+z_2e^{i\varphi})}d\varphi\right|\le\\&\le\frac{1}{(1-|z_1|)(1-|z_2|)}\frac{\sqrt{\pi}2^{\operatorname{Re}(c-b)}\left(1+e^{\frac{\pi}{2}|\operatorname{Im}(c+b-1)|}\right)}{\operatorname{Re}(c-b-1)}\frac{\Gamma\left(\frac{\operatorname{Re}(c-b)}{2}\right)}{\Gamma\left(\frac{\operatorname{Re}(c-b-1)}{2}\right)}.\end{aligned}$$

Тогда имеет место оценка

$$|\mathcal{R}_G^{(m,m)}| \le \frac{z_1^m z_2^m}{(1-|z_1|)(1-|z_2|)} \frac{(a_1)_m (a_2)_m}{(m!)^2} D_3(b,c), \tag{3.34}$$

где коэффициент $D_3(b,c)$ определен равенством

$$D_3(b,c) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \frac{2^{\mathrm{Re}(c-b-1)}\sqrt{\pi}}{\sin(\pi b)\mathrm{Re}(c-b-1)} \left(1 + e^{\frac{\pi}{2}|\mathrm{Im}(c+b-1)|}\right) \frac{\Gamma\left(\frac{\mathrm{Re}(c-b)}{2}\right)}{\Gamma\left(\frac{\mathrm{Re}(c-b-1)}{2}\right)}.$$

3.2.4. Сравнение величин $\mathcal{R}_G^{(m,0)}$, $\mathcal{R}_G^{(0,m)}$ и $\mathcal{R}_G^{(m,m)}$ при $m \to \infty$. Как и в пп. 2.2 вспомним, что согласно формуле Стирлинга [11] асимптотики величин $(a_j)_m/m!$ имеют степенной по m характер при $m \to \infty$. Поэтому, согласно формуле (3.11), суммируя выражения (3.27), (3.28) и (3.34) для $\mathcal{R}_G^{(0,m)}$, $\mathcal{R}_G^{(m,0)}$ и $\mathcal{R}_G^{(m,m)}$, получаем:

если $\{|z_1| < |z_2| < 1\}$, то

$$\mathcal{R}_G^{(m)}(z_1,z_2) = \mathcal{R}_G^{(0,m)}(z_1,z_2) + O(|z_2|^m), \qquad m \to \infty; \tag{3.35}$$

а если $\{|z_2| < |z_1| < 1\}$, то

$$\mathcal{R}_G^{(m)}(z_1,z_2) = \mathcal{R}_G^{(m,0)}(z_1,z_2) + O(|z_1|^m), \qquad m \to \infty. \tag{3.36}$$

Если же $|z_1| = |z_2| < 1$, то

$$\mathcal{R}_G^{(m)}(z_1,z_2) = \mathcal{R}_G^{(m,0)}(z_1,z_2) + \mathcal{R}_G^{(0,m)}(z_1,z_2) + O(|z_1|^{2m}), \qquad m \to \infty. \tag{3.37}$$

Учитывая формулы (3.35)–(3.37), а также (3.27)–(3.33), убеждаемся в справедливости утверждений (ii) теоремы 3, в том числе, асимптотик (3.7) и (3.9).

Теорема 3 доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Тарасов О.В.* Применение функциональных уравнений для вычисления фейнмановских интегралов // Теор. и матем. физ. 2019. Т. 200. № 2. С. 324–342.
2. *Власов В.И., Скороходов С.Л.* Аналитическое решение задачи о кавитационном обтекании клина. I // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. № 12. С. 2098–2121.
3. *Kalmykov M., Bytev V., Kniehl B., Moch S.-O., Ward B., Yost S.* Hypergeometric functions and Feynman diagrams. In: Blumlein J., Schneider C. (eds) Anti-Differentiation and the Calculation of Feynman Amplitudes. Texts & Monographs in Symbolic Computation (A Series of the Research Institute for Symbolic Computation, Johannes Kepler University, Linz, Austria). Springer, Cham, 2021.
4. *Bezrodnykh S.I., Vlasov V.I.* Asymptotics of the Riemann — Hilbert problem for the Somov model of magnetic reconnection of long shock waves // Матем. заметки. 2021. V. 110. № 6. P. 853–871,
5. *Шилин И. А., Чой Дж.* Метод континуальных теорем сложения и интегральные соотношения между функциями Кулона и функцией Аппеля F_1 // Ж. вычисл. мат. и матем. физики. 2022. Т. 62. № 9. С. 131–140.
6. *Karp D., Zhang Yi.* Convergent expansions and bounds for the incomplete elliptic integral of the second kind near the logarithmic singularity // Math. Comp. 2023. V. 92. № 344. P. 2769.
7. *Шилин И. А., Чой Дж.* Алгебры Ли и специальные функции, связанные с изотропным конусом // Итоги науки и техн. Соврем. мат. и ее прил. Темат. обз., 222, ВИНИТИ РАН, М., 2023, 141–152.
8. *Claude Duhr, Franziska Porkert* Feynman integrals in two dimensions and single-valued hypergeometric functions // J. High Energ. Phys. 2024. V. 2.
9. *Wei Fan.* Celestial conformal blocks of massless scalars and analytic continuation of the Appell function F1 // J. High Energ. Phys. 2024. V. 1.
10. *Appel P., Kampe de Feriet J.* Fonctions hypergeometriques et hyperspherique. Paris: Gauthier–Villars, 1926.

11. *Бейтмен Г., Эрдейи А.* Высшие трансцендентные функции. Гипергеометрическая функция. Функции Лежандра. М.: Наука, 1973.

12. *Exton H.* Multiple hypergeometric functions and application. New York: J. Willey & Sons inc, 1976.

13. *Erdelyi A.* Hypergeometric functions of two variables // Acta Mat. 1950. V. 83. Iss. 131. P. 131–164.

14. *Olsson O.M.* Integration of the partial differential equations for the hypergeometric function F_1 and F_D of two and more variables // J. Math. Phys. 1964. V. 5. № 420. P. 420–430.

15. *Безродных С.И.* Аналитическое продолжение функции Аппеля F_1 и интегрирование связанной с ней системы уравнений в логарифмическом случае // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 4. С. 555–587.

16. *Bezrodnykh S.I.* Analytic continuation of Lauricella's function $F_D^{(N)}$ for large in modulo variables near hyperplanes $\{z_j = z_l\}$ // Integral Transforms and Special Functions. 2022. V. 33. № 4. P. 276–291.

17. *Bezrodnykh S.I.* Analytic continuation of Lauricella's function $F_D^{(N)}$ for variables close to unit near hyperplanes $\{z_j = z_l\}$ // Integral Transforms and Special Functions. 2022. V. 33. № 5. P. 419–433.

18. *Colavecchia F.D., Gasaneo G., Miraglia J.E.* Numerical evaluation of Appell's F_1 hypergeometric function // Comput. Phys. Communicat. 2001. V. 138. P. 29–43.

19. *Colavecchia F.D., Gasaneo G.* f1: a code to compute Appell's F1 hypergeometric function // Comput. Phys. Communicat. 2004. V. 157. P. 32–38.

20. *Ananthanarayan B., Bera S., Friot S., Pathak T.* Olsson.wl & ROC2.wl: Mathematica packages for transformations of multivariable hypergeometric functions & regions of convergence for their series representations in the two variables case // Comput. Phys. Communicat. 2024. V. 300. 109162 crossref.

21. *Ananthanarayan B., Bera S., Friot S., Marichev O., Pathak T.* On the evaluation of the Appell F2 double hypergeometric function // Comput. Phys. Communicat. 2023. V. 284. 108589.

22. *Безродных С.И.* Гипергеометрическая функция Лауричеллы $F_D^{(N)}$, задача Римана–Гильберта и некоторые приложения // Успехи матем. наук. 2018. Т. 73. № 6 (444). С. 3–94.

23. *Безродных С.И.* Формулы для вычисления интегралов типа Эйлера и их приложение к задаче построения конформного отображения многоугольников // Ж. вычисл. матем. и матем. физ. 2023. V. 63. № 11. P. 1763–1798.

24. *Wong R.* Asymptotic approximations of integrals. Society for Industrial and Applied Mathematics, 2001.

ESTIMATION OF THE REMAINDER TERMS OF CERTAIN HORN HYPERGEOMETRIC SERIES

S. I. Bezrodnykh[a,*], O. V. Dunin-Barkovskaya[a,b,]**

[a] 119333 Moscow, Vavilov str., 44, Federal Research Center "Informatics and Management" of the Russian Academy of Sciences, Russia
[b] 119234 Moscow, Universitetsky pr., 13, Sternberg Astronomical Institute, Moscow State University, Russia
**e-mail: sbezrodnykh@mail.ru*
***e-mail: olga.ptitsyna@gmail.com*

Received: 20.04.2024
Revised: 31.05.2024
Accepted: 05.06.2024

Abstract. Integral representations and asymptotic estimates for remainder terms arising in the summation of the Appel hypergeometric F_1 series and its related series G_2, indicated in the Horn list of hypergeometric series of two variables, are constructed. The formulas found have an application to the development of algorithms for calculating the F_1 function using formulas of analytical continuation into the entire $\mathbb{C}^2$ space. The results can be applied in problems of mathematical physics and computational theory of function, including the construction of a conformal mapping of complex polygons based on the Schwarz–Christoffel integral.

Keywords: Appel and Horn hypergeometric functions, formulas of analytical continuation, effective calculation of hypergeometric functions.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2243–2261

УДК 519.642

БЫСТРОЕ ВЫЧИСЛЕНИЕ ИНТЕГРАЛЬНЫХ ОПЕРАТОРОВ ТИПА СВЕРТКИ В ЗАДАЧАХ ОЦЕНИВАНИЯ ОПЦИОНОВ В МОДЕЛЯХ ЛЕВИ[1)]

© 2024 г. А. С. Гречко[1,], О. Е. Кудрявцев[1,2,*]**

[1]344015, Ростов-на-Дону, Еременко 58/11, ООО НПФ "ИнВайз Системс", Россия
[2]344002, Ростов-на-Дону, Буденновский 20, Ростовский филиал Российской таможенной академии, Россия
**e-mail: koe@sfedu.ru*
***e-mail: alex@itparadigma.ru*

Поступила в редакцию 09.06.2024 г.
Переработанный вариант 01.08.2024 г.
Принята к публикации 23.08.2024 г.

Построен приближенный алгоритм вычисления интегральных операторов типа свертки, возникающих при оценивании барьерных опционов в моделях Леви методом Винера–Хопфа. Дополнительно исследован вопрос возможности применения методов машинного обучения (искусственных нейронных сетей) к аппроксимации специального вида интегралов, являющихся ключевым элементом в конструкции приближенных формул для рассматриваемых интегральных операторов Винера–Хопфа. Основная идея заключается в разложении функции цены в ряд Фурье и трансформации контура интегрирования для каждого слагаемого ряда Фурье. В результате мы получаем набор типовых интегралов, которые зависят от факторов Винера–Хопфа, но не зависят от функции цены, при этом наиболее затратная с вычислительной точки зрения часть численного метода сводится к вычислению указанных интегралов. Поскольку вычислять их нужно только один раз, а не на каждой итерации, как это было в стандартных реализациях метода Винера–Хопфа, то это существенно ускорит вычисления. Более того, для вычисления типовых интегралов можно обучить нейросеть. Предлагаемый подход особенно эффективен для спектрально односторонних процессов Леви, для которых известны явные формулы факторизации Винера–Хопфа. В этом случае мы получаем удобные для вычислений формулы путем интегрирования вдоль разреза. Главным преимуществом включения нейросетей в вычислительную схему является возможность проводить вычисления на неравномерной сетке. Такой гибридный численный метод сможет успешно конкурировать с классическими методами вычислений сверток в подобных задачах с помощью быстрого преобразования Фурье. Вычислительные эксперименты показывают, что нейросети с одним скрытым слоем из 20 нейронов способны эффективно справляться с задачами аппроксимации рассматриваемых вспомогательных интегралов. Библ. 25. Фиг. 2. Табл. 1.

Ключевые слова: факторизация Винера–Хопфа, процессы Леви, интегральные операторы свертки, интегральные преобразования, численные методы, машинное обучение, вычислительная финансовая математика.

DOI: 10.31857/S0044466924120027, **EDN**: KCYGDV

1. ВВЕДЕНИЕ

Последние 25 лет процессы Леви (подробнее, см., например, [1]–[25]), позволяющие моделировать скачки котировок акций, и позволяют получать более реалистичные оценки рисков по сравнению с гауссовыми моделями, вызывают интерес у исследователей в различных приложениях. Особую популярность модели Леви обрели в области вычисления цен финансовых опционов (см., например, [5], [10], [9], [15], [17], [19], [16], [24]), определяемых как математические ожидания функций, зависящих от процессов Леви и их экстремумов. Следует отметить, что разработка эффективных численных методов оценивания опционов является одной из ключевых задач вычислительной финансовой математики. Например, задача вычисления цен барьерных опционов сводится к решению начально-краевой задачи для интегродифференциального уравнения с частными производными, которое по своей сути является обратным уравнением Колмогорова. Наиболее популярными подходами к решению рассматриваемых задач являются: методы Монте-Карло, численные методы для вычисления математического ожидания, численные методы решения интегродифференциальных уравнений с частными производными. Отметим, что все эти подходы часто включают технику факторизации Винера–Хопфа для процессов Леви.

[1)] Работа выполнена при финансовой поддержке РНФ (проект 23-21-00474).

Подробный обзор можно найти в недавней книге [19], в которой представлен обширный список литературы по данной тематике. Несмотря на огромное количество статей по приложению процессов Леви к задачам вычислительной финансовой математики, исследования в этой области не теряют своей интенсивности. С одной стороны, это связано с тем, что разнообразие процессов Леви и их математическая сложность определяют необходимость разработки новых вычислительных методов, более доступных для практиков и учитывающих особенности конкретных моделей. С другой стороны, расширяется круг задач финансовой математики, где могут применяться модели Леви.

В настоящее время применение методов машинного обучения для решения задач финансовой математики вызывает активный интерес ученых. Вместе с тем, прямолинейное обучение искусственных нейросетей на рыночных данных стакивается с критикой из-за отсутствия теоретического обоснования адекватности таких моделей. С другой стороны, алгоритмы машинного обучения могут заменять некоторые рутинные элементы численных методов. Таким образом, создание гибридных методов, включающих в себя элементы "классических" методов вычислительной математики и элементов машинного обучения, можно считать наиболее перспективным направлением развития вычислительной финансовой математики, которая является одним из основных драйверов развития современных численных методов. Примеры гибридных методов вычислительной финансовой математики можно найти в работах [12], [14], [3], [22].

Целью настоящей статьи является построение приближенного алгоритма вычисления интегральных операторов типа свертки, возникающих при оценивании барьерных опционов в моделях Леви методом Винера–Хопфа. Основная идея заключается в разложении функции цены в ряд Фурье и трансформации контура интегрирования для каждого слагаемого ряда Фурье. В результате мы получаем набор типовых интегралов, которые зависят от факторов Винера–Хопфа, но не зависят от функции цены. Таким образом, наиболее затратная с вычислительной точки зрения часть численного метода сводится к вычислению этих интегралов. Поскольку вычислять их нужно только один раз, а не на каждой итерации, как это было в стандартных реализациях метода Винера–Хопфа, то это существенно ускорит вычисления. Особенно эффективен данный подход становится с применением явных формул символов факторов упрощенной факторизации Винера–Хопфа для умеренно устойчивых процессов Леви, разработанной в [18], [20]. В результате получается набор типовых интегралов, для вычисления которых можно обучить нейросеть. Главным преимуществом включения нейросетей в вычислительную схему является возможность проводить вычисления на неравномерной сетке. Такой гибридный численный метод сможет успешно конкурировать с классическими методами вычислений сверток в подобных задачах с помощью быстрого преобразования Фурье.

Статья имеет следующую структуру. Раздел 2 посвящен основным фактам из теории процессов Леви и вопросам, связанным с факторизацией Винера–Хопфа в моделях Леви. В разд. 3 построен алгоритм быстрого вычисления интегральных операторов свертки с ядром типа плотности вероятности случайных величин, принимающих только положительные или только отрицательные значения, рассмотрены возможности применения искусственных нейронных сетей для вычисления типовых интегралов, к вычислению которых сводится построенный метод. В последнем разделе приводятся вычислительные эксперименты, демонстрирующие возможности построенного метода для популярной модели финансовых рынков KoBoL (см. [5]), известной также как модель CGMY (см. [8]).

2. ПРОЦЕССЫ ЛЕВИ

2.1. Основные определения

Пусть $(X_t)_{t\geqslant 0}$ – процесс Леви на $\mathbf{R}$, тогда X_t полностью определяется своей характеристической экспонентой

$$\psi(\xi) := \frac{1}{t}\ln E[e^{i\xi X(t)}] = e^{-t\psi(\xi)},$$

согласно хорошо известной формуле Леви–Хинчина (см. [25])

$$\psi(\xi) = \frac{\sigma^2}{2}\xi^2 - i\gamma\xi + \int_{\mathbf{R}}(1 - e^{i\xi x} + i\xi x\mathbf{1}_{[-1,1]}(x))\Pi(dx), \tag{2.1}$$

где $\sigma \geqslant 0$ и $\gamma \in \mathbf{R}$ – константы, а Π – мера на $\mathbf{R}\setminus\{0\}$, удовлетворяющая свойству

$$\int_{\mathbf{R}} \min\{1, x^2\}\Pi(dx) < +\infty.$$

Параметр σ^2 называется *гауссовским коэффициентом*, мера Π – *мерой Леви* скачков, а тройка параметров (σ^2, γ, Π) – *порождающим триплетом*.

Если скачки процесса Леви имеют ограниченную вариацию (т.е. $\int_{\mathbf{R}} \min\{1,|x|\}\Pi(dx) < +\infty$) и $\sigma = 0$, то характеристическая экспонента имеет более простой вид

$$\psi(\xi) = -i\mu\xi + \int_{-\infty}^{+\infty} (1 - e^{i\xi y})\Pi(dy). \tag{2.2}$$

Регулярные процессы Леви экспоненциального типа были введены в [5]. Свойства таких процессов позволяют эффективно решать стандартные задачи оценивания опционов, сводящиеся к интегродифференциальным уравнениям.

Определение 2.1. Пусть $\lambda_- < 0 < \lambda_+$. Будем называть процесс Леви X *процессом Леви экспоненциального типа* $[\lambda_-, \lambda_+]$, если его мера Леви удовлетворяет следующему свойству:

$$\int_{-\infty}^{-1} e^{-\lambda_+ x}\Pi(dx) + \int_{1}^{+\infty} e^{-\lambda_- x}\Pi(dx) < \infty. \tag{2.3}$$

Лемма 2.1. *Пусть X – процесс Леви экспоненциального типа $[\lambda_-, \lambda_+]$. Тогда*

- *характеристическая экспонента ψ голоморфна в полосе $\Im\xi \in (\lambda_-, \lambda_+)$ и непрерывна вплоть до ее границы;*
- *существуют константы C и $\nu > 0$ такие, что для всех ξ в полосе $\Im\xi \in [\lambda_-, \lambda_+]$*

$$|\psi(\xi)| \leqslant C(1 + |\xi|)^{\nu}; \tag{2.4}$$

- *для любого $q > 0$ уравнение*

$$q + \psi(\xi) = 0 \tag{2.5}$$

имеет не более одного чисто мнимого корня в нижней (верхней) полуплоскости, назовем его в случае существования $-i\beta_+$ ($-i\beta_-$);

- *корень $-i\beta_\mp$ существует тогда и только тогда, когда*

$$q + \psi(i(\lambda_\pm \mp 0)) < 0, \tag{2.6}$$

и если он существует, то его кратность равна одному.

Доказательство леммы приведено в [5].

Пусть $(\Omega, \mathcal{F}, \mathbf{P})$ – вероятностное пространство, на котором определен одномерный процесс Леви X. Обозначим через $\mathcal{S}_t = \sup_{0\leqslant s\leqslant t} X_s$ и $\mathcal{I}_t = \inf_{0\leqslant s\leqslant t} X_s$. Процессы $\mathcal{S} = \{\mathcal{S}_t\}$ и $\mathcal{I} = \{\mathcal{I}_t\}$ называются процессами *супремума* и *инфимума* соответственно. Отметим, что для любого $t > 0$ имеют место следующие полезные соотношения (см. [25]):

$$\mathcal{S}_t \overset{d}{\sim} X_t - \mathcal{I}_t, \quad \mathcal{I}_t \text{ и } X_t - \mathcal{I}_t - \text{независимы}, \tag{2.7}$$

$$\mathcal{I}_t \overset{d}{\sim} X_t - \mathcal{S}_t, \quad \mathcal{S}_t \text{ и } X_t - \mathcal{S}_t - \text{независимы}. \tag{2.8}$$

Пусть $q > 0$, введем случайное время T_q, имеющее показательное распределение с параметром интенсивности q. Пусть существуют плотности вероятности случайных величин X_{T_q}, $\mathcal{S}_{T_q}$ и $\mathcal{I}_{T_q}$, обозначим их через p_q, p_q^+ и p_q^- соответственно. Отметим, что носитель p_q^+ расположен на правой полуоси, а носитель p_q^- – на левой.

Введем следующие интегральные операторы свертки с ядрами $p_q(-y)$, $p_q^+(-y)$ и $p_q^-(-y)$ соответственно:

$$\mathcal{E}_q u(x) = \int_{-\infty}^{+\infty} u(y)p_q(y-x)dy = \int_{-\infty}^{+\infty} u(x+y)p_q(y)dy = E[u(x+X_{T_q})], \tag{2.9}$$

$$\mathcal{E}_q^+ u(x) = \int_{-\infty}^{+\infty} u(y)p_q^+(y-x)dy = \int_{0}^{+\infty} u(x+y)p_q^+(y)dy = E[u(x+\mathcal{S}_{T_q})], \tag{2.10}$$

$$\mathcal{E}_q^- u(x) = \int_{-\infty}^{+\infty} u(y)p_q^-(y-x)dy = \int_{-\infty}^{0} u(x+y)p_q^-(y)dy = E[u(x+\mathcal{I}_{T_q})]. \tag{2.11}$$

Из определения операторов $\mathcal{E}_q$, $\mathcal{E}_q^+$, $\mathcal{E}_q^-$ в (2.9)–(2.11) и свойств (2.7), (2.8) следуют факторизационные тождества

$$\mathcal{E}_q = \mathcal{E}_q^+\mathcal{E}_q^- = \mathcal{E}_q^-\mathcal{E}_q^+, \tag{2.12}$$

которые играют ключевую роль в применении метода Винера–Хопфа при решении задач вычисления цен барьерных опционов в моделях Леви.

Определим преобразование Фурье функции u

$$\mathcal{F}u(\xi) = \int_{\mathbf{R}} e^{-ix\xi} u(x)dx \tag{2.13}$$

и характеристические функции распределений $\mathcal{S}_{T_q}$ и $\mathcal{I}_{T_q}$:

$$\phi_q^+(\xi) = E\left[e^{i\xi\mathcal{S}_{T_q}}\right] = \overline{(\mathcal{F}p_q^+)}(\xi), \tag{2.14}$$

$$\phi_q^-(\xi) = E\left[e^{i\xi\mathcal{I}_{T_q}}\right] = \overline{(\mathcal{F}p_q^-)}(\xi). \tag{2.15}$$

Используя функции $\phi_q^+(\xi)$ и $\phi_q^-(\xi)$, можно записать операторы свертки $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$ как интегральные операторы Фурье:

$$\mathcal{E}_q^+ u(x) = (2\pi)^{-1} \int_{\mathbf{R}} e^{ix\xi} \phi_q^+(\xi) \mathcal{F}u(\xi)d\xi, \tag{2.16}$$

$$\mathcal{E}_q^- u(x) = (2\pi)^{-1} \int_{\mathbf{R}} e^{ix\xi} \phi_q^-(\xi) \mathcal{F}u(\xi)d\xi. \tag{2.17}$$

Заметим, что в случае процессов Леви экспоненциального типа $[\lambda_-, \lambda_+]$ функции $\phi_q^+(\xi)$ ($\phi_q^-(\xi)$) допускают аналитическое продолжение в верхнюю полуплоскость $\Im\xi > \omega_-$, $\omega_- \geqslant \lambda_-$ (нижнюю полуплоскость $\Im\xi < \omega_+$, $\omega_+ \leqslant \lambda_+$).

Подробнее свойства процессов, применяемых при моделировании финансовых рынков, изложены в монографиях [5], [10].

2.2. Применение операторов $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$ при решении задач вычисления цен барьерных опционов

Пусть T, K, D, U, r — момент исполнения, цена исполнения, нижний и верхний барьеры и безрисковая процентная ставка соответственно, а цена акции $S_t = De^{X_t+x}$, $x \in \mathbf{R}$, — экспоненциальный процесс Леви при выбранной нейтральной по отношению к риску мере (подробнее см., например, [20]).

Обозначим через $h = \ln U/D$, а $G(x)$ – функцию выплат по опциону. Тогда выплата при исполнении составит $\mathbf{1}_{(0,h)}(x + X_T)G(x + X_T)$, а безарбитражная цена двухбарьерного опциона в начале рассматриваемого периода ($t = 0$) при $x \in (0, h)$ (т.е. начальная цена акции находится между барьерами D и U) определяется по формуле

$$V(T, x) = E\left[e^{-rT}\mathbf{1}_{x+\mathcal{I}_T>0}\mathbf{1}_{x+\mathcal{S}_T<h}G(x + X_T)\right]. \tag{2.18}$$

Решение задачи (2.18) численно сводится к последовательному вычислению следующих функций для $n = 1, 2, \ldots, N$:

$$v_n(q, x) = E\left[\frac{v_{n-1}(q, x + X_{T_{q+r}})}{(1 + r/q)}\mathbf{1}_{x+\mathcal{I}_{T_{q+r}}>0}\mathbf{1}_{x+\mathcal{S}_{T_{q+r}}<h}\right], \tag{2.19}$$

где $v_0(q, x) = G(x)\mathbf{1}_{(0,h)}(x)$ и $q = N/T$. При $N \to +\infty$ функция $v_N(q, x)$ стремится к $V(T, x)$ (см., например, [18], [20]).

Для однобарьерного случая ($h = +\infty$) можно применить метод Винера–Хопфа при $n = 1, \ldots, N$:

$$v_n(q, x) = \frac{1}{(1 + r/q)} E[v_{n-1}(q, x + \mathcal{I}_{T_{q+r}} + (X_{T_{q+r}} - \mathcal{I}_{T_{q+r}}))\mathbf{1}_{x+\mathcal{I}_{T_{q+r}}>h}] =$$

$$= \frac{1}{(1 + r/q)} \mathcal{E}_{q+r}^- \mathbf{1}_{(h,+\infty)} \mathcal{E}_{q+r}^+ v_{n-1}(q, x). \tag{2.20}$$

Заметим, что явные формулы для операторов $\mathcal{E}_q^-$ и $\mathcal{E}_q^+$ известны только для небольшого количества частных случаев: рациональной характеристической экспоненты, негауссовских процессов Леви с односторонними скачками. В общем случае необходимо аппроксимировать функции $\phi_q^\pm$ в формулах (2.16) и (2.17) с помощью специальных вычислительных техник. О приближенной факторизации Винера–Хопфа в задачах вычисления цен опционов можно подробно прочитать в [17], [2].

В работе [18] был предложен упрощенный вариант факторизации Винера–Хопфа для вычисления однобарьерных опционов, который значительно проще программно реализовать. Позднее, этот подход был распространен на двубарьерный случай (см. [20], [21]). Напомним основные шаги подхода [20] к вычислению (2.19)

для процессов Леви с характеристической экспонентой вида (2.2). Сначала мы представляем X_t как разность между двумя субординаторами X_t^+ и $-X_t^-$:

$$X_t = X_t^+ - (-X_t^-).$$

Напомним, что субординатор — это процесс Леви с траекториями, которые почти наверное неубывают. Согласно [10, предложение 3.10], субординатор не имеет диффузионной компоненты, только положительные скачки конечной вариации и неотрицательный снос.

Обозначим через $\psi_+(\xi)$ и $\psi_-(\xi)$ характеристические экспоненты X_t^+ и X_t^- соответственно. Тогда $\psi_+(\xi)$ и $\psi_-(\xi)$ имеют следующий вид:

$$\begin{aligned}\psi_+(\xi) &= -i\mu_+\xi + \int_0^{+\infty}(1-e^{i\xi y})F(dy), \mu_+ \geqslant 0,\\ \psi_-(\xi) &= -i\mu_-\xi + \int_{-\infty}^{0}(1-e^{i\xi y})F(dy), \mu_- \leqslant 0.\end{aligned}$$

Для процесса X_t^+ мы можем явным образом определить оператор $\mathcal{E}_q^+$ по формуле (2.10) с

$$p_q^+(x) = (2\pi)^{-1}\int_{\mathbf{R}} e^{ix\xi}q(q+\psi_+(\xi))^{-1}d\xi, x > 0, \tag{2.21}$$

а для процесса X_t^- – оператор $\mathcal{E}_q^-$ по формуле (2.11) с

$$p_q^-(x) = (2\pi)^{-1}\int_{\mathbf{R}} e^{ix\xi}q(q+\psi_-(\xi))^{-1}d\xi, x < 0. \tag{2.22}$$

Тогда (2.19) можно приближенно вычислить по формуле (см. детали в [20])

$$v_n(q,x) = \frac{\mathbf{1}_{(0,h)}(x)}{(1+r/q)}\mathcal{E}^+_{2(q+r)}\mathbf{1}_{(0,h)}\mathcal{E}^-_{q+r}\mathbf{1}_{(0,h)}\mathcal{E}^+_{2(q+r)}v_{n-1}(q,x). \tag{2.23}$$

Заметим, что как в случае (2.19), так и в случае (2.23), нам необходимо многократно вычислять действие интегральных операторов свертки $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$. Обычно данная вычислительная процедура реализуется с помощью быстрого преобразования Фурье. В следующем разделе мы рассмотрим альтернативный подход.

3. БЫСТРОЕ ВЫЧИСЛЕНИЕ ИНТЕГРАЛЬНЫХ ОПЕРАТОРОВ СВЕРТКИ С ЯДРОМ ТИПА ПЛОТНОСТИ ВЕРОЯТНОСТИ

3.1. Общие формулы вычисления действия операторов $\mathcal{E}^+$ и $\mathcal{E}^-$ на функции с компактным носителем

Пусть X – процесс Леви экспоненциального типа $[\lambda_-, \lambda_+]$, и пусть параметр $q > 0$ такой, что не выполнены условия (2.6). В этом случае функции $\phi_q^+(\xi)$ и $\phi_q^-(\xi)$ аналитичны в полуплоскостях $\Im\xi > \lambda_-$ и $\Im\xi < \lambda_+$ соответственно.

Применение метода Винера–Хопфа при вычислении цен опционов с выплатами, зависящими от траектории процесса Леви, обычно требует многократного применения операторов $\mathcal{E}^+$ и $\mathcal{E}^-$ (см., например, [17], [18], [21]). Отметим, что в случае, когда вычисляются цены двухбарьерных опционов, носители функций, на которые действуют операторы $\mathcal{E}^+$ и $\mathcal{E}^-$, ограничены барьерами сверху и снизу (см. [20], [21]).

Рассмотрим действие оператора $\mathcal{E}^+$ на функцию $u(x)$, носитель которой принадлежит интервалу (d,u). Пусть $M_0 > 1$. Рассмотрим разбиение интервала (d,u) точками $\{x_n\}_0^{M_0}$:

$$d = x_0 < x_1 < \cdots < x_{M_0} = u. \tag{3.1}$$

Пусть известны значения функции $u(x)$ в точках $\{x_n^\star\}_1^{M_0}$, где

$$x_{n-1} < x_n^\star < x_n \text{ для всех } 1 \leqslant n \leqslant M_0.$$

Тогда аппроксимируем преобразование Фурье функции $u(x)$ при $\Im\xi \neq 0$ следующим образом:

$$(\mathcal{F}u)(\xi) = \int_{-\infty}^{+\infty} e^{-ix\xi}u(x)dx = \int_d^u e^{-ix\xi}u(x)dx =$$

$$
\begin{aligned}
&= \int_{x_0}^{x_1} e^{-ix\xi}u(x)dx + \int_{x_1}^{x_2} e^{-ix\xi}u(x)dx + \cdots + \int_{x_{M_0-1}}^{x_{M_0}} e^{-ix\xi}u(x)dx \approx \\
&\approx \int_{x_0}^{x_1} e^{-ix\xi}u(x_1^\star)dx + \int_{x_1}^{x_2} e^{-ix\xi}u(x_2^\star)dx + \cdots + \int_{x_{M_0-1}}^{x_{M_0}} e^{-ix\xi}u(x_{M_0}^\star)dx = \\
&= u(x_1^\star)\int_{x_0}^{x_1} e^{-ix\xi}dx + u(x_2^\star)\int_{x_1}^{x_2} e^{-ix\xi}dx + \cdots + u(x_{M_0}^\star)\int_{x_{M_0-1}}^{x_{M_0}} e^{-ix\xi}dx = \\
&= u(x_1^\star)\frac{e^{-ix_1\xi} - e^{-ix_0\xi}}{-i\xi} + u(x_2^\star)\frac{e^{-ix_2\xi} - e^{-ix_1\xi}}{-i\xi} + \cdots + u(x_{M_0}^\star)\frac{e^{-ix_{M_0}\xi} - e^{-ix_{M_0-1}\xi}}{-i\xi} := \widetilde{u}(\xi). \qquad (3.2)
\end{aligned}
$$

Заметим, что $\widetilde{u}(\xi)$ можно доопределить в точке $\xi = 0$ по непрерывности.

Учитывая аналитичность функции $\phi_q^+(\xi)$ в полуплоскости $\Im\xi > \lambda_-$, мы можем сдвинуть прямую интегрирования в формуле (2.16):

$$
\mathcal{E}_q^+ u(x_k^\star) = (2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho} e^{ix_k^\star\xi}\phi_q^+(\xi)\mathcal{F}u(\xi)d\xi, \quad k = 1, \ldots, M_0, \qquad (3.3)
$$

где $\omega_- < \rho < 0$. Выбор ρ осуществляется так, чтобы сходился интеграл (2.13) (в нашем случае он сходится при любых ρ, но для применения нашей аппроксимации (3.2) $\rho = 0$ не подходит). Подробнее, о выборе параметра ρ в зависимости от вида функции выплат можно прочитать в монографиях [10], [5].

Применяя приближенную формулу (3.2), мы получаем в (3.3) для $k = M_0$

$$
\begin{aligned}
&\mathcal{E}_q^+ u(x_{M_0}^\star) \approx (2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho} e^{ix_{M_0}^\star\xi}\phi_q^+(\xi)\widetilde{u}(\xi)d\xi = \\
&= u(x_1^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0}^\star - x_1)\xi} - e^{i(x_{M_0}^\star - x_0)\xi}}{-i\xi}\phi_q^+(\xi)d\xi + \cdots + \qquad (3.4) \\
&+ u(x_{M_0-1}^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0}^\star - x_{M_0-1})\xi} - e^{i(x_{M_0}^\star - x_{M_0-2})\xi}}{-i\xi}\phi_q^+(\xi)d\xi + \qquad (3.5) \\
&+ u(x_{M_0}^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0}^\star - x_{M_0})} - e^{i(x_{M_0}^\star - x_{M_0-1})\xi}}{-i\xi}\phi_q^+(\xi)d\xi. \qquad (3.6)
\end{aligned}
$$

В силу того что $\xi = 0$ является устранимой точкой и $(x_{M_0}^\star - x_k) > 0$ при $k = 0, \ldots, M_0 - 1$, все интегралы от (3.4) до (3.5) обращаются в 0 по лемме Жордана. Разбивая последний интеграл (3.6) на два слагаемых и применяя в последнем теорему о вычетах, мы получаем

$$
\mathcal{E}_q^+ u(x_{M_0}^\star) = u(x_{M_0}^\star) - u(x_{M_0}^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0}^\star - x_{M_0})\xi}}{i\xi}\phi_q^+(\xi)d\xi, \rho = \lambda_-. \qquad (3.7)
$$

Пусть теперь $k = M_0 - 1$, тогда

$$
\begin{aligned}
&\mathcal{E}_q^+ u(x_{M_0-1}^\star) \approx (2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho} e^{ix_{M_0-1}^\star\xi}\phi_q^+(\xi)\widetilde{u}(\xi)d\xi = \\
&= u(x_1^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0-1}^\star - x_1)\xi} - e^{i(x_{M_0-1}^\star - x_0)\xi}}{-i\xi}\phi_q^+(\xi)d\xi + \cdots + \qquad (3.8) \\
&+ u(x_{M_0-2}^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0-1}^\star - x_{M_0-2})\xi} - e^{i(x_{M_0-1}^\star - x_{M_0-3})\xi}}{-i\xi}\phi_q^+(\xi)d\xi + \qquad (3.9) \\
&+ u(x_{M_0-1}^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0-1}^\star - x_{M_0-1})\xi} - e^{i(x_{M_0-1}^\star - x_{M_0-2})\xi}}{-i\xi}\phi_q^+(\xi)d\xi + \qquad (3.10) \\
&+ u(x_{M_0}^\star)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x_{M_0-1}^\star - x_{M_0})} - e^{i(x_{M_0-1}^\star - x_{M_0-1})\xi}}{-i\xi}\phi_q^+(\xi)d\xi. \qquad (3.11)
\end{aligned}
$$

Аналогично $(x_{M_0-1}^\star - x_k) > 0$ при $k = 0, \ldots, M_0 - 2$, и все интегралы от (3.8) до (3.9) обращаются в 0 по лемме Жордана. Разбивая предпоследний интеграл (3.10) на два слагаемых и применяя в последнем теорему о вычетах,

после перегруппировки с последним интегралом мы получаем

$$\begin{aligned}\mathcal{E}_q^+u(x^\star_{M_0-1}) &= u(x^\star_{M_0-1})+(u(x^\star_{M_0})-u(x^\star_{M_0-1}))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_{M_0-1}-x_{M_0-1})\xi}}{i\xi}\phi_q^+(\xi)d\xi-\\ &- u(x^\star_{M_0})(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_{M_0-1}-x_{M_0})\xi}}{i\xi}\phi_q^+(\xi)d\xi,\quad \rho=\lambda_-. \end{aligned}\tag{3.12}$$

Продолжая представленную процедуру для $k=M_0-2,M_0-3,...,1$, мы приходим к общей формуле

$$\begin{aligned}\mathcal{E}_q^+u(x^\star_k) &= u(x^\star_k)+(u(x^\star_{k+1})-u(x^\star_k))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_k-x_k)\xi}}{i\xi}\phi_q^+(\xi)d\xi+\cdots\\ \cdots &+ (u(x^\star_{M_0})-u(x^\star_{M_0-1}))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_k-x_{M_0-1})\xi}}{i\xi}\phi_q^+(\xi)d\xi-\\ &- u(x^\star_{M_0})(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_k-x_{M_0})\xi}}{i\xi}\phi_q^+(\xi)d\xi,\quad \rho=\lambda_-. \end{aligned}\tag{3.13}$$

По аналогии с предыдущими вычислениями мы можем аппроксимировать действие оператора $\mathcal{E}_q^-$ на функцию u. Теперь благодаря аналитичности функции $\phi_q^-(\xi)$ в полуплоскости $\Im\xi<\lambda_+$, мы сдвинем прямую интегрирования в формуле (2.17) вверх:

$$\mathcal{E}_q^-u(x^\star_k)=(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}e^{ix^\star_k\xi}\phi_q^-(\xi)\mathcal{F}u(\xi)d\xi,\quad k=1,\dots,M_0,\tag{3.14}$$

где $0<\rho<\omega_+$.

В результате проведенных выкладок для $k=1,...,M_0-1,M_0$, мы получаем общую формулу

$$\begin{aligned}\mathcal{E}_q^-u(x^\star_k) &= u(x^\star_k)+(u(x^\star_{k-1})-u(x^\star_k))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_k-x_{k-1})\xi}}{-i\xi}\phi_q^-(\xi)d\xi+\cdots\\ \cdots &+ (u(x^\star_1)-u(x^\star_2))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_k-x_1)\xi}}{-i\xi}\phi_q^-(\xi)d\xi-\\ &- u(x^\star_1)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(x^\star_k-x_0)\xi}}{-i\xi}\phi_q^-(\xi)d\xi,\quad \rho=\lambda_+. \end{aligned}\tag{3.15}$$

Если сетка $\{x_n\}_0^{M_0}$, определенная (3.1), является равномерной, т.е.

$$x_k=x_0+2hk,\qquad k=0,1\dots,M_0,$$

где $h=(u-d)/2M_0$, то определив

$$x^\star_k=x_k+h,\qquad k=1,\dots,M_0,\tag{3.16}$$

мы можем упростить формулы (3.13) и (3.15) следующим образом. Для каждого $k=1,\dots,M_0$

$$\begin{aligned}\mathcal{E}_q^+u(x^\star_k) &= u(x^\star_k)+(u(x^\star_{k+1})-u(x^\star_k))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{-ih\xi}}{i\xi}\phi_q^+(\xi)d\xi+\\ &+ (u(x^\star_{k+2})-u(x^\star_{k+1}))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{-i3h\xi}}{i\xi}\phi_q^+(\xi)d\xi+\cdots\\ \cdots &+ (u(x^\star_{M_0})-u(x^\star_{M_0-1}))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{-i(2M_0+3-2k)h\xi}}{i\xi}\phi_q^+(\xi)d\xi-\\ &- u(x^\star_{M_0})(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{-i(2M_0-1-2k)h\xi}}{i\xi}\phi_q^+(\xi)d\xi,\quad \rho=\lambda_-, \end{aligned}\tag{3.17}$$

и

$$\mathcal{E}_q^-u(x^\star_k) = u(x^\star_k)+(u(x^\star_{k-1})-u(x^\star_k))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{ih\xi}}{-i\xi}\phi_q^-(\xi)d\xi=$$

$$= (u(x^\star_{k-2}) - u(x^\star_{k-1}))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i3h\xi}}{-i\xi}\phi_q^-(\xi)d\xi + \cdots$$

$$\cdots + (u(x^\star_1) - u(x^\star_2))(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(2k-3)h\xi}}{-i\xi}\phi_q^-(\xi)d\xi -$$

$$- u(x^\star_1)(2\pi)^{-1}\int_{-\infty+i\rho}^{+\infty+i\rho}\frac{e^{i(2k-1)h\xi}}{-i\xi}\phi_q^-(\xi)d\xi, \quad \rho = \lambda_+. \tag{3.18}$$

Таким образом, для эффективной реализации формулы (3.17) и (3.18) нам для фиксированного $h > 0$ необходимо определить, соответственно, значения интегралов вида

$$(2\pi)^{-1}\int_{-\infty+i\lambda_-}^{+\infty+i\lambda_-}\frac{e^{-i(2k-1)h\xi}}{i\xi}\phi_q^+(\xi)d\xi, \quad k = 1,\ldots,M_0, \tag{3.19}$$

и

$$(2\pi)^{-1}\int_{-\infty+i\lambda_+}^{+\infty+i\lambda_+}\frac{e^{i(2k-1)h\xi}}{-i\xi}\phi_q^-(\xi)d\xi, \quad k = 1,\ldots,M_0. \tag{3.20}$$

С ростом k значения интегралов (3.19) и (3.20) стремятся к нулю, поэтому при выбранной точности вычислений в формулах (3.17) и (3.18) может понадобиться только часть слагаемых. Таким образом, в упрощенном виде формулы (3.19) и (3.20) с m слагаемыми можно переписать следующим образом:

$$\mathcal{E}_q^+ u(x^\star_k) = u(x^\star_k) + \frac{1}{2\pi}\sum_{j=1}^{m}(u(x^\star_{k+j}) - u(x^\star_{k+j-1}))\int_{-\infty+i\lambda_-}^{+\infty+i\lambda_-}\frac{e^{-i(2j-1)h\xi}}{i\xi}\phi_q^+(\xi)d\xi, \tag{3.21}$$

$$\mathcal{E}_q^- u(x^\star_k) = u(x^\star_k) + \frac{1}{2\pi}\sum_{j=1}^{m}(u(x^\star_{k-j}) - u(x^\star_{k-j+1}))\int_{-\infty+i\lambda_+}^{+\infty+i\lambda_+}\frac{e^{i(2j-1)h\xi}}{-i\xi}\phi_q^-(\xi)d\xi. \tag{3.22}$$

Назовем эту процедуру вычисления действия операторов $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$ методом суммирования интегралов (кратко – МСИ).

Заметим, что основным недостатками непосредственной реализации формулы (2.16) с помощью быстрого преобразования Фурье являются следующие условия применения этого алгоритма:

- сетка $\{x_k\}$ должна существенно выходить за пределы интервала (d, u) из-за периодичности дискретного преобразования Фурье;
- количество точек M в сетке $\{x_k\}$ должно совпадать с количеством точек при дискретизации интеграла Фурье (2.16) по ξ;
- количество точек M должно быть степенью 2;
- сетки по x и ξ обязательно должны быть равномерными;
- шаг дискретизации по x (Δx) и шаг дискретизации по ξ ($\Delta\xi$) должны удовлетворять соотношению

$$\frac{2\pi}{M_0} = \Delta x\Delta\xi,$$

которое при уменьшении Δx приводит к увеличению $\Delta\xi$ при фиксированном количестве точек и наоборот.

В предлагаемом нами подходе набор интегралов (3.19) и (3.20) можно однократно вычислить с необходимой точностью и применять при дальнейшей реализации численного метода.

Если сетка неравномерна, то нам нужно будет вычислять интегралы вида

$$(2\pi)^{-1}\int_{-\infty+i\lambda_-}^{+\infty+i\lambda_-}\frac{e^{-ix\xi}}{i\xi}\phi_q^+(\xi)d\xi, \quad k = 1,\ldots,M_0, \tag{3.23}$$

и

$$(2\pi)^{-1}\int_{-\infty+i\lambda_+}^{+\infty+i\lambda_+}\frac{e^{ix\xi}}{-i\xi}\phi_q^-(\xi)d\xi, \quad k = 1,\ldots,M_0, \tag{3.24}$$

для определенного набора значений параметра $x > 0$, фиксируемого нашей сеткой.

Отметим, что явные формулы для факторов $\phi_q^+(\xi)$ и $\phi_q^-(\xi)$ известны только для ограниченного количества моделей Леви, но вместе с тем существуют приближенные формулы (см., например, [17]). В недавних работах [18], [20], [21], разработан подход, в рамках которого вычисление цен барьерных опционов сводилось к последовательному применению операторов вида $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$, построенных для спектрально односторонних процессов Леви. В следующем разделе мы рассмотрим, как в этом случае формула (3.23) ((3.24)) может быть численно реализована особенно эффективно.

3.2. Явные формулы реализации операторов $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$ для спектрально односторонних умеренно устойчивых процессов Леви

Рассмотрим быстрое вычисление сверток для операторов $\mathcal{E}_q^+$, порождаемых чисто негауссовыми спектрально положительными процессами Леви X_t^+ с ограниченной вариацией скачков (см. [23]). Будем предполагать, что снос у процесса X_t^+ неотрицательный. В качестве базового примера X_t^+ рассмотрим умеренно устойчивый процесс Леви (TSL), как это делалось при вычислении барьерных опционов в [18], [20]. В этом случае характеристическая экспонента процесса будет иметь вид

$$\psi_+(\xi) = -i\mu\xi + c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-} - (-\lambda_- - i\xi)^{\nu_-}], \tag{3.25}$$

где $\nu_- \in (0,1)$, $c_- > 0$, $\mu \geqslant 0$ и $\lambda_- < -1$. При заданных ограничениях на параметр ν_- процесс Леви X_t^+ является субординатором. Пусть параметр $q > 0$. Соответственно, характеристическая функция $\phi_q^+(\xi) = q(q + \psi_+(\xi))^{-1}$ имеет аналитическое продолжение в верхнюю полуплоскость $\Im\xi > \lambda_-$ комплексной плоскости $\mathbf{C}$. Получим формулу для вычисления интегралов вида (3.23) в этом специальном случае. Обозначим через

$$I^+(q,\mu,c_-,\lambda_-,\nu_-,x) = \frac{1}{2\pi i}\int_{-\infty+i\lambda_-}^{+\infty+i\lambda_-} \frac{qe^{ix\xi}}{\xi(q+\psi_+(\xi))}d\xi. \tag{3.26}$$

Очевидно, что при $x > 0$ интеграл (3.26) обращается в 1 по лемме Жордана.

Отметим, что согласно лемме 2.1 уравнение

$$q + \psi_+(\xi) = 0 \tag{3.27}$$

может иметь максимум один чисто мнимый корень в нижней полуплоскости комплексной плоскости $\mathbf{C}$, при этом корень существует, когда

$$q + \psi_+(i(\lambda_- + 0)) < 0. \tag{3.28}$$

Для характеристической экспоненты $\psi_+(\xi)$ вида (3.25) условие (3.28) приобретает вид

$$q + \mu\lambda_- + c_-\Gamma(-\nu_-)(-\lambda_-)^{\nu_-} < 0. \tag{3.29}$$

Напомним, что $\Gamma(-\nu) < 0$ при $\nu \in (0,1)$. Поскольку обычно при реализации численных методов параметр q достаточно велик (q можно интерпретировать как $1/\Delta t$, где Δt – шаг дискретизации по времени), то будем считать, что

$$q + \mu\lambda_- + c_-\Gamma(-\nu_-)(-\lambda_-)^{\nu_-} >> 0. \tag{3.30}$$

Учитывая, что функция $\phi_q^+(\xi)$ имеет степенное убывание на бесконечности и аналитически продолжается в нижнюю полуплоскость $\mathbf{C}$ с разрезом $(-i\infty, i\lambda_-]$, при $x < 0$ мы можем трансформировать контур интегрирования и свести вычисление (3.26) к интегрированию вдоль разреза. Подсчитаем пределы подынтегральной функции на левом и правом краях разреза $[i\lambda_-, +i\infty)$, и перепишем формулу (3.26) для $x < 0$ в виде

$$I^+(q,\mu,c_-,\lambda_-,\nu_-,x) = \frac{1}{2\pi}\int_{-i\infty}^{i\lambda_-} \frac{qe^{ix\xi}}{i\xi(q+\psi_+(\xi-0))}d\xi - \frac{1}{2\pi}\int_{-i\infty}^{i\lambda_-} \frac{qe^{ix\xi}}{i\xi(q+\psi_+(\xi+0))}d\xi. \tag{3.31}$$

Сделаем замену $\xi = iz$, тогда мы получаем

$$I^+(q,\mu,c_-,\lambda_-,\nu_-,x) = \frac{1}{2\pi i}\int_{-\infty}^{\lambda_-} \frac{qe^{-xz}}{(-z)(q+\psi_+(iz+0))}dz - \frac{1}{2\pi i}\int_{-\infty}^{\lambda_-} \frac{qe^{-xz}}{(-z)(q+\psi_+(iz-0))}dz. \tag{3.32}$$

Имеем для $z < \lambda_-$

$$\frac{qe^{-xz}}{q+\psi_+(iz-0)} = \lim_{\epsilon\to 0}\frac{qe^{-xz}}{q - i\mu(iz-\epsilon) + c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-} - (-\lambda_- - i(iz-\epsilon))^{\nu_-}]} =$$

$$
\begin{aligned}
&= \lim_{\epsilon\to 0}\frac{qe^{-xz}}{q+\mu z+c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-}-(-\lambda_-+z+i\epsilon)^{\nu_-}]} = \\
&= \lim_{\epsilon\to 0}\frac{qe^{-xz}}{q+\mu z+c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-}-\exp\left(\nu_-\ln\left(-\lambda_-+z+i\epsilon\right)\right)]} = \\
&= \frac{qe^{-xz}}{q+\mu z+c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-}-\exp\left(i\pi\nu_-+\ln\left(\lambda_--z\right)\right)]} = \\
&= \frac{qe^{-xz}}{q+\mu z+c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-}-e^{i\pi\nu_-}(\lambda_--z)^{\nu_-}]}.
\end{aligned}
\tag{3.33}
$$

Напомним, что главное значение $\ln w$ многозначной функции $\operatorname{Ln} w$ определяется по формуле

$$\ln w=\ln|w|+i\arg w,\quad w\in\mathbf{C},$$

$\arg w$ – главное значение аргумента комплексного числа w.

Аналогично получаем, что при $z<\lambda_-$

$$
\begin{aligned}
\frac{qe^{-xz}}{q+\psi_+(iz+0)} &= \lim_{\epsilon\to 0}\frac{qe^{-xz}}{q-i\mu(iz+\epsilon)+c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-}-(-\lambda_--i(iz+\epsilon))^{\nu_-}]} = \\
&= \frac{qe^{-xz}}{q+\mu z+c_-\Gamma(-\nu_-)[(-\lambda_-)^{\nu_-}-e^{-i\pi\nu_-}(\lambda_--z)^{\nu_-}]}.
\end{aligned}
\tag{3.34}
$$

Принимая во внимание (3.33) и (3.34), мы преобразуем (3.32) к следующему виду:

$$
\begin{aligned}
I^+(q,\mu,c_-,\lambda_-,\nu_-,x) &= \frac{1}{2\pi i}\int_{-\infty}^{\lambda_-}\frac{qe^{-xz}\left(\psi_+(iz-0)-\psi_+(iz+0)\right)}{(-z)(q+\psi_+(iz+0))(q+\psi_+(iz-0))}dz = \\
&= \frac{1}{2\pi i}\int_{-\infty}^{\lambda_-}\frac{c_-\Gamma(-\nu_-)qe^{-xz}(\lambda_--z)^{\nu_-}\left(e^{-i\pi\nu_-}-e^{i\pi\nu_-}\right)}{(-z)(\Psi^+(\lambda_--z,\mu,c_1^-,c_2^-,\nu_-)+\Psi_0^+(\lambda_--z,c_3^-,\nu_-))}dz = \\
&= \frac{1}{\pi}\int_{-\infty}^{\lambda_-}\frac{c_3^-qe^{-xz}(\lambda_--z)^{\nu_-}}{(-z)(\Psi^+(\lambda_--z,\mu,c_1^-,c_2^-,\nu_-)+\Psi_0^+(\lambda_--z,c_3^-,\nu_-))}dz,
\end{aligned}
$$

где

$$
\begin{aligned}
&\Psi^+(w,\mu,c_1^-,c_2^-,\nu_-)=\left(c_1^--\mu w+c_2^-w^{\nu_-}\right)^2,\\
&c_1^-:=c_1^-(q,\mu,\lambda_-,c_-,\nu_-)=q+\mu\lambda_-+c_-\Gamma(-\nu_-)(-\lambda_-)^{\nu_-},\\
&c_2^-:=c_2^-(c_-,\nu_-)=-c_-\Gamma(-\nu_-)\cos(\pi\nu_-),\\
&\Psi_0^+(w,c_3^-,\nu_-)={c_3^-}^2w^{2\nu_-},\\
&c_3^-:=c_3^-(c_-,\nu_-)=-c_-\Gamma(-\nu_-)\sin(\pi\nu_-).
\end{aligned}
$$

Сделаем замену $w=\lambda_--z$, тогда интеграл (3.32) запишется в удобном для расчетов виде

$$
\begin{aligned}
I^+(q,\mu,c_-,\lambda_-,\nu_-,x) &= \frac{e^{-\lambda_-x}}{\pi}\int_0^{+\infty}\frac{c_3^-qe^{xw}w^{\nu_-}}{(w-\lambda_-)\left(\Psi^+(w,\mu,c_1^-,c_2^-,\nu_-)+\Psi_0^+(w,c_3^-,\nu_-)\right)}dw = \\
&= \frac{c_3^-qe^{-\lambda_-x}}{\pi}\int_0^{+\infty}\frac{e^{xw}w^{\nu_-}}{(w-\lambda_-)\left(\left(c_1^--\mu w+c_2^-w^{\nu_-}\right)^2+{c_3^-}^2w^{2\nu_-}\right)}dw.
\end{aligned}
$$

Теперь рассмотрим случай умеренно устойчивых процессов Леви X_t^- со спектрально негативными скачками ограниченной вариации и неположительным сносом:

$$\psi_-(\xi)=-i\mu\xi+c_+\Gamma(-\nu_+)[{\lambda_+}^{\nu_+}-(\lambda_++i\xi)^{\nu_+}],\tag{3.35}$$

где $\nu_+\in(0,1)$, $c_+>0$, $\mu\leqslant 0$ и $\lambda_+>0$. При заданных ограничениях на параметр ν_+ процесс Леви $-X_t^-$ является субординатором. Соответственно характеристическая функция $\phi_q^-(\xi)=q(q+\psi_-(\xi))^{-1}$ имеет аналитическое продолжение в нижнюю полуплоскость комплексной плоскости $\mathbf{C}$.

Интеграл (3.24) в данном случае обозначим через

$$I^-(q,\mu,c_+,\lambda_+,\nu_+,x)=-\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{qe^{ix\xi}}{\xi(q+\psi_-(\xi))}d\xi. \tag{3.36}$$

Очевидно, что при $x<0$ интеграл (3.36) обращается в 0 по лемме Жордана.

Отметим, что согласно лемме 2.1 уравнение

$$q+\psi_-(\xi)=0 \tag{3.37}$$

может иметь максимум один чисто мнимый корень в верхней полуплоскости комплексной плоскости $\mathbf{C}$, при этом корень существует, когда

$$q+\psi_-(i(\lambda_+-0))<0. \tag{3.38}$$

Для характеристической экспоненты $\psi_-(\xi)$ вида (3.35) условие (3.38) приобретает вид

$$q+\mu\lambda_++c_+\Gamma(-\nu_+){\lambda_+}^{\nu_+}<0. \tag{3.39}$$

По аналогии со спектрально положительным случаем будем считать, что

$$q+\mu\lambda_++c_+\Gamma(-\nu_+){\lambda_+}^{\nu_+}>>0. \tag{3.40}$$

Учитывая, что функция $\phi_q^-(\xi)$ имеет степенное убывание на бесконечности и аналитически продолжается в верхнюю полуплоскость $\mathbf{C}$ с разрезом $(i\lambda_+,+i\infty,)$, при $x>0$ мы трансформируем контур интегрирования и сводим вычисление (3.36) к интегрированию вдоль разреза. Как и выше, найдем пределы подынтегральной функции на левом и правом краях разреза $(i\lambda_+,+i\infty)$, и перепишем формулу (3.26) для $x>0$ в виде

$$I^-(q,\mu,c_+,\lambda_+,\nu_+,x)=\frac{1}{2\pi}\int_{i\lambda_+}^{+i\infty}\frac{qe^{ix\xi}}{(-i\xi)(q+\psi_-(\xi+0))}d\xi-\frac{1}{2\pi}\int_{i\lambda_+}^{+i\infty}\frac{qe^{ix\xi}}{(-i\xi)(q+\psi_-(\xi-0))}d\xi. \tag{3.41}$$

Сделаем замену $\xi=iz$, тогда мы получаем

$$I^-(q,\mu,c_+,\lambda_+,x)=\frac{1}{2\pi i}\int_{\lambda_+}^{+\infty}\frac{qe^{-xz}}{z(q+\psi_-(iz-0))}dz-\frac{1}{2\pi i}\int_{\lambda_+}^{+\infty}\frac{qe^{-xz}}{z(q+\psi_-(iz+0))}dz. \tag{3.42}$$

Имеем для $z>\lambda_+$

$$\begin{aligned}\frac{qe^{-xz}}{q+\psi_-(iz-0)} &= \lim_{\epsilon\to0}\frac{qe^{-xz}}{q-i\mu(iz-\epsilon)+c_+\Gamma(-\nu_+)[{\lambda_+}^{\nu_+}-(\lambda_++i(iz-\epsilon))^{\nu_+}]}=\\ &= \lim_{\epsilon\to0}\frac{qe^{-xz}}{q+\mu z+c_+\Gamma(-\nu_+)[{\lambda_+}^{\nu_+}-(\lambda_+-z-i\epsilon)^{\nu_+}]}=\\ &= \lim_{\epsilon\to0}\frac{qe^{-xz}}{q+\mu z+c_+\Gamma(-\nu_+)[({\lambda_+}^{\nu_+}-\exp(\nu_+\ln(\lambda_+-z-i\epsilon))]}=\\ &= \frac{qe^{-xz}}{q+\mu z+c_+\Gamma(-\nu_+)[({\lambda_+}^{\nu_+}-\exp(-i\pi\nu_++\ln(z-\lambda_+))]}=\\ &= \frac{qe^{-xz}}{q+\mu z+c_+\Gamma(-\nu_+)[{\lambda_+}^{\nu_+}-e^{-i\pi\nu_+}(z-\lambda_+)^{\nu_+}]}.\end{aligned} \tag{3.43}$$

Аналогично получаем, что при $z>\lambda_+$

$$\begin{aligned}\frac{qe^{-xz}}{q+\psi_-(iz+0)} &= \lim_{\epsilon\to0}\frac{qe^{-xz}}{q-i\mu(iz+\epsilon)+c_+\Gamma(-\nu_+)[{\lambda_+}^{\nu_+}-(\lambda_++i(iz+\epsilon))^{\nu_+}]}=\\ &= \frac{qe^{-xz}}{q+\mu z+c_+\Gamma(-\nu_+)[{\lambda_+}^{\nu_+}-e^{i\pi\nu_+}(z-\lambda_+)^{\nu_+}]}.\end{aligned} \tag{3.44}$$

Принимая во внимание (3.43) и (3.44), мы преобразуем (3.42) к следующему виду:

$$I^-(q,\mu,c_+,\lambda_+,\nu_+,x)=\frac{1}{2\pi i}\int_{\lambda_+}^{+\infty}\frac{qe^{-xz}\,(\psi_-(iz+0)-\psi_-(iz-0))}{z(q+\psi_-(iz+0))(q+\psi_-(iz-0))}dz=$$

$$= \frac{1}{2\pi i}\int_{\lambda_+}^{+\infty}\frac{c_+\Gamma(-\nu_+)qe^{-xz}(z-\lambda_+)^{\nu_+}\left(e^{-i\pi\nu_+}-e^{i\pi\nu_+}\right)}{z\left(\Psi^-(z-\lambda_+,\mu,c_1^+,c_2^+,\nu_+)+\Psi_0^-(z-\lambda_+,c_3^+,\nu_+)\right)}dz =$$
$$= \frac{1}{\pi}\int_{\lambda_+}^{+\infty}\frac{c_3^+qe^{-xz}(z-\lambda_+)^{\nu_+}}{z\left(\Psi^-(z-\lambda_+,\mu,c_1^+,c_2^+,\nu_+)+\Psi_0^-(z-\lambda_+,c_3^+,\nu_+)\right)}dz,$$

где

$$\begin{aligned}
\Psi^-(w,\mu,c_1^+,c_2^+,\nu_+) &= \left(c_1^+ + \mu w + c_2^+ w^{\nu_+}\right)^2,\\
c_1^+ &:= c_1^+(q,\mu,\lambda_+,c_+,\nu_+) = q + \mu\lambda_+ + c_+\Gamma(-\nu_+){\lambda_+}^{\nu_+},\\
c_2^+ &:= c_2^+(c_+,\nu_+) = -c_+\Gamma(-\nu_+)\cos(\pi\nu_+),\\
\Psi_0^-(w,c_3^+,\nu_+) &= {c_3^+}^2 w^{2\nu_+},\\
c_3^+ &:= c_3^+(c_+,\nu_+) = -c_+\Gamma(-\nu_+)\sin(\pi\nu_+).
\end{aligned}$$

Сделаем замену $w = z - \lambda_+$, тогда интеграл (3.42) запишется в удобном для расчетов виде

$$\begin{aligned}
I^-(q,\mu,c_+,\lambda_+,\nu_+,x) &= \frac{e^{-\lambda_+ x}}{\pi}\int_0^{+\infty}\frac{c_3^+qe^{-xw}w^{\nu_+}}{(w+\lambda_+)\left(\Psi^-(w,\mu,c_1^+,c_2^+,\nu_+)+\Psi_0^-(w,c_3^+,\nu_+)\right)}dw =\\
&= \frac{c_3^+qe^{-\lambda_+ x}}{\pi}\int_0^{+\infty}\frac{e^{-xw}w^{\nu_+}}{(w+\lambda_+)\left(\left(c_1^+ + \mu w + c_2^+ w^{\nu_+}\right)^2 + {c_3^+}^2 w^{2\nu_+}\right)}dw.
\end{aligned}$$

Таким образом, вычисление интегралов I^+ при $x<0$, $\mu\geqslant 0$ и I^- при $x>0$, $\mu\leqslant 0$ можно свести к вычислению одного и того же типа интеграла:

$$I(\mu,q,c,\lambda,\nu,x) = \int_0^{+\infty}\frac{qe^{-xw}w^{\nu}}{(w+\lambda)\left(\left(c_1 - \mu w + c_2 w^{\nu}\right)^2 + c_3 w^{2\nu}\right)}dw, \tag{3.45}$$

где

$$x>0,\quad q>0,\quad c>0,\quad \lambda>0,\quad \mu\geqslant 0,\quad 0<\nu<1,$$

$$\begin{aligned}
c_1 &:= c_1(q,\mu,\lambda,c,\nu) = q + \mu\lambda + c_+\Gamma(-\nu)\lambda^{\nu},\\
c_2 &:= c_2(c,\nu) = -c\Gamma(-\nu)\cos(\pi\nu),\\
c_3 &:= c_3(c,\nu) = \left(c\Gamma(-\nu)\sin(\pi\nu)\right)^2.
\end{aligned}$$

Типовые интегралы типа (3.45) можно заранее обучить вычислять методами машинного обучения. Используя, например, предобученные искусственные нейросети, вычисляющие интегралы (3.45), как вспомогательные элементы, можно строить гибридные численные методы вычисления цен опционов.

3.3. Возможности применения искусственных нейронных сетей для вычисления интегралов $I(\mu,q,c,\lambda,\nu,x)$

Напомним, что по своей сути искусственные нейронные сети (ИНС) представляют собой многопараметрические нелинейные отображения (подробнее, см., например, [7]). Примером ИНС является сеть прямого распространения – последовательность из композиции многомерных линейных преобразований и одномерных непрерывных нелинейных функций активации, которые обычно выполняют роль своего рода классификатора (см. [11]). Набор входных параметров называется входным слоем нейросети, конечный результат – выходным слоем, а промежуточные результаты вычислений нелинейных преобразований – скрытыми слоями.

Современные программные реализации искусственных нейронных сетей эффективно комбинированы с технологиями параллельных вычислений, что делает ИНС полезным инструментом для разработки гибридных численных методов. Согласно теореме Цыбенко (см. [11]), любую непрерывную функцию $f(x_1,\ldots,x_{d_0})$, заданную на компакте $C\subset R^{d_0}$, где $d_0\geqslant 1$, можно с любой степенью точности приблизить функцией вида

$$g(x_1,\ldots,x_{d_0}) = \sum_{j=1}^{M_0}\omega_j\sigma\Big(\sum_{k=1}^{d_0}\alpha_j^k x_k + \beta_j\Big),\quad \omega_j,\alpha_j^k,\beta_j\in\mathbf{R}, \tag{3.46}$$

где $\sigma(x)$ – сигмоидная функция активации (неубывающая непрерывная функция, которая стремится к 0 при $x \to -\infty$ и к 1 при $x \to +\infty$). Одной из наиболее популярных сигмоидных функций активации для ИНС является логистической функция, которая определяется формулой

$$\sigma(x) = \frac{e^x}{1+e^x}.$$

Напомним, что утверждения, обосновывающие возможности нейронных сетей различной архитектуры приближать функции, известны как универсальные теоремы аппроксимации. Наиболее известные результаты представлены в работах [11], [13]. Недавняя статья [22] предлагает вероятностную формулировку для универсальных теорем аппроксимации.

Заметим, что функция $g(x_1, \dots, x_{d_0})$ в (3.46) представляет собой простейшую нейросеть с d_0 нейронами на входном слое, одним нейроном на выходном слое и M_0 нейронами на скрытом слое. На практике для достижения заданной точности мы фиксируем число M_0 в архитектуре нашей ИНС, и с помощью одного из оптимизаторов (например, Adam) специализированной библиотеки (например, PyTorch) находим наилучший набор весов $\omega_j, \alpha_j^k, \beta_j$ для приближения функции $f(x_1, \dots, x_{d_0})$. Если достигнутой точности недостаточно, то следует увеличить число нейронов на скрытом слоем и повторить процесс обучения.

Проиллюстрируем возможности аппроксимации интегралов $I(\mu, q, c, \lambda, \nu, x)$ (3.45) с помощью простейших нейросетей прямого распространения с одним скрытым слоем. Отметим, что интеграл $I(\mu, q, c, \lambda, \nu, x)$ является функцией шести переменных. Как отмечалось выше, с ростом x значение интеграла стремится к 0, поэтому при стандартых значениях остальных параметров вычислять $I(\mu, q, c, \lambda, \nu, x)$ при x выше 0.5 обычно нецелесообразно. Поскольку $\Gamma(-\nu)$ принимает достаточно большие значения, когда ν близко к 0 или 1, то для упрощения вычислений будем рассматривать случай, когда $\nu \in [0.1, 0.9]$. Переформулируем теорему Цыбенко [11, theorem 2] для решения задачи аппроксимации интеграла (3.45) с учетом указанных ограничений на параметры x и ν.

Теорема 3.1. *Пусть σ – логистическая функция, и пусть заданы действительные положительные числа $\mu_{\max}$, $q_{\min} < q_{\max}$, $\lambda_{\min} < \lambda_{\max}$, $c_{\min} < c_{\max}$. Пусть функция $I(\mu, q, c, \lambda, \nu, x)$ определена формулой* (3.45) *с параметрами $\mu \in [0, \mu_{\max}]$, $q \in [q_{\min}, q_{\max}]$, $c \in [c_{\min}, c_{\max}]$, $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, $\nu \in [0.1, 0.9]$, $x \in [0, 0.5]$.*

Для сколь угодно малого $\epsilon > 0$ существует функция вида

$$G(y) = \sum_{j=1}^{M_0} \omega_j \sigma(\alpha_j \cdot y + \beta_j), \quad y \in \mathbf{R}^6, \tag{3.47}$$

где $\alpha \cdot y$ обозначает скалярное произведение векторов α и y, а веса $\alpha_j \in \mathbf{R}^6$, $\omega_j, \beta_j \in \mathbf{R}$ таковы, что

$$|I(y) - G(y))| < \epsilon \tag{3.48}$$

для всех $y \in [0, \mu_{\max}] \times [q_{\min}, q_{\max}] \times [c_{\min}, c_{\max}] \times [\lambda_{\min}, \lambda_{\max}] \times [0.1, 0.9] \times [0, 0.5]$.

С помощью теоремы 3.1 мы можем приблизить интеграл $I(\mu, q, c, \lambda, \nu, x)$ нейросетью (3.47) на замкнутом множестве параметров $(\mu, q, c, \lambda, \nu, x)$. Статистический анализ финансовых рынков (см., например, [8]) показывает, что параметры $\lambda_+, -\lambda_-$ обычно находятся в диапазоне $(5, 250)$, параметр c обычно не превышает 150, а параметр сноса μ по модулю обычно не превосходит 1. На основе этих эмпирических исследований можно определить граничные значения параметров μ, c, λ. Параметр q связан с дискретизацией времени и приблизительно представляет собой обратную величину к шагу по времени Δt (подробнее, см., например, [1], [17], [18]). Поэтому типовые значения q можно рассматривать в диапазоне от 10 до 10 000.

4. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Сравнение метода упрощенной факторизации Винера–Хопфа с другими численными методами, включающими в себя методы приближенной факторизации Винера–Хопфа, Монте-Карло и конечно-разностные схемы, проводилось в работах [18], [20], [21]. Построенная в разд. 3 численная процедура приближения действия операторов $\mathcal{E}^+$ и $\mathcal{E}^-$ не меняет общую итерационную схему (2.23) вычисления цен опционов с барьерами снизу D и сверху U, но дает альтернативный более эффективный подход реализации этих формул, чем традиционно используемый алгоритм Быстрого преобразования Фурье (кратко – БПФ). Учитывая тот факт, что формулу (2.23) нужно вычислять десятки или даже сотни раз, построенный метод МСИ (см. (3.21) и (3.22)) может существенно ускорить вычисления.

Отметим, что алгоритм БПФ для вычисления (3.21) ((3.22)) необходимо применять дважды для аппроксимации прямого и обратного преобразований Фурье. Каждый раз нам потребуется $O(M \log_2 M)$ операций, где

M – число точек расчетной сетки. Заметим, что в силу особенностей алгоритма БПФ область локализации для переменной x должна существенно перекрывать интервал $(0, \ln(U/D))$, где цены опциона отличны от нуля.

Заметим, что первые n интегралов в (3.21) и (3.22) нужно посчитать всего один раз. Когда значения необходимых интегралов (3.26) и (3.36) известны, то для численной реализации формулы (3.21) ((3.22)) потребуется $O(mM_0)$ операций, где M_0 (существенно меньше M) – число точек расчетной сетки, находящихся внутри интервала $(0, \ln(U/D))$, при этом вычислять значения тригонометрических функций, как в алгоритме БПФ, при данном подходе не требуется, как и не требуется отдельно применять индикатор-функцию интервала $(0, \ln(U/D))$, поскольку все вычисления проводятся исключительно для точек данного интервала. Таким образом, метод МСИ может быть в несколько раз быстрее.

Отметим, что наибольшую вычислительную трудность реализация формулы (2.23) вызывает на первом шаге, когда оператор $\mathcal{E}^+$ действует на функцию $v_0(q,x) := G(x)$, имеющую разрыв на левой границе интервала $(0, \ln(U/D))$. На каждом из следующих шагов операторы $\mathcal{E}_\pm$ сглаживают $v_n(q,x)$. Поэтому мы сравним алгоритмы БПФ и МСИ на примере вычисления $\mathcal{E}^+G(x)$, где $G(x)$ – функция выплат по опциону.

В качестве примера выберем процесс KoBoL (CGMY) с характеристической экспонентой

$$\psi(\xi) = -i\mu\xi + c\Gamma(-\nu)[\lambda_+^{\nu} - (\lambda_+ + i\xi)^{\nu} + (-\lambda_-)^{\nu} - (-\lambda_- - i\xi)^{\nu}],$$

где $\nu \in (0,1)$, $c > 0$, $\mu \in \mathbf{R}$ и $\lambda_- < -1 < 0 < \lambda_+$. Данная модель рассматривалась в работах [6], [20] при вычислении двубарьерных опционов.

Как и в [6], [20], мы выберем следующие параметры модели: $\nu = 0.5$, $\lambda_+ = 9$, $\lambda_- = -8$, $c = 1$, $\mu = -0.0423$, и рассмотрим опцион put с барьером снизу $D = 2800$, барьером сверху $U = 4200$ и ценой исполнения $K = 3500$. Время до исполнения опциона рассматривалось равным $T = 0.1$, а количество шагов по времени $N = 80$. Таким образом, параметр $q = N/T = 800$. Функция выплат в этом случае имеет вид $G(x) = \max(K - De^x, 0) \cdot \mathbf{1}_{(0,h)}(x)$, где $h = \ln(U/D)$. В реализации алгоритма (2.23) вычисления барьерных опционов нам необходимо многократно применять операторы свертки $\mathcal{E}_q^+$ и $\mathcal{E}_q^-$ вида (2.21) и (2.22) с соответствующими значениями параметра q. Так как снос μ отрицателен, то характеристическую экспоненту $\psi_+(\xi)$ в формуле для ядра (2.21) оператора $\mathcal{E}_q^+$ определим с помощью (3.25) с параметрами $\nu = 0.5$, $\lambda_- = -8$, $c = 1$, $\mu = 0$.

Вычислим действие оператора $\mathcal{E}^+$ на $G(x)$ для спектрально положительного умеренно устойчивого процесса Леви (2.14) с помощью приближенных формул (3.21), (3.26) и (3.45) методом МСИ и с помощью алгоритма БПФ. В качестве эталона для этих двух методов найдем сначала $\mathcal{E}^+G(x)$ в явном виде. На первом шаге определим преобразование Фурье функции выплат $G(x)$:

$$\begin{aligned}\hat{G}(\xi) &= \int_{-\infty}^{+\infty} e^{-ix\xi} G(x)dx = \int_0^{\ln(K/D)} e^{-ix\xi}(K - De^x)dx = \\ &= K\frac{1 - e^{-i\ln(K/D)\xi}}{i\xi} + \frac{Ke^{-i\ln(K/D)\xi} - D}{i\xi - 1}.\end{aligned}$$

Отметим, что в точках $\xi = 0$ и $\xi = -i$ функцию $\hat{G}(\xi)$ можно продолжить по непрерывности.

Далее при $x \geqslant \ln(K/D)$ получаем, что

$$\mathcal{E}_q^+G(x) = \int_0^{+\infty} G(x+y)p_q^+(y)dy = \int_0^{+\infty} \max(K - De^{x+y}, 0) \cdot \mathbf{1}_{(0,\ln(K/D))}(x+y)p_q^+(y)dy = 0.$$

При $x < \ln(K/D)$ запишем формулу для $\mathcal{E}_q^+G(x)$ в виде (2.16), сдвигая контур интегрирования и применяя теорему о вычетах, получим

$$\begin{aligned}&\mathcal{E}_q^+G(x) = \int_0^{+\infty} G(x+y)p_q^+(y)dy = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{ix\xi}\frac{qe^{ix\xi}}{\xi(q + \psi_+(\xi))}\hat{G}(\xi)d\xi = \\ &= K\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{qe^{ix\xi} - qe^{i(x-\ln(K/D))\xi}}{\xi(q+\psi_+(\xi))}d\xi + \frac{1}{2\pi i}\int_{-\infty}^{+\infty}\frac{Kqe^{i(x-\ln(K/D))\xi} - Dqe^{ix\xi}}{(\xi+i)(q+\psi_+(\xi))}d\xi = \\ &= K - D\frac{qe^x}{(q+\psi_+(-i))} - \frac{K}{2\pi i}\int_{-\infty+i\lambda_-}^{+\infty+i\lambda_-}\frac{qe^{i(x-\ln(K/D))\xi}}{\xi(q+\psi_+(\xi))}d\xi + \\ &+ \frac{K}{2\pi i}\int_{-\infty+i\lambda_-}^{+\infty+i\lambda_-}\frac{qe^{i(x-\ln(K/D))\xi}}{(\xi+i)(q+\psi_+(\xi))}d\xi.\end{aligned}$$

Два последних интеграла сведем к интегралам вида (3.45) с помощью техники, описанной в п. 3.2, и найдем их численно. Результаты вычисления $\mathcal{E}^+G(x)$ в явном виде с помощью методов МСИ и БПФ представлены

Таблица 1. Приближенные значения $\mathcal{E}^+G(x)$: метод МСИ против алгоритма БПФ

Метод	Явная	МСИ		МСИ		БПФ	
	формула	$h = 0.002$, $m = 5$		$h = 0.002$, $m = 10$		$h = 0.002$, $M = 2^{10}$	
$x = \ln(S/D)$	значение	значение	ошибка	значение	ошибка	значение	ошибка
0.002	692.322	693.475	0.17%	693.167	0.12%	686.046	–0.91%
0.022	635.639	636.779	0.18%	636.465	0.13%	634.518	–0.18%
0.042	577.819	578.938	0.19%	578.618	0.14%	577.103	–0.12%
0.062	518.842	519.929	0.21%	519.602	0.15%	518.250	–0.11%
0.082	458.686	459.727	0.23%	459.394	0.15%	458.136	–0.12%
0.102	397.332	398.310	0.25%	397.970	0.16%	396.785	–0.14%
0.122	334.760	335.652	0.27%	335.305	0.16%	334.193	–0.17%
0.142	270.956	271.727	0.28%	271.373	0.15%	270.353	–0.22%
0.162	205.908	206.512	0.29%	206.151	0.12%	205.258	–0.32%
0.182	139.615	139.979	0.26%	139.611	–0.00%	138.920	–0.50%
0.202	72.115	72.102	-0.02%	71.726	–0.54%	71.413	–0.97%
0.222	3.705	3.728	0.62%	3.728	0.62%	3.461	–6.59%

Параметры характеристической экспоненты ψ_+ в (3.25): $\nu = 0.5$, $\lambda_- = -8$, $c = 1$, $\mu = 0$.
Параметры опциона: $D = 2800$, $U = 4200$, $K = 3500$, $T = 0.1$, $x = \ln(S/D)$.
Параметры алгоритма БПФ: h – шаг расчетной сетки по x, M – количество точек расчетной сетки по x.
Параметры алгоритма МСИ: h – шаг расчетной сетки по x, m – количество слагаемых в формуле (3.21).

в табл. 1. Ошибки в табл. 1 показывают на сколько процентов отличаются значения, полученные методами МСИ и БПФ, от эталонных значений. Как видно из таблицы, даже с пятью слагаемыми метод МСИ дает значения близкие по точности с БПФ, а при $m = 10$ демонстрирует лучшую точность. При этом метод МСИ слегка переоценивает эталонные значения, в то время как БПФ недооценивает. Наиболее проблемная область находится около правого барьера, где значение оператора подходит к нулю.

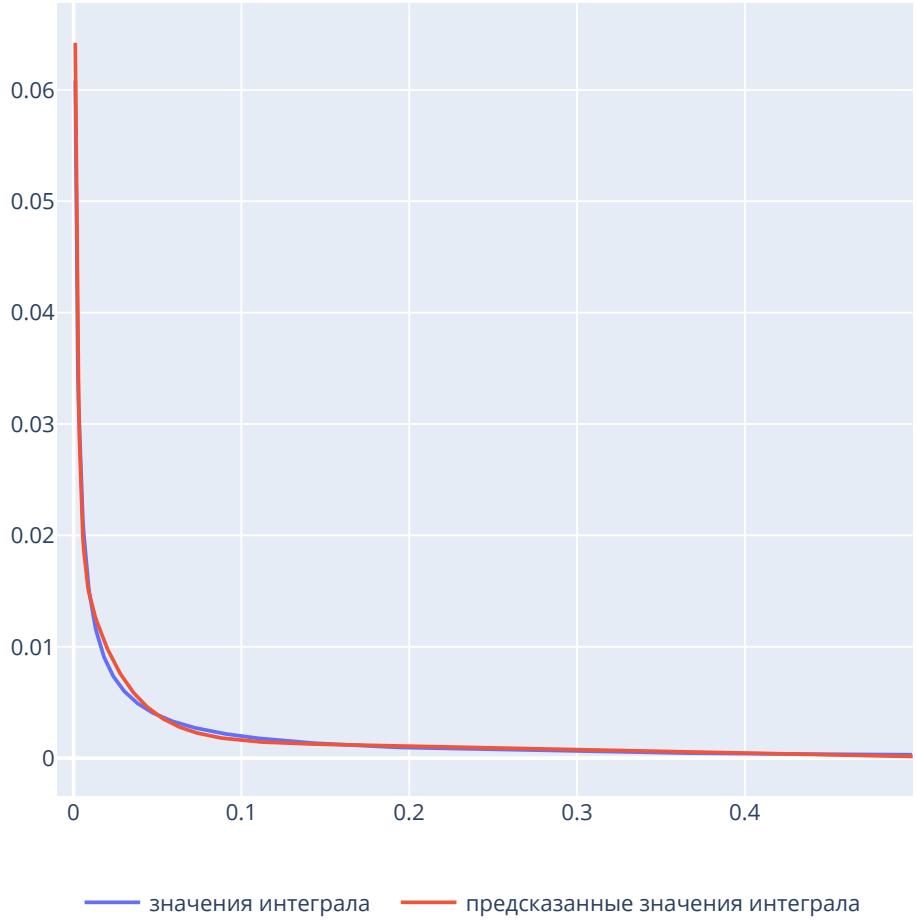
Заметим, что хранение интегралов $I(\mu, q, c, \lambda, \nu, x)$ в виде многомерной таблицы будет занимать значительные объемы памяти. Как правило, методы факторизации Винера–Хопфа требуют секунды или даже доли секунды на вычисление цены опциона. Таким образом, загрузка многомерной таблицы в память, поиск с последующей интерполяцией нужных интегралов будет представлять отдельную задачу, которая может привести к временным затратам, превышающим время работы самого численного метода.

Как отмечалось в п. 3.3 (см. теорему 3.1), теоретически можно обучить искусственную нейросеть, которая будет с необходимой точностью аппроксимировать интегралы $I(\mu, q, c, \lambda, \nu, x)$ вида (3.45). В этом случае можно будет проводить расчеты для вычисления интегральных слагаемых с помощью нейросети. Вместе с тем, архитектура такой нейросети будет включать в себя значительное число нейронов на скрытом слое, что может сделать вычислительную сложность такой аппроксимации интегралов сравнимой с непосредственным численным интегрированием. С другой стороны, нейросеть Цыбенко легко распараллеливается на графических процессорах.

Как правило, при вычислении цен опционов подбор шага h по x и количество шагов N подбирается адаптивно, путем последовательного уменьшения h при фиксированном N, а затем увеличением N и повторением подбора h. В этом случае для конкретной модели с фиксированными параметрами μ, c, λ_+, λ_- и ν и параметром алгоритма q может быть целесообразно обучить нейросеть с одним входным параметром x. Проиллюстрируем на примере, как в этом случае нейросеть будет аппроксимировать интеграл $I(\mu, q, c, \lambda, \nu, x)$.

Поскольку параметр сноса $\mu < 0$, для вычисления интегралов вида $I^+(q, \mu, c_-, \lambda_-, \nu_-, x)$ в (3.26) нам нужно выбрать следующие параметры в формуле (3.45): $\mu = 0$, $q = 800$, $c = 1$, $\lambda = -\lambda_- = 8$, $\nu = 0.5$.

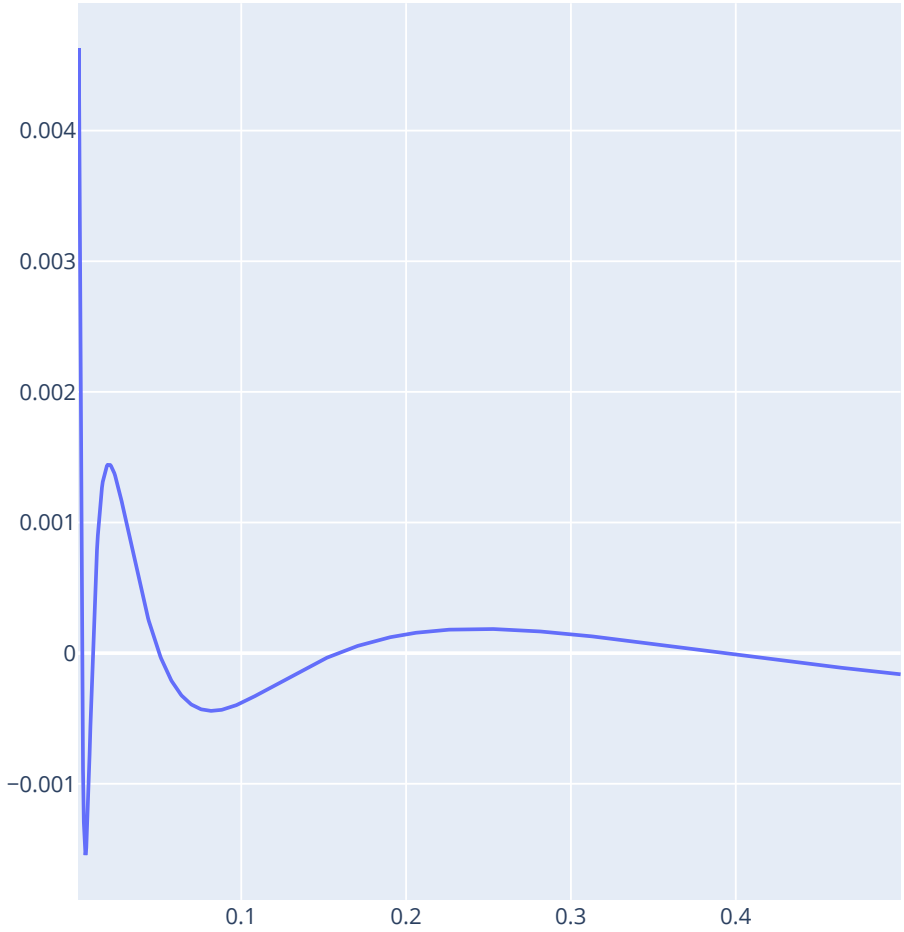
Для аппроксимации нашего интеграла мы выбрали нейросеть вида (3.46) с одним входным параметром $x_1 = x$ (т.е. $d_0 = 1$) и $N = 20$ нейронами на скрытом слое. Объем обучающей выборки составил 4096 значений интегралов, половина из которых вычислялась для выборки значений x, равномерно распределенных на от-



Фиг. 1. Аппроксимация $I(\mu, q, c, \lambda, \nu, x)$ нейросетью. Параметры характеристической экспоненты ψ_+ в (3.25): $\nu = 0.5$, $\lambda_- = -8$, $c = 1$, $\mu = 0$. Параметры оператора $\mathcal{E}_q^+$: $q = 800$. Ось абсцисс: $x = \ln(S/D)$. Ось ординат: значения интеграла $I(\mu, q, c, \lambda, \nu, x)$.

резке $[0.0005, 0.5]$, а вторая половина для выборки значений x из показательного распределения с параметром интенсивности равным 200. Вторая часть выборки нужна, чтобы учесть резкое изменение значений интеграла в окрестности 0. График значений интеграла можно увидеть на фиг. 1.

Нейросеть была обучена на языке программирования Python с помощью специализированной библиотеки PyTorch для глубоко обучения, которая содержит структуры данных для многомерных тензоров и эффективно реализует математические операции над этими тензорами на графических процессорах. В качестве функции потерь была выбрана функция L1Loss, которая измеряет среднее по обучающей выборке абсолютное отклонение предсказанного нейросетью результата от реального значения. Для минимизации функции потерь использовался оптимизатор Adam (англ. Adaptive Moment Estimation – Адаптивная оценка момента) со скоростью обучения $lr = 0.01$. За 80 000 эпох обучения функция потерь составила 0.00394. Результаты аппроксимации интеграла $I(\mu, q, c, \lambda, \nu, x)$ нейросетью представлены на фиг. 1, а график ошибки аппроксимации – на фиг. 2. Как следует их рисунков, нейросеть вполне хорошо справляется со своей задачей, а возможности применения параллельных вычисления с помощью библиотеки PyTorch делает применение таких нейросетей как элемента численного метода оправданным с вычислительной точки зрения.



Фиг. 2. Ошибка аппроксимации $I(\mu, q, c, \lambda, \nu, x)$ нейросетью. Параметры характеристической экспоненты ψ_+ в (3.25): $\nu = 0.5$, $\lambda_- = -8$, $c = 1$, $\mu = 0$. Параметры оператора $\mathcal{E}_q^+$: $q = 800$. Ось абсцисс: $x = \ln(S/D)$. Ось ординат: ошибка аппроксимации интеграла $I(\mu, q, c, \lambda, \nu, x)$.

ЗАКЛЮЧЕНИЕ

Представленный в работе алгоритм расчета интегральных операторов свертки с ядром типа плотности вероятности случайных величин, принимающих только положительные или только отрицательные значения, может успешно конкурировать с традиционным методом вычисления сверток – быстрым преобразование Фурье. Более того, наш подход предусматривает возможность встраивания нейросетей для аппроксимации интегралов, входящих в приближенную формулу.

Десятки лет назад для вычисления специальных функций использовались заранее заготовленные таблицы. В наше время для вычисления типовых интегралов можно использовать мощный аппарат искусственных нейросетей, что позволит создавать более эффективные гибридные численные методы, комбинирующие традиционные методы вычислительной математики (в идейной части алгоритма) и методы машинного обучения (в части выполнения повторяющихся рутинных элементов алгоритма).

С точки зрения практического применения, предлагаемый в работе подход естественным образом распараллеливается на базе nVidia CUDA API для ускоренного обучения нейросетей и подсчета приближенных значений интегралов на основе обученных нейросетей.

СПИСОК ЛИТЕРАТУРЫ

1. *Кудрявцев О.Е.* Эффективный численный метод решения специального класса интегро-дифференциальных уравнений, связанных с моделями Леви // Матем. моделирование. 2011. Т. 23. № 5. С. 95–104.

2. *Кудрявцев О.Е.* Приближенная факторизация Винера–Хопфа и метод Монте-Карло для процессов Леви // Теория вероятностей и ее применения. 2019. Т. 64. № 2. С. 228–257.

3. *Кудрявцев О.Е., Гречко А.С., Мамедов И.Э.* Метод Монте-Карло для вычисления цен опционов типа lookback в моделях Леви // Теория вероятностей и ее применения. 2024. Т. 69. № 2. С. 305–334.

4. *Ширяев А.Н.* Основы стохастической финансовой математики. Факты. Модели. М.: Фазис, 1998. Т. 1. 440 с.

5. *Boyarchenko S.I., Levendorskii S.Z.* Non-Gaussian Merton-Black-Scholes theory. New Jercey, London, Singapore, Hong Kong: World Sci., 2002, 420 p.

6. *Boyarchenko M., Levendorskii S.* Valuation of continuously monitored double barrier options and related securities // Math Financ. (2011) doi: 10.1111/j.1467-9965.2010.00469.x

7. *Dixon M.F., Halperin I., Bilokon P.* Machine Learning in Finance: From Theory to Practice. Germany: Springer Inter. Publ., 2020.

8. *Carr P., Geman H., Madan D.B., Yor M.* The fine structure of asset returns: an empirical investigation // J. Business. 2002. V. 75. № 2. P. 305–332.

9. *Cont R., Voltchkova E.* A finite difference scheme for option pricing in jump diffusion and exponential Levy models // SIAM J. Numer. Analys. 2005. V. 43. № 4. P. 1596–1626.

10. *Cont R., Tankov P.* Financial modelling with jump processes, 2nd Ed., Chapman & Hall/CRC Press, 2008. 606 p.

11. *Cybenko G.* Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals and Systems. 1989. 2. P. 303–314.

12. *Goudenège L., Molent A., Zanette A.* Machine learning for pricing American options in high-dimensional Markovian and non-Markovian models // Quantitative Finance. 2020. V. 20. № 4. P. 573–591.

13. *Hornik K.* Approximation capabilities of multilayer feedforward networks // Neural Networks. 1991. V. 4. № 2. P. 251–257.

14. *Huh J.* Pricing Options with Exponential Levy Neural Network // Expert Systems with Appl. 2019. V. 127. DOI: 10.1016/j.eswa.2019.03.008

15. *Itkin A.* Pricing Derivatives Under Levy Models: book, Birkhauser, 2017. 308 p.

16. *Kirkby J.L.* Robust barrier option pricing by frame projection under exponential Levy dynamics // Appl. Math. Finance. 2017. V. 24. № 4. P. 337–386.

17. *Kudryavtsev O. Levendorskii S.* Fast and accurate pricing of barrier options under Levy processes // Finance Stoch. 2009. V. 13. № 4. P. 531–562.

18. *Kudryavtsev O., Luzhetskaya P.* The Wiener-Hopf Factorization for Pricing Options Made Easy // Engineer. Lett. 2020. V. 28. № 4. P. 1310–1317.

19. *Kudryavtsev O., Zanette A.* Applications of Levy Processes // Math. Res. Developments, NY: Nova Science Publishers, Incorporated, 2021. ISBN: 978-1-53619-525 -5

20. *Kudryavtsev O.* A simple Wiener-Hopf factorization approach for pricing double-barrier options. In: Karapetyants A.N., Pavlov I.V., Shiryaev A.N. (Ed.) Operator Theory and Harmonic Analysis. OTHA 2020. Springer Proceedings in Mathematics & Statistics, Springer, Cham, 2021. V. 358, pp. 273–291.

21. *Kudryavtsev O.* A simplified Wiener–Hopf factorization method for pricing double barrier options under Levy processes // Comput. Manag. Sci. 2024. V. 21. P. 37.

22. *Kudryavtsev O., Danilova N.* Applications of artificial neural networks to simulating Levy processes // J. Math. Sci. 2023. V. 271. № 4. P. 421–433.

23. *Kyprianou A.E.* Introductory Lectures on Fluctuations of Levy Processes with Applications, 2006, Springer, Berlin.

24. *Phelan E., Marazzina D., Fusai G., and Germano G.* Fluctuation identities with continuous monitoring and their application to price barrier options // Europ. J. Operat. Res. 2018. V. 271. № 1. P. 210–223.

25. *Sato K.* Levy processes and infinitely divisible distributions. Cambridge: Cambridge Univ. Press, 1999. 486 p.

FAST CALCULATION OF INTEGRAL CONVOLUTION TYPE OPERATORS IN OPTION ESTIMATION PROBLEMS IN LEVY MODELS

A. S. Grechko[a,], O. E. Kudryavtsev[a,b*]**

[a] *344015 Rostov-on-Don, Yeremenko 58/11, Research and production company "InWise Systems" LLC, Russia*
[b] *344002 Rostov-on-Don, Budennovskiy 20, Rostov branch of the Russian Customs Academy, Russia*
**e-mail: koe@donrta.ru*
***e-mail: alex@itparadigma.ru*

Received: 09.06.2024
Revised: 01.08.2024
Accepted: 23.08.2024

Abstract. An approximate algorithm for calculating integral operators of the convolution type that arise when evaluating barrier options in Levy models by the Wiener–Hopf method is constructed. Additionally, the question of the possibility of applying machine learning methods (artificial neural networks) to the approximation of a special type of integrals, which are a key element in the construction of approximate formulas for the Wiener–Hopf integral operators under consideration, is investigated. The main idea is to decompose the price function into a Fourier series and transform the integration contour for each term of the Fourier series. As a result, we obtain a set of typical integrals that depend on Wiener–Hopf factors, but do not depend on the price function, while the most computationally expensive part of the numerical method is reduced to calculating these integrals. Since they need to be calculated only once, and not at each iteration, as was the case in standard implementations of the Wiener–Hopf method, this will significantly speed up calculations. Moreover, a neural network can be trained to calculate typical integrals. The proposed approach is especially effective for spectrally one-sided Levy processes, for which explicit Wiener–Hopf factorization formulas are known. In this case, we obtain computationally convenient formulas by integrating along the section. The main advantage of including neural networks in a computational scheme is the ability to perform calculations on an uneven grid. Such a hybrid numerical method will be able to successfully compete with classical methods of computing convolutions in similar tasks using the fast Fourier transform. Computational experiments show that neural networks with one hidden layer of 20 neurons are able to effectively cope with the tasks of approximating the auxiliary integrals under consideration.

Keywords: Wiener–Hopf factorization, Levy processes, integral convolution operators, integral transformations, numerical methods, machine learning, computational financial mathematics.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2262–2269

УДК 512.643

НЕВЫРОЖДЕННАЯ МАТРИЦА С ХОРОШО ОБУСЛОВЛЕННЫМ КОКВАДРАТОМ: КАК ПРИВЕСТИ ЕЕ К ДИАГОНАЛЬНОМУ ВИДУ ПОСРЕДСТВОМ КОНГРУЭНЦИИ

© 2024 г. Х. Д. Икрамов[1,*], А. М. Назари[2,]**

[1] *119992 Москва, Ленинские горы, МГУ, ВМК, Россия*
[2] *Faculty of Mathematics, Arak University, Arak, Islamic Republic Iran*
**e-mail: ikramov@cs.msu.su*
***e-mail: a-nazari@araku.ac.ir*

Поступила в редакцию 09.01.2024 г.
Переработанный вариант 18.04.2024 г.
Принята к публикации 31.05.2024 г.

Существуют эффективные программы для приведения диагонализуемой матрицы к диагональному виду посредством преобразования подобия. Юнитоидные матрицы являются аналогами диагонализуемых матриц в теории конгруэнтных преобразований. Однако, если исключить эрмитовы и, более общо, нормальные матрицы, то нет пользующихся общим признанием программ для приведения юнитоидной матрицы к диагональному виду посредством конгруэнций. Предложен алгоритм, способный выполнять эту задачу для специального класса юнитоидных матриц, а именно невырожденных матриц, коквадраты которых хорошо обусловлены в смысле полной проблемы собственных значений. Приведены примеры, иллюстрирующие работу этого алгоритма. Библ. 1.

Ключевые слова: *-конгруэнция, преобразование подобия, юнитоид, коквадрат, канонические углы, матрица с диагональным преобладанием, число обусловленности.

DOI: 10.31857/S0044466924120035, **EDN**: KCWHII

1. ВВЕДЕНИЕ

Квадратную комплексную матрицу A, имеющую базис из собственных векторов, называют диагонализуемой. Это означает, что для матрицы Q, составленной из собственных векторов матрицы A, произведение

$$B = Q^{-1}AQ \tag{1}$$

есть диагональная матрица. Переход от A к B называется преобразованием подобия.

В этой статье мы заинтересованы главным образом в матричных преобразованиях другого рода, а именно

$$A \to Q^*AQ. \tag{2}$$

Они называются преобразованиями эрмитовой конгруэнции (или *-конгруэнции). Мы будем называть их просто *конгруэнциями*.

Квадратную комплексную матрицу A, которая может быть приведена к диагональному виду посредством конгруэнции, называют юнитоидной матрицей или просто *юнитоидом*.

Предположим, что A — диагонализуемая матрица, хорошо обусловленная в смысле полной проблемы собственных значений. В этом случае ее собственные значения и собственные векторы могут быть эффективно и с хорошей точностью вычислены с помощью современного программного обеспечения. Ситуация с задачей приведения юнитоидной матрицы к диагональному виду совершенно иная. Насколько известно авторам, нет популярных программ, которые бы эффективно решали эту задачу. Единственным исключением является класс эрмитовых или, более общо, нормальных матриц.

Наша цель в этой статье — предложить алгоритм, способный решать эту задачу диагонализации для юнитоидных матриц одного специального класса. Описание этого класса дано в разд. 2. Наш алгоритм формулируется в разд. 4. Он основан на некоторой теории, излагаемой в разд. 3. Построение тестовых матриц для этого

алгоритма обсуждается в разд. 5. В заключительном разд. 6 приведены результаты численных экспериментов, выполненных с алгоритмом.

Все сведения о преобразованиях конгруэнции, нужные для понимания данного текста, можно найти в параграфе 4.5 книги [1].

2. ХОРОШО ОБУСЛОВЛЕННЫЕ ЮНИТОИДНЫЕ МАТРИЦЫ

Первое ограничение, которое мы накладываем на юнитоидную матрицу A, состоит в ее невырожденности. Это ограничение не является существенным. В самом деле, вырожденная юнитоидная матрица A может быть заменена невырожденной с помощью следующего простого алгоритма.

1. Найти базис $z_1, \ldots, z_d$ общего ядра матриц A и A^*.

2. Построить невырожденную $n \times n$-матрицу P, в которой векторы $z_1, \ldots, z_d$ являются последними d столбцами.

3. Выполнить конгруэнцию

$$A \to \tilde{A} = P^* A P.$$

Тогда $\tilde{A}$ — это блочно-диагональная матрица вида

$$\tilde{A} = \begin{pmatrix} A_r & 0 \\ 0 & 0 \end{pmatrix},$$

где $r = n - d$ и A_r — невырожденная матрица порядка r. Последующее приведение к диагональному виду выполняется для матрицы A_r.

Невырожденной матрице A можно сопоставить матрицу

$$\mathcal{C}_A = A^{-*} A, \tag{3}$$

называемую *коквадратом* матрицы A. Коквадрат юнитоида A является диагонализуемой матрицей. Кроме того, все собственные значения матрицы $\mathcal{C}_A$ унимодулярны:

$$\text{Spectrum of } \mathcal{C}_A = \{e^{i\phi_1}, e^{i\phi_2}, \ldots, e^{i\phi_n}\}. \tag{4}$$

Наше второе предположение относительно A заключается в том, что эти n чисел должны быть попарно различны. И это ограничение не является существенным; мы устраним его в наших последующих публикациях.

Третье ограничение, и на этот раз действительно важное: матрица $\mathcal{C}_A$ должна быть хорошо обусловлена в смысле полной проблемы собственных значений. Это означает прежде всего, что собственные значения в множестве (4) должны быть хорошо разделены. Как следствие, матрицу S, сформированную из собственных векторов коквадрата $\mathcal{C}_A$, можно выбрать так, чтобы ее число обусловленности

$$\text{cond}_2 S = \|S\|_2 \cdot \|S^{-1}\|_2 \tag{5}$$

имело умеренную величину.

Мы называем невырожденный юнитоид A *хорошо обусловленным*, если его коквадрат $\mathcal{C}_A$ удовлетворяет третьему ограничению.

3. ПРИВЕДЕНИЕ МАТРИЦЫ И ЕЕ КОКВАДРАТА К (БЛОЧНО) ДИАГОНАЛЬНОМУ ВИДУ

Теорема 1. *Пусть A — невырожденная матрица с коквадратом $\mathcal{C}_A$. Если A подвергнута конгруэнции*

$$A \to \tilde{A} = X^* A X, \tag{6}$$

то ее коквадрат претерпевает преобразование подобия, задаваемое той же матрицей X.

Доказательство. Действительно, из (6) следует, что

$$A^{-1} \to \tilde{A}^{-1} = X^{-1} A^{-1} X^{-*} \quad \text{и} \quad \mathcal{C}_A = A^{-*} A \to \mathcal{C}_{\tilde{A}} = X^{-1} \mathcal{C}_A X.$$

Следствие 1. Пусть A — невырожденный юнитоид. Предположим, что матрица X в соотношении (6) выбрана так, чтобы

$$D \equiv \tilde{A} = X^* A X$$

была диагональной матрицей. Тогда:

а) $\mathcal{C}_D$ — диагональная матрица с унимодулярными диагональными элементами $e^{i\phi_1}, e^{i\phi_2}, \dots, e^{i\phi_n}$;

б) эти числа $e^{i\phi_1}, e^{i\phi_2}, \dots, e^{i\phi_n}$ суть собственные значения матрицы $\mathcal{C}_A$;

в) запишем диагональные элементы матрицы D в полярной (или показательной) форме:

$$d_{jj} = \rho_j e^{i\alpha_j}, \qquad j = 1, 2, \dots, n; \tag{7}$$

тогда

$$\phi_j = 2\alpha_j, \qquad j = 1, 2, \dots, n.$$

Эти утверждения являются непосредственными следствиями теоремы 1 и определения коквадрата. Числа $\alpha_1, \alpha_2, \dots, \alpha_n$ называются *каноническими углами* матрицы A. В действительности это канонические углы всякой матрицы, конгруэнтной матрице A.

Ситуация, исследованная в теореме 1, может быть обращена следующим образом.

Теорема 2. *Пусть A — невырожденная матрица с коквадратом $\mathcal{C}_A$. Если $\mathcal{C}_A$ подвергнут подобию*

$$\mathcal{C}_A \to \mathcal{C} = X^{-1}\mathcal{C}_A X, \tag{8}$$

то матрица $\mathcal{C}$ есть коквадрат матрицы $B = X^*AX$.

В самом деле,

$$\mathcal{C}_B = B^{-*}B = X^{-1}A^{-*}X^{-*}X^*AX = X^{-1}(A^{-*}A)X = X^{-1}\mathcal{C}_A X = \mathcal{C}.$$

Теорема 3. *Предположим, что матрица X в соотношении* (8) *выбрана так, чтобы*

$$\mathcal{C} = X^{-1}(A^{-*}A)X = \begin{pmatrix} F & 0 \\ 0 & G \end{pmatrix}, \tag{9}$$

где F и G — квадратные матрицы порядков соответственно k и l ($k + l = n$). Пусть собственные значения блоков F и G удовлетворяют условиям

$$\overline{\lambda_r(F)}\lambda_s(G) \neq 1, \qquad r = 1, 2, \dots, k,\ s = 1, 2, \dots, l. \tag{10}$$

В этом случае конгруэнтное преобразование, задаваемое той же матрицей X, превращает A в прямую сумму

$$B = B_{11} \oplus B_{22},$$

а матрицы F и G в соотношении (8) *являются коквадратами соответственно блоков B_{11} и B_{22}.*

Доказательство. Перепишем соотношение (9) в виде

$$X^*AX = X^*A^*X \begin{pmatrix} F & 0 \\ 0 & G \end{pmatrix}. \tag{11}$$

Разобьем матрицу $B = X^*AX$ на блоки в соответствии с прямой суммой $F \oplus G$:

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}.$$

Из соотношения (11) вытекают следующие равенства для блоков B_{rs}:

$$B_{11} = B_{11}^*F, \qquad B_{22} = B_{22}^*G,$$

$$B_{12} = B_{21}^*G, \qquad B_{21} = B_{12}^*F.$$

Из двух последних равенств выводим

$$B_{12} - F^*B_{12}G = 0, \quad B_{21} - G^*B_{21}F = 0.$$

Тем самым блоки B_{12} и B_{21} являются решениями однородных матричных уравнений Стейна. Известно, что эти уравнения имеют только тривиальные решения $B_{12} = 0$ и $B_{21} = 0$, если собственные значения блоков F и G удовлетворяют условиям (10). Таким образом, мы имеем

$$B_{12} = 0,\ B_{21} = 0,\ F = B_{11}^{-*}B_{11},\ G = B_{22}^{-*}B_{22},$$

что доказывает теорему.

Следствие 2. Пусть A — невырожденная матрица. Предположим, что ее коквадрат $\mathcal{C}_A$ имеет унимодулярный спектр (4), где числа $\phi_1, \phi_2, \ldots, \phi_n$ принадлежат промежутку $[0, \pi)$ и попарно различны. Тогда:

а) A есть юнитоид;

б) матрица X в соотношении (6) может быть выбрана так, чтобы

$$D = X^* A X$$

была диагональной матрицей. В действительности, X — та же матрица, которая диагонализует коквадрат $\mathcal{C}_A$ посредством подобия;

в) числа

$$\alpha_j = \phi_j/2, \qquad j = 1, 2, \ldots, n,$$

являются каноническими углами матрицы A;

г) определим матрицу (см. (7))

$$R = \operatorname{diag}\left(\sqrt{\rho_1}, \sqrt{\rho_2}, \ldots, \sqrt{\rho_n}\right).$$

Тогда дополнительная конгруэнция

$$D \to R^{-1} D R^{-1}$$

превращает D в диагональную матрицу

$$\Sigma = \operatorname{diag}\left(e^{i\alpha_1}, e^{i\alpha_2}, \ldots, e^{i\alpha_n}\right).$$

Опять-таки, эти утверждения следуют непосредственно из теоремы 3. Матрица Σ называется *канонической формой* матрицы A.

4. АЛГОРИТМ

Пусть A — невырожденный юнитоид, коквадрат которого удовлетворяет условиям, изложенным в разд. 2. А именно, матрица $\mathcal{C}_A$ имеет простой унимодулярный спектр и хорошо обусловлена в смысле полной проблемы собственных значений. Выполним следующие действия.

1. Вычислить матрицу P, приводящую $\mathcal{C}_A$ к диагональному виду Λ посредством подобия:

$$P^{-1} \mathcal{C}_A P = \Lambda. \tag{12}$$

Приближение к P хорошего качества может быть получено применением к $\mathcal{C}_A$ Matlab процедуры eig или Maple процедуры Eigenvectors.

2. Выполнить конгруэнцию

$$A \to D = P^* A P.$$

Теоретически D должна быть диагональной матрицей, такой, что аргументы ее диагональных элементов являются каноническими углами матрицы A. Однако из-за ошибок округления мы можем рассчитывать лишь на получение хорошего приближения к D.

В следующих двух разделах мы обсудим, насколько хороши приближения, вычисляемые нашим алгоритмом.

5. ТЕСТОВЫЕ МАТРИЦЫ

Мы предварим построение тестовой матрицы A построением ее коквадрата. Для заданного порядка n выберем в промежутке $[0, \pi)$ псевдослучайные числа $\phi_1, \phi_2, \ldots, \phi_n$. Они определяют Λ, диагональную матрицу собственных значений коквадрата $\mathcal{C}_A$. Сам коквадрат определяется формулой (12), а именно

$$\mathcal{C}_A = P \Lambda P^{-1}. \tag{13}$$

Напомним, что P должна быть хорошо обусловленной матрицей. Чтобы удовлетворить это требование, построим P как матрицу с диагональным преобладанием. Кроме того, подчиним строчные множители преобладания

$$\alpha_j = \frac{\sum_{k \neq j} |p_{jk}|}{|p_{jj}|}, \qquad j = 1, 2, \ldots, n,$$

общей верхней границе, скажем, 0.5 или 0.8. Вдобавок потребуем, чтобы диагональ матрицы P, рассматриваемая как самостоятельная матрица, тоже была хорошо обусловлена. Это означает, что отношение

$$\frac{\max_j |p_{jj}|}{\min_j |p_{jj}|}$$

должно иметь умеренную величину. Все эти предосторожности обеспечивают матрице P хорошую обусловленность.

Имея теперь коквадрат $\mathcal{C}_A$, построим саму матрицу A. Сначала сформируем диагональную матрицу

$$D = \mathrm{diag}(e^{i\alpha_1}, e^{i\alpha_2}, \ldots, e^{i\alpha_n}),$$

где

$$\alpha_j = \phi_j/2, \qquad j = 1, 2, \ldots, n. \tag{14}$$

Согласно теореме 2, матрица

$$A = P^{-*} D P^{-1}$$

имеет $\mathcal{C}_A$ своим коквадратом. Итак, мы получили входную пару $(A, \mathcal{C}_A)$ для нашего алгоритма.

В действительности, у нас есть некоторая свобода в выборе матрицы A. Любое число α_j в конструкции, описанной выше (см. (14)), может быть заменено противоположным числом $-\alpha_j$. Полученная таким образом матрица $\hat{A}$ также имеет $\mathcal{C}_A$ своим коквадратом.

Проиллюстрируем эту конструкцию небольшим примером. Положим $n = 5$ и определим Λ как диагональную матрицу с унимодулярными диагональными элементами $0.39555 - 0.91844i$, $0.82987 - 0.55796i$, $0.69822 + 0.71588i$, $0.8555 - 0.51780i$ и $-0.67367 - 0.73904i$. Эта матрица является коквадратом диагональной матрицы

$$D = \mathrm{diag}(0.83533 - 0.54975\,i, 0.95652 - 0.29166\,i,$$

$$0.92147 + 0.38844\,i, -0.96320 + 0.26879\,i, 0.40394 - 0.91479\,i).$$

В качестве матрицы P возьмем матрицу

$$P = \begin{bmatrix} 0.52310 & 0.14009 & -0.023845 & -0.12072 & -0.14605 \\ -0.074516 & 0.36960 & -0.013413 & -0.056632 & -0.11475 \\ -0.11923 & -0.0029806 & 0.50522 & -0.026826 & 0.084948 \\ 0.064083 & 0.074516 & -0.032787 & 0.64232 & 0.040238 \\ 0.037258 & 0.014903 & 0.067064 & 0.049180 & 0.61401 \end{bmatrix}.$$

Это матрица со строгим диагональным преобладанием, имеющая спектральное число обусловленности 2.0151. Ее строчные множители преобладания равны приблизительно 0.82, 0.7, 0.46, 0.33, и 0.27.

Теперь построим нашу тестовую матрицу $A = P^{-*} D P^{-1}$ и ее коквадрат $\mathcal{C}_A = P\Lambda P^{-1}$:

$$A = \begin{bmatrix} 2.8503 - 1.6838\,i & -0.46875 + 0.47535\,i & 0.92323 + 0.25978\,i & 0.77833 - 0.33078\,i & 0.31452 - 0.10599\,i \\ -0.46875 + 0.47535\,i & 6.5125 - 2.1250\,i & -0.26455 - 0.14116\,i & 0.73500 - 0.17843\,i & 1.1075 - 0.28332\,i \\ 0.92323 + 0.25978\,i & -0.26455 - 0.14116\,i & 3.8312 + 1.5612\,i & 0.18418 + 0.13572\,i & -0.52631 + 0.16012\,i \\ 0.77833 - 0.33078\,i & 0.73500 - 0.17843\,i & 0.18418 + 0.13572\,i & -2.0899 + 0.55943\,i & 0.32784 + 0.072906\,i \\ 0.31452 - 0.10599\,i & 1.1075 - 0.28332\,i & -0.52631 + 0.16012\,i & 0.32784 + 0.072906\,i & 1.4123 - 2.5490\,i \end{bmatrix},$$

$$\mathcal{C}_A = \begin{bmatrix} 0.39753 & 0.17072 & -0.055836 & -0.096375 & 0.30075 \\ 0.032796 & 0.80977 & -0.035576 & -0.021665 & 0.29139 \\ 0.081672 & -0.028780 & 0.72674 & 0.021300 & -0.18109 \\ -0.041380 & 0.014573 & 0.023539 & 0.85848 & -0.11078 \\ 0.10361 & -0.0058126 & 0.19805 & 0.14534 & -0.68703 \end{bmatrix} +$$

$$+\begin{bmatrix} -0.90771 & 0.14475 & -0.077170 & -0.063877 & 0.0017931 \\ 0.037429 & -0.57369 & -0.038683 & -0.0029814 & 0.045353 \\ 0.35807 & -0.14298 & 0.75753 & 0.11993 & -0.15646 \\ -0.062564 & 0.018042 & -0.081500 & -0.53030 & -0.013913 \\ 0.029418 & -0.0077117 & 0.19960 & 0.031973 & -0.76319 \end{bmatrix} i.$$

К этой матричной паре A и $\mathcal{C}_A$ применим алгоритм, описанный в разд. 4. Сначала вычислим собственные векторы матрицы $\mathcal{C}_A$ с помощью Maple процедуры Eigenvectors. Из этих собственных векторов сформируем матрицу Q, играющую роль матрицы P из разд. 4:

$$Q = \begin{bmatrix} -0.046624 & -0.22527 & 0.95681 & -0.18334 & 0.34804 \\ -0.026226 & -0.17700 & -0.13630 & -0.086014 & 0.91824 \\ 0.98784 & 0.13102 & -0.21808 & -0.040743 & -0.0074052 \\ -0.064108 & 0.062064 & 0.11722 & 0.97558 & 0.18513 \\ 0.13113 & 0.94705 & 0.068149 & 0.074696 & 0.037026 \end{bmatrix} +$$

$$+\begin{bmatrix} -2.4286 \times 10^{-17} & 6.9389 \times 10^{-17} & 0.0 & -2.0691 \times 10^{-15} & -1.6653 \times 10^{-15} \\ 0.0 & -2.3592 \times 10^{-16} & 8.3267 \times 10^{-17} & -5.4395 \times 10^{-15} & 0.0 \\ 0.0 & 2.0817 \times 10^{-17} & 1.5266 \times 10^{-16} & -1.485 \times 10^{-16} & -1.5829 \times 10^{-16} \\ 4.1633 \times 10^{-17} & -2.0817 \times 10^{-17} & 3.4694 \times 10^{-17} & 0.0 & 8.7153 \times 10^{-15} \\ 1.3878 \times 10^{-17} & 0.0 & 3.2613 \times 10^{-16} & 8.0414 \times 10^{-17} & 7.8063 \times 10^{-16} \end{bmatrix} i.$$

Заметим, что порядок соответствующих собственных значений отличен от их порядка в исходной матрице Λ. Мы можем видеть это, рассматривая произведение $Q^{-1}\mathcal{C}_A Q$:

$$\begin{bmatrix} 0.69823 & -0.0000016227 & -0.0000020987 & 0.00000060427 & -0.0000002141 \\ -0.0000032202 & -0.67367 & -0.0000035373 & 0.0000045973 & -0.0000013728 \\ 0.00000026304 & 0.0000033418 & 0.39555 & -0.00000010912 & -0.0000023689 \\ -0.000000034412 & 0.0000027929 & 0.0000030551 & 0.85551 & 0.00000074351 \\ 0.00000063063 & -0.00000020583 & -0.0000022916 & 0.0000012487 & 0.82987 \end{bmatrix} +$$

$$+\begin{bmatrix} 0.71587 & -0.0000072692 & -0.0000064697 & 0.0000031102 & 0.0000014071 \\ 0.0000044353 & -0.73904 & -0.0000016337 & -0.000000624 & -0.0000006787 \\ 0.00000063385 & -0.0000010237 & -0.91844 & -0.0000026568 & 0.0000020208 \\ -0.00000055412 & 0.00000056946 & 0.0000017988 & -0.51779 & 0.0000017383 \\ 0.0000016017 & 0.00000010525 & 0.000000067107 & 0.00000038365 & -0.55796 \end{bmatrix} i.$$

Как и следовало ожидать, эта матрица почти диагональна: поскольку коквадрат $\mathcal{C}_A$ хорошо обусловлен, вычисленные собственные векторы являются хорошими приближениями к соответствующим собственным векторам.

Теперь вычислим произведение $Q^* A Q$:

$$\begin{bmatrix} 3.52292996754929 & -2.16493490 \times 10^{-15} & -1.28369537 \times 10^{-16} & -1.48839274 \times 10^{-15} & -2.55004351 \times 10^{-16} \\ -2.13717932 \times 10^{-15} & 0.960978339111032 & 1.26287869 \times 10^{-15} & 8.04911693 \times 10^{-16} & 1.76247905 \times 10^{-15} \\ 1.52655666 \times 10^{-16} & 4.99600361 \times 10^{-16} & 2.79472428991858 & 8.88178420 \times 10^{-16} & -1.545638612 \times 10^{-15} \\ -2.06085150 \times 10^{-15} & -2.77555756 \times 10^{-16} & 2.08166817 \times 10^{-17} & -2.22191653953022 & -8.95464258 \times 10^{-15} \\ 2.77555756 \times 10^{-17} & 2.22044604 \times 10^{-16} & 5.96744876 \times 10^{-16} & -1.87905247 \times 10^{-14} & 5.90405306809074 \end{bmatrix} +$$

$$+\begin{bmatrix} 1.48507456561132 & -8.32667269\times10^{-17} & 6.41847686\times10^{-16} & -9.40220124\times10^{-16} & 9.16801357\times10^{-16} \\ -6.10622664\times10^{-16} & -2.17629180 & -6.10622664\times10^{-16} & 4.99600361\times10^{-16} & 9.43689571\times10^{-16} \\ -3.5388359\times10^{-16} & -3.6082248\times10^{-16} & -1.839265501 & 7.63278330\times10^{-17} & -1.63064007\times10^{-15} \\ -2.029626467\times10^{-16} & -4.16333634\times10^{-17} & -8.56953397\times10^{-16} & 0.620050292285994 & 1.92376497\times10^{-14} \\ -1.38777878\times10^{-17} & -1.99840144\times10^{-15} & 1.83880689\times10^{-15} & -9.67281810\times10^{-15} & -1.800257000 \end{bmatrix} i.$$

Как мы и надеялись, эта матрица почти диагональна. Однако ее диагональные элементы совсем не похожи на канонические углы матрицы A (другими словами, на диагональные элементы исходной матрицы D). Это объясняется тем обстоятельством, что диагональная форма юнитоида не единственна. Выполним дополнительную конгруэнцию с диагональной матрицей преобразования, сконструированной так, чтобы сделать диагональные элементы унимодулярными числами. В результате получим матрицу F, показанную ниже.

$$F=\begin{bmatrix} 0.92147 & -7.1784\times10^{-16} & -3.5893\times10^{-17} & -5.0119\times10^{-16} & -5.2493\times10^{-17} \\ -7.0866\times10^{-16} & 0.40394 & 4.4764\times10^{-16} & 3.4359\times10^{-16} & 4.5994\times10^{-16} \\ 4.2685\times10^{-17} & 1.7709\times10^{-16} & 0.83532 & 3.1971\times10^{-16} & -3.4012\times10^{-16} \\ -6.9397\times10^{-16} & -1.1848\times10^{-16} & 7.4933\times10^{-18} & -0.96320 & -2.3731\times10^{-15} \\ 5.7137\times10^{-18} & 5.7943\times10^{-17} & 1.3132\times10^{-16} & -4.9798\times10^{-15} & 0.95652 \end{bmatrix}+$$

$$+\begin{bmatrix} 0.38845 & -2.761\times10^{-17} & 1.7947\times10^{-16} & -3.1660\times10^{-16} & 1.887300000\times10^{-16} \\ -2.0247\times10^{-16} & -0.91479 & -2.1644\times10^{-16} & 2.1326\times10^{-16} & 2.4626\times10^{-16} \\ -9.8948\times10^{-17} & -1.2789\times10^{-16} & -0.54976 & 2.7475\times10^{-17} & -3.5882\times10^{-16} \\ -6.8343\times10^{-17} & -1.7772\times10^{-17} & -3.0847\times10^{-16} & 0.26879 & 5.0983\times10^{-15} \\ -2.8568\times10^{-18} & -5.2150\times10^{-16} & 4.0464\times10^{-16} & -2.5634\times10^{-15} & -0.29167 \end{bmatrix} i.$$

С точностью до порядка расположения, диагональные элементы матрицы F те же, что у D.

6. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

Были построены тестовые матрицы порядков от 6 до 10 таким же образом, как это описано в разд. 5. К этим матрицам был применен алгоритм из разд. 4. Здесь мы представим полученные численные результаты.

Для каждого порядка n приводится следующая информация:

1) минимальное расстояние между диагональными элементами исходной матрицы Λ. Наряду с другими характеристиками, это важный показатель обусловленности коквадрата в смысле полной проблемы собственных значений;

2) диагональные элементы исходной матрицы D;

3) спектральное число обусловленности матрицы P и ее строчные множители преобладания;

4) диагональные элементы финальной матрицы A, полученной в результате конгруэнции $A\to Q^*AQ$ и дополнительной конгруэнции, превращающей диагональные элементы в унимодулярные числа; показан также модуль наибольшего внедиагонального элемента этой финальной матрицы.

Для $n=6$ имеем $\min_{i,j}|\lambda_i-\lambda_j|\approx 0.18916$. Диагональные элементы матрицы $D=[-0.99798+0.063579\,i,\ -0.19087-0.98162\,i,\ \ -0.93968-0.34204\,i,\ \ 0.68897+0.72479\,i,\ \ 0.28285+0.95916\,i,\ \ 0.29373-0.95589\,i]$, $\mathrm{cond}_2 P=2.5922$. Строчные множители преобладания матрицы $P=[0.92908,\ \ 0.44952,\ \ 0.54825,\ \ 0.59117,\ \ 0.43062,\ \ 0.41404]$. Диагональные элементы финальной матрицы $A=[-0.99798+0.063580\,i,\ \ -0.93969-$ $-0.34204\,i,\ \ 0.29373-0.95589\,i,\ \ 0.68895+0.72481\,i,\ \ -0.19087-0.98162\,i,\ \ 0.28284+0.95917\,i]$. Модуль наибольшего внедиагонального элемента матрицы $A=1.4041\times10^{-15}$.

Если $n=7$, то $\min_{i,j}|\lambda_i-\lambda_j|\approx 0.09006$. Диагональные элементы матрицы $D=[-0.33215-0.94323\,i,\ 0.46926-0.88306\,i,\ \ -0.85884-0.51225\,i,\ \ -0.60305+0.79771\,i,\ \ -0.99500-0.099839\,i,\ \ -0.64484-0.76432\,i,\ -0.98950-0.14454\,i]$, $\mathrm{cond}_2 P=4.6178$. Строчные множители преобладания матрицы $P=[0.97949,\ \ 0.62807,\ 0.62379,\ \ 0.57623,\ \ 0.53409,\ \ 0.42620,\ \ 0.44784]$. Диагональные элементы финальной матрицы $A=[0.46927-$ $-0.88306\,i,\ \ -0.60304+0.79771\,i,\ \ -0.33214-0.94323\,i,\ \ -0.64483-0.76433\,i,\ \ -0.85884-0.51225\,i,\ \ -0.98950-$ $-0.14454\,i,\ \ -0.99500-0.099841\,i]$. Модуль наибольшего внедиагонального элемента матрицы $A=5.18095\times$ $\times10^{-15}$.

Если $n = 8$, то $\min_{i,j}|\lambda_i - \lambda_j| \approx 0.067026$. Диагональные элементы матрицы $D = [-0.018714 + 0.99982\,i,\ -0.58544 + 0.81072\,i,\ \ -0.94184 + 0.33607\,i,\ \ -0.99183 + 0.12754\,i,\ \ 0.14779 - 0.98902\,i,\ \ 0.90666 + 0.42187\,i,\ -0.89201 - 0.45201\,i,\ \ 0.69023 + 0.72359\,i]$, $\mathrm{cond}_2 P = 3.3993$. Строчные множители преобладания матрицы $P = [0.66038,\ 0.71515,\ 0.59585,\ 0.50718,\ 0.50214,\ 0.51355,\ 0.53183,\ 0.45631]$. Диагональные элементы финальной матрицы $A = [-0.018714 + 0.99982\,i,\ 0.14779 - 0.98902\,i,\ -0.58543 + 0.81072\,i,\ -0.94184 + 0.33606\,i,\ -0.99183 + 0.12754\,i,\ \ 0.69022 + 0.72360\,i,\ \ -0.89201 - 0.45201\,i,\ \ 0.90666 + 0.42186\,i]$. Модуль наибольшего внедиагонального элемента матрицы $A = 4.0844 \times 10^{-15}$.

В случае $n = 9$ $\min_{i,j}|\lambda_i - \lambda_j| \approx 0.21319$. Диагональные элементы матрицы $D = [0.67771 + 0.73533\,i,\ 0.46537 - 0.88512\,i,\ \ -0.56004 + 0.82847\,i,\ \ -0.70299 + 0.71120\,i,\ \ -0.15475 - 0.98795\,i,\ \ 0.96553 + 0.26030\,i,\ -0.75223 - 0.65890\,i,\ \ -0.96317 + 0.26891\,i,\ \ 0.88759 + 0.46063\,i]$, $\mathrm{cond}_2 P = 3.2703$. Строчные множители преобладания матрицы $P = [0.85205,\ 0.72679,\ 0.65264,\ 0.56769,\ 0.57712,\ 0.49667,\ 0.43503,\ 0.49889,\ 0.42197]$. Диагональные элементы финальной матрицы $A = [-0.15475 - 0.98795\,i,\ \ 0.46538 - 0.88511\,i,\ \ -0.56005 + 0.82846\,i,\ \ -0.70299 + 0.71120\,i,\ \ -0.96317 + 0.26891\,i,\ \ 0.96553 + 0.26030\,i,\ \ 0.88759 + 0.46064\,i,\ \ 0.67770 + 0.73534\,i,\ \ -0.75224 - 0.65889\,i]$. Модуль наибольшего внедиагонального элемента матрицы $A = 4.4627 \times 10^{-15}$.

При $n = 10$ $\min_{i,j}|\lambda_i - \lambda_j| \approx 0.08846$. Диагональные элементы матрицы $D = [0.77102\ -0.63681\,i,\ 0.33951 + 0.94060\,i,\ \ 0.72532 + 0.68841\,i,\ \ -0.29731 - 0.95478\,i,\ \ -0.95444 - 0.29840\,i,\ \ 0.91527 + 0.40283\,i,\ 0.98347 - 0.18105\,i,\ \ -0.99052 + 0.13737\,i,\ \ -0.23408 + 0.97222\,i,\ \ 0.98242 + 0.18670\,i]$, $\mathrm{cond}_2 P = 2.9738$. Строчные множители преобладания матрицы $P = [0.75692,\ 0.66424,\ 0.77500,\ 0.67708,\ 0.55788,\ 0.53128,\ 0.66019,\ 0.49208,\ 0.51295,\ 0.49142]$. Диагональные элементы финальной матрицы $A = [0.77101\ -0.63682\,i,\ -0.23408 + 0.97222\,i,\ \ 0.98348 - 0.18104\,i,\ \ -0.99052 + 0.13737\,i,\ \ 0.72533 + 0.68840\,i,\ \ 0.98242 + 0.18670\,i,\ -0.95444 - 0.29840\,i,\ \ 0.91527 + 0.40284\,i,\ \ 0.33950 + 0.94060\,i,\ \ -0.29731 - 0.95478\,i]$. Модуль наибольшего внедиагонального элемента матрицы $A = 5.527 \times 10^{-15}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Хорн Р., Джонсон Ч.* Матричный анализ. М.: Мир, 1989.

A NONSINGULAR MATRIX WITH A WELL-CONDITIONED COSQUARE: HOW TO BRING IT TO DIAGONAL FORM BY A CONGRUENCE TRANSFORMATION

Kh. D. Ikramov[a,*], A. M. Nazari[b,]**

[a] *119992 Moscow, Leninskie Gory, Moscow State University, CMC Faculty, Russia*
[b] *Faculty of Mathematics, Arak University, Arak, Islamic Republic Iran*
**e-mail: ikramov@cs.msu.su*
***e-mail: a-nazari@araku.ac.ir*

Received: 09.01.2024
Revised: 18.04.2024
Accepted: 31.05.2024

Abstract. There exist efficient programs for bringing a diagonalizable matrix to diagonal form by a similarity transformation. In theory of congruence transformations, unitoid matrices are analogs of diagonalizable matrices. However, excepting Hermitian and, more generally, normal matrices, there are no recognized programs for bringing a unitoid matrix to diagonal form by a congruence transfor- mation. We propose an algorithm that is able to perform this task for a special class of unitoid matrices, namely, nonsingular matrices whose cosquares are well-conditioned with respect to the complete eigenproblem. Examples are presented to illustrate the performance of the algorithm.

Keywords: *-congruence transformation, similarity transformation, unitoid, cosquare, canonical angles, diagonally dominant matrix, condition number.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2270–2285

УДК 519.63

АПОСТЕРИОРНЫЕ ОЦЕНКИ ПОГРЕШНОСТИ ПРИБЛИЖЕННЫХ РЕШЕНИЙ ЭЛЛИПТИЧЕСКИХ КРАЕВЫХ ЗАДАЧ В ТЕРМИНАХ ЛОКАЛЬНЫХ НОРМ И ЦЕЛЕВЫХ ФУНКЦИОНАЛОВ

© 2024 г. А. В. Музалевский[2], С. И. Репин[1,2,*], М. Е. Фролов[2]

[1] 191023 Санкт-Петербург, ул. Фонтанка, 27, Санкт-Петербургское отделение Математического института им. В.А. Стеклова РАН, Россия

[2] 195251 Санкт-Петербург, ул. Политехническая, 29, Санкт-Петербургский Политехнический университет Петра Великого, Россия

**e-mail: repin@pdmi.ras.ru*

Поступила в редакцию 07.08.2024 г.

Переработанный вариант 07.08.2024 г.

Принята к публикации 27.08.2024 г.

Получены функциональные соотношения, которые позволяют оценивать точность приближенных решений в терминах мер, существенно отличных от энергетических норм, которые обычно используются для этих целей. В частности, они применимы к локальным нормам и мерам, построенным с помощью специально построенных линейных функционалов. Потребность в таких инструментах контроля точности возникает, если имеется особый интерес к поведению решения в некоторой подобласти или к специальным свойствам решения. Показано, что апостериорные оценки функционального типа, которые ранее использовались для глобальных оценок, могут быть адаптированы и для решения этой задачи. Получены функциональные тождества и оценки, позволяющие оценивать погрешность любых конформных аппроксимаций в терминах широкого класса мер, включающих локальные нормы и проблемно-ориентированные функционалы. Теоретические результаты проверены в серии примеров, которые подтверждают эффективность предлагаемого метода. Библ. 13. Фиг. 10. Табл. 10.

Ключевые слова: эллиптические краевые задачи, апостериорные оценки функционального типа, локальные оценки погрешности, проблемно-ориентировнный (goal-oriented) контроль точности.

DOI: 10.31857/S0044466924120042, **EDN**: KCLPCF

1. ВВЕДЕНИЕ

Апостериорные оценки дают вычисляемые оценки разности между решением некоторой задачи и его аппроксимацией, полученной с помощью какой-либо вычислительной технологии. Если речь идет о краевых задачах для уравнений в частных производных, то в большинстве случаев такие оценки используют глобальные (энергетические) нормы и оценивают точность в интегральном смысле. Однако такая информация может быть недостаточна. В инженерных и научных вычислениях часто интересуются специальными характеристиками точности: локальными ошибками и так называемыми проблемно-ориентированными функционалами ошибок (goal-oriented quantities). Для оценки последних был разработан особый метод, основанный на использовании краевой задачи, сопряженной к исходной (см., например, монографию [1] и статьи [2]–[5]). Есть и другие методы, использующие различные варианты постобработки (post-processing) приближенных решений, иерархии сеток и т.д., которые, однако, применимы только к галеркинским аппроксимациям и требуют выполнения ряда (иногда весьма обременительных) дополнительных условий. Как правило, они дают индикаторы погрешности, которые можно успешно использовать в адаптивных методах, но не гарантированные оценки погрешности приближенного решения.

В настоящей работе мы используем другой подход, который основан на апостериорных оценках функционального типа (см. [6], [7] и другие публикации, цитированные в этих работах). Здесь соответствующие оценки применимы к любым аппроксимациям из базового энергетического пространства независимо от способа их получения. В статье [8] обсуждается апостериорное тождество для норм отклонений от решения абстрактной эллиптической краевой задачи

$$\Lambda^* \mathcal{A} \Lambda u + \ell = 0, \tag{1.1}$$

где $\Lambda : V \to U$ – ограниченный линейный оператор, U – гильбертово пространство со скалярным произведением $(\cdot,\cdot)$, $\mathcal{A} \in \mathcal{L}(U,U)$ является линейным, ограниченным и самосопряженным оператором. Пространства U и V содержат функции, определенные в ограниченной области $\Omega \subset \mathbb{R}^d$ с липшицевой границей Γ, причем V – банахово пространство с нормой $\|\cdot\|_V$, а V_0 – его подпространство, V_0^* – пространство сопряженное к V_0, $\ell \in V_0^*$.

Сопряжение элементов V_0 и V_0^* обозначается $\langle v^*, v\rangle$. Оператор $\Lambda : V \to Y^*$ имеет сопряженный $\Lambda^* : Y^* \to \to V_0^*$, который определяется равенством

$$(y^*, \Lambda v) = \langle \Lambda^* y^*, v\rangle \qquad \forall v \in V, y^* \in Y^*. \tag{1.2}$$

Обобщенное решение задачи (1.1) определяется как функция $u \in V_0 + u_0$, удовлетворяющая интегральному тождеству

$$(\mathcal{A}\Lambda u, \Lambda w) + \langle \ell, w\rangle = 0 \qquad \forall w \in V_0. \tag{1.3}$$

Здесь

$$V_0 + u_0 := \{v \in V \mid v = w + v_0,\ w \in V_0\},$$

а $u_0 \in V$ – заданная функция, определяющая краевое условие. Функция $p^* := \mathcal{A}\Lambda u$ удовлетворяет уравнению $\Lambda^* p^* + \ell = 0$ и представляет собой решение соответствующей двойственной задачи. Пара u и p^* аппроксимируется функциями $v \in V_0 + u_0$ и $y^* \in Y^*$ соответственно. Нас интересуют функции

$$e := v - u \in V_0 \quad \text{и} \quad e^* = y^* - p^* \in Y^*,$$

которые представляют собой отклонения от искомых решений. Они удовлетворяют тождеству (см. [8])

$$\|\Lambda e\|^2_{\mathcal{A}} + \|e^*\|^2_{\mathcal{A}^{-1}} = \|\mathcal{A}\Lambda v - y^*\|^2_{\mathcal{A}^{-1}} + 2\langle \mathbf{r}(y^*), e\rangle, \tag{1.4}$$

где

$$\mathbf{r}(y^*) := \ell + \Lambda^* y^*, \quad \|y^*\|_{\mathcal{A}} := (\mathcal{A}y^*, y^*)^{1/2} \quad \text{и} \quad \|y^*\|_{\mathcal{A}^{-1}} := (\mathcal{A}^{-1}y^*, y^*)^{1/2}.$$

Это тождество является основой для получения гарантированных апостериорных оценок в предположении, что Λ и $\mathcal{A}$ удовлетворяют условиям

$$c_1\|y^*\|^2 \le (\mathcal{A}y^*, y^*) \le c_2\|y^*\|^2 \qquad \forall y^* \in U,$$
$$\|\Lambda w\| \ge C_\Lambda \|w\|_V \quad \forall w \in V,$$

с некоторыми положительными постоянными c_1, c_2 и C_Λ. Заметим, что аналогичные тождества имеют место и для многих нелинейных эллиптических краевых задач. Методы получения таких тождеств и соответствующие оценки подробно исследованы в монографии [9].

Нетрудно видеть, что левая часть (1.4) представляет собой комбинированную меру отклонения от решения, выраженную через энергетические нормы. Естественно возникает вопрос о том, можно ли получать подобные оценки в терминах других мер. В отношении мер более сильных, чем энергетическая, он рассмотрен в работе [10] (конечно, при этом предполагается, что и решение, и его аппроксимация обладают требуемой регулярностью).

В настоящей статье мы изучаем в некотором смысле противоположную проблему: как получить гарантированные и полностью вычисляемые оценки в терминах более слабых мер. В частности, речь идет о локальных нормах, связанных с областью $\omega \subset \Omega$, где Ω – ограниченная область с липшицевой границей, в которой рассматривается краевая задача, а также о мерах, определенных с помощью заданного линейного функционала ξ.

На первый взгляд имеется простое решение, действительно,

$$|\langle \xi, v - u\rangle| \le [\![\, \xi \,]\!]\, \|v - u\|_V, \tag{1.5}$$

где $[\![\, \xi \,]\!]$ обозначает норму $\xi \in V_0^*$. Таким образом, имея оценку в норме пространства V всегда можно получить оценку для более слабой меры в левой части (1.5). Однако, как правило, оценка такого типа является сильно завышенной и малополезной на практике. Ниже предлагается единообразный подход к получению полностью гарантированных оценок ошибок в терминах мер такого типа. Он использует некоторые результаты, которые были ранее получены в [7], [11]. Подчеркнем, что предлагаемый метод не использует никаких специальных свойств решений и аппроксимаций. Соответствующие обоснования, представленные в разд. 3, используют только обобщенные формулировки краевых задач и некоторые функциональные неравенства. В разд. 4 на нескольких примерах показаны возможности таких специализированных функциональных оценок как для эффективного контроля точности в локальных нормах, так и для получения хороших индикаторов ошибок.

2. ДВУСТОРОННИЕ ОЦЕНКИ ПОГРЕШНОСТИ В ГЛОБАЛЬНЫХ НОРМАХ

Задача (1.1) соответствует широкому классу уравнений и систем уравнений дивергентного типа. Вопрос о получении гарантированных двусторонних оценок для норм $\|\Lambda e\|_{\mathcal{A}}$ и $\|e^*\|_{\mathcal{A}^{-1}}$ для краевых задач такого типа хорошо изучен (см. [7], [9] и цитированную там литературу). Напомним некоторые результаты этой теории, которые потребуются в дальнейшем. Прежде всего необходимо ввести некоторые дополнительные понятия и обозначения. В дополнение к пространствам V и U мы введем Гильбертово пространство $\mathcal{V}$ такое, что $V \subset \mathcal{V} \subset V^*$. Пространства V, $\mathcal{V}$ и V^* образуют так называемый триплет Гельфанда, если скалярное произведение $(\cdot,\cdot)_{\mathcal{V}}$ согласовано с пространствами V и V^* таким образом, что $\langle v^*, v\rangle = (v^*, v)_{\mathcal{V}}$, если $v^* \in \mathcal{V}$.

Мы предполагаем, что ℓ имеет представление в виде $\langle \ell, w\rangle = (g, w)_{\mathcal{V}}$, где $g \in \mathcal{V}$, и определим два подпространства пространства U:

$$Q^* := \{y^* \in U \mid \Lambda^* y^* \in \mathcal{V}\} \quad \text{и} \quad Q^*_\ell := \{y^* \in U \mid \Lambda^* y^* + \ell = 0 \in \mathcal{V}\}.$$

Пространство Q^* является Гильбертовым относительно скалярного произведения

$$(q, y)_{Q^*} := (q^*, y^*) + (\Lambda^* q^*, \Lambda^* y^*)_{\mathcal{V}}.$$

Нетрудно видеть, что из (1.4) следует соотношение

$$\|\Lambda e\|_{\mathcal{A}} \le \|\mathcal{A}\Lambda v - q^*\|_{\mathcal{A}^{-1}} \qquad \forall q^* \in Q^*_\ell, \tag{2.1}$$

которое можно рассматривать как обобщенную форму так называемой оценки метода гиперциклов. В контексте задачи линейной упругости оно было получено в статье [12], а для вариационной задачи с квадратичным функционалом в [13]. Оценка (2.1) переходит в равенство, если взять инфимум по q^*, т.е.

$$\|\Lambda e\|_{\mathcal{A}} = \inf_{q^* \in Q^*_\ell} \|\mathcal{A}\Lambda v - q^*\|_{\mathcal{A}^{-1}}. \tag{2.2}$$

Если $q^* \in Q^*_\ell$, то из (1.4) также следует оценка

$$\|q^* - p^*\|_{\mathcal{A}^{-1}} = \|e^*\|_{\mathcal{A}^{-1}} \le \|\mathcal{A}\Lambda v - q^*\|_{\mathcal{A}^{-1}} \qquad \forall v \in V_0 \tag{2.3}$$

и равенство

$$\|e^*\|_{\mathcal{A}^{-1}} = \inf_{\eta \in V_0} \|\mathcal{A}\Lambda \eta - q^*\|_{\mathcal{A}^{-1}}. \tag{2.4}$$

Инфимум в (2.2) и (2.4) достигается на функциях $q^* = p^*$ и $v = u$ соответственно.

Условие $q^* \in Q^*_\ell$, которое присутствует в (2.2) и (2.4) является неудобным с практической точки зрения. Его можно значительно ослабить, допуская для этой переменной значительно более широкое множество Q^*, которое позволяет использовать простые аппроксимации (этот вопрос подробно обсуждается в гл. 3 [7]). Пусть

$$d_{Q^*_\ell}(y^*) := \inf_{q^* \in Q^*_\ell} \|y^* - q^*\|_{\mathcal{A}^{-1}}.$$

обозначает расстояние между $y^* \in Q^*$ и множеством Q^*_ℓ. Тогда вместо (2.2) мы имеем

$$\|\Lambda e\|_{\mathcal{A}} = \inf_{y^* \in Q^*} M_\oplus(v, y^*), \quad \text{где} \quad M_\oplus(v, y^*) := \|\mathcal{A}\Lambda v - y^*\|_{\mathcal{A}^{-1}} + d_{Q^*_\ell}(y^*). \tag{2.5}$$

Соответствующее расстояние можно оценить

$$d_{Q^*_\ell}(y^*) \le \frac{1}{C_\Lambda}\|\ell + \Lambda^* y^*\| \le C\|\ell + \Lambda^* y^*\|_{\mathcal{A}^{-1}}, \quad C = c_1^{-1} C_\Lambda{}^{-1}. \tag{2.6}$$

Также тождество (2.2) дает полностью вычисляемую оценку снизу нормы $\|\Lambda e\|_{\mathcal{A}}$. Действительно,

$$\begin{aligned} \frac{1}{2}\|\Lambda e\|^2_{\mathcal{A}} &= \inf_{y^* \in Q^*} \sup_{w \in V_0} \left\{ \frac{1}{2}\|\mathcal{A}\Lambda v - y^*\|^2_{\mathcal{A}^{-1}} + \langle \Lambda^* y^* + \ell, w\rangle \right\} \ge \\ &\ge \sup_{w \in V_0} \inf_{y^* \in Q^*} \left\{ \frac{1}{2}\|\mathcal{A}\Lambda v - y^*\|^2_{\mathcal{A}^{-1}} + (y^*, \Lambda w) + \langle \ell, w\rangle \right\}. \end{aligned} \tag{2.7}$$

Инфимум по y^* несложно вычислить, действительно он достигается при $y^* = A\Lambda(v + w)$, так что

$$\inf_{y^* \in Q^*} \Big(\frac{1}{2}\|y^*\|^2_{\mathcal{A}^{-1}} - (y^*, \Lambda v) + (y^*, \Lambda w) \Big) = -\frac{1}{2}\|\Lambda(v + w)\|^2_{\mathcal{A}}.$$

Подстановка в (2.7) приводит к неравенству

$$\frac{1}{2}\|\Lambda e\|_{\mathcal{A}}^2 \geq \sup_{w\in V_0}\left\{\frac{1}{2}\|\mathcal{A}\Lambda v\|_{\mathcal{A}^{-1}}^2 - \frac{1}{2}\|\Lambda(v+w)\|_{\mathcal{A}}^2 + \langle \ell, w\rangle\right\} = \sup_{w\in V_0}\left\{-\frac{1}{2}\|\Lambda w\|_{\mathcal{A}}^2 - (\Lambda v, \Lambda w) + \langle \ell, w\rangle\right\}.$$

Следовательно, для любой функции $w \in V_0$ имеет место оценка

$$\|\Lambda e\|_{\mathcal{A}} \geq M_{\ominus}(v, w), \tag{2.8}$$

где

$$M_{\ominus}^2(v, w) := -\|\Lambda w\|_{\mathcal{A}}^2 - 2(\Lambda v, \Lambda w) + 2\langle \ell, w\rangle.$$

С помощью (2.4) мы получаем оценки для двойственной переменной:

$$\|y^* - p^*\|_{\mathcal{A}^{-1}} \leq \|p^* - q^*\|_{\mathcal{A}^{-1}} + \|y^* - q^*\|_{\mathcal{A}^{-1}} = \inf_{v\in V_0}\|\mathcal{A}\Lambda v - q^*\|_{\mathcal{A}^{-1}} + \|y^* - q^*\|_{\mathcal{A}^{-1}} \leq$$
$$\leq \inf_{v\in V_0}\|\mathcal{A}\Lambda v - y^*\|_{\mathcal{A}^{-1}} + 2d_{Q_\ell^*}(y).$$

Таким образом,

$$\|y^* - p^*\|_{\mathcal{A}^{-1}} \leq M_{\oplus}^*(v, y) := \|\mathcal{A}\Lambda v - y^*\|_{\mathcal{A}^{-1}} + 2d_{Q_\ell^*}(y^*) \qquad \forall\, v \in V_0 \tag{2.9}$$

и

$$\|y^* - p^*\|_{\mathcal{A}^{-1}} \geq \|p^* - q^*\|_{\mathcal{A}^{-1}} - \|y^* - q^*\|_{\mathcal{A}^{-1}} \geq \inf_{v\in V_0}\|\mathcal{A}\Lambda v - q^*\|_{\mathcal{A}^{-1}} - 2d_{Q_\ell^*}(y^*).$$

Оценки (2.5), (2.8) и (2.9) позволяют контролировать погрешность аппроксимаций в глобальных нормах. Эффективность этих оценок обсуждается в разд. 4, где они используются для анализа стандартных конечно-элементных аппроксимаций. Вполне естественно, что их точность существенно зависит от выбора y^* в (2.5) и w в (2.8). Чем точнее двусторонние оценки, тем большие вычислительные затраты необходимы для нахождения функций y^* и w. Далее мы покажем, что эти затраты являются оправданными, поскольку позволяют получить также двусторонние гарантированные оценки для широкого набора специальных мер, а также очень эффективные индикаторы распределения ошибок по области.

3. ДВУСТОРОННИЕ ОЦЕНКИ ДЛЯ СПЕЦИАЛЬНЫХ МЕР ОШИБОК

3.1. Основные оценки

Мы рассматриваем класс мер для функций e и e^*, которые задаются неотрицательными функционалами

$$\Phi : V \to \mathbb{R}_{\geq 0} \quad \text{и} \quad \Phi^* : U \to \mathbb{R}_{\geq 0},$$

обращающимися в нуль, если $e = 0$ или $e^* = 0$. Пусть Φ удовлетворяют условиям

$$\Phi(e_1 + e_2) \leq \Phi(e_1) + \Phi(e_2) \qquad \forall e_1, e_2 \in V \tag{3.1}$$

и

$$\Phi(e) \leq C_\Phi\|\Lambda e\|_{\mathcal{A}} \qquad \forall e \in V. \tag{3.2}$$

Аналогично Φ^* удовлетворяет условиям

$$\Phi^*(e_1^* + e_2^*) \leq \Phi^*(e_1^*) + \Phi^*(e_2^*) \qquad \forall e^* \in Y^* \tag{3.3}$$

и

$$\Phi^*(e^*) \leq C_{\Phi^*}\|e^*\|_{\mathcal{A}^{-1}} \qquad \forall e_1^*, e_2^* \in Y^*. \tag{3.4}$$

Условия (3.1)–(3.4) охватывают весьма широкий класс функционалов, используемых для контроля приближенных решений. В частности, $\Phi(e)$ и $\Phi^*(e^*)$ можно определить как нормы $\|\Lambda e\|_{\mathcal{A},\omega}$ и $\|e^*\|_{\mathcal{A}^{-1},\omega}$, связанные с некоторой локальной областью $\omega \subset \Omega$. В качестве локальных норм другого рода можно взять

$$\|e\|_{s,\omega} := \left(\int_\omega |e|^s d\Omega\right)^{1/s}, \quad \|e^*\|_{s,\omega} := \left(\int_\omega |e^*|^s d\Omega\right)^{1/s}, \ s \geq 1,$$

или нормы, определенные на многообразиях меньшей размерности (например, на части границы области). Также часто используются функционалы, порожденные некоторыми линейными функционалами, ориентированными на специфические формы контроля (так называемые "goal oriented quantities"). Их мы отдельно рассмотрим в п. 3.2.

Условия (3.1) и (3.3) означают, что меры удовлетворяют неравенству треугольника. Их можно сделать несколько более общими, но для краткости изложения удобно оставить их в такой форме. Для определения констант C_Φ и C_{Φ^*} необходимо использовать подходящие функциональные неравенства. Например, пусть $s = 1$, V – это пространство Соболева $\overset{\circ}{H}{}^1(\Omega)$, а Λ – оператор градиента. Тогда для локальной L_1-нормы имеет место оценка

$$\int\limits_\omega |e|d\Omega \le |\omega|^{1/2}\|e\|_{2,\omega} \le |\omega|^{1/2} C_F \|\nabla e\|_\Omega \le \left(\frac{|\omega|}{c_1}\right)^{1/2} C_F \|\nabla e\|_{\mathcal{A}},$$

где C_F обозначает постоянную в неравенстве Фридрихса для функций, обращающихся в нуль на границе $\partial\Omega$. Эта оценка показывает, что условие (3.2) выполнено.

Нижеприведенная теорема устанавливает функциональные соотношения, которые выполняются для ошибок, оцениваемых в терминах мер $\Phi(e)$ и $\Phi^*(e^*)$. Эти соотношения позволяют в дальнейшем получать легко вычисляемые оценки для этих мер.

Теорема 1. *1. Если* Φ *удовлетворяет* (3.1) *и* (3.2)*, то*

$$\Phi(e) = \inf_{\eta\in V}\inf_{q^*\in Q_\ell^*}\Big\{\Phi(\eta) + C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\Big\} = \sup_{\eta\in V}\sup_{q^*\in Q_\ell^*}\Big\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\Big\}, \tag{3.5}$$

причем

$$|\Phi(e) - \Phi(\eta)| \le C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}} \qquad \forall \eta \in V, q^* \in Q_\ell^*. \tag{3.6}$$

2. Если Φ^* *удовлетворяет* (3.3) *и* (3.4)*, то*

$$\Phi^*(e^*) = \inf_{\tau^*\in Q_0^*}\inf_{\eta\in V}\Big\{\Phi^*(\tau^*) + C_{\Phi^*}\|\mathcal{A}\Lambda\eta - q^* - \tau^*\|_{\mathcal{A}^{-1}}\Big\} = \sup_{\tau^*\in Q_0^*}\sup_{\eta\in V}\Big\{\Phi^*(\tau^*) - C_{\Phi^*}\|\mathcal{A}\Lambda\eta - q^* - \tau^*\|_{\mathcal{A}^{-1}}\Big\}, \tag{3.7}$$

причем

$$|\Phi^*(e^*) - \Phi^*(\tau^*)| \le C_{\Phi^*}\|\mathcal{A}\Lambda\eta - q^* - \tau^*\|_{\mathcal{A}^{-1}} \qquad \forall \eta \in V,\ \tau^* \in Q_0^*. \tag{3.8}$$

Доказательство. Пусть $\eta \in V$. Используя (3.1) и (3.2), мы заключаем, что

$$\Phi(e) \le \Phi(\eta) + \Phi(e-\eta) \le \Phi(\eta) + C_\Phi\|\Lambda(e-\eta)\|_{\mathcal{A}} \tag{3.9}$$

и

$$\Phi(e) \ge \Phi(\eta) - \Phi(\eta-e) \ge \Phi(\eta) - C_\Phi\|\Lambda(e-\eta)\|_{\mathcal{A}}. \tag{3.10}$$

Неравенство (3.6) следует из (3.9) и (3.9).

Учитывая (2.2), оценка (3.9) приобретает вид

$$\Phi(e) \le \Phi(\eta) + C_\Phi \inf_{q^*\in Q_\ell^*}\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}} \le \Phi(\eta) + C_\Phi\|\mathcal{A}\Lambda(v-\eta) - p^*\|_{\mathcal{A}^{-1}}.$$

Таким образом,

$$\inf_{\eta\in V}\inf_{q^*\in Q_\ell^*}\{\Phi(\eta) + C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\} \ge \Phi(e).$$

При этом

$$\inf_{\eta\in V}\inf_{q^*\in Q_\ell^*}\{\Phi(\eta) + C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\} \le \Phi(v-u) + C_\Phi\|\Lambda u - p^*\|_{\mathcal{A}^{-1}} = \Phi(e),$$

и мы приходим к тождеству (3.5).

Аналогичным образом (2.2) и (3.10) дают оценку

$$\Phi(e) \ge \Phi(\eta) - C_\Phi \inf_{q^*\in Q_\ell^*}\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}} =$$
$$= \sup_{q^*\in Q_\ell^*}\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\} \ge \Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - p^*\|_{\mathcal{A}^{-1}}.$$

Неравенство верно для любого $\eta \in V$, так что

$$\Phi(e) \geq \sup_{\eta\in V}\sup_{q^*\in Q^*_\ell}\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\}.$$

Поскольку

$$\sup_{\eta\in V}\sup_{q^*\in Q^*_\ell}\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}}\} \geq \sup_{\eta\in V}\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - p^*\|_{\mathcal{A}^{-1}}\} \geq \Phi(e),$$

приходим к (3.5).

Для доказательства (3.7) используем оценки

$$\Phi^*(e^*) \leq \Phi^*(\tau^*) + \Phi^*(e^* - \tau^*) \leq \Phi^*(\tau^*) + C_{\Phi^*}\|e^* - \tau^*\|_{\mathcal{A}^{-1}},$$
$$\Phi^*(e^*) \geq \Phi^*(\tau^*) - \Phi^*(\tau^* - e^*) \geq \Phi^*(\tau^*) - C_{\Phi^*}\|e^* - \tau^*\|_{\mathcal{A}^{-1}},$$

которые выполняются для любого $\tau^* \in U$. Если $\tau^* \in Q^*_0$, то $q^* - \tau^* \in Q^*_\ell$, и в соответствии с (2.4) имеет место оценка

$$\|e^* - \tau^*\|_{\mathcal{A}^{-1}} = \|q^* - \tau^* - p^*\|_{\mathcal{A}^{-1}} = \inf_{\eta\in\in V_0}\|\mathcal{A}\Lambda\eta - q^* + \tau^*\|_{\mathcal{A}^{-1}}.$$

Следовательно, имеет место (3.8), и для любого $\tau^* \in Q^*_0$

$$\Phi^*(e^*) \leq \Phi^*(\tau^*) + C_{\Phi^*}\inf_{\eta\in\in V_0}\|\mathcal{A}\Lambda\eta - q^* + \tau^*\|_{\mathcal{A}^{-1}} \leq \Phi^*(\tau^*) + C_{\Phi^*}\|p^* - q^* + \tau^*\|_{\mathcal{A}^{-1}}.$$

Поскольку $q^* - p^* \in Q^*_\ell$,

$$\inf_{\tau^*\in Q^*_\ell}\inf_{\eta\in\in V_0}\{\Phi^*(\tau^*) + \|\mathcal{A}\Lambda\eta - q^* + \tau^*\|_{\mathcal{A}^{-1}}\} = \Phi^*(e^*).$$

Аналогично

$$\Phi^*(e^*) \geq \Phi^*(\tau^*) - C_{\Phi^*}\sup_{\eta\in\in V_0}\|\mathcal{A}\Lambda\eta - q^* + \tau^*\|_{\mathcal{A}^{-1}} \geq \Phi^*(\tau^*) - C_{\Phi^*}\|p^* - q^* + \tau^*\|_{\mathcal{A}^{-1}},$$

откуда следует (3.7).

Следствие 1. Определим $\ker\Phi := \{v \in V \mid \Phi(v) = 0\}$. В соответствии с (3.5)

$$\Phi(e) \leq C_\Phi\inf_{\eta\in\ker\Phi}\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}} \qquad \forall q^* \in Q^*_\ell. \tag{3.11}$$

Эту оценку можно рассматривать как обобщение неравенства (2.1) на введенный класс мер ошибок. Легко видеть, что (2.1) следует из (3.11). Действительно, если $\Phi(e) = \|\Lambda e\|_{\mathcal{A}}$, то $\ker\Phi$ содержит только нулевой элемент, а $C_\Phi = 1$. Оценка (3.11) показывает, что мажоранта $\Phi(e)$ может быть получена с помощью дополнительной минимизации на множестве $\ker\Phi$. Далее мы покажем, что, по крайней мере, для одного вида Φ эта оценка точна.

Соотношения (3.5) и (3.7) в определенном отношении аналогичны равенствам (2.2) и (2.4) для глобальных норм и обладают тем же самым недостатком: они определены на весьма узком классе функций Q^*_ℓ. Поскольку

$$\inf_{q^*\in Q^*_\ell}\|\mathcal{A}\Lambda(v-\eta) - q^*\|_{\mathcal{A}^{-1}} \leq \left\{\|\mathcal{A}\Lambda(v-\eta) - y^*\|_{\mathcal{A}^{-1}} + d_{Q^*_\ell}(y^*)\right\}$$

и имеет место (3.5), получаем оценку

$$\Phi(e) \leq \inf_{\eta\in V_0}\left\{\Phi(\eta) + C_\Phi\|\mathcal{A}\Lambda(v-\eta) - y^*\|_{\mathcal{A}^{-1}} + d_{Q^*_\ell}(y^*)\right\},$$

где $y^* \in Q^*$, а равенство достигается $y^* = p^*$. Таким образом,

$$\Phi(e) = \inf_{y^*\in Q^*}\inf_{\eta\in V_0}\left\{\Phi(\eta) + C_\Phi\|\mathcal{A}\Lambda(v-\eta) - y^*\|_{\mathcal{A}^{-1}} + d_{Q^*_\ell}(y^*)\right\}.$$

Из (3.5) также следует, что

$$\Phi(e) \geq \sup_{\eta\in V_0}\sup_{q^*\in Q^*_\ell}\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - y^*\|_{\mathcal{A}^{-1}} - \|y^* - q^*\|_{\mathcal{A}^{-1}}\} \geq$$
$$\geq \sup_{\eta\in V_0}\left\{\Phi(\eta) - C_\Phi\|\mathcal{A}\Lambda(v-\eta) - y^*\|_{\mathcal{A}^{-1}} - d_{Q^*_\ell}(y^*)\right\}.$$

Эта миноранта точна, и равенство достигается при $\eta = v - u$ и $y^* = p^*$.

3.2. Меры ошибок, индуцированные линейными функционалами

В работах, посвященных апостериорному контролю точности решений дифференциальных уравнений, часто используются функционалы вида

$$\Phi(e) = |\langle \xi, e\rangle|,\ \xi \in V^* \quad \text{и} \quad \Phi^*(e^*) = |(\kappa, e^*)|,\ \kappa \in U. \tag{3.12}$$

Например, если $\xi(x)$ это неотрицательная функция в Ω, то величина $\left|\int\limits_\Omega \xi e dx\right|$ дает некоторую характеристику ошибки e, которая, в первую очередь, ориентирована на те части области, где ξ имеет большие значения. Функционал такого типа не является нормой и, вообще говоря, порождает довольно слабую меру. Для того чтобы частично компенсировать этот недостаток, иногда предлагается использовать несколько таких функционалов и задать полунорму вида

$$|[v]| := \sup_{k=1,\dots,M} |\langle \xi_k, v\rangle|, \qquad k = 1, 2, \dots, M.$$

Рассмотрим приложение вышеприведенных результатов к случаю (3.12). Пусть

$$V_\xi := \ker \Phi = \{\varphi \in V_0 \mid \langle \xi, \varphi\rangle = 0\} \quad \text{и} \quad [\![\, \xi \,]\!] := \sup_{w \in V_0} \frac{|\,\langle \xi, w\rangle\,|}{\|\Lambda w\|_{\mathcal{A}}}.$$

Тогда

$$|\,\langle \xi, e\rangle\,| \le [\![\, \xi \,]\!]\, \|\Lambda e\|_{\mathcal{A}}, \tag{3.13}$$

так что в этом случае в качестве C_Φ можно взять $[\![\, \xi \,]\!]$ или оценку сверху этой величины. Покажем, что

$$|\,\langle \xi, e\rangle\,| = [\![\, \xi \,]\!] \inf_{\varphi \in V_\xi} \|\Lambda(e - \varphi)\|_{\mathcal{A}}. \tag{3.14}$$

Из (3.13) следует, что

$$|\,\langle \xi, e\rangle\,| \le [\![\, \xi \,]\!] \inf_{\varphi \in V_\xi} \|\Lambda(e - \varphi)\|_{\mathcal{A}}. \tag{3.15}$$

Пусть $u_\xi \in V$ является решением задачи

$$(\mathcal{A}\Lambda u_\xi, \Lambda w) = \langle \xi, w\rangle \qquad \forall w \in V, \tag{3.16}$$

где $\xi \in V^*$. Функция $\widehat{u}_\xi$ существует и единственна, причем $\|\Lambda u_\xi\|_{\mathcal{A}} \le [\![\, \xi \,]\!]$. С другой стороны,

$$[\![\, \xi \,]\!] = \sup_{w \in V_0} \frac{(\mathcal{A}\Lambda u_\xi, \Lambda w)}{\|\Lambda w\|_{\mathcal{A}}} \le \|\Lambda u_\xi\|_{\mathcal{A}},$$

так что $\|\Lambda u_\xi\|_{\mathcal{A}} = [\![\, \xi \,]\!]$.

Положим $\widehat{\varphi} = e - \frac{\langle \xi, e\rangle}{[\![\, \xi \,]\!]^2} u_\xi$. Тогда

$$\langle \xi, \widehat{\varphi}\rangle = \langle \xi, e\rangle \left(1 - \frac{\langle \xi, u_\xi\rangle}{[\![\, \xi \,]\!]^2}\right) = 0$$

и, следовательно, $\widehat{\varphi} \in V_\xi$. Поэтому

$$\inf_{\varphi \in V_\xi} \|\Lambda(e - \varphi)\|_{\mathcal{A}} \le \|\Lambda(e - \widehat{\varphi})\|_{\mathcal{A}} = |\langle \xi, e\rangle| \frac{\|\Lambda u_\xi\|_{\mathcal{A}}}{[\![\, \xi \,]\!]^2} = \frac{1}{[\![\, \xi \,]\!]} |\langle \xi, e\rangle|,$$

левая часть (3.15) совпадает с правой, что и доказывает (3.14).

Используем (2.5) и (2.6):

$$|\,\langle \xi, e\rangle\,| \le M_{\oplus,\xi}(v, y^*, \varphi) := [\![\, \xi \,]\!]\, M_\oplus(v - \varphi, y^*). \tag{3.17}$$

Минимизация правой части (3.17) по ϕ и y^* дает точное значение $\Phi(e)$.

4. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

В процессе численных экспериментов вычислялись апостериорные оценки ошибки (мажоранта и миноранта) для решений различных краевых задач вида $\operatorname{div}(A\nabla u) + f = 0$, $u = 0$ на Γ. В качестве приближенных решений используются стандартные конечноэлементные аппроксимации, так что v представляет собой непрерывную кусочно-аффинную функцию, удовлетворяющую краевому условию. Для контроля ошибок глобальных норм используем оценки (2.5) и (2.8). При этом

$$M_{\oplus}(v, y) = ||A\nabla v - y^*||_{A^{-1},\Omega} + C_F||\operatorname{div} y^* + f||_{\Omega}$$

и

$$M^2_{\ominus}(v, w) = -||\nabla w||^2_{A,\Omega} - 2\int\limits_{\Omega} A\nabla v \cdot \nabla w \, dx + 2\int\limits_{\Omega} fw \, dx.$$

Функция y^* получается с помощью реконструкции вычисленного потока $A\nabla v$ и, затем, уточняется в процессе минимизации мажоранты. Функция w строится на подпространстве непрерывных кусочно-квадратичных функций. Как правило, процесс получения точных двусторонних оценок требует более значительных вычислительных затрат, чем те, что требуются для вычисления v. Это естественно, поскольку построение решения с гарантированной точностью должно требовать больших затрат машинного времени, чем просто получение некоторого приближенного решения. Более того, приведенные далее примеры показывают, что дополнительные затраты можно эффективно использовать для анализа точности с помощью различных функционалов Φ. В частности, мы без всяких дополнительных затрат можем использовать оценку

$$\Phi(e) \leqslant \Phi(w) + C_{\Phi}||A\nabla(v + w) - y^*||_{A^{-1},\Omega} + C_{\Phi}C_F||\operatorname{div} y^* + f||_{\Omega} \tag{4.1}$$

с самыми разными Φ, которые удовлетворяют (3.1), (3.2).

4.1. Пример 1

Решается задача $\Delta u + f = 0$ в области $\Omega = [0, 1]^2$ с нулевыми граничными условиями, $f = -2(x_1^2 + x_2^2 - -x_1 - x_2)$. Здесь известно точное решение $u = x_1(1-x_1)x_2(1-x_2)$. В качестве приближенного решения берется галеркинская аппроксимация на разных сетках (N_p – количество узлов).

Результаты использования глобальных оценок (2.5) и (2.8) показаны (в процентах) в левой половине табл. 1, где видно, что энергетическая норма ошибки (нормированная относительно нормы решения) очень точно оценивается как сверху так и снизу.

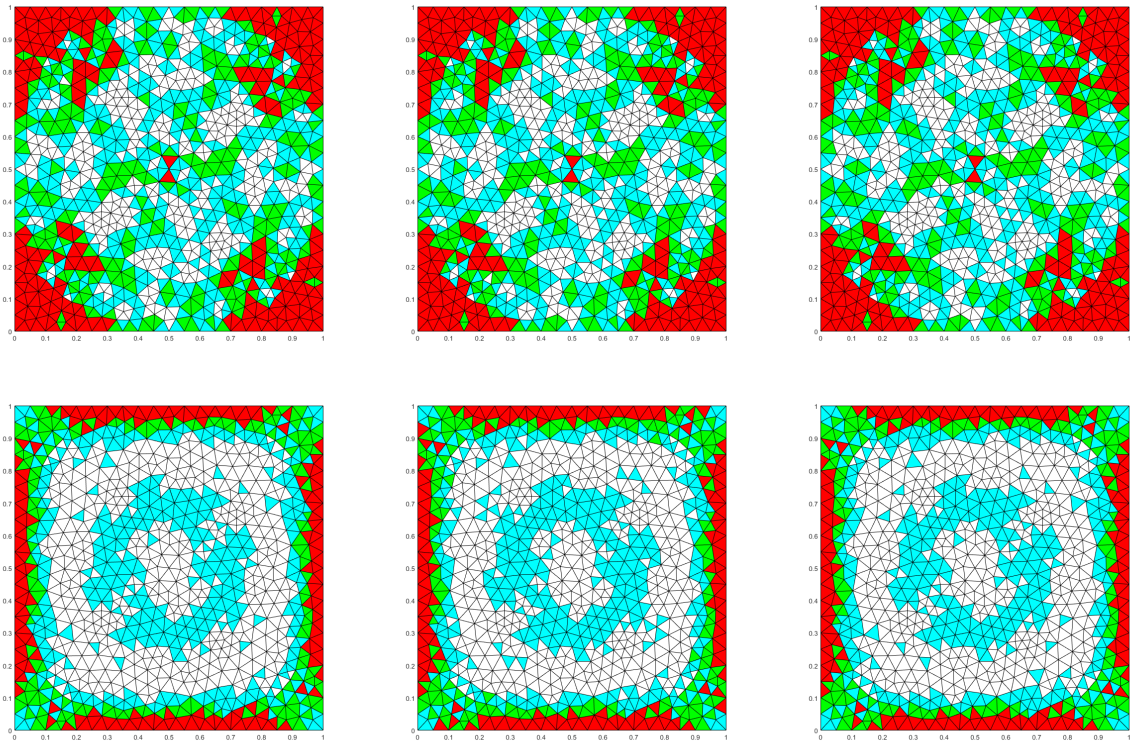
При этом первое слагаемое $M_{\oplus}$ дает прекрасный индикатор распределения ошибок по элементам. Три картинки верхнего ряда фиг. 1 показывают это распределение, где элементы, окрашенные в красный цвет содержат 50% суммарной ошибки, красные и зеленые – 70%, а красные, голубые и зеленые – 90%. Есть и другой способ построить индикатор этого распределения. Для этого надо положить $\Phi(e) = ||\nabla e||_T$, где T – это треугольник из конечноэлементного разбиения Ω. Если w – это функция, обеспечивающая достаточно точную оценку снизу (как, например в табл. 1), то $\Phi(w)$ является таким индикатором. На фиг. 1 показаны соответствующие результаты. Рисунок слева построен по истинному распределению ошибки, индикатор в центре построен по $M_{\oplus}$, а индикатор справа построен по $\Phi(w)$. Нетрудно видеть, что эти индикаторы (которые играют важную роль при использовании адаптивных численных методов) воспроизводятся с высокой точностью, построенные с помощью мажоранты и локальной ошибки почти не отличаются от отметок индикатора, построенного по точной ошибке.

Теперь мы покажем, что используя уже вычисленные функции w и y^*, можно без особого труда получать оценки и индикаторы для других функционалов ошибок. Начнем с примера, когда такой функционал представляет собой L_2 норму, т.е.

$$\Phi(e) = ||e||_{\Omega}.$$

Таблица 1. Двусторонние оценки глобальной ошибки в примере 1

N_p	$\frac{M_{\ominus}}{\|\|\nabla u\|\|}$	$\frac{\|\|\nabla(u-v)\|\|}{\|\|\nabla u\|\|}$	$\frac{M_{\oplus}}{\|\|\nabla u\|\|}$	N_p	$\frac{M_{\ominus}}{\|\|\nabla u\|\|}$	$\frac{\|\|\nabla(u-v)\|\|}{\|\|\nabla u\|\|}$	$\frac{M_{\oplus}}{\|\|\nabla u\|\|}$
177	10.67	10.68	10.85	177	18.68	18.69	18.94
727	5.10	5.11	5.14	727	15.90	15.90	16.11
2810	2.56	2.56	2.57	2810	15.24	15.24	15.46



Фиг. 1. Индикаторы распределения ошибок в примере 1.

Таблица 2. Оценки L_2 нормы ошибки

N_p	$\frac{\Phi(e)}{\|\|u\|\|}$,%	$\frac{S}{\|\|u\|\|}$,%	$\frac{S_1}{\|\|u\|\|}$,%	$\frac{S_2}{\|\|u\|\|}$,%	$\frac{S_3}{\|\|u\|\|}$,%
177	1.3792	2.0452	1.3778	0.4942	0.1732
727	0.3152	0.4622	0.3151	0.1085	0.0385
2810	0.0786	0.1152	0.0786	0.0270	0.0096

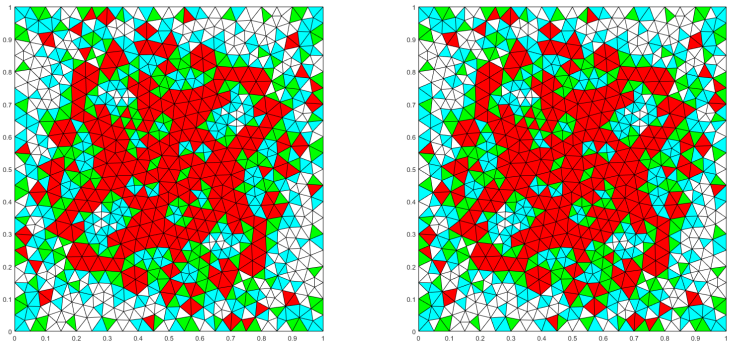
Результаты представлены в табл. 2. Здесь (и далее везде) $S := S_1+S_2+S_3$, а компоненты этой суммы соответствуют трем слагаемым мажоранты (4.1). В данном случае $S_1 = \Phi(w)$, $S_2 = C_F||\nabla(v+w)-y||_\Omega$ и $S_3 = C_F^2||\mathrm{div} y+f||_\Omega$. Видно, что сумма этих величин мажорирует $\Phi(e)$, а S_1 практически совпадает с $\Phi(e)$. Более того, с помощью S_1 можно построить очень хороший индикатор, который показывает значения $\Phi(e)$, соответствующие отдельным элементам. На фиг. 2 слева маркировка элементов осуществлена по функции $\Phi(e)$, а справа – по S_1. Эти маркировки практически не отличаются друг от друга.

В табл. 3 представлены результаты для другого функционала: $\Phi(e) = \left|\int\limits_\Omega e\,dx\right|$. Здесь

$$S_1 = \Phi(w), \quad S_2 = C_F|\Omega|^{1/2}||\nabla(v+w)-y||_\Omega \quad \text{и} \quad S_3 = C_F^2|\Omega|^{1/2}||\mathrm{div}\, y+f||_\Omega.$$

Несмотря на то что этот функционал сильно отличается от энергетической нормы (и является лишь весьма слабой полунормой), соответствующая мажоранта дает весьма точную оценку сверху (например, при $N_p = 2810$ величина $\Phi(e) = 0.0681$ оценивается как 0.1047).

Теперь обратимся к оценкам локального типа, когда функционал $\Phi(e)$ ассоциирован с некоторой подобластью ω. Пусть, например, эта подобласть расположена в левом нижнем углу Ω и $\Phi(e) = \left|\int\limits_\omega e\,dx\right|$, где $\omega = [0,1/2]^2$. В табл. 4 представлены соответствующие результаты. Здесь гарантированная верхняя граница в несколько раз превышает истинное значение ошибки, но значения $\Phi(w)$ по-прежнему очень близки к $\Phi(e)$. Таким образом, предлагаемый метод для $N_p = 2810$ дает следующий результат: величина $\Phi(e)$ не превосходит 0.0526 и близка



Фиг. 2. Индикатор ошибки в норме L_2.

Таблица 3. Оценка для $\Phi(e) = \left|\int\limits_{\Omega} e\,dx\right|$

N_p	$\frac{\Phi(e)}{\|\|u\|\|}$,%	$\frac{S}{\|\|u\|\|}$,%	$\frac{S_1}{\|\|u\|\|}$,%	$\frac{S_2}{\|\|u\|\|}$,%	$\frac{S_3}{\|\|u\|\|}$,%
177	1.1901	1.8552	1.1879	0.4942	0.1732
727	0.2737	0.4206	0.2735	0.1085	0.0385
2810	0.0681	0.1047	0.0681	0.0270	0.0096

Таблица 4. Оценка для $\Phi(e) = \left|\int\limits_{\omega} e\,dx\right|$

N_p	$\frac{\Phi(e)}{\|\|u\|\|}$,%	$\frac{S}{\|\|u\|\|}$,%	$\frac{S_1}{\|\|u\|\|}$,%	$\frac{S_2}{\|\|u\|\|}$,%	$\frac{S_3}{\|\|u\|\|}$,%
177	0.2415	0.9083	0.2410	0.4942	0.1732
727	0.0627	0.2097	0.0627	0.1085	0.0385
2810	0.0160	0.0526	0.0160	0.0270	0.0096

к 0.0160, т.е. гарантированная оценка превосходит точное значение примерно в 3 раза, а индикатор (вероятное значение) практически совпадает с точным.

Рассмотрим локальные оценки другого рода. Разобьем Ω на 16 одинаковых квадратных подобластей

$$\omega_{ij} := [\frac{i}{N}, \frac{i+1}{N}] \times [\frac{j}{N}, \frac{j+1}{N}], \quad i,j = 1,2,3,4, \quad N = 4,$$

и пусть

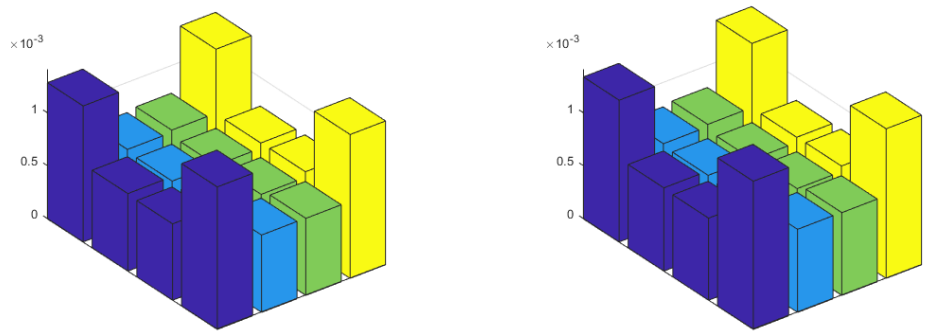
$$\Phi_{ij}(e) = ||\nabla e||_{\omega_{ij}}.$$

Оценки сверху для локальных норм вычисляются согласно (4.1), где

$$S_1 = \Phi(w), \quad S_2 = ||\nabla(v+w) - y||_{\Omega}, \quad S_3 = C_F||\mathrm{div} y + f||_{\Omega}.$$

На фиг. 3 показаны истинные значения локальных ошибок (слева) и их мажоранты, вычисленные согласно предлагаемому нами методу. При $N_p = 2810$ абсолютные значения локальных ошибок колеблются в диапазоне $10^{-3} - 10^{-4}$, а их переоценка мажорантой составляет не более 5–7%.

Отметим, что обсуждаемый нами метод позволяет получать как глобальные, так и локальные оценки не только для галеркинских, но и для любых конформных аппроксимаций. Например, в качестве аппроксимации решения можно взять кусочно-аффинный интерполянт функции $v = \frac{1}{16}\sin \pi x \sin \pi y$ на соответствующей сетке. Такая функция может рассматриваться как некоторая конечноэлементная аппроксимация решения, но, конечно, она не является галеркинской. Для нее глобальные оценки (2.5) и (2.8) показаны (в процентах) в правой половине табл. 1. Видно, что и для такой не-галеркинской аппроксимации энергетическая норма ошибки очень точно оценивается как сверху, так и снизу. Так же, как и ранее, можно построить индикаторы распределения ошибок по области. Эти результаты показаны в нижнем ряду фиг. 1. Рисунок слева построен по истинному



Фиг. 3. Распределение локальных ошибок.

распределению ошибки, индикатор в центре построен по $M_{\oplus}$, а индикатор справа построен с помощью функции w. Обратимся к локальным ошибкам для подобластей ω_{ij}, аналогичных тем, что были рассмотрены ранее. В табл. 5 представленные 16 значений $\|\nabla e\|_{\omega_{ij}}$ (верхняя часть) сравниваются с теми локальными оценками, что были вычислены согласно предлагаемой методике (нижняя часть). Нетрудно видеть, что эти значения очень

Таблица 5. Локальные ошибки и их оценки для не-галеркинской аппроксимации

5.5672e-03	5.2766e-03	5.2254e-03	5.7631e-03
5.2976e-03	3.2929e-03	3.2620e-03	5.4771e-03
5.2635e-03	3.2441e-03	3.2525e-03	5.4045e-03
5.7214e-03	5.4481e-03	5.4768e-03	5.8414e-03
6.8223e-03	6.5317e-03	6.4805e-03	7.0182e-03
6.5526e-03	4.5480e-03	4.5171e-03	6.7322e-03
6.5186e-03	4.4992e-03	4.5076e-03	6.6596e-03
6.9765e-03	6.7032e-03	6.7319e-03	7.0965e-03

близки, а переоценка колеблется в пределах 1.2–1.4. Таким образом, для данного приближенного решения метод работает ничуть не хуже, чем для галеркинской аппроксимации.

4.2. Пример 2

В L-образной области $\Omega_L := [0,1]^2 \setminus [1/2,1]^2$ решается задача

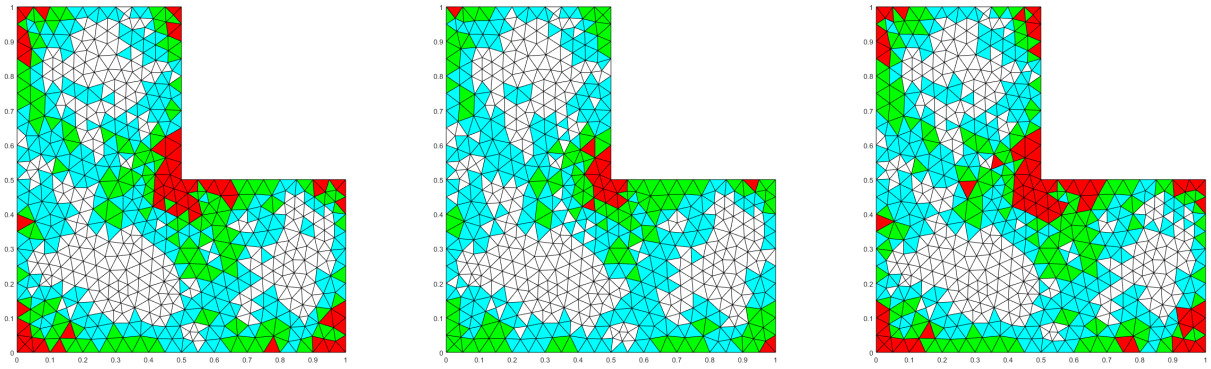
$$\Delta u + 1 = 0, \qquad u = 0 \ \text{ на } \ \partial\Omega.$$

В качестве точного решения u используется так называемое референсное решение (referenced solution), построенное на очень тонкой сетке, содержащей 226 837 узла. Результаты, относящиеся к двусторонним оценкам энергетической нормы ошибки приведены (в процентах) в табл. 6.

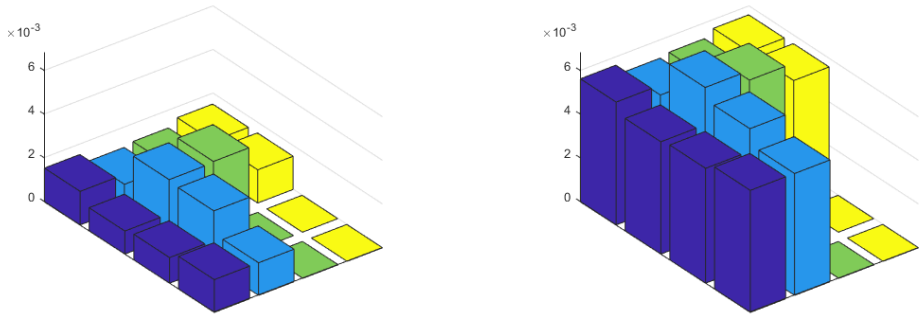
С помощью фиг. 4 можно сравнить работу индикаторов, генерируемых мажорантой (рисунок в центре) и минорантой (рисунок справа). Слева показано истинное распределение ошибки по области.

Таблица 6. Двусторонние оценки глобальной ошибки в примере 2

N_p	$\frac{M_{\ominus}}{\|\|\nabla u\|\|}$	$\frac{\|\|\nabla(u-v)\|\|}{\|\|\nabla u\|\|}$	$\frac{M_{\oplus}}{\|\|\nabla u\|\|}$
148	19.56	20.15	21.80
535	10.43	10.77	11.95
2102	5.80	6.10	6.80



Фиг. 4. Индикаторы распределения ошибок в примере 2.



Фиг. 5. Распределение локальных ошибок (регулярная сетка).

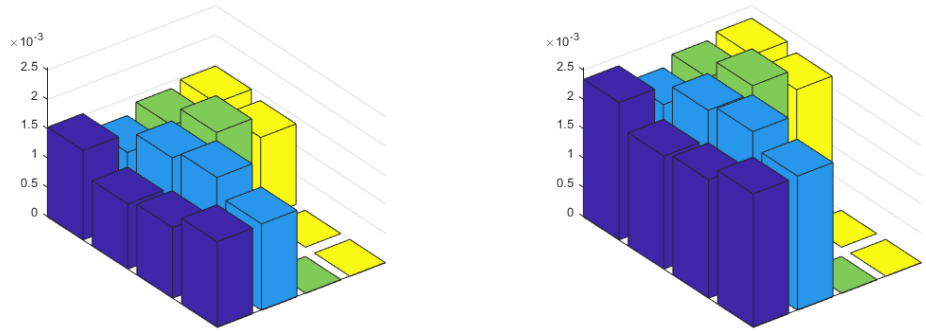
Обратимся к локальным ошибкам для подобластей ω_{ij}, аналогичных тем, что были рассмотрены в предыдущем примере (разница только в том, что теперь этих подобластей 12, а не 16). Определим значения $\|\nabla e\|_{\omega_{ij}}$ для конечноэлементного решения, построенного на сетке, составленной из примерно одинаковых треугольников с 2102 узлами (фиг. 7 слева). Такую квазиоднородную сетку будем называть регулярной. Гистограмма локальных ошибок показана в левой части фиг. 5. Значения ошибок, полученные с помощью мажоранты, показаны в правой части фиг. 5.

Видно, что в этом случае оценка превышает настоящее значение локальной ошибки в несколько раз. Индексы эффективности, которые показывают отношение мажоранты к норме ошибки, колеблются от 2.5 до 5.

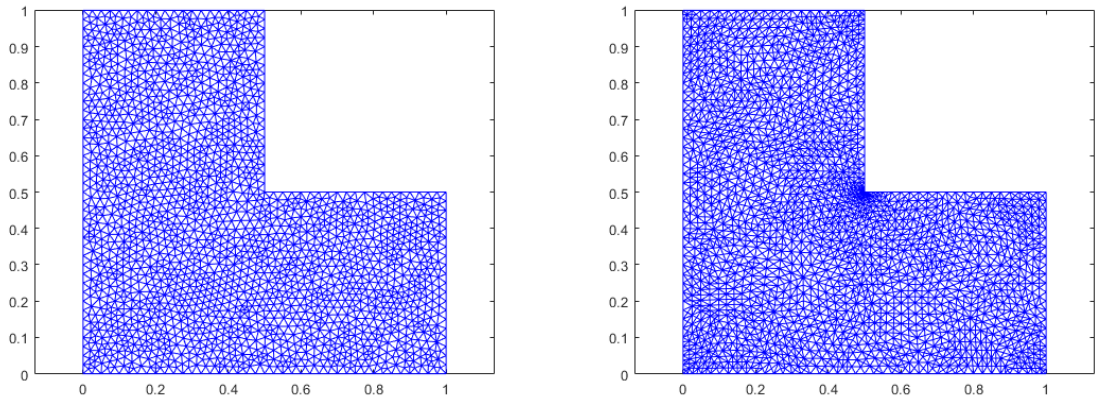
Этот эксперимент показывает, что для задач с особенностями в углах возможна значительная переоценка точной ошибки, если функция y^* определена недостаточно точно. Однако, надо заметить, что использование регулярной сетки в такой задаче довольно неестественно, и любой эффективный численный алгоритм будет строить сетки, адаптированные к сингулярности во внутреннем углу области. Задача была решена на одной из таких сеток, состоящей из 2696 узлов (фиг. 7 справа). Величины локальных ошибок и их оценки изображены в виде гистограмм на фиг. 6. Для адаптированной сетки точность локальных оценок значительно выше, чем для квазиоднородной, а соответствующие индексы эффективности незначительно колеблются около 1.5 для всех подобластей ω_{ij}. Вероятно, это связано с тем, что на адаптированной сетке функции w и y^* удалось построить значительно лучше.

4.3. Пример 3

В области Ω_L решается задача $\Delta u + 8\pi^2 \sin 2\pi x \sin 2\pi y = 0$ с нулевыми граничными условиями. В этом примере известно точное решение $u = \sin 2\pi x \sin 2\pi y$. Для мажоранты и миноранты энергетической нормы ошибки результаты (в процентах) приведены в табл. 7. В отличие от примера 2, здесь во внутреннем углу у решения особенности нет. Поэтому даже для решения, построенного на регулярной сетке (содержащей 2102 узлов) локальные оценки дают значения, приемлемые с точки зрения индекса эффективности. Рассмотрим локальные ошибки для подобластей ω_{ij}, что были рассмотрены в предыдущем примере. Вычислив значения $\|\nabla e\|_{\omega_{ij}}$ по



Фиг. 6. Распределение локальных ошибок (адаптированная сетка).



Фиг. 7. Конечноэлементные сетки в примере 3.

Таблица 7. Двусторонние оценки глобальной ошибки в примере 3

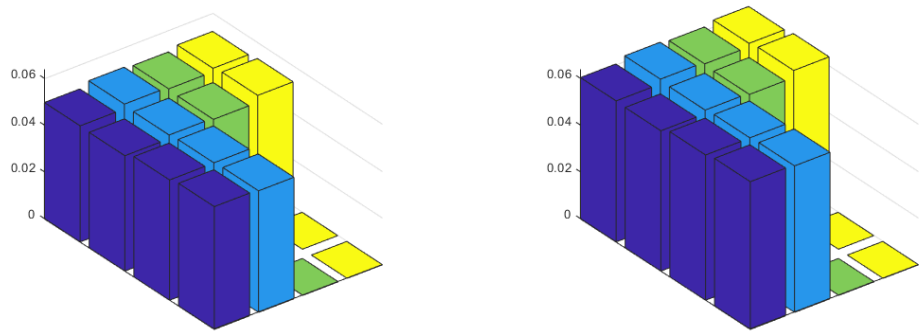
N_p	$\frac{M_\ominus}{\|\|\nabla u\|\|}$	$\frac{\|\|\nabla(u-v)\|\|}{\|\|\nabla u\|\|}$	$\frac{M_\oplus}{\|\|\nabla u\|\|}$
148	19.86	19.95	22.59
535	9.74	9.75	10.41
2102	4.87	4.87	5.04

конечноэлементному решению, построенному на равномерной сетке с 2102 узлами, получим гистограмму в левой части фиг. 8. Значения, полученные с помощью оценок локальных ошибок, показаны в правой части фиг. 8. Соответствующие индексы эффективности колеблются около 1.2.

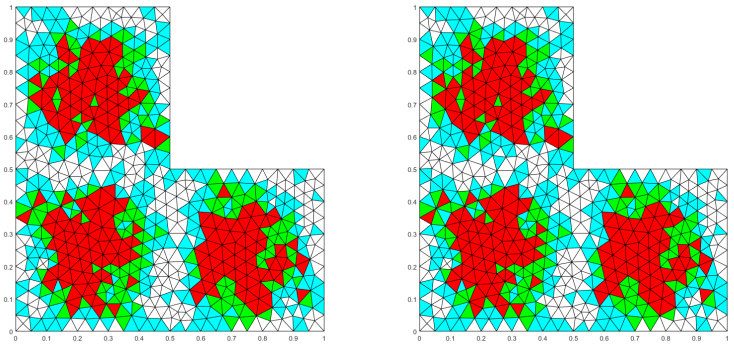
В табл. 8 представлены результаты для случая

$$\Phi(e) = \int\limits_\Omega |e|\, dx.$$

Здесь $S_1 = \Phi(w)$, $S_2 = C_F|\Omega|^{1/2}||\nabla(v+w) - y||_\Omega$, $S_3 = C_F^2|\Omega|^{1/2}||\mathrm{div}y + f||_\Omega$. На фиг. 9 слева маркировка элементов осуществлена по функции $\Phi(e)$, а справа – по S_1. Эти маркировки практически не отличаются друг от друга. Определенная симметрия картинки объясняется периодичностью решения, которая индуцирует аналогичную периодичность в разнице между решением и галеркинской аппроксимацией, если ошибка оценивается с помощью нормы, не содержащей производных.



Фиг. 8. Распределение локальных ошибок.



Фиг. 9. Индикатор ошибки по L_1-норме.

Таблица 8. Оценка для $\Phi(e) = \int\limits_\Omega |e|\,dx$

N_p	$\frac{\Phi(e)}{\|\|u\|\|},\%$	$\frac{S}{\|\|u\|\|},\%$	$\frac{S_1}{\|\|u\|\|},\%$	$\frac{S_2}{\|\|u\|\|},\%$	$\frac{S_3}{\|\|u\|\|},\%$
148	4.0386	13.5253	4.0097	3.8425	5.6731
535	0.9880	3.2617	0.9862	0.9208	1.3548
2102	0.2426	0.7995	0.2425	0.2178	0.3392

4.4. Пример 4

В заключение рассмотрим пример, где матрица A отлична от единичной и имеет существенно разные собственные значения. Здесь рассматривается задача

$$\operatorname{div}(A\nabla u) + 1 = 0 \quad \text{в} \quad \Omega = [0,1]^2$$

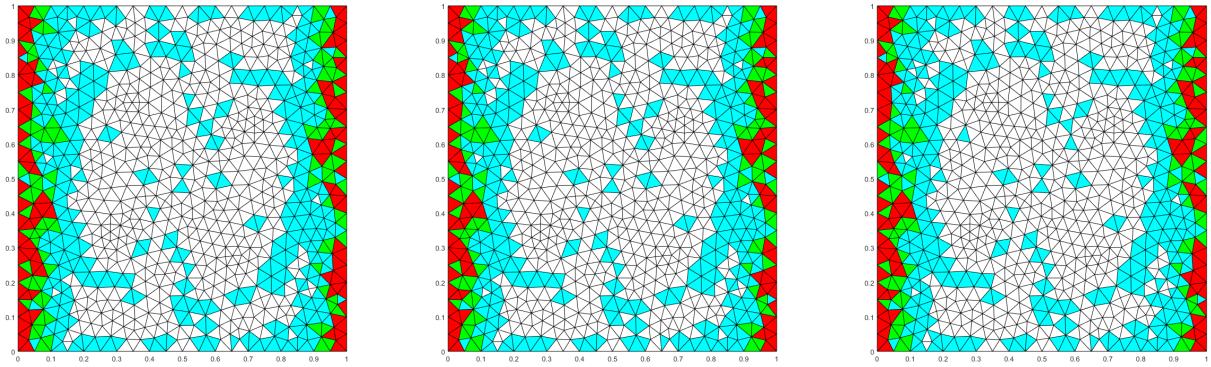
с нулевыми граничными условиями и $A = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$. В качестве точного решения u для вычислений используется референсное решение, т.е. галеркинская аппроксимация, построенная на симплициальной сетке с 299 313 узлами. Для мажоранты и миноранты энергетической нормы ошибки результаты (в процентах) приведены в табл. 9. Как и в предыдущих примерах, мажоранта и миноранта генерируют очень хорошие индикаторы распределения ошибки, выраженной в энергетической норме $\|\nabla e\|_A$ (см. фиг. 10). Рисунок слева отражает истинное распределение ошибки, рисунок в центре построен по $M_\oplus$, а справа – по $M_\ominus$.

В табл. 10 представлены результаты для

$$\Phi(e) = \left| \int\limits_\Omega \phi(x) e\,dx \right|, \qquad \phi(x) = \begin{cases} 1, & \text{если } x \in \omega, \\ 0, & \text{если } x \notin \omega, \end{cases}$$

Таблица 9. Двусторонние оценки глобальной ошибки в примере 4

N_p	$\frac{M_\ominus}{\|\|\nabla u\|\|}$	$\frac{\|\|\nabla(u-v)\|\|_A}{\|\|\nabla u\|\|}$	$\frac{M_\oplus}{\|\|\nabla u\|\|}$
177	33.00	33.18	33.22
727	17.22	17.25	17.25
2810	8.45	8.45	8.45



Фиг. 10. Индикаторы распределения ошибки в примере 4.

Таблица 10. Оценка для $\Phi(e) = \left|\int\limits_\Omega \phi(x)e\,dx\right|$

N_p	$\frac{\Phi(e)}{\|\|u\|\|}$,%	$\frac{S}{\|\|u\|\|}$,%	$\frac{S_1}{\|\|u\|\|}$,%	$\frac{S_2}{\|\|u\|\|}$,%	$\frac{S_3}{\|\|u\|\|}$,%
177	0.4395	2.4647	0.4349	2.0276	0.0022
727	0.1329	0.7012	0.1325	0.5686	0.0002
2810	0.0325	0.1807	0.0325	0.1482	0.0000

где $\omega = [0, 1/2]^2$.

Здесь, $S_1 = \Phi(w)$, $S_2 = C_F|\omega|^{1/2}||A\nabla(v+w) - y||_{A^{-1},\Omega}$ и $S_3 = C_F^2|\omega|^{1/2}||\mathrm{div}y + f||_\Omega$. Гарантированная оценка сверху переоценивает $\Phi(e)$ примерно в 5 раз, но индикатор (величина S_1) практически совпадает с $\Phi(e)$.

Суммируя результаты приведенных экспериментов, можно сказать, что апостериорные оценки функционального типа позволяют уверенно контролировать точность приближенного не только в энергетической норме, но и в терминах локальных норм и специально сконструированных функционалов.

СПИСОК ЛИТЕРАТУРЫ

1. *Bangerth W., Rannacher R.* Adaptive finite element methods for differential equations. Berlin: Birkhauser, 2003.
2. *Johnson C., Hansbo P.* Adaptive finite elements in computational mechanics // Comput. Methods Appl. Mech. Engrg. 1992. V. 101. P. 143–181.
3. *Johnson C., Szepessy A.* Adaptive finite element methods for conservation laws based on a posteriori error estimates // Commun. Pure and Appl. Math. 1995. V. XLVIII. P. 199–234.
4. *Mommer M.S., Stevenson R.* A goal-oriented adaptive finite element method with convergence rates // SIAM J. Numer. Anal. 2009. V. 47. № 2. P. 861—886.
5. *Stein E., Ruter M., Ohnimus S.* Error-controlled adaptive goal-oriented modeling and finite element approximations in elasticity // Comput. Meth. Appl. Mech. Engrg. 2007. V. 196. № 37—40. P. 3598—3613.
6. *Repin S. I.* A posteriori error estimation for variational problems with uniformly convex functionals // Math. Comput. 2000. V. 69. № 230. P. 481–500.

7. *Repin S.* A posteriori estimates for partial differential equations. Berlin: Walter de Gruyter GmbH & Co. KG, 2008.

8. *Репин С.И.* Тождество для отклонений от точного решения задачи $\Lambda^* \mathcal{A} \Lambda u + \ell = 0$ и его следствия // Ж. вычисл. матем. и матем. физ. 2021. Т. 61. № 12. С. 22–45.

9. *Repin S., Sauter S.* Accuracy of Mathematical Models. Dimension Reduction, Simplification, and Homogenization. EMS Tracts in Mathematics. Vol. 33. 2020.

10. *Репин С.И.* Оценки отклонения от точных решений краевых задач в мерах более сильных, чем энергетическая норма // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. № 5. С. 767–783

11. *Repin S.I.* A posteriori estimates in local norms // J.Math. Sci. 2004. V. 124 № 3. P. 5026–5035.

12. *Prager W., Synge J.L.* Approximations in elasticity based on the concept of functions space // Quart. Appl. Math. 1947. V. 5. P. 241–269.

13. *Mikhlin S.G.* Variational Methods in Mathematical Physics. Oxford: Pergamon Press, 1964.

A POSTERIORI ERROR ESTIMATES FOR APPROXIMATE SOLUTIONS OF ELLIPTIC BOUNDARY VALUE PROBLEMS IN TERMS OF LOCAL NORMS AND OBJECTIVE FUNCTIONALS

A. V. Muzalevsky[b], S. I. Repin[a,b,*], M. E. Frolov[b]

[a] *191023 St. Petersburg, Fontanka str., 27, St. Petersburg Department of the V.A. Steklov Mathematical Institute of the Russian Academy of Sciences, Russia*

[b] *195251 St. Petersburg, Polytechnic Str., 29, Peter the Great St. Petersburg Polytechnic University, Russia*

**e-mail: repin@pdmi.ras.ru*

Received: 07.08.2024
Revised: 07.08.2024
Accepted: 27.08.2024

Abstract. Functional relations have been obtained that allow us to evaluate the accuracy of approximate solutions in terms of measures significantly different from the energy norms that are usually used for these purposes. In particular, they are applicable to local norms and measures constructed using specially built linear functionals. The need for such precision control tools arises if there is a special interest in the behavior of the solution in some subdomain or in the special properties of the solution. It is shown that a posteriori functional-type estimates, which were previously used for global estimates, can be adapted to solve this problem. Functional identities and estimates are obtained that allow estimating the error of any conformal approximations in terms of a wide class of measures, including local norms and problem-oriented functionals. The theoretical results are verified in a series of examples that confirm the effectiveness of the proposed method.

Keywords: elliptic boundary value problems, a posteriori functional type estimates, local error estimates, goal-oriented accuracy control.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2286–2302

УДК 519.622

ОБОБЩЕНИЯ СТАДИЙНОГО ПОРЯДКА МЕТОДОВ РУНГЕ–КУТТЫ

© 2024 г. Л. М. Скворцов[1,*]

[1] 127051 Москва, Трубная ул., 51, корп. 1, ООО "ЗВ Сервис", Россия
**e-mail: lm.skvo@gmail.com*

Поступила в редакцию 27.05.2024 г.
Переработанный вариант 02.07.2024 г.
Принята к публикации 23.08.2024 г.

Рассматривается применение методов Рунге–Кутты для решения жестких систем обыкновенных дифференциальных уравнений и дифференциально-алгебраических уравнений . При решении таких задач часто проявляется эффект снижения порядка, когда при заданной точности реальный порядок метода оказывается ниже классического порядка, что неизбежно приводит к повышению вычислительных затрат. Чтобы избежать снижения порядка, метод должен иметь достаточно высокий стадийный порядок. Однако методы, обеспечивающие наиболее удобную и эффективную реализацию, имеют низкий стадийный порядок. Поэтому актуальна задача построения методов, которые при низком стадийном порядке обладают свойствами методов более высокого стадийного порядка. Настоящая статья посвящена построению методов такого типа. Рассматриваются однократно диагонально-неявные, явные и обратные к явным методы. Приведены результаты решения тестовых задач. Библ. 44. Фиг. 3. Табл. 4.

Ключевые слова: методы Рунге–Кутты, жесткие и дифференциально-алгебраические задачи, стадийный порядок, псевдостадийный порядок, слабый стадийный порядок, квазистадийный порядок.

DOI: 10.31857/S0044466924120055, **EDN**: KCDGDO

1. ВВЕДЕНИЕ

Один шаг численного решения задачи Коши для системы обыкновенных дифференциальных уравнений (ОДУ)

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0$$

выполняется согласно следующим формулам:

$$\mathbf{F}_i = \mathbf{f}\left(t_0 + c_i h,\ \mathbf{y}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{F}_j\right), \quad i = 1, \ldots, s, \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i \mathbf{F}_i, \tag{1.1}$$

где h – размер шага, s – число стадий. Коэффициенты метода обычно представляют в виде таблицы Бутчера

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^{\mathrm{T}} \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \cdots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}.$$

В методах с автоматическим выбором размера шага дополнительно вычисляют вектор

$$\hat{\mathbf{y}}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} \hat{b}_i \mathbf{F}_i, \tag{1.2}$$

позволяющий получить оценку локальной ошибки в виде нормы вектора $\mathbf{y}_1 - \hat{\mathbf{y}}_1$. Формулу (1.2) называют вложенной, а вместе с (1.1) – вложенной парой методов Рунге–Кутты. В таблице Бутчера коэффициенты $\hat{b}_i$ вложенной пары приводят под коэффициентами b_i.

Метод является явным, если $a_{ij} = 0$ при $j \geq i$, в противном случае он неявный. Неявный метод, имеющий $a_{ij} = 0$ при $j > i$, называется диагонально-неявным (DIRK). Среди методов DIRK обычно применяют однократно диагонально-неявные (SDIRK), у которых все диагональные элементы a_{ii} равны между собой, а также аналогичные методы с явной первой стадией (ESDIRK). Достоинством методов DIRK является то, что среди неявных методов они наиболее просто и эффективно реализуются.

Неявный метод называется жестко точным, если $b_i = a_{si}$, $i = 1, \ldots, s$. Стадийный порядок (СП) метода определяется как наибольшее целое q, для которого выполняются равенства

$$k\,\mathbf{A}\mathbf{c}^{k-1} = \mathbf{c}^k, \quad k\mathbf{b}^{\mathrm{T}}\mathbf{c}^{k-1} = 1, \qquad k = 1, \ldots, q$$

(здесь и далее предполагаем покомпонентное выполнение операций умножения и возведения в степень для векторов). СП явных методов и SDIRK не может быть больше 1, а методов ESDIRK – больше 2.

Основным показателем, характеризующим точность методов решения задачи Коши, является порядок сходимости p. При $h \to 0$ глобальная ошибка пропорциональна h^p. Для нежестких задач ошибки различных компонент решения, соответствующих элементарным дифференциалам, определяются через коэффициенты погрешности, задаваемые в виде

$$e(T_{ij}) = 1 - \gamma(T_{ij})\mathbf{b}^{\mathrm{T}}\,\Phi(T_{ij}),$$

где T_{ij} – корневое дерево i-го порядка (существует взаимно однозначное соответствие между корневыми деревьями и элементарными дифференциалами). Например, $e(T_{32}) = 1 - 6\mathbf{b}^{\mathrm{T}}\mathbf{A}\mathbf{c}$. Значения $\gamma(T_{ij})$ и $\Phi(T_{ij})$ для деревьев до 5-го порядка включительно приведены в [1], [2].

Для решения жестких задач обычно применяют неявные методы, которые обеспечивают устойчивость численного решения при достаточно большом размере шага. Но при этом реальный порядок сходимости может быть ниже классического порядка, что приводит к увеличению вычислительных затрат для достижения заданной точности. Чтобы избежать снижения порядка, следует использовать жестко точные методы, имеющие достаточно высокий СП (снижение порядка отсутствует при $q = p$ и небольшое при $q = p - 1$). Феномен снижения порядка впервые был объяснен с помощью уравнения Протеро–Робинсона

$$y' = \lambda(y - \varphi(t)) + \varphi'(t), \quad y(t_0) = \varphi(t_0), \tag{1.3}$$

(см. [3]), имеющего решение $y(t) = \varphi(t)$.

Исследование сходимости численного решения с помощью уравнения (1.3) выполнялось в [2]–[6] и многих других работах, где отмечалась важность высокого СП для эффективного решения жестких задач. В [7]–[24] предлагались методы, построенные на основе минимизации ошибки решения этого и некоторых других простейших уравнений. Такой подход позволяет уменьшить снижение порядка. Уравнение (1.3) является линейным, что не всегда позволяет объяснить поведение ошибки при решении нелинейных задач. В [2], [9], [11] приведены результаты решения жестких задач, которые невозможно объяснить ни в рамках классической теории, ни с помощью уравнения Протеро–Робинсона. В [2], [8], [11] были рассмотрены простейшие модельные уравнения, позволившие объяснить поведение ошибки в этих случаях. На их основе сформированы функции погрешности $e_{ij}(z)$, определенные как локальные относительные ошибки решения соответствующих модельных уравнений в зависимости от значения $z = h\lambda$. При заданном i число таких функций n_i равно числу различных корневых деревьев порядка i для методов Рунге–Кутты, применяемых к дифференциально-алгебраическим уравнениям (ДАУ) индекса 1 (см. [2], [5]). При $i = 2$ имеется одна такая функция, при $i = 3$ их две, а при $i = 4$ – пять. Функции погрешности задаются в виде

$$e_{ij}(z) = z\mathbf{b}^{\mathrm{T}}\,(\mathbf{I} - z\mathbf{A})^{-1}\,\mathbf{d}_{ij} + 1 - i\mathbf{b}^{\mathrm{T}}\mathbf{c}^{i-1}, \qquad j = 1, \ldots, n_i\,,$$

где

$$\mathbf{d}_{ij} = \frac{\gamma\,(T_{i+1,j})}{i+1}\mathbf{\Phi}\,(T_{i+1,j}) - i\mathbf{A}\mathbf{c}^{\,i-1}.$$

Жесткие модельные уравнения и соответствующие им значения $\mathbf{d}_{ij}$ для $i \leq 4$ приведены в табл. 1, где $\mathbf{e} = (1, \ldots, 1)^{\mathrm{T}}$, а переменные x_{ij} задаются следующими уравнениями:

$$x_1' = 1, \quad x_{21}' = x_1, \quad x_{31}' = x_1^2, \quad x_{32}' = x_{21}, \quad x_{ij}\,(0) = 0.$$

Используемые на практике методы удовлетворяют условию 1-го СП, поэтому уравнения относительно x_1 и y_1 решаются точно, а все уравнения относительно y_{i1} – это фактически уравнения (1.3) при $\varphi(t) = t^i$. Функции

Таблица 1. Жесткие модельные уравнения и значения $\mathbf{d}_{ij}$

i	j	Уравнение	$y_{ij}(t)$	$\mathbf{d}_{ij}$
1	1	$y'_1 = \lambda(y_1 - x_1) + 1$	t	$\mathbf{d}_1 = \mathbf{c} - \mathbf{Ae}$
2	1	$y'_{21} = \lambda(y_{21} - x_1^2) + 2t$	t^2	$\mathbf{d}_{21} = \mathbf{c}^2 - 2\mathbf{Ac}$
3	1	$y'_{31} = \lambda(y_{31} - x_1^3) + 3t^2$	t^3	$\mathbf{d}_{31} = \mathbf{c}^3 - 3\mathbf{Ac}^2$
	2	$y'_{32} = \lambda(y_{32} - x_1 x_{21}) + 3t^2/2$	$t^3/2$	$\mathbf{d}_{32} = 2\mathbf{c}(\mathbf{Ac}) - 3\mathbf{Ac}^2$
4	1	$y'_{41} = \lambda(y_{41} - x_1^4) + 4t^3$	t^4	$\mathbf{d}_{41} = \mathbf{c}^4 - 4\mathbf{Ac}^3$
	2	$y'_{42} = \lambda(y_{42} - x_1^2 x_{21}) + 2t^3$	$t^4/2$	$\mathbf{d}_{42} = 2\mathbf{c}^2(\mathbf{Ac}) - 4\mathbf{Ac}^3$
	3	$y'_{43} = \lambda(y_{43} - x_1 x_{31}) + 4t^3/3$	$t^4/3$	$\mathbf{d}_{43} = 3\mathbf{c}(\mathbf{Ac}^2) - 4\mathbf{Ac}^3$
	4	$y'_{44} = \lambda(y_{44} - x_1 x_{32}) + 2t^3/3$	$t^4/6$	$\mathbf{d}_{44} = 6\mathbf{c}(\mathbf{A}^2\mathbf{c}) - 4\mathbf{Ac}^3$
	5	$y'_{45} = \lambda(y_{45} - x_{21}^2) + t^3$	$t^4/4$	$\mathbf{d}_{45} = 4(\mathbf{Ac})^2 \breve{} 4\mathbf{Ac}^3$

$e_{i1}(z)$ получаем также и при разложении локальной ошибки численного решения уравнения (1.3) в ряд Тейлора. В [20]–[24] предлагались методы, удовлетворяющие условиям

$$e_{i1}(z) \equiv 0, \quad i = 1, \ldots, \tilde{q}, \quad \tilde{q} \leqslant p,$$

при $\tilde{q} > q$. Значение $\tilde{q}$ определено как слабый (weak) СП (ССП). Методы, удовлетворяющие в общем случае более строгим условиям

$$e_{ij}(z) \equiv 0, \quad i = 1, \ldots, \bar{q}, \quad j = 1, \ldots, n_i, \tag{1.4}$$

при $\bar{q} > q$, рассматривались в [2], [9]–[16], при этом значение $\bar{q}$ названо псевдостадийным порядком (ПСП). Справедливо неравенство $q \leqslant \bar{q} \leqslant \tilde{q} \leqslant p$ (в [2], [16] доказано, что из (1.4) следует $p \geqslant \bar{q}$). Мы рассматриваем методы, имеющие $\tilde{q}$, $\bar{q} > q$, и исследуем их свойства в сравнении с методами, имеющими такое же, как $\tilde{q}$ или $\bar{q}$, значение q. При построении таких методов удобно вместо условий $e_{ij}(z) \equiv 0$ использовать эквивалентные условия

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}^k\mathbf{d}_{ij} = 0, \quad k = 0, \ldots, s-1; \quad 1 - i\mathbf{b}^{\mathrm{T}}\mathbf{c}^{i-1} = 0, \tag{1.5}$$

а если метод жестко точный, то вместо (1.5) можно использовать условия

$$\mathbf{e}_s^{\mathrm{T}}\mathbf{A}^k\mathbf{d}_{ij} = 0, \quad \mathbf{e}_s = (0, \ldots, 0, 1)^{\mathrm{T}}, \quad k = 0, \ldots, s-1.$$

При $i \leq q$ имеем $e_{ij}(z) \equiv 0$, тогда соответствующие модельные уравнения решаются точно, а все функции $e_{q+1,\,j}(z)$ равны между собой. Поэтому если $\tilde{q} = q + 1$, то $\bar{q} = \tilde{q}$. В этом случае поведение ошибки решения жестких задач соответствует более высокому (на 1) СП. Методы, имеющие $\tilde{q} > \bar{q}$, наиболее эффективны при решении линейных задач, но они могут быть заметно менее эффективными при решении нелинейных задач.

Часто уравнения бывают представлены не в нормальной форме Коши, а в виде системы ДАУ, содержащей как дифференциальные, так и алгебраические уравнения. Будем рассматривать ДАУ, представленные в полуявной форме (форме Хессенберга), т.е. в виде

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{0} = \mathbf{g}(\mathbf{y}, \mathbf{z}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \quad \mathbf{z}(t_0) = \mathbf{z}_0.$$

Согласно приведенному в [5] определению, индекс дифференцирования системы ДАУ есть наименьшее число аналитических дифференцирований, требующихся для того, чтобы из исходных уравнений путем алгебраических преобразований можно было получить систему ОДУ в форме Коши. При этом каждое дифференцирование понижает индекс на 1.

Высокий СП важен и для эффективного решения ДАУ высших индексов (см. [2], [5], [11], [25]–[29]). Систему ДАУ индекса 2 (ДАУ2) можно представить в виде

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{0} = \mathbf{g}(\mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \quad \mathbf{z}(t_0) = \mathbf{z}_0,$$

где матрица $\mathbf{g_y f_z}$ обратима в окрестности решения. Предполагаем, что начальные условия согласованы, т. е. удовлетворяют алгебраической части и уравнению $\mathbf{0} = \mathbf{g_y f}(\mathbf{y}, \mathbf{z})$, полученному в результате дифференцирования алгебраической части. Систему ДАУ индекса 3 (ДАУ3) можно представить в виде

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, \mathbf{z}), \quad \mathbf{z}' = \mathbf{k}(\mathbf{y}, \mathbf{z}, \mathbf{u}), \quad \mathbf{0} = \mathbf{g}(\mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \quad \mathbf{z}(t_0) = \mathbf{z}_0, \quad \mathbf{u}(t_0) = \mathbf{u}_0,$$

где матрица $\mathbf{g_y f_z k_u}$ обратима в окрестности решения. Предполагаем, что начальные условия согласованы, т. е. удовлетворяют алгебраической части и уравнениям, полученным в результате однократного и двухкратного дифференцирования алгебраической части.

Предположим, что метод жестко точный и имеет обратимую матрицу $\mathbf{A}$. Сходимость решения ДАУ2 и ДАУ3 такими методами исследовалась в [5], [25], [26], где получены порядки сходимости для различных компонент. Аналогичные результаты получены в [27] для случая решения ДАУ2 жестко точными методами вида

$$\begin{array}{c|cc} 0 & 0 & \mathbf{0}^{\mathrm{T}} \\ \tilde{\mathbf{c}} & \tilde{\mathbf{a}} & \tilde{\mathbf{A}} \\ \hline & b_1 & \tilde{\mathbf{b}} \end{array}, \tag{1.6}$$

где матрица $\tilde{\mathbf{A}}$ обратима (к таким методам относятся ESDIRK). В табл. 2 приведены порядки сходимости различных компонент ДАУ, полученные как следствия теорем из этих работ при $q = 1,\ 2,\ 3$. При наличии некоторых дополнительных условий порядки могут быть выше. Например, если выполняется упрощающее предположение $D(\zeta)$, $\zeta \geq 1$, то значение p_y может быть больше, чем указано в табл. 2 (см. [26], [27]). Но для всех рассмотренных здесь методов это предположение не выполняется. Нам неизвестны теоретические результаты о сходимости методов вида (1.6) при решении ДАУ3, но численные эксперименты дают оценки порядков как в табл. 2.

Таблица 2. Порядки сходимости компонент ДАУ2 и ДАУ3

q	ДАУ2		ДАУ3		
	p_y	p_z	p_y	p_z	p_u
1	$\min(p,\ 2)$	1	—	—	—
2	$\min(p,\ 3)$	2	2	2	1
3	$\min(p,\ 4)$	3	3	3	2

В общем случае жестко точные методы, имеющие 1-й СП и обратимую матрицу $\mathbf{A}$, не обеспечивают сходимости при решении ДАУ3. В [2], [11] на основе модельных уравнений для ДАУ3 были получены необходимые условия сходимости таких методов:

$$\mathbf{e}_s^{\mathrm{T}}\mathbf{A}^{-2}\mathbf{c}^2 = 2, \quad \mathbf{e}_s^{\mathrm{T}}\mathbf{A}^{-2}\mathbf{c} = 0; \tag{1.7}$$

$$\mathbf{e}_s^{\mathrm{T}}\mathbf{A}^{-2}\mathbf{c}^2 = 2, \quad \mathbf{b}^{\mathrm{T}}\big(\mathbf{A}^{-2}\mathbf{c}^2\big)^j = 2^j, \quad j = 1,\ 2,\ 3,\ \ldots \tag{1.8}$$

(необходимо выполнение (1.7) либо (1.8)). Численные эксперименты показали, что выполнение (1.7) либо (1.8) обеспечивает сходимость всех компонент с 1-м порядком, а для сходимости со 2-м порядком y- и z-компонент необходимо также выполнение условия

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}\big(\mathbf{A}^{-2}\mathbf{c}^2\big)^j = 2^{j-1}, \quad j = 1,\ 2,\ 3,\ \ldots \tag{1.9}$$

В [2], [11], [29] были получены также и некоторые другие условия, необходимые для повышения порядков сходимости компонент ДАУ, но у нас нет доказательства, что эти условия (вместе с условиями порядка для ОДУ) также и достаточные.

Условия (1.8), (1.9) выполняются для методов 2-го СП, но 2-й ПСП не гарантирует их выполнения. В разд.3 статьи рассмотрен квазистадийный порядок, который позволяет обеспечить выполнение этих и некоторых других условий сходимости численного решения ДАУ.

В статье рассмотрены однократно диагонально-неявные, явные и обратные к явным методы, которые, имея 1-й СП, позволяют избежать снижения порядка при решении жестких задач. Эти методы имеют также повышенные порядки сходимости при решении ДАУ индексов 2 и 3. Для явных и обратных к ним методов мы приводим коэффициенты вложенной формулы. Условимся говорить, что метод имеет ПСП $\bar{q}$, если $\tilde{q} = \bar{q} > q$. А если $\tilde{q} > \bar{q} > q$, то будем указывать значения ПСП и ССП ($\bar{q}$ и $\tilde{q}$).

2. МЕТОДЫ SDIRK

При построении всех методов предполагаем, что они удовлетворяют условию 1-го СП: $\mathbf{Ae} = \mathbf{c}$. Жесткоточные методы SDIRK имеют таблицу Бутчера:

$$\begin{array}{c|ccccc} \gamma & \gamma & & & & \\ c_2 & a_{21} & \gamma & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_{s-1} & a_{s-1,1} & a_{s-1,2} & \cdots & \gamma & \\ 1 & b_1 & b_2 & \cdots & b_{s-1} & \gamma \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & \gamma \end{array}.$$

Рассмотрим сначала методы, имеющие $p = \bar{q} = 2$. Для них должны выполняться условия

$$\mathbf{e}_s^{\mathrm{T}}\mathbf{A}^i\left(\mathbf{c}^2 - 2\mathbf{Ac}\right) = 0, \quad i = 0,\ \ldots,\ s-1. \tag{2.1}$$

Примем $\bar{\mathbf{A}} = \mathbf{A} - \gamma\mathbf{I}$, $\mathbf{I} = \operatorname{diag}(\mathbf{e})$. Тогда условия (2.1) эквивалентны условиям

$$\mathbf{e}_s^{\mathrm{T}}\bar{\mathbf{A}}^i\left(\mathbf{c}^2 - 2\mathbf{Ac}\right) = 0, \quad i = 0,\ \ldots,\ s-1. \tag{2.2}$$

При $i = s - 1$ получаем

$$\mathbf{e}_s^{\mathrm{T}}\bar{\mathbf{A}}^{s-1}\left(\mathbf{c}^2 - 2\mathbf{Ac}\right) = -b_{s-1}a_{s-1,s-2}\cdots a_{21}\gamma^2 = 0, \tag{2.3}$$

откуда следует, что один из сомножителей в (2.3) должен быть равен 0. Это не может быть b_{s-1} или a_{21}, поскольку в этом случае одна из стадий становится лишней и ее можно исключить. Поэтому число стадий методов, удовлетворяющих условиям (2.2), должно быть не менее 4-х.

При $s = 4$ условия (2.2) могут быть выполнены, если только $\gamma = 1 \mp \sqrt{2}/2$ (как у 2-стадийного SDIRK 2-го порядка), при этом меньшее значение γ обеспечивает лучшую точность. При свободных коэффициентах c_2 и c_3 остальные коэффициенты находим по формулам

$$\gamma = 1 \mp \frac{\sqrt{2}}{2}, \quad a_{32} = 0, \quad b_2 = \frac{\gamma^2(1-\gamma)}{(c_3-\gamma)(c_3-c_2)}, \quad b_3 = \frac{\gamma^2(1-\gamma)}{(c_2-\gamma)(c_2-c_3)},$$
$$a_{21} = c_2 - \gamma, \quad a_{31} = c_3 - \gamma, \quad b_1 = 1 - b_2 - b_3 - \gamma.$$

Потребовав также выполнения условий (1.8), (1.9), получим метод

$$\begin{array}{c|cccc} \gamma & \gamma & & & \\ 0 & -\gamma & \gamma & & \\ 2\gamma & \gamma & 0 & \gamma & \\ 1 & 0 & \beta & \beta & \gamma \\ \hline 1 & 0 & \beta & \beta & \gamma \end{array}, \quad \gamma = 1 - \frac{\sqrt{2}}{2}, \quad \beta = \frac{1-\gamma}{2}. \tag{2.4}$$

По своим свойствам он очень похож на метод TR-BDF2:

$$\begin{array}{c|ccc} 0 & 0 & & \\ 2\gamma & \gamma & \gamma & \\ 1 & \beta & \beta & \gamma \\ \hline 1 & \beta & \beta & \gamma \end{array}$$

(см. [2], [30]). Оба метода имеют одинаковую функцию устойчивости и одинаковые коэффициенты погрешности 3го порядка. Отличие в том, что метод (2.4) имеет обратимую матрицу $\mathbf{A}$ и все его внутренние стадии L-устойчивы, тогда как 2-я стадия TR-BDF2 только A-устойчива. При решении жестких задач оба метода свободны от снижения порядка и показывают близкие результаты, а при решении ДАУ2 и ДАУ3 обеспечивают сходимость с порядками $p_y = p_z = 2, p_u = 1$.

При $s = 5$ можно построить метод 3-го порядка, имеющий $\bar{q} = 2$, если γ – корень многочлена $1 - 9z + 18z^2 - 6z^3$. Метод будет L-устойчивым, если $\gamma = 0.435866\ldots$ При заданном γ такие методы образуют 4-параметрическое семейство со свободными коэффициентами c_2, c_3, c_4, b_4. Остальные коэффициенты находим по формулам

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \gamma & c_2 & c_3 \\ \gamma^2 & c_2^2 & c_3^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 - b_4 - \gamma \\ 1/2 - b_4c_4 - \gamma \\ 1/3 - b_4c_4^2 - \gamma \end{bmatrix}, \quad a_{42} = \frac{\gamma^2\left(1 - 4\gamma + 2\gamma^2\right)}{2b_4(c_2-\gamma)(c_2-c_3)}, \tag{2.5}$$
$$a_{43} = a_{42}\frac{\gamma - c_2}{c_3 - \gamma}, \quad a_{32} = 0, \quad a_{21} = c_2 - \gamma, \quad a_{31} = a_3 - \gamma, \quad a_{41} = c_4 - a_{42} - a_{43} - \gamma.$$

Подбором c_2 и b_4 можно обеспечить 3-й ССП метода, тогда коэффициенты c_3 и c_4 остаются свободными. Мы задали их исходя из минимизации величины $\max|a_{ij}|$ при ограничениях $0 \le c_i \le 1$. В результате получили

$$c_3 = 1, \quad c_4 = 0, \quad c_2 = 4\gamma - c_3, \quad b_4 = \frac{2\left(1 - 8\gamma + 12\gamma^2\right) + 3c_2c_3\left(1 - 4\gamma + 2\gamma^2\right)}{6\left(c_4 - \gamma\right)\left(c_4 - c_2\right)\left(c_4 - c_3\right)}. \tag{2.6}$$

Метод, задаваемый формулами (2.5), (2.6), свободен от снижения порядка при решении линейных жестких задач, но при решении нелинейных задач возможно небольшое снижение порядка.

Метод (2.5), (2.6) не обеспечивает сходимости при решении ДАУ3, поскольку условия (1.8), (1.9) не выполняются. Эти условия будут выполняться, если задать $c_2 = 0$, $c_3 = 2\gamma$ и выбрать b_4 из условия $b_1 = 0$. Коэффициент c_4 задаем таким, чтобы выполнялось первое из условий

$$\mathbf{b}^{\mathrm{T}}\left(\mathbf{c}\left(\mathbf{A}^{-1}\mathbf{c}^2 - 2\mathbf{A}\mathbf{c}\right)\right) = 2/3, \quad \mathbf{b}^{\mathrm{T}}\left(\mathbf{A}^{-1}\mathbf{c}^2 - 2\mathbf{A}\mathbf{c}\right)^2 = 4/3, \tag{2.7}$$

необходимых для сходимости укомпоненты ДАУ2 с 3-м порядком (см. [2], [11], [29]) . В результате получен метод 2-го ПСП с коэффициентами (2.5),

$$c_2 = 0, \quad c_3 = 2\gamma, \quad c_4 = \frac{2 - 9\gamma + 6\gamma^2}{3\left(1 - 4\gamma + 2\gamma^2\right)}, \quad b_4 = \frac{1 - 6\gamma + 6\gamma^2}{3c_4\left(c_4 - 2\gamma\right)}, \tag{2.8}$$

который можно использовать для решения ДАУ3. Численные эксперименты показали, что при решении ДАУ3 он обеспечивает сходимость с порядками $p_y = p_z = 2$, $p_u = 1$ (такими же, как и методы 2-го СП). При решении ДАУ2 также получаем $p_y = p_z = 2$, хотя методы 2-го СП при $p \ge 3$ обеспечивают $p_y = 3$. Этот факт объясняется невыполнением 2-го условия в (2.7).

Таким образом, повышение ПСП приводит к эффекту, аналогичному повышению СП при решении жестких ОДУ, но для решения ДАУ2 и ДАУ3 необходимо выполнение дополнительных условий. Нам не удалось построить метод SDIRK, обладающий всеми свойствами методов 3-го СП. Методы ESDIRK 2-го СП с такими свойствами были построены в [2], [15]. В этой статье мы специально ограничились рассмотрением методов 1-го СП, для которых проще и удобнее показать основные принципы построения методов повышенного ПСП.

3. КВАЗИСТАДИЙНЫЙ ПОРЯДОК

В [31] было введено понятие квазистадийного порядка (КСП), согласно которому метод имеет прямой (forward) КСП $\hat{q}$, если при $q \ge 1$ выполняются условия

$$\mathbf{b}\left(\mathbf{c}^i - i\mathbf{A}\mathbf{c}^{i-1}\right) = (0, \ldots, 0)^{\mathbf{T}}, \quad i = 2, \ldots, \hat{q}, \tag{3.1}$$

и обратный (inverse) КСП $\hat{q}$ при выполнении условий

$$\mathbf{b}\left(\mathbf{A}^{-1}\mathbf{c}^i - i\mathbf{c}^{i-1}\right) = (0, \ldots, 0)^{\mathbf{T}}, \quad i = 2, \ldots, \hat{q} \tag{3.2}$$

(умножение векторов выполняем покомпонентно). Эти условия означают, что если $b_j \ne 0$, то j-й компонент вектора в скобках равен 0. Повышенный ($\hat{q} > 1$) прямой КСП обычно используется в явных методах Рунге–Кутты высоких порядков, упрощая их построение. Например, метод Дорманда–Принса 5-го порядка имеет 3-й прямой КСП. В [31] построен 4-стадийный метод SDIRK 3го порядка, удовлетворяющий условиям (3.1), (3.2) при $\hat{q} = 2$.

В [2], [11], [29] на основе модельных уравнений были получены необходимые условия для повышения p_y, p_z на 1 при решении ДАУ2 методами 1-го СП. При выполнении условий (3.1), (3.2) для $\hat{q} = 2$ эти условия также будут выполняться. Эксперименты показали, что метод из [31] действительно обеспечивает сходимость как у методов 2-го СП при решении ДАУ2, но при решении ДАУ3 сходимость отсутствует.

На основе условий, аналогичных (3.1), (3.2), рассмотрим построение методов SDIRK 3-го порядка, которые обеспечивают сходимость как у методов 2-го СП также и для ДАУ3. Для этого потребуем выполнения условий

$$\mathbf{b}\left(\mathbf{A}^k\mathbf{c}^2 - 2\mathbf{A}^{k+1}\mathbf{c}\right) = (0, \ldots, 0)^{\mathrm{T}}, \quad k = 0, -1, -2; \tag{3.3а}$$

$$1 - 3\mathbf{b}^{\mathrm{T}}\mathbf{c}^2 = 0. \tag{3.3б}$$

Условия (3.3а) можно рассматривать как обобщение условий 2-го КСП для ДАУ3. При $k = 0$ условия (3.3) обеспечивают 3-й порядок метода. При $k = -1$ обеспечиваются условия (2.7), а при $k = -2$ – условия (1.8). Для обеспечения 2-го порядка y- и z-компонент ДАУ3 дополнительно необходимо выполнение условий (1.9).

При $s = 5$ мы нашли пять методов SDIRK, удовлетворяющих условиям (3.3), (1.9), все они имеют $\gamma = 1/4$, $c_4 = 1/3$, $\mathbf{b} = (0,\ 0,\ 0,\ 3/4,\ 1/4)^{\mathrm{T}}$. По результатам экспериментов мы убедились, что эти методы обеспечивают такие же порядки сходимости компонент ДАУ2 и ДАУ3, как методы 2-го СП. Приводим один из них, который является L-устойчивым и имеет малые значения коэффициентов ($\max|a_{ij}| = 0.75$):

$$\begin{array}{c|ccccc} 1/4 & 1/4 & & & & \\ 1/2 & 1/4 & 1/4 & & & \\ 31/40 & 63/400 & 147/400 & 1/4 & & \\ 1/3 & 25/189 & 1/12 & -25/189 & 1/4 & \\ 1 & 0 & 0 & 0 & 3/4 & 1/4 \\ \hline & 0 & 0 & 0 & 3/4 & 1/4 \end{array} \tag{3.4}$$

Более значимых результатов удалось добиться для жестко точных методов ESDIRK, таблица Бутчера которых имеет вид (1.6). Такие методы имеют 2-й СП, а условия 3-го КСП для ДАУ3 запишутся в виде

$$\tilde{\mathbf{b}}\left(\tilde{\mathbf{A}}^k\tilde{\mathbf{c}}^3 - 3\tilde{\mathbf{A}}^{k-1}\tilde{\mathbf{c}}^2\right) = (0,\ \ldots,\ 0)^{\mathrm{T}}, \quad k = 0, -1, -2. \tag{3.5}$$

На этот раз выполнение (3.5) гарантирует также и выполнение условий

$$\tilde{\mathbf{b}}^{\mathrm{T}}\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{c}}^3 = 3, \quad \tilde{\mathbf{b}}^{\mathrm{T}}\tilde{\mathbf{A}}^{-3}\tilde{\mathbf{c}}^3 = 6, \quad \tilde{\mathbf{b}}^{\mathrm{T}}\left(\tilde{\mathbf{c}}\left(\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{c}}^3\right)\right) = 2, \quad \tilde{\mathbf{b}}^{\mathrm{T}}\left(\tilde{\mathbf{A}}^{-2}\tilde{\mathbf{c}}^3\right)^2 = 12, \quad \tilde{\mathbf{b}}^{\mathrm{T}}\left(\tilde{\mathbf{c}}\left(\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{c}}^3\right)\right) = 3/4,$$

необходимых (наряду с условиями классического порядка) для обеспечения таких же порядков сходимости компонент ДАУ2 и ДАУ3, как у методов 3-го СП. В [2], [29] рассмотрено 1параметрическое семейство 6стадийных методов ESDIRK 4-го порядка, удовлетворяющих условиям (3.5). Там же приведены результаты решения тестовых задач, подтверждающие заявленные порядки сходимости.

4. ЯВНЫЕ МЕТОДЫ

Классические явные методы малопригодны для решения жестких задач, но они могут успешно решать умеренно жесткие задачи. В настоящее время известны также специальные явные методы, которые позволяют эффективно решать задачи с вещественным жестким спектром матрицы Якоби. К ним относятся стабилизированные методы с расширенными областями устойчивости (см. [2], [5], [12], [32], [33]), а также адаптивные нелинейные методы, осуществляющие самонастройку расчетной схемы на основе полученных на предварительных стадиях оценок больших по модулю собственных чисел матрицы Якоби (см. [2], [13]).

Явные методы Рунге–Кутты имеют только 1-й СП, что затрудняет построение методов повышенной точности для жестких и умеренно жестких задач. Но можно повысить точность, используя методы, обладающие свойствами методов более высокого СП. Впервые явные методы Рунге–Кутты, имеющие повышенный ПСП ($\bar{q} = 2,\ 3$), были предложены в [9]. На их основе в [2], [12], [13], [15] построены явные стабилизированные и адаптивные методы 3-го и 4-го порядков. В [23] построены явные методы, имеющие значения p и $\tilde{q}$ вплоть до 5, но при этом $\bar{q} = 2$.

Построение методов, имеющих $\tilde{q} > 1$, основано на соотношениях

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}^k(\mathbf{c}^i - i\mathbf{A}\mathbf{c}^{i-1}) = 0 \quad \text{при} \quad k < p - i, \tag{4.1а}$$

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}^l\,\mathbf{c}^j = 0 \quad \text{при} \quad l > s - 2, \tag{4.1б}$$

где (4.1а) следует из условий порядка, а (4.1б) – из структуры $\mathbf{A}$ и $\mathbf{c}$. Из (1.5), (4.1) следует, что методы, имеющие $\tilde{q} > 1$, должны удовлетворять условиям

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}^k\left(\mathbf{c}^i - \mathbf{A}\mathbf{c}^{i-1}\right) = 0, \quad k = p - i,\ \ldots,\ s - 2, \quad i = 1,\ \ldots,\ \tilde{q}. \tag{4.2}$$

Найдем минимальное число стадий, необходимое (но необязательно достаточное) для построения метода порядка p при $\tilde{q} = \bar{q} = 2$. При $s = p$ таких методов нет, поскольку $\mathbf{b}^{\mathrm{T}}\mathbf{A}^{s-2}\,(\mathbf{c}^2 - 2\mathbf{A}\mathbf{c}) = c_2\,/\,p! \neq 0$. При $s = p + 1$ из (4.2) получаем условия

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}^k\left(\mathbf{c}^2 - \mathbf{A}\mathbf{c}\right) = 0, \quad k = s - 3,\ s - 2.$$

С учетом (4.1б) эти условия запишутся в виде

$$\begin{aligned} &\mathbf{b}^{\mathrm{T}}\mathbf{A}^{s-3}\left(\mathbf{c}^2 - 2\mathbf{A}\mathbf{c}\right) = \mathbf{b}^{\mathrm{T}}\mathbf{A}^{s-3}\mathbf{c}^2 = b_{s-1}a_{s-1,s-2}\cdots a_{32}c_2^2 + b_s a_{s,s-1}\cdots a_{54}\left(a_{42}c_2^2 + a_{43}c_3^2\right) = 0, \\ &\mathbf{b}^{\mathrm{T}}\mathbf{A}^{s-2}\left(\mathbf{c}^2 - 2\mathbf{A}\mathbf{c}\right) = \mathbf{b}^{\mathrm{T}}\mathbf{A}^{s-2}\mathbf{c}^2 = b_s a_{s,s-1}\cdots a_{32}c_2^2 = 0. \end{aligned} \tag{4.3}$$

Учтем также одно из условий порядка $p = s - 1$:

$$\mathbf{b}^{\mathrm{T}}\mathbf{A}^{s-3}\mathbf{c} = b_{s-1}a_{s-1,s-2}\cdots a_{32}c_2 + b_s a_{s,s-1}\cdots a_{54}\,(a_{42}c_2 + a_{43}c_3) = 1/p! \neq 0. \tag{4.4}$$

Из (4.3), (4.4) видно, что эти условия будут выполняться, только если

$$a_{32} = 0,\ \ a_{42}c_2^2 + a_{43}c_3^2 = 0 \tag{4.5}$$

(при $s = 3$ принимаем $a_{42} = b_2$, $a_{43} = b_3$). В [9] приведены методы порядков 3, 4 и 5, имеющие $s = p + 1$, $\bar{q} = 2$.

Рассмотрим построение методов, имеющих заданные значения s, p и $\tilde{q}$. В [23] доказано, что ССП таких методов удовлетворяет неравенству $\tilde{q} \leqslant s - p + 1$, а при $\tilde{q} = s - p + 1$ функция устойчивости метода имеет вид $R(z) = 1 + z + \cdots + z^p/p!$. В [23] приведены также некоторые варианты структуры матрицы $\mathbf{A}$ при заданном $\tilde{q}$. Мы используем структуру, основанную на выполнении равенств

$$\begin{aligned} &a_{ij} = 0, \quad i = 3,\ \ldots,\ \tilde{q}+1, \quad j = 2,\ \ldots,\ i-1,\\ &\sum_{j=2}^{i-1} a_{ij}c_j^k = 0, \quad i = \tilde{q}+2,\ \ldots,\ \tilde{q}+k, \quad k = 2,\ \ldots,\ \tilde{q} \end{aligned} \tag{4.6}$$

(если $i = s + 1$, то принимаем $a_{ij} = b_j$). При $\tilde{q} > 2$ построенный на основе (4.6) метод в общем случае имеет $\bar{q} < \tilde{q}$. Чтобы повысить значение $\bar{q}$, следует обеспечить выполнение некоторых дополнительных условий.

Рассмотрим сначала построение метода, имеющего $p = s - 1 = 3$ и $\bar{q} = 2$. В этом случае, кроме (4.5), должны выполняться условия 3-го порядка:

$$b_2c_2 + b_3c_3 + b_4c_4 = 1/2, \quad b_2c_2^2 + b_3c_2^2 + b_4c_4^2 = 1/3, \quad b_4\,(a_{42}c_2 + a_{43}c_3) = 1/6. \tag{4.7}$$

При свободных коэффициентах c_2, c_3, c_4, b_4 коэффициенты a_{42}, a_{43}, b_2, b_3 находим из (4.5), (4.7), а коэффициенты a_{i1}, b_1 находим из условия 1-го СП.

Построим два метода, удовлетворяющих (4.5), (4.7). В первом из них потребуем, чтобы метод имел вложенную формулу 2-го порядка, совпадающую с 4-й стадией и имеющую $\bar{q} = 2$. Тогда к условиям (4.7) добавляются соотношения $c_4 = 1$, $a_{42}c_2 + a_{43}c_3 = 1/2$, откуда, учитывая (4.7), получаем $b_4 = 1/3$. Исходя из условия минимизации коэффициентов погрешности, принимаем также $c_2 = 1/2$, $c_3 = 1$, тогда $e(T_{4i}) = (0,\ -1/3,\ 1,\ 1)$. В результате получен метод

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1 & 1 & 0 & & \\ 1 & -1/2 & 2 & -1/2 & \\ \hline b_i & 1/6 & 2/3 & -1/6 & 1/3 \\ \hat{b}_i & -1/2 & 2 & -1/2 & 0 \end{array}. \tag{4.8}$$

Если принять $c_4 = 0$, то получим метод

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1 & 1 & 0 & & \\ 0 & -3/2 & 2 & -1/2 & \\ \hline b_i & -1/6 & 2/3 & 1/6 & 1/3 \\ \hat{b}_i & -1/2 & 2 & -1/2 & 0 \end{array}, \tag{4.9}$$

имеющий $e_{32}(z) \equiv 0$ и $e(T_{4i}) = (0,\ 1,\ 1,\ 1)$. Т. е. выполняется условие (1.5) 3-го ПСП ($i = 3$) для $j = 2$ (но не для $j = 1$). Благодаря этому методу (4.9) может иметь преимущество по сравнению с (4.8) при решении нелинейных задач.

Рассмотрим теперь методы, имеющие $p = \bar{q} = 3$. В этом случае минимальное число стадий $s = p + 2 = 5$. Такие методы имеют свободные коэффициенты c_2, c_3, c_4, b_5, а остальные коэффициенты находим по формулам

$$\begin{aligned} &a_{32} = a_{42} = a_{43} = 0, \quad c_5 = 0,\\ &\begin{bmatrix} a_{52} \\ a_{53} \\ a_{54} \end{bmatrix} = \mathbf{C}^{-1}\begin{bmatrix} 1/(6b_5) \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} b_2 \\ b_3 \\ b_4 \end{bmatrix} = \mathbf{C}^{-1}\begin{bmatrix} 1/2 \\ 1/3 \\ 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} c_2 & c_3 & c_4 \\ c_2^2 & c_3^2 & c_4^2 \\ c_2^3 & c_3^3 & c_4^3 \end{bmatrix} \end{aligned} \tag{4.10}$$

и из условия 1-го СП. Приняв $c_2 = 1/3$, $c_3 = 2/3$, $c_4 = b_5 = 1$, получим метод

$$\begin{array}{c|ccccc} 0 & & & & & \\ 1/3 & 1/3 & & & & \\ 2/3 & 2/3 & 0 & & & \\ 1 & 1 & 0 & 0 & & \\ 0 & -11/12 & 3/2 & -3/4 & 1/6 & \\ \hline b_i & 1/4 & -3 & 15/4 & -1 & 1 \\ \hat{b}_i & 1/12 & -7/4 & 9/2 & 1/2 & 0 \end{array}. \tag{4.11}$$

Этот метод имеет $e(T_{4i}) = (1,\ 1,\ 1,\ 1)$.

Повысим порядок метода до 4-го при $\bar{q} = 3$, тогда $s = p + 2 = 6$. При свободных коэффициентах c_2, c_3, c_4, c_6, a_{65}, b_6 остальные коэффициенты вычисляем по формулам

$$a_{32} = a_{42} = a_{43} = 0, \quad c_5 = 0, \quad b_5 = b_6 a_{65}\,(4 - 3/c_6)\,,$$
$$\begin{bmatrix} a_{52} \\ a_{53} \\ a_{54} \end{bmatrix} = \mathbf{C}^{-1} \begin{bmatrix} 1/(24 b_6 a_{65}) \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} a_{62} \\ a_{63} \\ a_{64} \end{bmatrix} = \mathbf{C}^{-1} \begin{bmatrix} 1/(8 b_6 c_6) \\ 1/(12 b_6) \\ 0 \end{bmatrix}, \quad \begin{bmatrix} b_2 \\ b_3 \\ b_4 \end{bmatrix} = \mathbf{C}^{-1} \begin{bmatrix} 1/2 - b_6 c_6 \\ 1/3 - b_6 c_6^2 \\ 1/4 - b_6 c_6^3 \end{bmatrix}, \tag{4.12}$$

где матрица $\mathbf{C}$ такая же, как в (4.10). Приведем один из методов этого семейства:

$$\begin{array}{c|cccccc} 0 & & & & & & \\ 1/3 & 1/3 & & & & & \\ 2/3 & 2/3 & 0 & & & & \\ 1 & 1 & 0 & 0 & & & \\ 0 & -11/12 & 3/2 & -3/4 & 1/6 & & \\ 1 & 1/4 & -3 & 15/4 & -1 & 1 & \\ \hline b_i & -1/8 & 3/8 & 3/8 & -1/8 & 1/4 & 1/4 \\ \hat{b}_i & 1/4 & -3 & 15/4 & -1 & 1 & 0 \end{array}. \tag{4.13}$$

Отметим, что вложенный метод в (4.13) совпадает с основным методом в (4.11).

Наконец, при $s = 7$ построим метод 4-го порядка, имеющий $\bar{q} = 3$, $\tilde{q} = 4$ (значения $\bar{q} = 4$ нам добиться не удалось). В этом случае свободными являются коэффициенты c_2, c_3, c_4, c_5, c_7, a_{76}, b_7, а остальные коэффициенты вычисляем по формулам

$$a_{32} = a_{42} = a_{43} = a_{52} = a_{53} = a_{54} = 0, \quad c_6 = 0, \quad b_6 = b_7 a_{76}\,(4 - 3/c_7)\,,$$
$$\begin{bmatrix} a_{62} \\ a_{63} \\ a_{64} \\ a_{65} \end{bmatrix} = \bar{\mathbf{C}}^{-1} \begin{bmatrix} 1/(24 b_7 a_{76}) \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} a_{72} \\ a_{73} \\ a_{74} \\ a_{75} \end{bmatrix} = \bar{\mathbf{C}}^{-1} \begin{bmatrix} 1/(8 b_7 c_7) \\ 1/(12 b_7) \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \bar{\mathbf{C}}^{-1} \begin{bmatrix} 1/2 - b_7 c_7 \\ 1/3 - b_7 c_7^2 \\ 1/4 - b_7 c_7^3 \\ -b_7 c_7^4 \end{bmatrix}, \tag{4.14}$$
$$\bar{\mathbf{C}} = [c_{ij}], \quad c_{ij} = c_{j+1}^i, \quad i, j = 1,\ \ldots,\ 4.$$

При любых допустимых значениях свободных коэффициентов полученного метода все элементы векторов $\mathbf{d}_{21}$, $\mathbf{d}_{31}$, $\mathbf{d}_{32}$, $\mathbf{d}_{41}$ из табл. 1 равны 0, а ненулевые элементы векторов $\mathbf{d}_{42},\ \ldots\ ,\ \mathbf{d}_{45}$ имеют следующие значения:

$$d_{421} = d_{431} = d_{441} = c_7/4, \quad d_{451} = \frac{c_7\,(4c_7 - 3) + 9a_{76}}{144 b_7 c_7^2 a_{76}}, \quad d_{452} = \frac{1}{144 b_7 a_{76}}.$$

Выбрав $a_{76} = c_7\,(3 - 4c_7)/9$, мы обнулим d_{451}. Уменьшая c_7 и увеличивая b_7, можно добиться сколь угодно малых значений оставшихся ненулевых элементов, но при этом получаем очень большие коэффициенты метода, что недопустимо. Поэтому мы задали компромиссные значения $c_7 = 1/8$, $b_7 = 8$. В результате получен метод

$$\begin{array}{c|ccccccc} 0 & & & & & & & \\ 1/4 & 1/4 & & & & & & \\ 1/2 & 1/2 & 0 & & & & & \\ 3/4 & 3/4 & 0 & 0 & & & & \\ 1 & 1 & 0 & 0 & 0 & & & \\ 0 & -5/4 & 12/5 & -9/5 & 4/5 & -3/20 & & \\ 1/8 & -17/24 & 23/18 & -17/24 & 5/18 & -7/144 & 5/144 & \\ \hline b_i & 21/16 & 5/36 & -199/24 & 257/36 & -251/144 & -50/9 & 8 \\ \hat{b}_i & 7/2 & -136/9 & 58/3 & -88/9 & 35/18 & 10/9 & 0 \end{array}. \tag{4.15}$$

Полученные методы позволяют обеспечить плотную выдачу (dense output) результатов в промежуточных точках с порядком аппроксимации, совпадающим с порядком метода. Для плотной выдачи используем формулу

$$\mathbf{y}\left(t_0+\theta h\right)=\mathbf{y}_0+h\sum_{i=1}^{s}\beta_i\left(\theta\right)\mathbf{F}_i. \tag{4.16}$$

Условия порядка p этой формулы запишутся в виде

$$\left[\beta_1(\theta),\ldots,\beta_s(\theta)\right]\boldsymbol{\Phi}\left(T_{ij}\right)=\theta^i/\gamma\left(T_{ij}\right),\quad i=1,\ldots,p,\quad j=1,\ldots,\nu_i, \tag{4.17}$$

где ν_i – число различных корневых деревьев порядка i. На основе этих соотношений можно получить линейные алгебраические уравнения относительно $\beta_i(\theta)$, решая которые, находим искомые зависимости в виде многочленов от θ. Приведем формулы расчета $\beta_i(\theta)$ для рассмотренных методов, при этом принимаем $\beta_1(\theta)=\theta-\beta_2(\theta)-\cdots-\beta_s(\theta)$.

Для семейства 4-стадийных методов 3-го порядка, задаваемого формулами (4.5), (4.7), получаем:

$$\beta_4\left(\theta\right)=\theta^3 b_4,\quad \begin{bmatrix}\beta_2\left(\theta\right)\\ \beta_3\left(\theta\right)\end{bmatrix}=\begin{bmatrix}c_2 & c_3\\ c_2^2 & c_3^2\end{bmatrix}^{-1}\begin{bmatrix}\theta^2/2-\beta_4\left(\theta\right)c_4\\ \theta^3/3-\beta_4\left(\theta\right)c_4^2\end{bmatrix}. \tag{4.18}$$

Для методов, задаваемых формулами (4.10):

$$\beta_5\left(\theta\right)=\theta^3 b_5,\quad \begin{bmatrix}\beta_2\left(\theta\right)\\ \beta_3\left(\theta\right)\\ \beta_4\left(\theta\right)\end{bmatrix}=\mathbf{C}^{-1}\begin{bmatrix}\theta^2/2\\ \theta^3/3\\ 0\end{bmatrix}. \tag{4.19}$$

Для методов, задаваемых формулами (4.12):

$$\beta_6\left(\theta\right)=\theta^4 b_6,\ \ \beta_5\left(\theta\right)=\theta^3 b_6 a_{65}\left(4-3\theta/c_6\right),\quad \begin{bmatrix}\beta_2\left(\theta\right)\\ \beta_3\left(\theta\right)\\ \beta_4\left(\theta\right)\end{bmatrix}=\mathbf{C}^{-1}\begin{bmatrix}\theta^2/2-\beta_6\left(\theta\right)c_6\\ \theta^3/3-\beta_6\left(\theta\right)c_6^2\\ \theta^4/4-\beta_6\left(\theta\right)c_6^3\end{bmatrix}. \tag{4.20}$$

Для методов, задаваемых формулами (4.14):

$$\beta_7\left(\theta\right)=\theta^4 b_7,\ \ \beta_6\left(\theta\right)=\theta^3 b_7 a_{76}\left(4-3\theta/c_7\right),\quad \begin{bmatrix}\beta_2\left(\theta\right)\\ \beta_3\left(\theta\right)\\ \beta_4\left(\theta\right)\\ \beta_5\left(\theta\right)\end{bmatrix}=\bar{\mathbf{C}}^{-1}\begin{bmatrix}\theta^2/2-\beta_7\left(\theta\right)c_7\\ \theta^3/3-\beta_7\left(\theta\right)c_7^2\\ \theta^4/4-\beta_7\left(\theta\right)c_7^3\\ -\beta_7\left(\theta\right)c_7^4\end{bmatrix}. \tag{4.21}$$

Среди рассмотренных методов особый интерес представляют методы (4.8) и (4.13), в которых вложенная формула совпадает с последней стадией. В [8] был предложен способ построения явных стабилизированных методов с расширенными областями устойчивости, основанный на методах такого типа. В [2], [12] на основе этих методов построены стабилизированные методы 3-го и 4-го порядков, имеющие ПСП 2 и 3.

5. МЕТОДЫ, ОБРАТНЫЕ К ЯВНЫМ МЕТОДАМ

Обратный метод для метода (1.1) получим, поменяв местами $\mathbf{y}_0$ и $\mathbf{y}_1$, заменив h на $-h$, а затем вместо t_0 подставив t_0+h. Обратный метод обладает тем свойством, что если сделать один шаг прямого метода в соответствии с (1.1), а затем сделать шаг обратным методом в обратном направлении (поменяв h на $-h$), то получим исходный вектор $\mathbf{y}_0$. Отметим, что термин обратный (inverse), используемый для обозначения таких методов в [34] и некоторых других работах, не является общепринятым. Наряду с ним используют термин присоединенный (adjoint) [1] и отраженный (reflected) [6]. В [1], [6] изложены основные свойства обратных методов, из которых следует, что порядок, СП, ССП и ПСП обратного метода совпадают с одноименными значениями исходного метода, а главный член погрешности отличается множителем $(-1)^p$.

В [34] было предложено использовать обратные к явным методы Рунге–Кутты для решения жестких задач. Такие методы обладают рядом полезных свойств. Они жестко точные, удобны для реализации, поскольку являются мононеявными, и имеют высокий порядок L-затухания (согласно [35], метод имеет порядок L-затухания μ, если $R(z)=O(z^{-\mu})$ при $z\to\infty$). В то же время эти методы имеют существенный недостаток – 1-й СП, что приводит к снижению точности и реального порядка при решении жестких и дифференциально-алгебраических уравнений. Использование методов с повышенными значениями ПСП и ССП позволяет избавиться от этого недостатка.

Пусть a_{ij}, b_j, c_i – коэффициенты явного s-стадийного метода Рунге–Кутты. Тогда формулы обратного метода запишутся в виде

$$\mathbf{F}_i = \mathbf{f}\left(t_0 + (1-c_i)\,h,\ \mathbf{y}_1 - h\sum_{j=1}^{i-1} a_{ij}\mathbf{F}_j\right), \quad i = 1,\ldots,s, \quad \mathbf{y}_1 = \mathbf{y}_0 + h\sum_{i=1}^{s} b_i\mathbf{F}_i. \tag{5.1}$$

Зная $\mathbf{y}_1$, можно найти все значения $\mathbf{F}_i$ непосредственно по формулам (5.1). Это позволяет свести систему алгебраических уравнений (5.1) к уравнению относительно только вектора $\mathbf{y}_1$, что обеспечивает эффективную реализацию метода. В [36] такие методы были названы мононеявными (они рассматривались также в [2], [14], [37], [38]). Реализация методов, обратных к явным методам, обсуждалась в [14], [34].

Полученные в результате обращения неявные методы позволяют задать хорошее начальное приближение для итераций при решении алгебраических уравнений (5.1). Формулы (4.16)–(4.21) справедливы также и для обратных методов, поэтому их можно использовать не только для плотной выдачи, но и для прогноза решения на следующем шаге. Пусть выполнен очередной шаг $\mathbf{y}_0 \to \mathbf{y}_1$ размером h, а размер следующего шага выбран равным wh. Тогда прогноз решения на следующем шаге можно получить по формуле (4.16), где $\theta = 1 + w$, при этом порядок прогноза совпадает с порядком метода. На первом шаге в качестве прогноза можно использовать явный метод с достаточно малым размером шага.

При исследовании обратных методов бывает удобно иметь их коэффициенты в более привычном виде (как в таблице Бутчера). В таком случае коэффициенты обратных методов получаем в виде

$$c_i^* = 1 - c_{s+1-i}, \quad a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j}, \quad b_j^* = b_{s+1-j}.$$

6. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Все методы, которые мы исследовали, имеют 1-й СП. Примем следующие обозначения для методов: МЕТОД$sp\bar{q}$, если $\tilde{q} = \bar{q}$, и МЕТОД$sp\bar{q}(\tilde{q})$, если $\tilde{q} > \bar{q}$ (s – число стадий, p – порядок). Если $\tilde{q} = \bar{q} = 1$, то указываем только s и p. Через SDIRK33 обозначим жестко точный метод 3-го порядка из [39]. Через ERK33 обозначим явный метод 3-го порядка из [40] (он получен исходя из минимизации коэффициентов погрешности, и реализован во вложенной паре Богацки–Шампайна [41]). Через ERK44 обозначим классический явный метод Рунге–Кутты. Рассмотренные в этой статье методы обозначим в виде (приводим номера формул):

SDIRK532 – (2.5), (2.8)	ERK432b – (4.9)
SDIRK532(3) – (2.5), (2.6)	ERK533 – (4.11)
SDIRK53 – (3.4)	ERK643 – (4.13)
ERK432 – (4.8)	ERK743(4) – (4.15)

Названия обратных к явным методов начинаем с буквы I (IERK432 и т. д.).

Исследуем влияние жеткости системы ОДУ на точность численного решения. Для этого используем две задачи: нелинейную и линейную. Нелинейная задача (задача Капса) задается уравнениями

$$\begin{aligned} y_1' &= -(\mu+2)y_1 + \mu\, y_2^2, \quad y_1(0) = 1, \\ y_2' &= y_1 - y_2 - y_2^2, \quad y_2(0) = 1, \quad 0 \leqslant t \leqslant 1, \end{aligned} \tag{6.1}$$

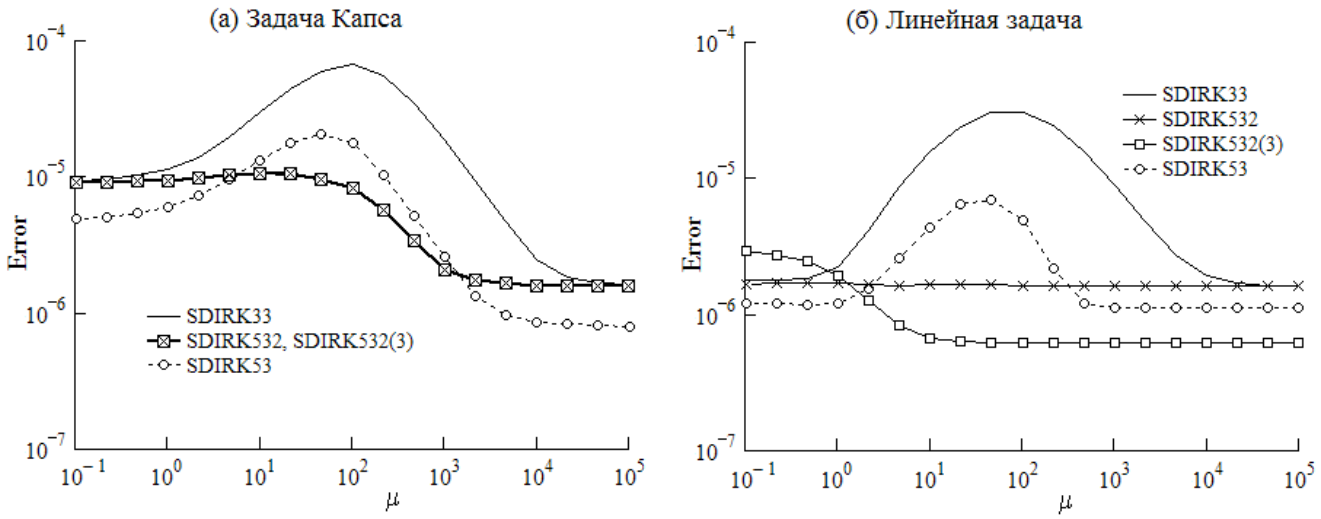
и имеет точное решение $y_1(t) = \exp(-2t)$, $y_2(t) = \exp(-t)$. Собственные числа матрицы Якоби мало изменяются на траектории решения и при больших μ в начальной точке примерно равны $-(\mu + 4)$ и -1.

Линейная задача запишется в виде

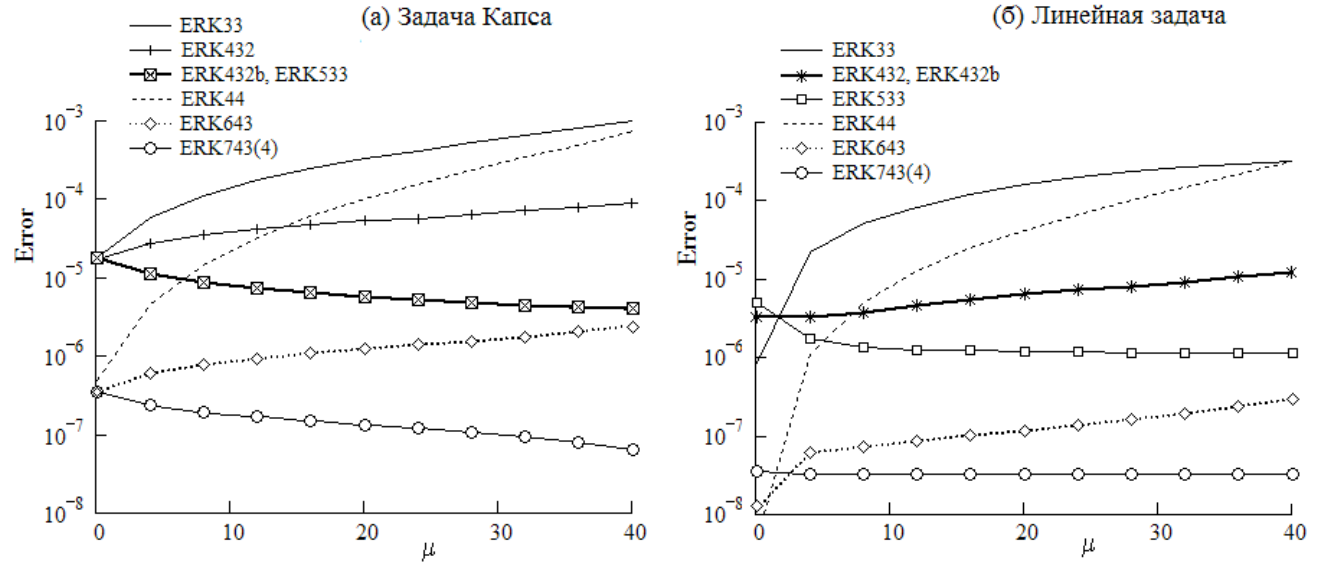
$$\begin{aligned} &\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \begin{bmatrix} a & b \\ b & a \end{bmatrix} \begin{bmatrix} y_1 - \sin(t) \\ y_2 - \cos(t) \end{bmatrix} + \begin{bmatrix} \cos(t) \\ -\sin(t) \end{bmatrix}, \quad \mathbf{y}(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\ &a = -(\mu+1)/2, \quad b = -(\mu-1)/2, \quad 0 \leqslant t \leqslant 1. \end{aligned} \tag{6.2}$$

Она имеет решение $y_1(t) = \sin(t)$, $y_2(t) = \cos(t)$, а собственные числа матрицы Якоби равны $-\mu$ и -1.

Обе задачи решались с размером шага $h = 1/20$. В каждой точке решения вычисляем евклидову норму ошибки, а в качестве ошибки решения задачи принимаем максимальную среди этих норм. Зависимости ошибки от показателя жесткости μ приведены на фиг. 1–3. Интервал значений μ для явных методов выбран таким, чтобы все методы были устойчивы.



Фиг. 1. Зависимости ошибки от показателя жесткости для методов SDIRK.

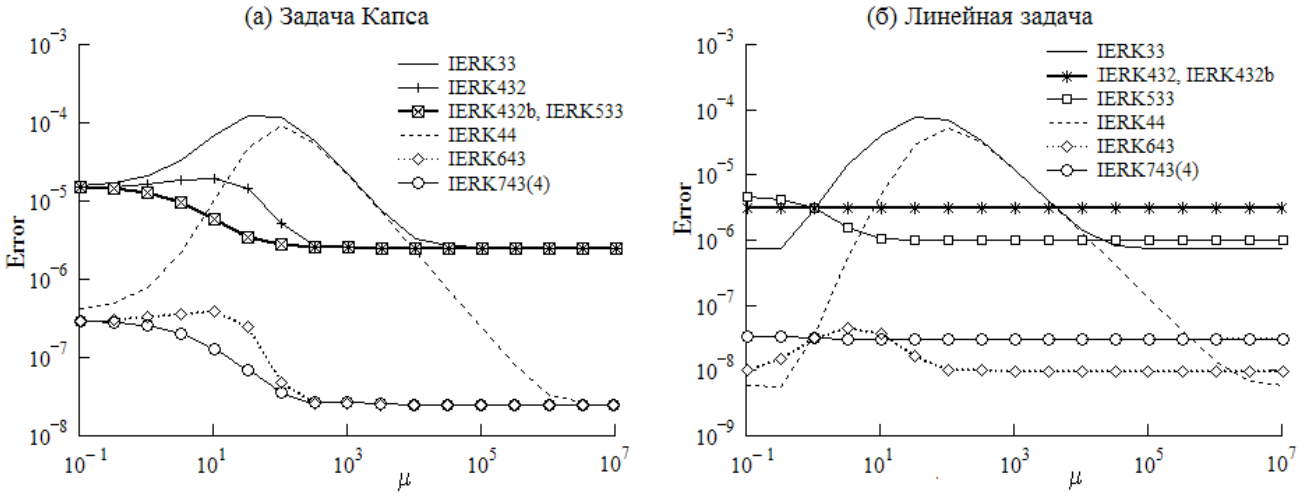


Фиг. 2. Зависимости ошибки от показателя жесткости для явных методов.

Снижение порядка неявных методов SDIRK33, SDIRK53, IERK33, IERK44 проявляется как заметное увеличение ошибки при умеренных значениях μ. Оценка порядка метода на таких участках снижается практически до $q = 1$ (см. [2], [14], [15]). При дальнейшем увеличении μ ошибка уменьшается, что объясняется жесткой точностью методов. Другие неявные методы имеют $p - \bar{q} \leqslant 1$, поэтому снижение точности и порядка невелико или совсем отсутствует. Явные методы ERK33 и ERK44 демонстрируют заметное снижение точности при увеличении μ, но при этом оценка порядка не уменьшается, а немного увеличивается.

Приведенные на фиг. 1 результаты решения задачи Капса методами SDIRK532 и SDIRK532(3) практически совпадают, что можно объяснить доминированием нелинейной составляющей ошибки. Но при решении линейной жесткой задачи метод SDIRK532(3) демонстрирует заметное преимущество, поскольку имеет 3-й ССП ($e31(z) \equiv 0$). Аналогичное замечание справедливо для методов ERK432b, ERK533 (фиг. 2) и методов IERK432b, IERK533 (фиг. 3).

Практически совпадают также приведенные на фиг. 2 результаты решения линейной задачи методами ERK432 и ERK432b. Отличие этих методов в том, что ERK432 имеет $e32(z) \neq 0$, тогда как ERK432b имеет $e32(z) \equiv 0$. Функция $e32(z)$ описывает нелинейные составляющие ошибки, поэтому значения этой функции не влияют на решение линейных задач, что приводит к совпадению результатов. Но при решении нелинейной



Фиг. 3. Зависимости ошибки от показателя жесткости для методов IERK

задачи Капса метод ERK432b имеет преимущество. Все сказанное справедливо также для результатов методов IERK432 и IERK432b на фиг. 3.

Таким образом, жесткие модельные уравнения и полученные на их основе функции погрешности $e_{ij}(z)$ позволили объяснить поведение ошибки численного решения в тех случаях, когда никак иначе это поведение объяснить невозможно. Что касается поведения ошибки при малых значениях μ (нежесткие задачи), то оно вполне объяснимо с помощью коэффициентов погрешности.

Были также проведены эксперименты по исследованию сходимости различных компонент решения ДАУ. Для этого использовались тестовые задачи из [2], [26], [27], [42], [43]. Приведем результаты решения двух задач. Первая задача – ДАУ индекса 2 – взята из [27] и имеет вид

$$\begin{aligned} &y_1' = y_1 y_2^2 z^2, \quad y_2' = y_1^2 y_2^2 - 3y_2^2 z, \quad 0 = y_1^2 y_2 - 1, \\ &y_1(0) = y_2(0) = z(0) = 1, \quad 0 \leqslant t \leqslant 0.1. \end{aligned} \tag{6.3}$$

Она имеет точное решение $y_1(t) = \exp(t)$, $y_2(t) = \exp(-2t)$, $z(t) = \exp(2t)$. Вторая задача – ДАУ индекса 3 – взята из [26] и имеет вид

$$\begin{aligned} &y_1' = 2y_1 y_2 z_1 z_2, \quad y_2' = -y_1 y_2 z_2^2, \quad z'_1 = (y_1 y_2 + z_1 z_2)\, u, \quad z'_2 = -y_1 y_2^2 z_2^3 u^2, \quad 0 = y_1 y_2^2 - 1, \\ &y_1(0) = y_2(0) = z_1(0) = z_2(0) = u(0) = 1, \quad 0 \leqslant t \leqslant 0.1. \end{aligned} \tag{6.4}$$

Она имеет точное решение $y_1(t) = z_1(t) = \exp(2t)$, $y_2(t) = z_2(t) = \exp(-t)$, $u(t) = \exp(t)$.

Результаты решения задачи (6.3) приведены в табл. 3, а задачи (6.4) – в табл. 4, где через e_y, e_z и e_u обозначены ошибки решения соответствующих компонент при $h = 0.01$ (вычисляем их аналогично ошибкам решения ОДУ (6.1) и (6.2)). В табл. 3, 4 приведены также оценки порядков сходимости $\tilde{p}_y$, $\tilde{p}_z$, $\tilde{p}_u$ соответствующих компонент, которые вычисляем по формуле

$$\tilde{p} = \frac{\ln(e(h)/e(h/2))}{\ln(2)},$$

где $e(h)$ – ошибка при размере шага h (размер шага последовательно уменьшаем до тех пор, пока искомая оценка практически не перестает изменяться). При решении некоторых других задач индексов 2 и 3 были получены более низкие оценки, чем для задач (6.3) и (6.4), которые приведены в скобках. Методы SDIRK33 и SDIRK532(3) не обеспечивают сходимости при решении ДАУ3, поэтому соответствующие результаты не приводим.

Рассмотрим результаты методов SDIRK. Метод SDIRK33 дает оценки порядков сходимости при решении ДАУ2, которые соответствуют теоретическим результатам в табл. 2. Оценки сходимости других методов SDIRK можно объяснить выполнением необходимых условий, рассмотренных в разд. 1–3 статьи (более подробно см. [2], [11], [29]).

Оценки сходимости методов IERK можно объяснить, если принять предположение, что метод, имеющий $q = 1, \bar{q} > 1$, обеспечивает такие же порядки сходимости, как метод, имеющий $q = \bar{q}$. При этом более высокие

Таблица 3. Результаты решения ДАУ индекса 2

Метод	e_y	e_z	$\tilde{p}_y$	$\tilde{p}_z$
SDIRK33	2.48×10^{-4}	1.27×10^{-2}	2	1
SDIRK532	4.78×10^{-5}	1.17×10^{-2}	2	2
SDIRK532(3)	7.11×10^{-5}	2.56×10^{-3}	2	2
SDIRK53	4.25×10^{-6}	1.40×10^{-3}	3	2
IERK432	7.03×10^{-7}	1.21×10^{-4}	3	2
IERK432b	3.84×10^{-6}	1.17×10^{-4}	3	2
IERK533	2.13×10^{-7}	2.63×10^{-7}	3	3
IERK643	2.82×10^{-9}	3.49×10^{-9}	4	4(3)
IERK743(4)	1.37×10^{-8}	1.69×10^{-8}	4	4

Таблица 4. Результаты решения ДАУ индекса 3

Метод	e_y	e_z	e_u	$\tilde{p}_y$	$\tilde{p}_z$	$\tilde{p}_u$
SDIRK532	7.55×10^{-6}	1.10×10^{-4}	2.75×10^{-2}	2	2	1
SDIRK53	3.33×10^{-6}	1.24×10^{-5}	4.35×10^{-2}	2	2	1
IERK432	2.11×10^{-7}	3.78×10^{-5}	1.78×10^{-4}	2	2	2(1)
IERK432b	2.15×10^{-7}	3.78×10^{-5}	8.10×10^{-3}	2	2	1
IERK533	6.24×10^{-8}	1.06×10^{-7}	2.04×10^{-5}	3	3	2
IERK643	8.17×10^{-10}	6.85×10^{-10}	2.04×10^{-5}	4	4(3)	2
IERK743(4)	1.11×10^{-9}	1.01×10^{-9}	1.86×10^{-5}	3	3	2

оценки (например, оценка $\tilde{p}_z = 4$ метода IERK643 или оценка $\tilde{p}_u = 2$ метода IERK432) можно объяснить особенностью решаемой задачи, а более низкие оценки позволили бы опровергнуть это предположение.

7. ЗАКЛЮЧЕНИЕ

В работе рассмотрены методы Рунге–Кутты, которые, имея 1-й СП, обладают свойствами методов более высокого СП. К ним относятся методы, имеющие повышенные (больше 1) значения ПСП либо КСП. При решении жестких ОДУ методы с повышенным ПСП позволяют избежать снижения порядка либо сделать его незначительным. А при решении ДАУ индексов 2 и 3 методы с повышенным ПСП или КСП позволяют повысить порядки сходимости p_y, p_z, p_u.

Методы с повышенным ПСП требуют выполнения дополнительных $\tilde{q} - 1$ стадий на каждом шаге. Однако многократное повышение точности этих методов по сравнению с обычными методами может компенсировать эти затраты и обеспечить заметное преимущество при решении жестких ОДУ и ДАУ высших индексов. Среди рассмотренных методов наибольший практический интерес представляют явные и обратные к ним методы. Отметим, что стадии $2, \ldots, \tilde{q} + 1$ явных методов могут быть распараллелены. Отметим также, что на основе явных методов можно построить стабилизированные методы (см. [2], [12]), для которых дополнительные стадии составляют малую часть общих вычислительных затрат.

Обратные методы являются мононеявными, что обеспечивает более эффективную реализацию по сравнению с неявными методами общего вида. Дополнительное преимущество дает прогноз начального значения для итераций, порядок которого совпадает с порядком метода (см. разд. 5). В этом случае допустимо выполнять только одну ньютоновскую итерацию для решения алгебраической системы (5.1), в результате можно получить двухшаговый линейно неявный метод (одношаговые методы такого типа рассматривались в [44]).

Еще одно преимущество рассмотренных здесь обратных методов – высокий порядок L-затухания, равный p для основного и $p - 1$ для вложенного метода. Этот факт имеет существенное значение при решении ДАУ индексов 2 и 3. В [43] на простейшем примере было показано, что процедура контроля ошибки при решении

ДАУ может оказаться неустойчивой и привести к аварийной остановке. Чтобы этого не происходило, рекомендуют исключить из процедуры переменные высших индексов (в (6.3), (6.4) это переменные z, z_1, z_2, u). Однако это приводит к ослаблению контроля ошибки, к тому же индексы переменных могут быть неизвестны. Альтернативой может быть использование вложенных пар, в которых основной и вложенный методы имеют порядок L-затухания не меньше 2. При $p \geq 3$ все методы IERK обладают этим свойством.

СПИСОК ЛИТЕРАТУРЫ

1. *Хайрер Э., Нёрсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990.
2. *Скворцов Л.М.* Численное решение обыкновенных дифференциальных и дифференциально-алгебраических уравнений. М.: ДМК Пресс, 2022.
3. *Prothero A., Robinson A.* On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations // Math. Comput. 1974. V. 28. № 125. P. 145–162.
4. *Деккер К., Вервер Я.* Устойчивость методов Рунге–Кутты для жестких нелинейных дифференциальных уравнений. М.: Мир, 1988.
5. *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. М.: Мир, 1999.
6. *Butcher J.C.* Numerical methods for ordinary differential equations. Chichester: John Wiley and Sons, 2008.
7. *Скворцов Л.М.* Диагонально неявные FSAL-методы Рунге–Кутты для жестких и дифференциально-алгебраических систем // Матем. моделирование. 2002. Т. 14. № 2. С. 3–17.
8. *Скворцов Л. М.* Точность методов Рунге Кутты при решении жестких задач // Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 9. С. 1374–1384.
9. *Скворцов Л.М.* Явные методы Рунге–Кутты для умеренно жестких задач // Ж. вычисл. матем. и матем. физ. 2005. Т. 45. № 11. 2017–2030.
10. *Скворцов Л.М.* Диагонально неявные методы Рунге–Кутты для жестких задач // Ж. вычисл. матем. и матем. физ. 2006. Т. 46. № 12. С. 2209–2222.
11. *Скворцов Л.М.* Модельные уравнения для исследования точности методов Рунге–Кутты // Матем. моделирование. 2010. Т. 22. № 5. С. 146–160.
12. *Скворцов Л.М.* Явные стабилизированные методы Рунге–Кутты // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. № 7. С. 1236–1250.
13. *Скворцов Л.М.* Явные адаптивные методы Рунге–Кутты для жестких и колебательных задач // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. № 8. С. 1434–1448.
14. *Скворцов Л.М.* О неявных методах Рунге–Кутты, полученных в результате обращения явных методов // Матем. моделирование. 2017. Т. 29. № 1. С. 3–19.
15. *Скворцов Л.М.* Как избежать снижения точности и порядка методов Рунге–Кутты при решении жестких задач // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 7. С. 1126–1141.
16. *Скворцов Л.М.* Неявные методы Рунге–Кутты с явными внутренними стадиями // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 3.
17. *Rang J.* An analysis of the Prothero–Robinson example for constructing new DIRK and ROW methods // J. Comput. Appl. Math. 2014. V. 262. P. 105–114.
18. *Rang J.* An analysis of the Prothero–Robinson example for constructing new adaptive ESDIRK methods of order 3 and 4 // Appl. Numer. Math. 2015. V. 94. P. 75–87.
19. *Rang J.* The Prothero and Robinson example: Convergence studies for Runge–Kutta and Rosenbrock–Wanner methods // Appl. Numer. Math. 2016. V. 108. P. 37–56.

20. *Ketcheson D.I., Seibold B., Shirokoff D., Zhou D.* DIRK schemes with high weak stage order // Lecture Notes in Comput. Science and Engng. V. 134. Spectral and High Order Methods for Partial Differential Equations. Springer, 2020. P. 453–463.

21. *Biswas A., Ketcheson D.I., Seibold B., Shirokoff D.* Design of DIRK schemes with high weak stage order // Commun. Appl. Math. Comput. Sci. 2023. V. 18. P. 1–28.

22. *Biswas A., Ketcheson D.I., Seibold B., Shirokoff D.* Algebraic structure of the weak stage order conditions for Runge–Kutta methods. // SIAM J. Numer. Anal. 2024. V. 62. № 1. P. 48–72.

23. *Biswas A., Ketcheson D.I., Roberts S., Seibold B., Shirokoff D.* Explicit Runge–Kutta methods that alleviate order reduction. https://arxiv.org/abs/2310.02817.

24. *Rosales R.R., Seibold B., Shirokoff D., Zhou D.* Spatial manifestations of order reductions in Runge–Kutta methods for initial boundary value problems // Commun. Math. Sci. 2024. V. 22. № 3. P. 613–653.

25. *Hairer E., Lubich Ch., Roche M.* The numerical solution of differential-algebraic systems by Runge–Kutta methods. Berlin: SpringerVerlag, 1989.

26. *Jay L.* Convergence of Runge-Kutta methods for differential-algebraic systems of index 3 // Appl. Numer. Math. 1995. V. 17. № 2. P. 97–118.

27. *Jay L.* Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2 // BIT. 1993. V. 33. № 1. P. 137–150.

28. *Бояринцев Ю.Е., Чистяков В.Ф.* Алгебро-дифференциальные системы. Методы решения и исследования. Новосибирск: Наука, 1998.

29. *Скворцов Л.М.* Диагонально-неявные методы Рунге–Кутты для дифференциально-алгебраических уравнений индексов 2 и 3 // Ж. вычисл. матем. и матем. физ. 2010. Т. 50. № 6. С. 1047–1059.

30. *Hosea M.E., Shampine L.F.* Analysis and implementation of TR-BDF2 // Appl. Numer. Math. 1996. V. 20. № 1–2. P. 21–37.

31. *Cameron F., Palmroth M., Piche R.* Quasi stage order conditions for SDIRK methods // Appl. Numer. Math. 2002. V. 42. № 1–3. P. 61–75.

32. *Лебедев В.И.* Как решать явными методами жесткие системы дифференциальных уравнений // Вычисл. процессы и системы. М.: Наука, 1991. Вып. 8. С. 237–291.

33. *Новиков Е.А.* Явные методы для жестких систем. Новосибирск: Наука, 1997.

34. *Калиткин Н.Н., Пошивайло И.П.* Вычисления с использованием обратных схем Рунге–Кутты // Матем. моделирование. 2013. Т. 25. № 10. С. 79–96.

35. *Кочетков К.А., Ширков П.Д.* L-затухающие ROW-методы третьего порядка точности // Ж. вычисл. матем. и матем. физ. 1997. Т. 37. № 6. С. 699–710.

36. *Cash J.R., Singhal A.* Mono-implicit Runge–Kutta formulae for the numerical integration of stiff differential systems // IMA J. Numer. Anal. 1982. V. 2. P. 211–227.

37. *Kulikov G.Yu., Shindin S.K.* Adaptive nested implicit Runge–Kutta formulas of Gauss type // Appl. Numer. Math. 2009. V. 59. № 3–4. P. 707–722.

38. *Куликов Г.Ю.* Вложенные симметричные неявные гнездовые методы Рунге–Кутты типов Гаусса и Лобатто для решения жестких обыкновенных дифференциальных уравнений и гамильтоновых систем // Ж. вычисл. матем. и матем. физ. 2015. Т. 55. № 6. С. 986–1007.

39. *Alexander R.* Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s // SIAM J. Numer. Anal. 1977. V. 14. № 6. P. 1006–1021.

40. *Ralston A.* Runge–Kutta methods with minimal error bounds // Math. Comput. 1962. V. 16. P. 431–437.

41. *Bogacki P., Shampine L.F.* A 3(2) pair of Runge–Kutta formulas // Appl. Math. Lett. 1989. V. 2. № 4. P. 321–325.

42. *Булатов М.В., Соловарова Л.С.* О потере L-устойчивости неявного метода Эйлера для одной линейной задачи // Изв. Иркутского гос. ун-та. Сер. Математика. 2015. Т. 12. С. 3–11.

43. *Скворцов Л.М.* Методы ESDIRK третьего и четвертого порядков для жестких и дифференциально-алгебраических задач // Ж. вычисл. матем. и матем. физ. 2022. Т. 62. № 5. С. 790–808.

44. *Зубанов А.М., Кутрухин Н.Н., Ширков П.Д.* О построении линейно неявных схем, LN-эквивалентных неявным методам Рунге–Кутты // Компьютерные исследования и моделирование. 2012. Т. 4.№ 3. С. 483–496.

GENERALIZATIONS OF THE STAGE ORDER OF RUNGE–KUTTA METHODS

L. M. Skvortsov*

127051 Moscow, Trubnaya str., 51, bldg.. 1, 3B Service LLC, Russia
**e-mail: lm.skvo@gmail.com*

Received: 27.05.2024
Revised: 02.07.2024
Accepted: 31.08.2024

Abstract. The application of Runge–Kutta methods for solving rigid systems of ordinary differential equations and differential algebraic equations is considered. When solving such problems, the effect of reducing the order is often manifested, when, with a given accuracy, the real order of the method turns out to be lower than the classical order, which inevitably leads to increased computational costs. To avoid reducing the order, the method must have a sufficiently high stage order. However, the methods that provide the most convenient and efficient implementation have a low stage order. Therefore, the task of constructing methods that, at a low stage order, have the properties of methods of a higher stage order is relevant. This article is devoted to the construction of methods of this type. Singly diagonal-implicit, explicit methods and those inverse to the explicit ones are considered. The results of solving test problems are presented.

Keywords: Runge–Kutta methods, rigid and differential algebraic problems, stage order, pseudo-stage order, weak stage order, quasi-stage order.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2303–2311

УДК 517.38

КВАДРАТУРНЫЕ ФОРМУЛЫ ДЛЯ СИНГУЛЯРНЫХ ИНТЕГРАЛОВ, СОДЕРЖАЩИЕ ЗНАЧЕНИЯ ФУНКЦИИ И ЕЕ ПРОИЗВОДНЫХ

© 2024 г. Ш.С. Хубежты[1,2,3,*], Л.Ю. Плиева[1,2,3,]**

[1]362025 Владикавказ, ул. Ватутина, 44-46, Северо-Осетинский государственный университет им. К. Л. Хетагурова, Республика Северная Осетия – Алания

[2]362025 Владикавказ, ул. Ватутина, 53, Южный математический институт – филиал ВНЦ РАН, Республика Северная Осетия – Алания

[3]362002 Владикавказ, ул. Молодежная, 7, Финансовый университет при правительстве РФ, Владикавказский филиал, Республика Северная Осетия – Алания

**e-mail: shalva57@rambler.ru*

***e-mail: plieva-21@mail.ru*

Поступила в редакцию 18.04.2024 г.

Переработанный вариант 18.04.2024 г.

Принята к публикации 23.08.2024 г.

Строятся квадратурные формулы для сингулярных интегралов на отрезке интегрирования $[-1, 1]$ с определенными весовыми функциями $p(t)$. При построении используются значения функции и ее производных в нулях многочлена Чебышёва. Полученные формулы являются квадратурными формулами интерполяционного типа и имеют алгебраическую степень точности $2m-1$. Оценка погрешности приводится. Библ. 4.

Ключевые слова: сингулярные интегралы, квадратурная формула, узлы, производная функции, весовая функция, интерполяционный многочлен, остаточный член.

DOI: 10.31857/S0044466924120064 , **EDN**: KCAUSX

ВВЕДЕНИЕ

При решении сингулярных интегральных уравнений встречаются случаи, когда требуется найти не только значения функции, но и значения ее производных. Для численного решения таких уравнений естественно требуется использовать квадратурные формулы, содержащие эти два значения. С использованием этих данных еще и повышается степень точности вычисления.

Целью настоящей статьи является построение квадратурных формул для сингулярных интегралов со значениями функции и ее производными.

Итак, пусть заданы узлы $x_1, x_2, \dots, x_m$, значения функции $\varphi(x_1), \varphi(x_2), \dots, \varphi(x_m)$ и ее производных $\varphi'(x_1), \varphi'(x_2), \dots, \varphi'(x_m)$. Задача ставиться следующим образом: построить квадратурную формулу вида

$$\int_a^b p(t)\frac{\varphi(t)}{t-x}dt \approx \sum_{i=1}^{m}\left(A_i(x)\varphi\left(x_i\right)+B_i(x)\varphi'\left(x_i\right)\right), \quad a<x<b. \tag{1}$$

Очевидно, что для таких квадратурных формул придется использовать интерполяционный многочлен Эрмита, построенный по заданным узлам $x_1, x_2, \dots, x_m$ и значениями $\varphi(x_1), \varphi(x_2), \dots, \varphi(x_m)$, $\varphi'(x_1), \varphi'(x_2), \dots, \varphi'(x_m)$.

Такой многочлен имеет вид [1, гл. 1, § 13]

$$P(x)=\sum_{i=1}^{m}\frac{\omega^2(x)}{\omega'^2(x_i)(x-x_i)^2}\left[\varphi(x_i)\left(1-\frac{\omega''(x_i)}{\omega'(x_i)}(x-x_i)\right)+\varphi'(x_i)(x-x_i)\right], \tag{2}$$

где

$$\omega(x)=\prod_{j=1}^{m}\left(x-x_j\right), \quad \omega'(x)=\sum_{i=1}^{m}\prod_{\substack{j=1\\ j\neq i}}^{m}\left(x-x_j\right), \quad \omega''(x)=\sum_{i=1}^{m}\sum_{\substack{k=1\\ k\neq i}}^{m}\prod_{\substack{j=1\\ j\neq i,k}}^{m}\left(x-x_j\right).$$

Остаточный член интерполяционного многочлена (2) выражается формулой

$$R_m(\varphi, x) = \frac{\varphi^{(2m)}(\xi)}{(2m)!}\omega^2(x), \quad a \leqslant \xi \leqslant b. \tag{3}$$

Подставляя многочлен (2) вместо $\varphi(t)$ в сингулярный интеграл

$$\int_a^b p(t)\frac{\varphi(t)}{t-x}dt,$$

получим

$$\begin{aligned}\int_a^b p(t)\frac{\varphi(t)}{t-x}dt \approx \int_a^b p(t)\frac{1}{t-x}\left\{\sum_{i=1}^m\left(\frac{\omega(t)}{\omega'(x_i)(t-x_i)}\right)^2\left[\varphi(x_i)\left(1-\frac{\omega''(x_i)}{\omega'(x_i)}(t-x_i)\right)+\varphi'(x_i)(t-x_i)\right]\right\}dt = \\ = \sum_{i=1}^m\left(A_i(x)\varphi(x_i)+B_i(x)\varphi'(x_i)\right),\end{aligned} \tag{4}$$

где

$$A_i(x) = \int_a^b p(t)\frac{1}{t-x}\left(\frac{\omega(t)}{\omega'(x_i)(t-x_i)}\right)^2\left(1-\frac{\omega''(x_i)}{\omega'(x_i)}(t-x_i)\right)dt, \tag{5}$$

$$B_i(x) = \int_a^b p(t)\frac{1}{t-x}\left(\frac{\omega(t)}{\omega'(x_i)(t-x_i)}\right)^2(t-x_i)dt, \quad i = 1, 2, \ldots, m. \tag{6}$$

В дальнейшем будем рассматривать отрезок $[a, b] = [-1, 1]$ и часто встречающиеся случаи весовой функции

$$p(t) = \frac{1}{\sqrt{1-t^2}}, \quad p(t) = \sqrt{1-t^2}, \quad p(t) = \sqrt{\frac{1+t}{1-t}}, \quad p(t) = \sqrt{\frac{1-t}{1+t}}. \tag{7}$$

Будем также использовать преобразование вида

$$\frac{1}{(t-x)(t-x_i)} = \frac{1}{x-x_i}\left(\frac{1}{t-x}-\frac{1}{t-x_i}\right), \tag{8}$$

и равенство вида:

$$\int_{-1}^1 p(t)P_n(t)Q(t)dt = 0, \tag{9}$$

где $P_n(t), n = 0, 1, \ldots,$ ортогональные многочлены по весу $p(t)$, а $Q(t)$ произвольный многочлен степени меньше n.

1. КВАДРАТУРНЫЕ ФОРМУЛЫ ДЛЯ СИНГУЛЯРНОГО ИНТЕГРАЛА $\int_{-1}^1 \frac{1}{\sqrt{1-t^2}}\frac{\varphi(t)}{t-x}dt$

Известно, что на отрезке $[-1, 1]$ при весовой функции $p(t) = \frac{1}{\sqrt{1-t^2}}$, ортогональными многочленами являются многочлены Чебышёва I рода $T_m(t) = \cos(m\arccos t)\,(m = 0, 1, \ldots)$, с корнями $x_k = \cos\frac{2k-1}{2m}\pi$, $k = 1, 2, \ldots, m$. Также справедливы формулы [2, гл. 2, §3]

$$\int_{-1}^1 \frac{1}{\sqrt{1-t^2}}T_m(t)T_n(t)dt = \begin{cases}0 & \text{при } m \neq n,\\ \frac{\pi}{2} & \text{при } m = n.\end{cases} \tag{1.1}$$

Вычислим теперь коэффициенты квадратурной формулы (1) по формулам (5) и (6). При вычислении этих коэффициентов мы будем пользоваться следующими формулами [2, гл. 7, §3], [2, гл. 2, §3], [3, гл. 2, § 2.10]

$$\int_{-1}^1 \frac{1}{\sqrt{1-t^2}}f(t)dt \approx \frac{\pi}{m}\sum_{k=1}^m f(x_k), \quad x_k = \cos\frac{2k-1}{2m}\pi, \tag{1.2}$$

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}T_m(t)Q(t)dt=0, \tag{1.3}$$

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)}{t-x}dt=\pi U_{m-1}(x),\quad U_{m-1}(x)=\frac{\sin\left(m\arccos x\right)}{\sqrt{1-x^2}}. \tag{1.4}$$

В этом случае

$$\omega(t)=\frac{1}{2^{m-1}}T_m(t). \tag{1.5}$$

Вычислим коэффициенты $B_i(x)\,(i=1,2,\ldots,m)$

$$B_i(x)=\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{1}{t-x}\left(\frac{T_m(t)}{T_m'\left(x_i\right)\left(t-x_i\right)}\right)^2(t-x_i)dt=$$

$$=\frac{1}{\left(T_m'\left(x_i\right)\right)^2\left(x-x_i\right)}\left\{\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m^2(t)}{t-x}dt-\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m^2(t)}{t-x_i}dt\right\}=$$

$$=\frac{1}{\left(T_m'\left(x_i\right)\right)^2\left(x-x_i\right)}\left\{\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{\left(T_m(t)-T_m(x)\right)T_m(t)}{t-x}dt+\right.$$

$$\left.+T_m(x)\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)}{t-x}dt-\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{\left(T_m(t)-T_m\left(x_i\right)\right)T_m(t)}{t-x_i}dt\right\}=$$

$$=\frac{1}{\left(T_m'\left(x_i\right)\right)^2\left(x-x_i\right)}\left\{0+\pi T_m(x)U_{m-1}(x)-0\right\}=\frac{\pi T_m(x)U_{m-1}(x)}{\left(T_m'\left(x_i\right)\right)^2\left(x-x_i\right)}.$$

Таким образом, имеем

$$B_i(x)=\frac{\pi T_m(x)U_{m-1}(x)}{\left(T_m'\left(x_i\right)\right)^2\left(x-x_i\right)}. \tag{1.6}$$

Перейдем к вычислению коэффициентов $A_i(x)$:

$$A_i(x)=\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{1}{t-x}\left(\frac{T_m(t)}{T_m'\left(x_i\right)\left(t-x_i\right)}\right)^2\left(1-\frac{T_m^{''}\left(x_i\right)}{T_m'\left(x_i\right)}(t-x_i)\right)dt=$$

$$=\frac{1}{\left(T_m'\left(x_i\right)\right)^2}\left[\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m^2(t)dt}{\left(t-x\right)\left(t-x_i\right)^2}-\frac{T_m^{''}\left(x_i\right)}{T_m'\left(x_i\right)}\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m^2(t)dt}{\left(t-x\right)\left(t-x_i\right)}\right]=$$

$$=\frac{1}{\left(T_m'\left(x_i\right)\right)^2}\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m^2(t)}{\left(t-x\right)\left(t-x_i\right)^2}dt-\frac{T_m^{''}\left(x_i\right)}{T_m'\left(x_i\right)}B_i(x).$$

Рассмотрим интеграл

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m^2(t)}{\left(t-x\right)\left(t-x_i\right)^2}dt=\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)}{\left(t-x\right)\left(t-x_i\right)}\frac{T_m(t)-T_m\left(x_i\right)}{t-x_i}dt=$$

$$=\frac{1}{x-x_i}\left[\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)\left(T_m(t)-T_m\left(x_i\right)\right)}{\left(t-x\right)\left(t-x_i\right)}dt-\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)\left(T_m(t)-T_m\left(x_i\right)\right)}{\left(t-x_i\right)^2}dt\right]=$$

$$=\frac{1}{x-x_i}\left[\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)-T_m(x)}{t-x}\frac{T_m(t)-T_m\left(x_i\right)}{t-x_i}dt+T_m(x)\times\right.$$

$$\times\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{1}{t-x}\frac{T_m(t)-T_m(x_i)}{t-x_i}dt-\frac{\pi}{m}\sum_{k=1}^{m}\frac{T_m(x_k)}{x_k-x_i}\frac{T_m(x_k)-T_m(x_i)}{x_k-x_i}\Bigg]=\frac{1}{x-x_i}\times$$

$$\times\Bigg[\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)-T_m(x)}{t-x}\frac{T_m(t)-T_m(x_i)}{t-x_i}dt+\frac{T_m(x)}{x-x_i}\left(\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)-T_m(x_i)}{t-x}dt-$$

$$-\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{T_m(t)-T_m(x_i)}{t-x_i}dt\right)-\frac{\pi}{m}(T'_m(x_i))^2\Bigg]=\frac{1}{x-x_i}\Bigg[\frac{\pi}{m}\sum_{k=1}^{m}\frac{T_m(x_k)-T_m(x)}{x_k-x}\times$$

$$\times\frac{T_m(x_k)-T_m(x_i)}{x_k-x_i}-\frac{\pi}{m}(T'_m(x_i))^2+\frac{T_m(x)}{x-x_i}\Big[\pi U_{m-1}(x)-\pi U_{m-1}(x_i)\Big]\Bigg]=$$

$$=\frac{1}{x-x_i}\left[\frac{\pi}{m}\left(\frac{T_m(x)T'_m(x_i)}{x-x_i}-(T'_m(x_i))^2\right)+\frac{T_m(x)}{x-x_i}\Big[\pi U_{m-1}(x)-\pi U_{m-1}(x_i)\Big]\right]=$$

$$=\frac{\pi}{x-x_i}\left(-mU^2_{m-1}(x_i)+\frac{T_m(x)U_{m-1}(x)}{x-x_i}\right).$$

Окончательно получим

$$A_i(x)=\frac{\pi}{(x-x_i)(T'_m(x_i))^2}\left\{-mU^2_{m-1}(x_i)+\frac{T_m(x)U_{m-1}(x)}{x-x_i}-\frac{T''_m(x_i)}{T'_m(x_i)}T_m(x)U_{m-1}(x)\right\}. \tag{1.7}$$

Для частного случая $m=1$ получим

$$A_1(x)=0,\quad B_1(x)=\pi,$$

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{\varphi(t)}{t-x}dt\approx\pi\varphi'(0).$$

Для случая $m=2$ имеем

$$A_1(x)=\frac{\pi}{\sqrt{2}}\left(1-x^2\right);\quad A_2(x)=\frac{\pi}{\sqrt{2}}\left(x^2-1\right);$$

$$B_1(x)=\frac{\pi}{2}\left(x^2+\frac{x}{\sqrt{2}}\right);\quad B_2(x)=\frac{\pi}{2}\left(x^2-\frac{x}{\sqrt{2}}\right).$$

Квадратурная формула для случая $m=2$ будет иметь вид

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{\varphi(t)}{t-x}dt\approx\frac{\pi}{\sqrt{2}}\left(1-x^2\right)\varphi\left(\frac{1}{\sqrt{2}}\right)+\frac{\pi}{\sqrt{2}}\left(x^2-1\right)\varphi\left(-\frac{1}{\sqrt{2}}\right)+$$

$$+\frac{\pi}{2}\left(x^2+\frac{x}{\sqrt{2}}\right)\varphi'\left(\frac{1}{\sqrt{2}}\right)+\frac{\pi}{2}\left(x^2-\frac{x}{\sqrt{2}}\right)\varphi'\left(-\frac{1}{\sqrt{2}}\right).$$

2. КВАДРАТУРНЫЕ ФОРМУЛЫ ДЛЯ СИНГУЛЯРНОГО ИНТЕГРАЛА $\int_{-1}^{1}\sqrt{1-t^2}\frac{\varphi(t)}{t-x}dt$

В случае весовой функции $p(t)=\sqrt{1-t^2}$ на отрезке $[-1,1]$ ортогональными многочленами являются многочлены Чебышёва II рода $U_m(x)=\frac{\sin((m+1)\arccos x)}{\sqrt{1-x^2}}$. В этом случае коэффициенты квадратурной формулы (1) подставляя $\omega(t)=\frac{1}{2^m}U_m(t)$, вычисляются по формулам:

$$A_i(x)=\int\limits_{-1}^{1}\sqrt{1-t^2}\frac{1}{t-x}\left(\frac{U_m(t)}{U'_m(x_i)(t-x_i)}\right)^2\left(1-\frac{U''_m(x_i)}{U'_m(x_i)}(t-x_i)\right)dt, \tag{2.1}$$

$$B_i(x)=\int\limits_{-1}^{1}\sqrt{1-t^2}\frac{1}{t-x}\frac{U^2_m(t)}{(U'_m(x_i))^2(t-x_i)^2}(t-x_i)dt, \tag{2.2}$$

$$x_i = cos\frac{\pi i}{m+1}, \quad (i = 1, 2, \ldots, m). \tag{2.3}$$

Для вычисления этих коэффициентов мы будем пользоваться формулами [2, гл.7, §3], [2, гл. 2, §3], [3, гл. 2, § 2.10]

$$\int\limits_{-1}^{1} \sqrt{1-t^2} f(t)dt \approx \frac{\pi}{m+1} \sum_{k=1}^{m} \left(1-x_k^2\right) f\left(x_k\right), x_k = cos\frac{\pi k}{m+1}. \tag{2.4}$$

$$\int\limits_{-1}^{1} \sqrt{1-t^2} U_m(t) Q(t)dt = 0, \tag{2.5}$$

$$\int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m(t)}{t-x}dt = -\pi T_{m+1}(x). \tag{2.6}$$

Вычислим коэффициенты $B_i(x)\,(i = 1, 2, \ldots, m)$:

$$\begin{aligned} B_i(x) &= \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{1}{t-x} \frac{U_m^2(t)}{\left(U_m'(x_i)\right)^2 (t-x_i)}dt = \frac{1}{\left(U_m{}'(x_i)\right)^2 (x-x_i)} \times \\ &\times \left(\int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m^2(t)}{t-x}dt - \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m^2(t)}{t-x_i}dt \right) = \frac{1}{\left(U_m{}'(x_i)\right)^2 (x-x_i)} \times \\ &\times \left(\int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m(t)-U_m(x)}{t-x} U_m(t)dt + U_m(x) \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m(t)}{t-x}dt - \right. \\ &\left. - \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m(t)-U_m\left(x_i\right)}{t-x_i} U_m(t)dt \right) = \frac{1}{\left(U_m{}'(x_i)\right)^2 (x-x_i)} \times \\ &\times \left(0 + U_m(x)\left(-\pi T_{m+1}(x)\right) - 0\right) = -\frac{\pi U_m(x) T_{m+1}(x)}{\left(U_m{}'(x_i)\right)^2 (x-x_i)}. \end{aligned} \tag{2.7}$$

Перейдем к вычислению коэффициентов $A_i(x)$:

$$\begin{aligned} A_i(x) &= \frac{1}{\left(U_m{}'(x_i)\right)^2} \left(\int\limits_{-1}^{1} \frac{\sqrt{1-t^2}}{t-x} \left(\frac{U_m(t)-U_m\left(x_i\right)}{t-x_i} \right)^2 dt - \right. \\ &\left. - \frac{U_m''(x_i)}{U_m'(x_i)} \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{1}{t-x} \frac{U_m(t)-U_m(x_i)}{t-x_i} U_m(t)dt \right) = \\ &= \frac{1}{\left(U_m{}'(x_i)\right)^2 (x-x_i)} \left(\int\limits_{-1}^{1} \sqrt{1-t^2} \frac{1}{t-x} \frac{U_m(t)-U_m\left(x_i\right)}{t-x_i} U_m(t)dt - \right. \\ &\left. - \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{1}{t-x_i} \frac{U_m(t)-U_m(x_i)}{t-x_i} U_m(t)dt \right) - \frac{U_m''(x_i)}{U_m'(x_i)} B_i(x). \end{aligned}$$

Рассмотрим отдельно выражение в скобках

$$\begin{aligned} &\int\limits_{-1}^{1} \sqrt{1-t^2} \frac{1}{t-x} \frac{U_m(t)-U_m\left(x_i\right)}{t-x_i} U_m(t)dt - \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{1}{t-x_i} \frac{U_m(t)-U_m\left(x_i\right)}{t-x_i} U_m(t)dt = \\ &= \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m(t)-U_m(x)}{t-x} \frac{U_m(t)-U_m\left(x_i\right)}{t-x_i}dt + U_m(x) \int\limits_{-1}^{1} \sqrt{1-t^2} \frac{U_m(t)-U_m\left(x_i\right)}{(t-x)\left(t-x_i\right)}dt - \end{aligned}$$

$$-\int\limits_{-1}^{1}\sqrt{1-t^2}\frac{1}{t-x_i}\frac{U_m(t)-U_m(x_i)}{t-x_i}U_m(t)dt=\frac{\pi}{m+1}\sum_{k=1}^{m}\left(1-x_k^2\right)\frac{U_m(x_k)-U_m(x)}{x_k-x}\frac{U_m(x_k)-U_m(x_i)}{x_k-x_i}+$$

$$+\frac{U_m(x)}{x-x_i}\left(\int\limits_{-1}^{1}\sqrt{1-t^2}\frac{U_m(t)}{t-x}dt-\int\limits_{-1}^{1}\sqrt{1-t^2}\frac{U_m(t)}{t-x_i}dt\right)-\frac{\pi}{m+1}\sum_{k=1}^{m}\left(1-x_k^2\right)\left(\frac{U_m(x_k)-U_m(x_i)}{x_k-x_i}\right)^2=$$

$$=\frac{\pi}{m+1}\left(1-x_i^2\right)\frac{U_m(x)}{x-x_i}U_m'(x_i)+\frac{U_m(x)}{x-x_i}\left(-\pi T_{m+1}(x)+\pi T_{m+1}(x_i)\right)-\frac{\pi}{m+1}\left(1-x_i^2\right)\left(U_m'(x_i)\right)^2.$$

Окончательно для $A_i(x)$ с учетом равенства

$$\frac{1}{m+1}\left(1-x_i^2\right)U_m'(x_i)+T_{m+1}(x_i)=0 \tag{2.8}$$

получим

$$A_i(x)=\frac{\pi}{\left(U_m'(x_i)\right)^2(x-x_i)}\left(-\frac{\left(1-x_i^2\right)\left(U_m'(x_i)\right)^2}{m+1}-\frac{U_m(x)T_{m+1}(x)}{x-x_i}+\frac{U_m''(x_i)}{U_m'(x_i)}U_m(x)T_{m+1}\right). \tag{2.9}$$

Рассмотрим частные случаи, когда например $m=1$ и $m=2$.

При $m=1$

$$A_1(x)=-\pi x;B_1(x)=\frac{\pi}{2}-\pi x^2,$$

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{\varphi(t)}{t-x}dt\approx-\pi x\varphi(0)+\left(\frac{\pi}{2}-\pi x^2\right)\varphi'(0).$$

При $m=2$ имеем

$$A_1(x)=\frac{\pi}{4}\left(8x^4-10x^2-2x+2\right);\quad A_2(x)=\frac{\pi}{4}\left(-8x^4+10x^2-2x-2\right);$$

$$B_1(x)=-\pi\left(x^4+\frac{1}{2}x^3-\frac{3}{4}x^2-\frac{3}{8}x\right);\quad B_2(x)=-\pi\left(x^4-\frac{1}{2}x^3-\frac{3}{4}x^2+\frac{3}{8}x\right).$$

$$\int\limits_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\frac{\varphi(t)}{t-x}dt\approx\frac{\pi}{4}\left(8x^4-10x^2-2x+2\right)\varphi\left(\frac{1}{2}\right)+\frac{\pi}{4}\left(-8x^4+10x^2-2x-2\right)\varphi\left(-\frac{1}{2}\right)-$$

$$-\pi\left(x^4+\frac{1}{2}x^3-\frac{3}{4}x^2-\frac{3}{8}x\right)\varphi'\left(\frac{1}{2}\right)-\pi\left(x^4-\frac{1}{2}x^3-\frac{3}{4}x^2+\frac{3}{8}x\right)\varphi'\left(-\frac{1}{2}\right).$$

Эти формулы для функций $\varphi(t)=1$ и $\varphi(t)=t$ дают точные результаты.

3. КВАДРАТУРНЫЕ ФОРМУЛЫ ДЛЯ СИНГУЛЯРНОГО ИНТЕГРАЛА $\int_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{\varphi(t)}{t-x}dt$

Многочленами ортогональными по весу $p(t)=\sqrt{\frac{1+t}{1-t}}$ на отрезке $[-1,1]$ являются многочлены $C_m(x)=\frac{\cos\left(\frac{2m+1}{2}\arccos x\right)}{\cos\left(\frac{1}{2}\arccos x\right)}$ с корнями $x_k=\cos\frac{2k-1}{2m+1}\pi$, $(k=1,2,\ldots,m)$.

В этом случае коэффициенты квадратурной формулы (1) $A_i(x)$ и $B_i(x)$ $(i=1,2,\ldots,m)$ подставляя вычисляются по формулам

$$A_i(x)=\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{1}{t-x}\left(\frac{C_m(t)}{C_m'(x_i)(t-x_i)}\right)^2\left(1-\frac{C_m''(x_i)}{C_m'(x_i)}(t-x_i)\right)dt, \tag{3.1}$$

$$B_i(x)=\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{1}{t-x}\left(\frac{C_m(t)}{C_m'(x_i)(t-x_i)}\right)^2(t-x)dt. \tag{3.2}$$

Следуя предыдущим рассуждениям мы для вычисления коэффициентов (3.1) и (3.2) будем пользоваться формулами [3, гл. 2, § 2.10]

$$\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)}{t-x}dt=\pi S_m(x),\quad x\in(-1;1)\,, \tag{3.3}$$

где

$$S_m(x)=\frac{\sin\left(\frac{2m+1}{2}\arccos x\right)}{\sin\left(\frac{1}{2}\arccos x\right)}, \tag{3.4}$$

$$\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}f(t)dt\approx\frac{2\pi}{2m+1}\sum_{k=1}^{m}(1+x_k)\,f\,(x_k)\,,\quad x_k=\cos\frac{2k-1}{2m+1}\pi, \tag{3.5}$$

$$\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}C_m(t)Q(t)dt=0. \tag{3.6}$$

Вычислим коэффициенты $B_i(x)$ $(i=1,2,\ldots,m)$:

$$B_i(x)=\frac{1}{\left(C'_m\,(x_i)\right)^2}\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m^2(t)}{(t-x)\,(t-x_i)}dt=\frac{1}{\left(C'_m\,(x_i)\right)^2(x-x_i)}\times$$

$$\times\left(\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m^2(t)}{t-x}dt-\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m^2(t)}{t-x_i}dt\right)=\frac{1}{\left(C'_m\,(x_i)\right)^2(x-x_i)}\times$$

$$\times\left(\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m(x)}{t-x}C_m(t)dt+C_m(x)\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)}{t-x}dt-\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m\,(x_i)}{t-x_i}C_m(t)dt\right)=$$

$$=\frac{1}{\left(C'_m\,(x_i)\right)^2(x-x_i)}\,(0+\pi C_m(x)S_m(x)-0)=\frac{\pi C_m(x)S_m(x)}{\left(C'_m\,(x_i)\right)^2(x-x_i)}. \tag{3.7}$$

Коэффициенты $A_i(x)$ будут равны:

$$A_i(x)=\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{1}{t-x}\left(\frac{C_m(t)}{C'_m\,(x_i)\,(t-x_i)}\right)^2dt-\frac{C''_m\,(x_i)}{C'_m\,(x_i)}B_i(x).$$

Рассмотрим отдельно интеграл

$$\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{1}{t-x}\left(\frac{C_m(t)-C_m\,(x_i)}{t-x_i}\right)^2dt=\frac{1}{x-x_i}\left[\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m\,(x_i)}{t-x}\frac{C_m(t)-C_m\,(x_i)}{t-x_i}dt-\right.$$

$$\left.-\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\left(\frac{C_m(t)-C_m\,(x_i)}{t-x_i}\right)^2dt\right]=\frac{1}{x-x_i}\left[\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m\,(x_i)}{t-x}\frac{C_m(t)-C_m\,(x_i)}{t-x_i}dt+\right.$$

$$\left.+C_m(x)\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m\,(x_i)}{t-x_i}\frac{1}{t-x}dt-\frac{2\pi}{2m+1}\sum_{k=1}^{m}(1+x_k)\left(\frac{C_m(x_k)-C_m\,(x_i)}{x_k-x_i}\right)^2\right]=$$

$$=\frac{1}{x-x_i}\left[\frac{2\pi}{2m+1}\sum_{k=1}^{m}(1+x_i)\frac{C_m(x_k)-C_m\,(x)}{x_k-x}\frac{C_m(x_k)-C_m\,(x_i)}{x_k-x_i}+\frac{C_m(x)}{x-x_i}\left\{\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m\,(x_i)}{t-x}dt-\right.\right.$$

$$\left.\left.-\int\limits_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{C_m(t)-C_m\,(x_i)}{t-x_i}dt\right\}-\frac{2\pi}{2m+1}\,(1+x_i)\,(C'_m\,(x_i))^2\right]=\frac{1}{x-x_i}\left[\frac{2\pi}{2m+1}\,(1+x_i)\,\frac{C_m(x)}{x-x_i}C'_m\,(x_i)+\right.$$

$$+\frac{C_m(x)}{x-x_i}\left(\pi S_m(x)-\pi S_m(x_i)-\frac{2\pi}{2m+1}(1+x_i)\left(C_m'(x_i)\right)^2\right)\Bigg]=$$
$$=\frac{\pi C_m(x)}{(x-x_i)^2}\left(\frac{2}{2m+1}(1+x_i)C_m'(x_i)+S_m(x)-S_m(x_i)\right).$$

Окончательно получим

$$A_i(x)=\frac{\pi}{\left(C_m'(x_i)\right)^2(x-x_i)}\left\{\frac{C_m(x)S_m(x)}{x-x_i}-\frac{2}{2m+1}(1+x_i)\left(C_m'(x_i)\right)^2-\frac{C_m''(x_i)}{C_m'(x_i)}C_m(x)S_m(x)\right\}. \tag{3.8}$$

При $m=1$ имеем

$$A_1(x)=\pi,\quad B_1(x)=\pi\left(x+\frac{1}{2}\right),$$
$$\int_{-1}^{1}\sqrt{\frac{1+t}{1-t}}\frac{\varphi(t)}{t-x}dt\approx\pi\varphi\left(\frac{1}{2}\right)+\pi\left(x+\frac{1}{2}\right)\varphi'\left(\frac{1}{2}\right).$$

4. КВАДРАТУРНЫЕ ФОРМУЛЫ ДЛЯ СИНГУЛЯРНОГО ИНТЕГРАЛА $\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}\frac{\varphi(t)}{t-x}dt$

Многочленами ортогональными по весу $p(t)=\sqrt{\frac{1-t}{1+t}}$ на отрезке $[-1,1]$ являются многочлены $S_m(x)=\frac{\sin\left(\frac{2n+1}{2}\arccos x\right)}{\sin\left(\frac{1}{2}\arccos x\right)}$ с корнями $x_k=\cos\frac{2\pi k}{2m+1},(k=1,2,\ldots,m)$.

В этом случае коэффициенты квадратурной формулы (1) $A_i(x)$ и $B_i(x)$ $(i=1,2,\ldots,m)$ подставляя $\omega(t)=\frac{S_m(t)}{2^m}$, вычисляются по формулам

$$A_i(x)=\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}\frac{1}{t-x}\left(\frac{S_m(t)}{S_m'(x_i)(t-x_i)}\right)^2\left(1-\frac{S_m''(x_i)}{S_m'(x_i)}(t-x_i)\right)dt, \tag{4.1}$$

$$B_i(x)=\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}\frac{1}{t-x}\left(\frac{S_m(t)}{S_m'(x_i)(t-x_i)}\right)^2(t-x_i)dt. \tag{4.2}$$

Пользуясь формулами [2, гл.7, §3], [3, гл. 2, § 2.10]

$$\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}f(t)dt\approx\frac{2\pi}{2m+1}\sum_{k=1}^{m}(1-x_k)f(x_k),\quad x_k=\cos\frac{2\pi k}{2m+1}, \tag{4.3}$$

$$\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}S_m(t)Q(t)dt=0, \tag{4.4}$$

$$\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}\frac{S_m(t)}{t-x}dt=-\pi C_m(x),\quad x\in(-1;1), \tag{4.5}$$

получим следующие выражения для коэффициентов $A_i(x)$ и $B_i(x)$:

$$B_i(x)=\frac{\pi C_m(x)S_m(x)}{\left(S_m{}'(x_i)\right)^2(x-x_i)}, \tag{4.6}$$

$$A_i(x)=\frac{\pi}{\left(S_m{}'(x_i)\right)^2(x-x_i)}\left\{-\frac{C_m(x)S_m(x)}{x-x_i}-\frac{2}{2m+1}(1-x_i)\left(S_m{}'(x_i)\right)^2+\frac{S_m''(x_i)}{S_m'(x_i)}C_m(x)S_m(x)\right\}. \tag{4.7}$$

При $m=1$ имеем

$$A_1(x)=-\pi,\quad B_1(x)=\pi\left(\frac{1}{2}-x\right),$$

$$\int_{-1}^{1}\sqrt{\frac{1-t}{1+t}}\frac{\varphi(t)}{t-x}dt\approx-\pi\varphi\left(-\frac{1}{2}\right)+\pi\left(\frac{1}{2}-x\right)\varphi'\left(-\frac{1}{2}\right).$$

5. ЗАКЛЮЧЕНИЕ

Построенные квадратурные формулы можно эффективно использовать для численного решения сингулярных интегральных уравнений методом квадратур.

Остаточный член построенных квадратурных формул выражается формулой

$$R_m(\varphi,x)=\int_{-1}^{1}\frac{p(t)}{t-x}\frac{\varphi^{(2m)}(\xi)}{(2m)!}\omega^2(t)dt,\quad \xi,x\in(-1;1). \tag{5.1}$$

СПИСОК ЛИТЕРАТУРЫ

1. *Гончаров В. Л.* Теория интерполирования и приближения функции. М-Л, РТТН, 1954. 328 с.
2. *Крылов В. И.* Приближенное вычисление интегралов. М.: Наука, 1967. 500 с.
3. *Хубежты Ш. С.* Квадратурные формулы для сингулярных интегралов и некоторые их применения. Владикавказ: ЮМИ ВНЦ РАН, 2011. 235 с.
4. *Лифанов И. К.* Метод сингулярных интегральных уравнений и численный эксперимент. М.: ТОО “Янус”. 1995. – 520 с.

QUADRATURE FORMULAS FOR SINGULAR INTEGRALS CONTAINING THE VALUES OF A FUNCTION AND ITS DERIVATIVES

Sh. S. Khubezhty[a,b,c,*], L. Yu. Plieva[a,b,c,]**

[a] 362025 Vladikavkaz, 44-46 Vatutina St., K. L. Khetagurov North Ossetian State University, Republic of North Ossetia – Alania

[b] 362025 Vladikavkaz, 53 Vatutina str., Southern Mathematical Institute, branch of Vladikavkaz Scientific Center, RAS, Republic of North Ossetia – Alania,

[c] 362002 Vladikavkaz, Molodezhnaya str., 7, Financial University under the Government of the Russian Federation, Vladikavkaz branch, Republic of North Ossetia – Alania

**e-mail: shalva57@rambler.ru*

***e-mail: plieva-21@mail.ru*

Received: 18.04.2024
Revised: 18.04.2024
Accepted: 23.08.2024

Abstract. Quadrature formulas for singular integrals on the integration interval $[-1,1]$ with certain weight functions $p(t)$ are constructed. The values of the function and its derivatives in zeros of the Chebyshev polynomial are used in the construction. The resulting formulas are quadrature formulas of the interpolation type and have an algebraic degree of accuracy of $2m-1$. The error estimate is given.

Keywords: singular integrals, quadrature formula, nodes, derivative of a function, weight function, interpolation polynomial, residual term.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2312–2331

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 517.9

МЕТОД ВОЗМУЩЕНИЙ И РЕГУЛЯРИЗАЦИЯ ПРИНЦИПА ЛАГРАНЖА В НЕЛИНЕЙНЫХ ЗАДАЧАХ НА УСЛОВНЫЙ ЭКСТРЕМУМ[1)]

© 2024 г. М. И. Сумин[1,*]

[1]392000 Тамбов, ул. Интернациональная, 33, ТГУ им. Г.Р. Державина, Россия
**e-mail: m.sumin@mail.ru*

Поступила в редакцию 08.07.2024 г.
Переработанный вариант 08.07.2024 г.
Принята к публикации 23.08.2024 г.

Рассматривается регуляризация принципа Лагранжа (ПЛ) в недифференциальной форме в нелинейной (невыпуклой) задаче на условный экстремум с операторным ограничением-равенством в гильбертовом пространстве. Множество ее допустимых элементов принадлежит полному метрическому пространству, существование решения задачи априори не предполагается. Ограничение-равенство содержит аддитивно входящий в него параметр, что обеспечивает возможность применения для исследования задачи "нелинейного варианта" метода возмущений. Основное предназначение регуляризованного ПЛ — устойчивое генерирование обобщенных минимизирующих последовательностей (ОМП) в рассматриваемой нелинейной задаче. Его можно трактовать как ОМП-образующий (регуляризирующий) оператор, ставящий в соответствие каждому набору исходных данных задачи субминималь (минималь) ее отвечающего этому набору регулярного модифицированного функционала Лагранжа (МФЛ), двойственная переменная в котором генерируется в соответствии с процедурой стабилизации по Тихонову двойственной задачи. Конструкция МФЛ полностью определяется видом "нелинейных" субдифференциалов полунепрерывной снизу и, вообще говоря, невыпуклой функции значений как функции параметра задачи. В качестве таких субдифференциалов используются хорошо известные в негладком (нелинейном) анализе проксимальный субградиент и субдифференциал Фреше. Регуляризованный ПЛ "преодолевает" свойства некорректности классического аналога и может трактоваться как регуляризирующий алгоритм, составляя тем самым теоретическую основу для создания устойчивых методов практического решения нелинейных задач на условный экстремум. Библ. 32.

Ключевые слова: нелинейная задача на условный экстремум, операторное ограничение-равенство, правило множителей Лагранжа в недифференциальной форме, регуляризация, метод возмущений, функция значений, проксимальный субградиент, субдифференциал Фреше, модифицированная двойственная задача, обобщенная минимизирующая последовательность, регуляризирующий алгоритм.

DOI: 10.31857/S0044466924120076, **EDN**: KBZJNN

ВВЕДЕНИЕ

Опубликованное в 1788 г. классическое правило множителей Лагранжа (см. [1]), которое по сложившейся традиции называют также принципом Лагранжа (ПЛ), можно трактовать как сформулированный классиком результат, являющийся естественным итогом многолетних усилий ведущих математиков того времени, искавших ответ на естественный вопрос о том, как решать задачи на максимум и минимум, возникающие в самых различных ситуациях, связанных с человеческой деятельностью, в том числе, и сугубо практического характера (см. [2]–[4]). За прошедшие два с лишним столетия теория ПЛ получила фундаментальное развитие и стала основой таких важнейших современных математических дисциплин, как теория экстремальных задач, теория оптимального управления (см. [2]–[4]). ПЛ лежит также в основе современных численных методов в теории условной оптимизации и стало быть имеет непосредственное отношение к практическому решению самых различных оптимизационных задач (см. [5]). Вместе с тем, при решении большого числа важных задач условной оптимизации, возникающих, в частности, при исследовании проблем современного естествознания, непосредственное применение этого классического результата сталкивается с трудностями принципиального характера,

[1)]Результаты исследований автора, представленные в разд. 1, 3, получены при финансовой поддержке РНФ № 23-11-20020, https://rscf.ru/project/23-11-20020/; результаты исследований, представленные в разд. 2, получены при финансовой поддержке Минобрнауки Тамбовской области № 2-ФП-2023.

вызванными хорошо известными свойствами некорректности таких задач (см. [5]), а также связанными с ними свойствами некорректности самого ПЛ (см. [6]–[9]). Мы говорим здесь о непосредственном применении ПЛ, если в нашем распоряжении имеется тот или иной алгоритм (оператор), который позволяет выделять удовлетворяющие ему допустимые элементы из всей совокупности составляющих ПЛ соотношений.

Кратко проиллюстрируем сказанное относительно свойств некорректности ПЛ на примере "простейшей" задачи на условный экстремум с операторным (т.е. задаваемым оператором с бесконечномерным образом) ограничением-равенством в гильбертовом пространстве

$$(P) \qquad \|z\|^2 \to \inf, \quad Az = h, \quad z \in \mathcal{D} \subset Z,$$

где $A : Z \to H$ — линейный ограниченный оператор, $h \in H$ — заданный элемент, $\mathcal{D}$ — выпуклое замкнутое множество, Z, H — гильбертовы пространства. Применительно к задаче (P) выделим в данном контексте три основные позиции, связанные с такими свойствами некорректности ПЛ, как невыполнимость и неустойчивость (см. [6]–[9]). Напомним, что о невыполнимости правила множителей мы говорим тогда, когда известно, что это классическое правило в той или иной задаче на условный экстремум не может быть записано. В свою очередь, неустойчивость классического правила мы понимаем в том смысле, что неустойчиво по отношению к возмущению исходных данных экстремальной задачи ведут себя допустимые элементы, удовлетворяющие всем составляющим правило множителей соотношениям.

1. Когда можно записать соотношения ПЛ, выполнимость ПЛ. Известные подходы к выводу ПЛ (см. [2], [4]) требуют замкнутости образа оператора A. При этом, как отмечено в [2, разд. 3.2.4, с. 260], невыполнение условия замкнутости может приводить к тому, что ПЛ вовсе невозможно записать (см. также соответствующие примеры в [6]–[8]). Это условие не выполняется, например, в случае вполне непрерывного оператора A (см. [10, с. 225, теорема 1]), что является естественным для задач оптимизации, связанных с дифференциальными уравнениями (оптимальное управление, обратные задачи). Подход к выводу ПЛ для задачи (P) с помощью метода возмущений (см., например, [2, разд. 3.3.2]), использующий включение этой задачи в семейство аналогичных задач, зависящих от параметра $p \in H$, вида

$$(P_p) \qquad \|z\|^2 \to \inf, \quad Az = h + p, \quad z \in \mathcal{D} \subset Z,$$

предполагает жесткую связь соотношений ПЛ с субдифференциальными свойствами функции значений этой задачи. Именно, как показано в [6, теорема 2.1], [11, теорема 1.1], этот подход позволяет формально получить невырожденный (регулярный или нерегулярный) ПЛ в задаче $(P) = (P_0)$ тогда и только тогда, когда имеет место хотя бы одно из двух соотношений, $\partial\beta(0) \neq \emptyset$ или $\partial^\infty\beta(0) \neq \{0\}$, где $\partial\beta(0)$ и $\partial^\infty\beta(0)$ — субдифференциал и сингулярный (асимптотический) субдифференциал (в смысле выпуклого анализа) в нуле выпуклой полунепрерывной снизу функции значений $\beta(p) \equiv \min\limits_{z \in \mathcal{D},\, Az = h+p} \|z\|^2$, $p \in H$. Однако, к сожалению, проверка выполнимости нужных субдифференциальных свойств функции значений сама по себе является сложной математической задачей.

2. Когда неизвестно, можно ли записать соотношения ПЛ или нет, возможная невыполнимость ПЛ. Применительно к задаче (P) классический ПЛ для гладких задач с равенствами из книги [2, с. 253, 254] (см. также [4, с. 12, следствие 1]) в случае $\mathcal{D} = Z$, когда Z, H — гильбертовы пространства, формулируется как следствие выполнимости определенного набора соответствующих условий. Если же указанные условия не выполняются, то и о выполнимости ПЛ соответствующие утверждения [2, с. 253, 254], [4, с. 12, следствие 1] никакой информации не несут. Хорошо известен пример невыполнимости ПЛ в [2, разд. 3.2.4], примеры невыполнимости ПЛ применительно к задаче (P) можно найти в [6]–[9].

3. Неустойчивость ПЛ. Наконец, если задача (P) такова, что в ней все же "можно записать" ПЛ, то "практическое" использование классического результата (например, при нахождении приближений к решению задачи) неизбежно наталкивается на проблему его неустойчивости по отношению к возмущению ее исходных данных (см. [6]–[9]).

Приведенный выше анализ примера (P), анализ других подобных примеров [6]–[9] дают основание рассматривать классический ПЛ как математический объект с присущими ему от природы свойствами некорректности. Такая аргументация, подобная той, которая принята в теории некорректных задач [12], явилась причиной, в соответствии с которой в [6], [7] (см. также [8]) было предложено регуляризовать ПЛ в недифференциальной форме, взяв за основу связанный двойственностью подход к регуляризации (см. [13]–[15]). В этом случае центральная роль в задачах на условный экстремум естественным образом переходит (подробности см. в [6]–[9]) от классического понятия оптимального элемента к понятию обобщенной минимизирующей последовательности (ОМП) (см. определение в разд. 1). В математическом программировании такие последовательности часто называют оптимальными обобщенными планами [16], в оптимальном управлении — мини-

мизирующими приближенными решениями (см. [17]). В [6], [7], [9], [18] такой подход был реализован применительно к выпуклой задаче на условный экстремум с операторным ограничением-равенством и конечным числом функциональных ограничений-неравенств (в [9] рассматривалась непосредственно задача (P)). Здесь и ниже под операторным понимается ограничение, задаваемое оператором с бесконечномерным образом. Ограничения задач в [6], [7], [9], [18] содержат аддитивно входящие в них параметры, что позволило в указанных работах воспользоваться так называемым методом возмущений [2, разд. 3.3.2]. Регуляризованные ПЛ в недифференциальной форме (см. [6], [7], [9], [18]): 1) формулируются как теоремы существования ОМП в исходной задаче, состоящих из минималей регулярных функционалов Лагранжа, двойственные переменные для которых генерируются в соответствии с выбранной процедурой регуляризации двойственных задач; 2) обобщают правило множителей, приводят к нему "в пределе" и сохраняют общую структуру классического аналога; 3) являются условиями обычной оптимальности, но выраженными в секвенциальной форме в терминах регулярных функционалов Лагранжа; 4) "преодолевают" свойства некорректности своего классического аналога и представляют собой универсальные ОМП-образующие (регуляризирующие) алгоритмы в смысле [19], [20] для решения задач на условный экстремум; 5) характеризуются жесткой связью свойств сходимости (по основной и двойственной переменным) с субдифференциальными (в смысле выпуклого анализа) свойствами их функций значений как функций параметров задач.

Настоящая работа, продолжающая линию работ [6], [7], [9], [18], непосредственно опирается на "нелинейную" версию основанной на двойственности регуляризации [15] (см. также [14]). Присутствие словосочетания "регуляризация принципа Лагранжа" в навании статьи связано с тем, что, в отличие от [15], регуляризация классического правила ниже проводится не только в случае регулярности (случай **А)**, см. п. 2.6) нелинейной (невыпуклой) задачи, но и в случае, когда такая регулярность отсутствует (случай **Б)**, см. п. 2.6). Здесь и ниже, как и в [15], регулярность понимается в смысле существования в задаче обобщенного вектора Куна–Таккера (см. п. 2.6). В работе показывается, как регуляризация ПЛ в недифференциальной форме может быть организована для нелинейной (невыпуклой) задачи на условный экстремум с допустимым множеством из полного метрического пространства и с операторным ограничением-равенством в гильбертовом пространстве. Рассматриваемая ниже задача так же, как и в [6], [7], [9], [18], содержит аддитивно входящий в ограничение-равенство бесконечномерный параметр, что, во-первых, объясняет наличие словосочетания "метод возмущений" в названии статьи и, во-вторых, предопределяет существенную опору при регуляризации ПЛ на "нелинейный вариант" метода возмущений и конструкции современного негладкого анализа (см. [21]–[24]). В отличие от случая выпуклой задачи (см. [6], [7], [9], [18]), в котором центральная роль принадлежит субдифференциалам (в смысле выпуклого анализа) выпуклой полунепрерывной снизу функции значений, ниже эту роль в нелинейной (вообще говоря, невыпуклой) задаче играют уже субдифференциалы в смысле негладкого (нелинейного) анализа ее нелинейной (вообще говоря, невыпуклой) полунепрерывной снизу функции значений. В качестве указанных "нелинейных" субдифференциалов выступают так называемые проксимальный субградиент (см. [21]–[23]) и субдифференциал Фреше (см. [21], [24], см. разд. 2.1). Важнейшее значение при этом имеет тот факт, что как проксимальный субградиент, так и субдифференциал Фреше "имеют смысл" в точках плотного множества в эффективном множестве функции значений нелинейной задачи (так называемая плотность обобщенной субдифференцируемости, см. п. 2.2). Это, в свою очередь, влечет выполнимость характеристического свойства рассматриваемой нелинейной параметрической задачи, в соответствии с которым можно утверждать, что регулярность задачи на условный экстремум "реализуется достаточно часто", а именно, множество всех соответствующих регулярным задачам значений параметра всюду плотно в эффективном множестве ее функции значений (см. п. 2.6). Условия обычной оптимальности в недифференциальной форме для рассматриваемой ниже задачи в двух случаях ее регулярности (проксимальный субградиент непуст (см. [11, теорема 2.1]), субдифференциал Фреше непуст (см. [11, теорема 2.2])), т.е. существования соответствующего обобщенного вектора Куна–Таккера можно найти в [11]. Работа призвана, в частности, показать, что большая часть перечисленных в [6], [7], [9], [18] характеристических свойств регуляризованного ПЛ в случае выпуклых задач (см. предыдущий абзац) сохраняется и в задачах нелинейных.

Итак, основные усилия в работе направлены на получение регуляризованного ПЛ в недифференциальной форме в нелинейной (невыпуклой) задаче на условный экстремум. Его главное предназначение — устойчивое генерирование ОМП из субминималей (минималей) ее функции Лагранжа, а точнее, ее модифицированной функции Лагранжа (МФЛ), взятой при значениях двойственной переменной, вырабатываемых соответствующей процедурой регуляризации двойственной задачи. Естественность используемых ниже конструкций МФЛ в нелинейной задаче достаточно подробно обсуждается в п. 2.3. Указанное генерирование обеспечивается в нелинейной задаче в общей ситуации, если только ее функция значений конечна в соответствующей этой задаче точке, т.е. вне зависимости от свойств ее "нелинейной" субдифференцируемости в этой точке. Оператор, ставящий в соответствие набору возмущенных исходных данных указанные субминимали (минимали) мы называем ОМП-образующим (см. определение 1.1), так как последовательно взятые соответствующие им значе-

ния целевого функционала и задающего ограничение оператора, при стремлении возмущения задачи к нулю, обеспечивают приближение к ее нижней грани (сходимость по функции) и выполнение в пределе операторного равенства (сходимость "по ограничениям"). В некоторых частных случаях такая одновременная "сходимость по функции и по ограничениям" влечет и сходимость по аргументу к решению задачи (см. замечания 3.1, 3.2). Здесь существенным является то, что вид используемых в этих случаях МФЛ полностью определяется видом используемых соответствующих "нелинейных" субдифференциалов полунепрерывной снизу функции значений задачи — проксимального субградиента и субдифференциала Фреше (см. разд. 2.3). Следует отметить, что различные конструкции МФЛ хорошо известны в научной литературе (см., например, книги [25]–[27] и их библиографию). Однако, в отличие от [25]–[27], ниже в статье эти конструкции полностью определяются соответствующими конструкциями "нелинейных" субдифференциалов полунепрерывной снизу функции значений нелинейной задачи в бесконечномерном гильбертовом пространстве (подробности см. в п. 2.3).

Конструированием ОМП из субминималей (минималей) МФЛ излагаемая ниже регуляризация ПЛ в нелинейной задаче на условный экстремум существенно отличается от регуляризации по Тихонову нелинейных задач условной оптимизации [5, гл. 9], при которой минимизирующая последовательность в нелинейной задаче с конечным числом функциональных ограничений типа равенства и неравенстваа строится из субминималей (минималей) соответствующих функционалов Тихонова. Одновременно, в отличие от регуляризации по Тихонову задач условной оптимизации [5, гл. 9], где регуляризация происходит по функции (задачи первого типа [5, гл. 9]) или по аргументу (задачи второго типа [5, гл. 9]), здесь регуляризация идет одновременно по функции и "по ограничениям", что обеспечивает во многих частных случаях регуляризацию и по аргументу. Существование решения рассматриваемой ниже нелинейной задачи априори не предполагается. Однако в случае его существования (см. замечания 1.1, 3.1) конструируемая ОМП, в зависимости от свойств задачи, в том или ином смысле может к нему сходиться (см. замечание 3.2). Получаемый ниже регуляризованный ПЛ естественно трактовать как теоретическую базу для создания устойчивых методов практического решения нелинейных задач условной оптимизации. Эффективность его практической реализации напрямую определяется "качеством минимизации" МФЛ, которое, в свою очередь, существенно связано, как это обычно бывает в "нелинейных ситуациях", со спецификой конкретной практической задачи. Заметим, наконец, что регуляризация ПЛ в регулярном случае была рассмотрена ранее применительно к нелинейной задаче оптимального управления в работе [28]. В свою очередь, в [29] аналогичные вопросы были рассмотрены для регулярной нелинейной задачи условной оптимизации с операторным ограничением-равенством и конечным числом функциональных ограничений-неравенств.

Статья состоит из введения и трех основных разделов, первый из которых посвящен постановке нелинейной (невыпуклой) задачи условной оптимизации в полном метрическом пространстве и фомулировке необходимых результатов и понятий, в частности, центрального для всей работы понтия ОМП-образующего оператора. Во втором разделе приводятся основные вспомогательные факты, связанные с "нелинейными" субдифференциалами полунепрерывных снизу функций в гильбертовом пространстве. Далее здесь показывается как эти субдифференциалы "порождают" естественным образом соответствующие им МФЛ и, как следствие, связанную с ними модифицированную двойственную задачу. Наконец, третий раздел посвящен регуляризации ПЛ в рассматриваемой нелинейной задаче в случае, когда регуляризация двойственной к ней задачи, являющейся с точностью до знака целевого функционала выпуклой задачей минимизации общего вида (т.е. без ограничений типа равенства и неравенства, см., например, задачу (3) в [5, гл. 9, §2]), организуется в соответствии с привычной схемой стабилизации по Тихонову (см., например, [5, гл. 9, §4]). Здесь же обсуждается связь основанной на двойственности регуляризации с классическим методом Тихонова, а также указывается на ее обусловленное опорой на теорию двойственности существенное отличие от последнего.

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим параметрическую (т.е. зависящую от параметра) нелинейную (невыпуклую) задачу минимизации

$$(P_p) \qquad f(z) \to \inf, \quad g(z) = p, \quad z \in \mathcal{D} \subset Z,$$

где $f : \mathcal{D} \to R^1$ – непрерывный функционал, $g : \mathcal{D} \to H$ – непрерывный оператор с компактной в H областью значений $g(\mathcal{D}) \equiv \{y = g(z) \in H : z \in \mathcal{D}\}$, $\mathcal{D} \subset Z$ — замкнутое ограниченное множество в полном метрическом пространстве Z с метрикой ρ, H — гильбертово пространство, $p \in H$ — параметр. Здесь и ниже компактность множества понимается в том смысле, что из каждой бесконечной последовательности его элементов можно извлечь сходящуюся подпоследовательность (см., например, [10, разд. 19.3]) (во многих случаях в подобной

ситуации используется термин предкомпактность). Будем также считать, что

$$|f(z_1) - f(z_2)| \le L\rho(z_1, z_2),\ \|g(z_1) - g(z_2)\| \le L\rho(z_1, z_2)\ \forall\, z_1,\ z_2 \in \mathcal{D}, \tag{1.1}$$

где $L > 0$ не зависит от $z_1,\ z_2 \in \mathcal{D}$.

Пусть F – множество всевозможных наборов исходных данных $\mathsf{f} \equiv \{f,\ g\}$, каждый из которых состоит из непрерывного на $\mathcal{D}$ функционала f, непрерывного оператора g с компактной в H областью значений $g(\mathcal{D})$ и с указанными выше свойствами (1.1) с независящей от набора постоянной L. Определим наборы невозмущенных $\mathsf{f}^0 \in \mathsf{F}$ и возмущенных $\mathsf{f}^\delta \in \mathsf{F}$ исходных данных соответственно: $\mathsf{f}^0 \equiv \{f^0,\ g^0\}$ и $\mathsf{f}^\delta \equiv \{f^\delta,\ g^\delta\}$, $\delta \in (0, \delta_0]$, $\delta_0 > 0$, – некоторое число. Будем считать, что выполняются следующие оценки:

$$|f^\delta(z)|,\ |g^\delta(z)| \le N,\ \ |f^\delta(z) - f^0(z)| \le K\delta,\ \ \|g^\delta(z) - g^0(z)\| \le K\delta\ \ \forall\, z \in \mathcal{D}, \tag{1.2}$$

где $N > 0,\ K > 0$ — некоторые не зависящие от δ постоянные.

Обозначим задачу (P_p), функционал f, оператор g, соответствующие набору исходных данных f^δ, $\delta \in [0, \delta_0]$, через (P_p^δ), f^δ, g^δ соответственно. Обозначим также $\mathcal{D}_p^{\delta,\epsilon} \equiv \{z \in \mathcal{D} :\ \|g^\delta(z) - p\| \le \epsilon\},\ \epsilon \ge 0,\ \mathcal{D}_p^{0,0} \equiv \mathcal{D}_p^0$. Определим обобщенную функцию значений (S-функцию) $\beta :\ H \to R^1 \cup \{+\infty\}$ задачи (P_p^0)

$$\beta(p) \equiv \lim_{\epsilon \to +0} \beta_\epsilon(p),\ \ \beta_\epsilon(p) \equiv \inf_{z \in \mathcal{D}_p^{0,\epsilon}} f^0(z), \qquad \beta_\epsilon(p) \equiv +\infty,\ \text{если}\ \mathcal{D}_p^{0,\epsilon} = \emptyset.$$

Очевидно, в общей ситуации $\beta(p) \le \beta_0(p)$, где $\beta_0(p) \equiv \inf\limits_{z \in \mathcal{D}_p^0} f^0(z)$ — классическое значение задачи (P_p^0). Справедлива следующая важная для дальнейших построений лемма.

Лемма 1.1. *Функция значений $\beta :\ H \to R^1 \cup \{+\infty\}$ является полунепрерывной снизу.*

Лемма доказывается точно так же, как и в случае выпуклой задачи в [11, лемма 1.2].

Определим обобщенную минимизирующую последовательность (ОМП) в задаче (P_p^0) как последовательность элементов $z^i \in \mathcal{D},\ i = 1, 2, \ldots$, такую, что $f^0(z^i) \le \beta(p) + \delta^i,\ z^i \in \mathcal{D}_p^{0,\epsilon^i}$ для некоторых последовательностей сходящихся к нулю неотрицательных чисел $\delta^i,\ \epsilon^i,\ i = 1, 2, \ldots$. Можно утверждать, что ввиду ограниченности $\mathcal{D}$ величина $\beta(p)$ конечна тогда и только тогда, когда в задаче (P_p^0) существует "допустимая" последовательность $z^i \in \mathcal{D}_p^{0,\epsilon^i},\ i = 1, 2, \ldots$, где $\epsilon^i,\ i = 1, 2, \ldots$, — некоторая последовательность сходящихся к нулю неотрицательных чисел.

С учетом приближенного задания исходных данных мы имеем вместо задачи (P_p^0) семейство зависящих от характеризующей ошибку их задания величины δ задач

$$(P_p^\delta) \qquad\qquad f^\delta(z) \to \inf,\ \ g^\delta(z) = p,\ z \in \mathcal{D}.$$

Верхний индекс δ в исходных данных задачи (P_p^δ) означает, что они либо заданы точно ($\delta = 0$), либо являются возмущенными ($\delta > 0$), т.е. задаются с ошибкой, величину которой и характеризует параметр $\delta \in [0, \delta_0]$, $\delta_0 > 0$ — некоторое фиксированное число.

Определим ОМП-образующий оператор (см. [28], [29]) (определение ОМП-образующего оператора для выпуклых задач см. в [19], [20]), обсуждение взаимосвязи этого понятия с классическим понятием регуляризирующего оператора [12] в случае задачи частного вида (P) см. в [9].

Определение 1.1. Пусть $\delta^k \in (0, \delta_0),\ k = 1, 2, \ldots$, — сходящаяся к нулю последовательность положительных чисел. Зависящий от $\delta^k,\ k = 1, 2, \ldots$, вообще говоря, многозначный оператор $R_p(\cdot, \delta^k)$, ставящий в соответствие каждому набору исходных данных f^{δ^k}, удовлетворяющих оценкам (1.2) при $\delta = \delta^k$, множество $R_p(\mathsf{f}^{\delta^k}, \delta^k) \equiv \mathcal{W}_p^{\delta^k} \subset \mathcal{D}$, называется *ОМП-образующим* в задаче (P_p^0), если любая последовательность $z^{\delta^k} \in \mathcal{W}_p^{\delta^k},\ k = 1, 2, \ldots$, есть ОМП в этой задаче.

Замечание 1.1. Обсуждение условий, при которых задача (P_p^0) разрешима см. ниже в замечании 3.1. В свою очередь, в замечании 3.2 можно найти обсуждение условий, при которых конструируемые ниже ОМП обладают свойствами слабой или сильной сходимости к решениям задачи (P_p^0) в случае гильбертова пространства Z.

2. СУБДИФФЕРЕНЦИАЛЫ ПОЛУНЕПРЕРЫВНЫХ СНИЗУ ФУНКЦИЙ, МФЛ И ДВОЙСТВЕННЫЕ ЗАДАЧИ

При получении основных результатов центральную роль будут играть конструкции МФЛ. Эти конструкции естественным образом порождаются соответствующими конструкциями обобщенных субдифференциалов

полунепрерывной снизу функции значений рассматриваемой оптимизационной задачи. Обобщенные субдифференциалы являются в последние десятилетия предметом интенсивного изучения в негладком анализе (см. [21]–[24]). Ниже нам понадобятся два важнейших в негладком анализе понятия нормалей к замкнутым множествам и соответствующие понятия субдифференциалов (субградиентов) полунепрерывных снизу функций в банаховых пространствах.

2.1. Субдифференциалы полунепрерывных снизу функций

Первым из используемых ниже двух понятий субдифференциалов является понятие проксимального субградиента полунепрерывной снизу функции (см., например, [21]–[23]) в гильбертовом пространстве. Напомним кратко необходимые факты, связанные с этим понятием. Для этого сначала напомним понятие проксимальной нормали (см., например, [22, гл. 4, 5], [23, гл. 1, разд. 1, 2]).

Определение 2.1. (а) Пусть H — гильбертово пространство, $S \subset H$ — замкнутое множество, $\bar{s} \in S$. Вектор $\zeta \in H$ называется *проксимальной нормалью* к множеству S в точке $\bar{s} \in S$, если существует постоянная $M > 0$ такая, что

$$\langle \zeta, s - \bar{s} \rangle \le M\|s - \bar{s}\|^2 \ \ \forall\, s \in S.$$

Множество всех таких векторов ζ, представляющее собой конус, обозначается через $\hat{N}_S(\bar{s})$ и называется *проксимальным нормальным конусом*.

(б) Пусть $f : H \to R^1 \cup \{+\infty\}$ полунепрерывная снизу функция и $\bar{x} \in \operatorname{dom} f$. Вектор $\zeta \in H$ называется *проксимальным субградиентом* функции f в точке $\bar{x}$, если $(\zeta, -1) \in \hat{N}_{\operatorname{epi} f}(\bar{x}, f(\bar{x}))$. Множество всех таких векторов ζ обозначается через $\partial^P f(\bar{x})$ и называется проксимальным субградиентом f в точке $\bar{x}$.

Замечание 2.1. Можно показать, что ζ есть проксимальная нормаль к S в $\bar{s}$ тогда и только тогда, когда $\bar{s}$ есть ближайшая в S точка к некоторой точке вида $\bar{s} + t\zeta,\ t > 0$.

Напомним, наконец, критерий того, что данный вектор является проксимальным субградиентом полунепрерывной снизу функции в заданной точке (см., например, [22, утверждение 4A.3]).

Лемма 2.1. *Пусть H – гильбертово пространство, $f : H \to R^1 \cup \{+\infty\}$ – полунепрерывная снизу функция и $\bar{x} \in \operatorname{dom} f$. Вектор $\zeta \in H$ является проксимальным субградиентом функции f в точке $\bar{x}$, т.е. $\zeta \in \partial^P f(x)$, тогда и только тогда, когда существуют постоянные $R > 0$ и $\delta > 0$ такие, что*

$$\langle \zeta, x - \bar{x} \rangle \le f(x) - f(\bar{x}) + R\|x - \bar{x}\|^2 \ \ \forall\, x \in S_\delta(\bar{x}) \equiv \{x' \in H : \|x' - \bar{x}\| < \delta\}$$

или

$$f(\bar{x}) - \langle \zeta, \bar{x} \rangle \le f(x) - \langle \zeta, x \rangle + R\|x - \bar{x}\|^2 \ \ \forall\, x \in S_\delta(\bar{x}).$$

Напомним далее понятие нормали Фреше к замкнутому множеству в банаховом пространстве, а также соответствующее понятие субдифференциала Фреше полунепрерывной снизу функции [21, 24]. Следующие два определения, а также замечание могут быть найдены в [24, разд. 1.1.1, 1.3.2].

Определение 2.2. Пусть Ω – непустое множество банахова пространства X. Пусть $x \in cl\,\Omega$ и $u \xrightarrow{\Omega} x$ означает, что $u \to x$ с $u \in \Omega$. Тогда непустое множество

$$\hat{N}(x; \Omega) \equiv \{x^* \in X^* : \limsup_{u \xrightarrow{\Omega} x} \frac{\langle x^*, u - x \rangle}{\|u - x\|} \le 0\},$$

являющееся конусом, называется *нормальным конусом Фреше* к Ω в точке x. При $x \notin cl\,\Omega$ полагается $\hat{N}(x; \Omega) = \emptyset$.

Определение 2.3. Пусть $f : X \to R^1 \cup \{+\infty\}$ – полунепрерывная снизу функция, определенная на банаховом пространстве X, $\bar{x} \in \operatorname{dom} f$. Множество

$$\hat{\partial} f(\bar{x}) \equiv \{x^* \in X^* : (x^*, -1) \in \hat{N}((\bar{x}, f(\bar{x})); \operatorname{epi} f)\},$$

называется *субдифференциалом Фреше* функции f в точке $\bar{x}$. При этом полагается $\hat{\partial} f(\bar{x}) = \emptyset$ в случае $x \notin \operatorname{dom} f$.

Справедливо следующее

Замечание 2.2. Субдифференциал $\hat{\partial} f(\bar{x})$ может быть записан в виде

$$\hat{\partial} f(\bar{x}) = \{x^* \in X^* : \liminf_{u \to x} \frac{f(u) - f(x) - \langle x^*, u - x \rangle}{\|u - x\|} \ge 0\}.$$

Справедлива также следующая [24, утверждение 1.84] лемма.

Лемма 2.2. *Пусть $f : X \to R^1 \cup \{+\infty\}$ — полунепрерывная снизу функция, определенная на банаховом пространстве X, $x \in \operatorname{dom} f$. Тогда $x^* \in \hat{\partial} f(x)$ в том и только в том случае, если для любого $\epsilon > 0$ существует окрестность X_ϵ точки x такая, что*

$$f(x') - f(x) - \langle x^*, x' - x \rangle + \epsilon\|x' - x\| \ge 0 \ \ \forall\, x' \in X_\epsilon$$

или

$$f(x) - \langle x^*, x\rangle \le f(x') - \langle x^*, x'\rangle + \epsilon\|x' - x\| \ \ \forall\, x' \in X_\epsilon.$$

2.2. Плотность обобщенной субдифференцируемости

Важнейшим свойством полунепрерывных снизу функций $f : X \to R^1 \cup \{+\infty\}$ является то, что как множество $\partial^P f(x)$ в случае гильбертова пространства X, так и множество $\hat\partial f(x)$ в случае пространства X из достаточно обширного класса банаховых пространств (подробности см., например, в [21]–[24]) не пусто для плотного в $\operatorname{dom} f$ множества точек x. В настоящей работе в качестве пространства X выступает гильбертово пространство H, для которого указанные выше свойства заведомо справедливы (см., например, [23, теорема 3.1], [24, следствие 2.29]).

2.3. Субдифференциалы функции значений и порождаемые ими модифицированные функции Лагранжа

Как показано в [14], [15], [30], для задачи (P_p) с полунепрерывной снизу функцией значений $\beta : H \to \to R^1 \cup \{+\infty\}$ являются естественными две конструкции МФЛ в зависимости от субдифференциальных свойств β в фиксированной индивидуальной точке p.

2.3.1. Первая МФЛ. В соответствии с п. 2.2 множество $\{p \in \operatorname{dom}\beta : \partial^P\beta(p) \ne \emptyset\}$ всюду плотно в $\operatorname{dom}\beta$. Если точка $p \in \operatorname{dom}\beta$ такова, что $\partial^P\beta(p) \ne \emptyset$ и $\zeta \in \partial^P\beta(p)$, то для задачи (P_p) естественной является конструкция МФЛ вида

$$L^2_{p,c}(z,\zeta) \equiv f(z) - \langle \zeta, g(z) - p\rangle + c\|g(z) - p\|^2, \ \ z \in \mathcal{D},$$

с некоторым достаточно большим коэффициентом $c > 0$, зависящим, вообще говоря, от $\zeta \in \partial^P\beta(p)$, которому естественно и удобно придать смысл коэффициента штрафа. Покажем это, используя условие компактности образа $g(\mathcal{D})$.

Действительно, в этом случае из леммы 2.1 следует, что существуют постоянные $R > 0$ и $\delta > 0$ (зависящие от точки p и элемента ζ) такие, что

$$\beta(p) - \langle \zeta, p\rangle \le \beta(p') - \langle \zeta, p'\rangle + R\|p' - p\|^2 \ \ \forall\, p' \in S_\delta(p). \tag{2.1}$$

Так как в силу ограниченности множества $\mathcal{D}$ эффективное множество $\operatorname{dom}\beta$ ограничено и функция β ограничена на множестве $\operatorname{dom}\beta$, то в силу неравенства (2.1) можем записать для некоторой постоянной $c = c(p,\zeta) > 0$

$$\beta(p) - \langle \zeta, p\rangle \le \beta(p') - \langle \zeta, p'\rangle + c\|p' - p\|^2 \ \ \forall\, p' \in H \tag{2.2}$$

или

$$\beta(p) - \langle \zeta, p\rangle < \beta(p') - \langle \zeta, p'\rangle + \hat c\|p' - p\|^2 \ \ \forall\, p' \in H, \ \ p' \ne p, \qquad \hat c > c,$$

откуда в силу строгого неравенства $\hat c > c$, полунепрерывности снизу функции значений β и ограниченности множества $\operatorname{dom}\beta$ следует, что минимизирующей последовательностью в задаче минимизации

$$\beta(p') - \langle \zeta, p'\rangle + \hat c\|p' - p\|^2 \to \inf, \ \ p' \in H,$$

является лишь любая последовательность p^k, $k = 1, 2, \ldots$, сходящаяся к точке p такая, что $\beta(p^k) \to \beta(p)$, $k \to \infty$, и никакая другая последовательность. Отсюда следует, что в задаче минимизации МФЛ

$$L^2_{p,\hat c}(z,\zeta) \equiv f(z) - \langle \zeta, g(z) - p\rangle + \hat c\|g(z) - p\|^2 \to \inf, \ \ z \in \mathcal{D}, \tag{2.3}$$

минимизирующей является лишь последовательность z^k, $k = 1, 2, \ldots$, такая, что $f(z^k) \to \beta(p)$, $g(z^k) \to p$, $k \to \infty$, и никакая другая последовательность. При этом справедливо равенство

$$\inf_{z\in\mathcal{D}}(f(z) - \langle \zeta, g(z) - p\rangle + \hat c\|g(z) - p\|^2) = \beta(p). \tag{2.4}$$

Покажем это. Пусть $\bar z^k$, $k = 1, 2, \ldots$, — минимизирующая последовательность в задаче минимизации (2.3). Тогда благодаря ограниченности $\mathcal{D}$, ограниченности значений функции f и компактности области значений оператора g без ограничения общности считаем, что $f(\bar z^k) \to \bar f$, $g(\bar z^k) \to \bar p$, $k \to \infty$, где $\bar f \in R^1$, $\bar p \in H$ — некоторые элементы. Тогда рассмотрим два возможных случая.

Во-первых, если $\bar p = p$, но $\bar f \ne \beta(p)$, то в силу последних предельных соотношений, с учетом определения величины $\beta(p)$, можем записать

$$f(\bar z^k) + \langle -\zeta, g(\bar z^k) - p\rangle + \hat c\|g(\bar z^k) - p\|^2 \to \bar f > \beta(p).$$

Во-вторых, если же $p \neq \bar{p}$, то в силу тех же предельных соотношений, определения обобщенной нижней грани $\beta(\bar{p})$ и соотношений (2.2), можем записать одновременно $\bar{f} \geq \beta(\bar{p})$,

$$f(\bar{z}^k) + \langle -\zeta, g(\bar{z}^k) - p\rangle + \hat{c}\|g(\bar{z}^k) - p\|^2 \to \bar{f} + \langle -\zeta, \bar{p} - p\rangle + \hat{c}\|\bar{p} - p\|^2 \geq$$

$$\geq \beta(\bar{p}) + \langle -\zeta, \bar{p} - p\rangle + \hat{c}\|\bar{p} - p\|^2 > \beta(p).$$

Полученные в обоих случаях строгие неравенства вступают в противоречие с предположением о том, что последовательность $\bar{z}^k$, $k = 1, 2, \ldots$, — минимизирующая, так как для последовательности z^k, $k = 1, 2, \ldots$, указанного выше вида ($f(z^k) \to \beta(p)$, $g(z^k) \to p$, $k \to \infty$) мы можем записать

$$f(z^k) + \langle -\zeta, g(z^k) - p\rangle + \hat{c}\|g(z^k) - p\|^2 \to \beta(p).$$

Далее, можно утверждать, что если для некоторых $\zeta \in H$, $\hat{c} > 0$ выполняется равенство (2.4), то выполняется и неравенство (2.2) при $c = \hat{c}$, а следовательно, и неравенство (2.1) при $R = \hat{c}$ и любом $\delta > 0$, т.е. $\zeta \in \partial^P \beta(p)$. Для того чтобы это показать, перепишем равенство (2.4) в виде

$$\beta(p) - \langle \zeta, p\rangle \leq f(z) - \langle \zeta, g(z)\rangle + \hat{c}\|g(z) - p\|^2 \ \forall\, z \in \mathcal{D}. \tag{2.5}$$

Пусть $z^k \in \mathcal{D}$, $k = 1, 2, \ldots$, $\bar{p} \in \operatorname{dom}\beta$ — такие произвольные последовательность и точка, что $f(\bar{z}^k) \to$ $\to \beta(\bar{p})$, $g(\bar{z}^k) \to \bar{p}$, $k \to \infty$. Тогда с учетом неравенства (2.5) получаем

$$\beta(p) - \langle \zeta, p\rangle \leq \beta(\bar{p}) - \langle \zeta, \bar{p}\rangle + \hat{c}\|\bar{p} - p\|^2,$$

т.е., действительно, неравенство (2.2) выполняется.

Более того, как показано в [30], если коэффициент штрафа $c = c(\zeta)$ можно взять независимым от $\zeta \in \partial^P \beta(p)$, то этот коэффициент можно считать столь большим, что в задаче минимизации МФЛ

$$L^2_{p,c}(z, \zeta) \equiv f(z) - \langle \zeta, g(z) - p\rangle + c\|g(z) - p\|^2 \to \inf, \ \ z \in \mathcal{D},$$

равенство

$$\inf_{z \in \mathcal{D}} (f(z) - \langle \zeta, g(z) - p\rangle + c\|g(z) - p\|^2) = \beta(p)$$

выполняется лишь при всех $\zeta \in \partial^P \beta(p)$, и минимизирующей в ней является лишь последовательность z^k, $k = 1, 2, \ldots$, такая, что $f(z^k) \to \beta(p)$, $g(z^k) \to p$, $k \to \infty$ и никакая другая последовательность.

2.3.2. Вторая МФЛ. В соответствии с п. 2.2 множество $\{p \in \operatorname{dom}\beta : \hat{\partial}\beta(p) \neq \emptyset\}$ всюду плотно в $\operatorname{dom}\beta$. Если точка $p \in \operatorname{dom}\beta$ такова, что $\hat{\partial}\beta(p) \neq \emptyset$ и $\zeta \in \hat{\partial}\beta(p)$, то для задачи (P_p) естественной является конструкция МФЛ вида

$$L^1_{p,c}(z, \zeta) \equiv f(z) - \langle \zeta, g(z) - p\rangle + c\|g(z) - p\|, \ \ z \in \mathcal{D},$$

с некоторым достаточно большим коэффициентом $c > 0$, зависящим, вообще говоря, от $\zeta \in \hat{\partial}\beta(p)$, который, как и в предыдущем случае, будем называть коэффициентом штрафа. Рассуждения, аналогичные тем, что и при получении первой МФЛ, опять же с опорой на условие компактности образа $g(\mathcal{D})$, можно провести и в случае использования понятия субдифференциала Фреше. В этом случае из леммы 2.2 следует, что существуют постоянные $R > 0$ и $\delta > 0$ такие, что

$$\beta(p) - \langle \zeta, p\rangle \leq \beta(p') - \langle \zeta, p'\rangle + R\|p' - p\| \ \ \forall\, p' \in S_\delta(p). \tag{2.6}$$

Так как в силу ограниченности множества $\mathcal{D}$ эффективное множество $\operatorname{dom}\beta$ ограничено и функция β ограничена на множестве $\operatorname{dom}\beta$, то в силу неравенства (2.6) можем записать для некоторой постоянной $c = c(p, \zeta) > 0$

$$\beta(p) - \langle \zeta, p\rangle \leq \beta(p') - \langle \zeta, p'\rangle + c\|p' - p\| \ \ \forall\, p' \in H$$

или

$$\beta(p) - \langle \zeta, p\rangle < \beta(p') - \langle \zeta, p'\rangle + \hat{c}\|p' - p\| \ \ \forall\, p' \in H, \ \ p' \neq p, \qquad \hat{c} > c,$$

откуда в силу полунепрерывности снизу функции значений β и ограниченности множества $\operatorname{dom}\beta$ следует, что минимизирующей последовательностью в задаче минимизации

$$\beta(p') - \langle \zeta, p'\rangle + \hat{c}\|p' - p\| \to \inf, \ \ p' \in H,$$

является лишь любая последовательность p^k, $k = 1, 2, \ldots$, сходящаяся к точке p такая, что $\beta(p^k) \to \beta(p)$, $k \to \infty$, и никакая другая последовательность. Отсюда, как и в случае первой МФЛ, следует, что в задаче минимизации второй МФЛ

$$L^1_{p,\hat{c}}(z,\zeta) \equiv f(z) - \langle \zeta, g(z) - p\rangle + \hat{c}\|g(z) - p\| \to \inf, \;\; z \in \mathcal{D}, \tag{2.7}$$

минимизирующей является лишь последовательность z^k, $k = 1, 2, \ldots$, такая, что $f(z^k) \to \beta(p)$, $g(z^k) \to p$, $k \to \infty$ и никакая другая последовательность. При этом справедливо равенство

$$\inf_{z\in\mathcal{D}}(f(z) - \langle \zeta, g(z) - p\rangle + \hat{c}\|g(z) - p\|) = \beta(p).$$

2.3.3. Комбинированная МФЛ. Далее определим, с учетом конструкций МФЛ в (2.3), (2.7) (см. также [25]–[27]), комбинированную (смешанную) МФЛ задачи (P^δ_p) (см. [14], [15], [30])

$$L^\delta_{p,c}(z,\lambda) \equiv \frac{1}{2}L^{\delta,1}_{p,2c}(z,\lambda) + \frac{1}{2}L^{\delta,2}_{p,2c}(z,\lambda) \equiv$$

$$\equiv f^\delta(z) + \langle \lambda, g^\delta(z) - p\rangle + c\psi(\|g^\delta(z) - p\|), z \in \mathcal{D},\; \lambda \in H,\; c \geq 0,$$

где функция $\psi:\; R^1_+ \to R^1_+$ определяется формулой

$$\psi(t) \equiv l_1 t + l_2 t^2,\; t \in R^1_+,$$

в которой весовые множители $l_1,\, l_2 \in \{0, 1\}$ — фиксированные числа. Функции ψ удобно придать смысл функции штрафа.

2.4. Модифицированная двойственная задача

Определим, в свою очередь, и модифицированную двойственную задачу

$$V^\delta_{p,c}(\lambda) \to \sup, \quad \lambda \in H, \quad V^\delta_{p,c}(\lambda) \equiv \inf_{z\in\mathcal{D}} L^\delta_{p,c}(z,\lambda).$$

Условия на исходные данные задачи (P^0_p) таковы, что вогнутая функция $V^\delta_{p,c}(\lambda),\; \lambda \in H$ при $\delta \in [0, \delta_0]$ является определенной (конечной) при любом $c \in R^1$ для любой точки $\lambda \in H$ и ограниченной на любом множестве вида $\{\lambda \in H :\; \|\lambda\| < C\}$, а значит, и локально липшицевой (см. [31, определение 2.3]) на $\operatorname{dom} V^\delta_{p,c} = H$ (см. [31, теорема 2.1, следствие 2.3]). Кроме того, справедлива оценка

$$|V^\delta_{p,c}(\lambda) - V^0_{p,c}(\lambda)| \leq C\delta(1 + \|\lambda\| + |c|) \;\; \forall \lambda \in H,\; c \geq 0, \tag{2.8}$$

где $C > 0$ — независящая от δ постоянная. Для ее доказательства предположим без ограничения общности рассуждений, что $V^\delta_{p,c}(\lambda) \geq V^0_{p,c}(\lambda)$. Тогда можем записать следующую цепочку равенств и неравенств:

$$|V^\delta_{p,c}(\lambda) - V^0_{p,c}(\lambda)| = V^\delta_{p,c}(\lambda) - V^0_{p,c}(\lambda) =$$

$$= V^\delta_{p,c}(\lambda) - \inf_{z\in\mathcal{D}}(L^0_{p,c}(z,\lambda) - L^\delta_{p,c}(z,\lambda) + L^\delta_{p,c}(z,\lambda)) \leq$$

$$\leq V^\delta_{p,c}(\lambda) - V^\delta_{p,c}(\lambda) - \inf_{z\in\mathcal{D}}(L^0_{p,c}(z,\lambda) - L^\delta_{p,c}(z,\lambda)) =$$

$$= -\inf_{z\in\mathcal{D}}(L^0_{p,c}(z,\lambda) - L^\delta_{p,c}(z,\lambda)) \leq \sup_{z\in\mathcal{D}}|L^\delta_{p,c}(z,\lambda) - L^0_{p,c}(z,\lambda)|,$$

очевидным следствием которой, с учетом оценок (1.2), и является доказываемая оценка (2.8).

2.5. Супердифференциал функции цели модифицированной двойственной задачи

Ниже важное значение для нас будет иметь лемма, в которой устанавливается выражение для супердифференциала $\partial V^\delta_{p,c}$ вогнутой функции значений $V^\delta_{p,c}$ при условии компактности образа оператора $g^\delta:\; \mathcal{D} \to H$. Здесь под супердифференциалом вогнутой функции $V^\delta_{p,c}$, как и ранее, понимается субдифференциал (в смысле выпуклого анализа) с обратным знаком выпуклой функции $-V^\delta_{p,c}$. Утверждение этой формулируемой ниже леммы является следствием утверждения, доказательство которого см. в [13, лемма 2]. Строго говоря, лемма 2 в [13] доказывается в случае гильбертова пространства Z, однако это доказательство практически дословно повторяется в рассматриваемом здесь случае, когда Z — полное метрическое пространство.

Лемма 2.3. *Супердифференциал (в смысле выпуклого анализа)* $\partial V^{\delta}_{p,c}(\lambda)$, $\delta \in [0, \delta_0]$, *вогнутой функции* $V^{\delta}_{p,c}$ *в точке* $\lambda \in H$ *при любом* $c \in R^1$ *выражается формулой*

$$\partial V^{\delta}_{p,c}(\lambda) = \partial_C V^{\delta}_{p,c}(\lambda) = \operatorname{cl}\operatorname{conv}\{\lim_{i\to\infty}(g^{\delta}(z^i) - p) : \tag{2.9}$$

$$z^i \in \mathcal{D},\ \ L^{\delta}_{p,c}(z^i, \lambda) \to \inf_{z\in\mathcal{D}} L^{\delta}_{p,c}(z, \lambda),\ i \to \infty\} \equiv \operatorname{cl}\operatorname{conv} Q^{\delta}_{p,c}(\lambda),$$

где $\partial_C V^{\delta}_{p,c}(\lambda)$ *— обобщенный градиент Кларка функции* $V^{\delta}_{p,c}$ *в точке* λ, $\operatorname{cl}\operatorname{conv} A$ *— замыкание выпуклой оболочки множества* A.

2.6. Обобщенный вектор Куна–Таккера

Введем также соответствующее понятие обобщенного вектора Куна–Таккера задачи (P^0_p) Возможны две и только две ситуации для исходной задачи (P^0_p):

А) в задаче существует обобщенный вектор Куна–Таккера, т.е. вектор $\lambda \in H$, для которого $\beta(p) \le \inf_{z\in\mathcal{D}} L^0_{p,c}(z, \lambda)$ для некоторого $c > 0$;

Б) в задаче не существует обобщенного вектора Куна–Таккера в указанном смысле.

В соответствии со сказанным в пп. 2.2, 2.3 можно утверждать, что:

I. Множество тех $p \in \operatorname{dom}\beta$, для которых задача (P^0_p) обладает обобщенным вектором Куна–Таккера, является всюду плотным в $\operatorname{dom}\beta$.

II. Обобщенный вектор Куна–Таккера в задаче (P^0_p) заведомо существует тогда, когда имеет место, по крайней мере, одно из следующих двух условий: 1. $\partial^P\beta(p) \neq \emptyset$; 2. $\hat{\partial}\beta(p) \neq \emptyset$. При этом в случае существования обобщенного вектора Куна–Таккера λ в задаче (P^0_p), когда штрафной множитель $l_1 = 0$, выполняется и включение $\lambda \in \partial^P\beta(p)$.

Существование вектора Куна–Таккера в указанном обобщенном смысле эквивалентно тому, что при некотором $c > 0$ целевая функция $V^0_{p,c}(\lambda),\ \lambda \in H$, модифицированной двойственной задачи

$$V^0_{p,c}(\lambda) \to \sup,\quad \lambda \in H,$$

достигает значения $\beta(p)$ в некоторой точке $\lambda^0 \in H$. Действительно, пусть, с одной стороны, λ^0 — вектор Куна–Таккера в указанном смысле. Для любой ОМП $z^i \in \mathcal{D},\ i = 1, 2, \ldots$, задачи (P^0_p) можем записать очевидное предельное соотношение

$$L^0_{p,c}(z^i, \lambda) \to \beta(p),\ i \to \infty \quad \forall\, \lambda \in H.$$

Отсюда имеем очевидное неравенство $V^0_{p,c}(\lambda) \le \beta(p)\ \forall\, \lambda \in H$. Поэтому, так как

$$\beta(p) \le \inf_{z\in\mathcal{D}} L^0_{p,c}(z, \lambda^0) \le L^0_{p,c}(z^i, \lambda^0),$$

то

$$\beta(p) = \inf_{z\in\mathcal{D}} L^0_{p,c}(z, \lambda^0) = V^0_{p,c}(\lambda^0).$$

Так как, к тому же, $V^0_{p,c}(\lambda) \le \beta(p)\ \forall\, \lambda \in H$, то получаем

$$\sup_{\lambda\in H} V^0_{p,c}(\lambda) = V^0_{p,c}(\lambda^0) = \beta(p).$$

С другой стороны, если $V^0_{p,c}$ достигает равного $\beta(p)$ максимума в точке $\lambda^0 \in H$, т.е.

$$\sup_{\lambda\in H} V^0_{p,c}(\lambda) = V^0_{p,c}(\lambda^0) = \beta(p) = \inf_{z\in\mathcal{D}} L^0_{p,c}(z, \lambda^0),$$

то, очевидно, λ^0 – обобщенный вектор Куна–Таккера.

3. РЕГУЛЯРИЗАЦИЯ ПЛ В НЕЛИНЕЙНОЙ ЗАДАЧЕ НА УСЛОВНЫЙ ЭКСТРЕМУМ

В основе регуляризации ПЛ в нелинейной задаче (P^0_p) лежит процедура основанной на двойственности регуляризации [15] (см. также [14]).

3.1. Двойственная регуляризация в нелинейной задаче на условный экстремум

Лемма 2.3 дает возможность организовать поиск максимума в задаче максимизации при каждом $c > 0$ сильно вогнутого функционала $R^{\delta,\alpha}_{p,c}(\lambda) \equiv V^{\delta}_{p,c}(\lambda) - \alpha\|\lambda\|^2, \lambda \in H, \delta \geq 0, \alpha > 0$. На этом пути с целью конструирования ОМП в исходной задаче (P^0_p) будем рассматривать при некотором достаточно большом $c > 0$ задачу

$$R^{\delta,\alpha}_{p,c}(\lambda) \to \max, \quad \lambda \in \Lambda_c \equiv \{\lambda \in H : \|\lambda\| \leq c\}. \tag{3.1}$$

Обозначим через $\lambda^{\delta,\alpha}_{p,c}$ единственную в Λ_c точку, дающую на Λ_c максимум функционалу $R^{\delta,\alpha}_{p,c}$ (она, очевидно, существует). Покажем, что при условии согласования

$$\frac{\delta}{\alpha(\delta)} \to 0, \ \alpha(\delta) \to 0, \ \delta \to 0, \tag{3.2}$$

стабилизация (регуляризация) по Тихонову (3.1) процесса поиска максимума в модифицированной двойственной задаче

$$V^0_{p,c}(\lambda) \to \sup, \quad \lambda \in \Lambda_c, \tag{3.3}$$

конструктивно порождает ОМП $z^i \in \mathcal{D}, \ i = 1, 2, \ldots$, в задаче (P^0_p), т.е. $f^0(z^i) \to \beta(p), \ \ z^i \in \mathcal{D}^{0,\epsilon^i}_p, \ \ \epsilon^i \to 0, i \to \infty$. При этом в случае **А)** величина c может быть взята равной любому фиксированному достаточно большому положительному числу. В случае же **Б)** штрафной коэффициент c необходимо стремить к $+\infty$ согласованно со стремлением к нулю δ.

3.1.1. Случай А). Итак, предполагаем, что задача (P^0_p) обладает вектором Куна–Таккера в указанном обобщенном смысле и хотя бы один из весовых коэффициентов l_1, l_2 больше нуля. Будем опираться в этом случае на соответствующие результаты [15]. Рассматриваем задачу (3.1) при произвольном достаточно большом фиксированном $c > 0$. Замкнутое выпуклое множество всех точек λ, доставляющих равное $\beta(p)$ максимальное значение функции $V^0_{p,c}$ на H, обозначим через $K_{p,c}$. Пусть далее $c > 0$ столь велико, что $K_{p,c} \cap \Lambda_c \neq \emptyset$.

3.1.1.1. Точки максимума целевой функции модифицированной двойственной задачи и ОМП в исходной задаче условной минимизации. Как показано в [15, разд. 3, с. 958], имеет место следующее утверждение о связи точек максимума функции $V^0_{p,c}$, в которых принимается значение $\beta(p)$, с ОМП в исходной задаче (P^0_p).

Утверждение 3.1. *Пусть $\tilde{c} > c$. Если $\lambda \in K_{p,c}$, то для любой минимизирущей в задаче*

$$L^0_{p,\tilde{c}}(z, \lambda) \to \inf, \ z \in \mathcal{D},$$

последовательности $z^i \in \mathcal{D}$, $i = 1, 2, \ldots$, справедливы предельные соотношения

$$L^0_{p,\tilde{c}}(z^i, \lambda) \to \beta(p) = \inf_{z \in \mathcal{D}} L^0_{p,\tilde{c}}(z, \lambda), \ \ f^0(z^i) \to \beta(p), \ \ g^0(z^i) - p \to 0, \ i \to \infty, \tag{3.4}$$

что обеспечивает равенства (см. равенство (2.9))

$$\{0\} = Q^0_{p,\tilde{c}}(\lambda) = \partial V^0_{p,\tilde{c}}(\lambda),$$

и говорит о том, что каждая из указанных последовательностей $z^i \in \mathcal{D}$, $i = 1, 2, \ldots$, является ОМП в задаче (P^0_p).

Доказательство. Так как $\lambda \in K_{p,c} \subset H$, то можем записать $\beta(p) = \inf\limits_{z \in \mathcal{D}} L^0_{p,c}(z, \lambda)$, а, следовательно, и $\beta(p) = \inf\limits_{z \in \mathcal{D}} L^0_{p,\tilde{c}}(z, \lambda)$, так как любая ОМП в задаче (P^0_p) является минимизирующей последовательностью как для функционала $L^0_{p,c}(z, \lambda), \ z \in \mathcal{D}$, так и для функционала $L^0_{p,\tilde{c}}(z, \lambda), \ z \in \mathcal{D}$. Предположим, что существует такая последовательность

$$z^i \in \mathcal{D}, \ i = 1, 2, \ldots, \ L^0_{p,\tilde{c}}(z^i, \lambda) \to \inf_{z \in \mathcal{D}} L^0_{p,\tilde{c}}(z, \lambda) = \beta(p), \ i \to \infty,$$

что

$$\lim_{i \to \infty} (g^0(z^i) - p) = \tilde{g} \neq 0. \tag{3.5}$$

Но тогда можем записать

$$L^0_{p,\tilde{c}}(z^i, \lambda) = f^0(z^i) + \langle \lambda, g^0(z^i) - p \rangle + \tilde{c}\psi(\|g^0(z^i) - p\|) =$$

$$= f^0(z^i) + \langle \lambda, g^0(z^i) - p \rangle + c\psi(\|g^0(z^i) - p\|) + (\tilde{c} - c)\psi(\|g^0(z^i) - p\|) =$$

$$= L^0_{p,c}(z^i, \lambda) + (\tilde{c} - c)\psi(\|g^0(z^i) - p\|) \to \beta(p), \ i \to \infty,$$

откуда ввиду соотношений (3.5) получаем, что $\limsup\limits_{i\to\infty} L^0_{p,c}(z^i,\lambda) < \beta(p)$, что, в свою очередь, влечет строгое неравенство $\inf\limits_{z\in\mathcal{D}} L^0_{p,c}(z,\lambda) < \beta(p)$, которое противоречит равенству $\inf\limits_{z\in\mathcal{D}} L^0_{p,c}(z,\lambda) = V^0_{p,c}(\lambda) = \beta(p)$. Таким образом, первое утверждение доказано. Справедливость второго утверждения вытекает из уже доказанного первого с учетом определения множества $Q^0_{p,\tilde{c}}(\lambda)$ (см. выражение (2.9) для супердифференциала целевой функции модифицированной двойственной задачи).

3.1.1.2. Стабилизация по Тихонову модифицированной двойственной задачи. В силу оценки (2.8), условия согласования (3.2) и теоремы о сходимости метода стабилизации Тихонова (см., например, [5, гл. 9, § 4, теорема 2]) можно утверждать, что справедливо следующее.

Утверждение 3.2. *Если $K_{p,c}\cap\Lambda_c=\emptyset$, то*

$$\|\lambda^{\delta,\alpha(\delta)}_{p,c} - \lambda^0_{p,c}\| \to 0,\ \ \delta\to 0, \tag{3.6}$$

где $\lambda^0_{p,c}$ — минимальное по норме решение задачи (3.3). Если же $K_{p,c}\cap\Lambda_c\neq\emptyset$, то точка $\lambda^0_{p,c}$ в предельном соотношении (3.6) одновременно является и минимальной по норме точкой во множестве $K_{p,c}$.

3.1.1.3. Конструирование ОМП в исходной задаче условной минимизации. Итак, пусть $K_{p,c}\cap\Lambda_c\neq\emptyset$. Пусть далее $\delta^s,\ s=1,2,\ldots,$ — произвольная сходящаяся к нулю последовательность положительных чисел. Рассмотрим в этом случае последовательность $z^{c,\delta^s,i}_\kappa,\ i=1,2,\ldots,$ являющуюся минимизирующей для функционала $L^{\delta^s,\alpha(\delta^s)}_{p,c+\kappa}(z,\lambda^{\delta^s,\alpha(\delta^s)}_{p,c})$, $z\in\mathcal{D}$, $s=1,2,\ldots,$ где $\kappa>0$ – не зависящая от $s=1,2,\ldots$ постоянная. Примем при этом обозначение $z^{c,\delta^s,i}_0\equiv z^{c,\delta^s,i}$ при $\kappa=0$. Тогда можем записать неравенства

$$f^{\delta^s}(z^{c,\delta^s,i}_\kappa) + \langle\lambda^{\delta^s,\alpha(\delta^s)}_{p,c}, g^{\delta^s}(z^{c,\delta^s,i}_\kappa)-p\rangle + (c+\kappa)\psi(\|g^{\delta^s}(z^{c,\delta^s,i}_\kappa)-p\|) \le \tag{3.7}$$

$$\le V^{\delta^s}_{p,c+\kappa}(\lambda^{\delta^s,\alpha(\delta^s)}_{p,c}) + \epsilon^{c,\delta^s,i},\ \ i=1,2,\ldots,\ \ \epsilon^{c,\delta^s,i}\to 0,\ \ i\to\infty,\ \ i=1,2,\ldots.$$

В силу ограниченности множества Λ_c и оценок (1.2), (2.8), а также выбора подпоследовательности $i(c,s)$, $s=1,2,\ldots,$ последовательности $i=1,2,\ldots$ такой, что $\epsilon^{c,\delta^s,i(c,s)}\to 0,\ \ s\to\infty$, из (3.7) выводим, обозначив $z^{c,\delta^s,i(c,s)}_\kappa\equiv z^s_\kappa$, что

$$f^0(z^s_\kappa) + \langle\lambda^{\delta^s,\alpha(\delta^s)}_{p,c}, g^0(z^s_\kappa)-p\rangle + (c+\kappa)\psi(\|g^0(z^s_\kappa)-p\|) \le$$

$$\le V^0_{p,c+\kappa}(\lambda^{\delta^s,\alpha(\delta^s)}_{p,c}) + \epsilon^{c,\delta^s,i(c,s)} + \gamma^s,\ \ s=1,2,\ldots,$$

где $\gamma^s\to 0,\ s\to\infty$. В свою очередь, из этой оценки в силу предельного соотношения (3.6) и непрерывности функционала $V^0_{p,c+\kappa}$ (см. п. 2.4), получаем

$$f^0(z^s_\kappa) + \langle\lambda^0_{p,c}, g^0(z^s_\kappa)-p\rangle + (c+\kappa)\psi(\|g^0(z^s_\kappa)-p\|) \le$$

$$\le V^0_{p,c+\kappa}(\lambda^0_{p,c}) + \tilde{\gamma}^s,\ \ s=1,2,\ldots,\ \ \tilde{\gamma}^s\to 0,\ \ s\to\infty.$$

Последнее неравенство говорит о том, что последовательность $z^s_\kappa\equiv z^{c,\delta^s,i(c,s)}_\kappa$, $s=1,2,\ldots,$ является минимизирующей в задаче

$$L^0_{p,c+\kappa}(z,\lambda^0_{p,c})\to\inf,\ \ z\in\mathcal{D}.$$

Но такой последовательностью в силу включения $\lambda^0_{p,c}\in K_{p,c}$, строгого неравенства $\kappa>0$ и утверждения 3.1 может быть лишь последовательность с указанными свойствами (3.4) при $\tilde{c}=c+\kappa$. Таким образом, сконструированная выше последовательность элементов $z^s_\kappa\equiv z^{c,\delta^s,i(c,s)}_\kappa$, $s=1,2,\ldots,$ будет представлять собой ОМП в задаче (P^0_p).

Итак, в общей ситуации для построения ОМП в задаче (P^0_p) в случае **А)** требуется в регуляризованном процессе максимизации модифицированной двойственной задачи решать задачу минимизации МФЛ при двух значениях штрафного коэффициента c: c и $c+\kappa$. В то же время, как показано в [15], [30], во многих важных частных случаях, когда известна дополнительная информация о субдифференциальных свойствах функции значений β в конкретной точке $p\in\operatorname{dom}\beta$, ОМП в задаче (P^0_p) будет построенная выше последовательность $z^s_\kappa\equiv z^{c,\delta^s,i(c,s)}_\kappa$, $s=1,2,\ldots,$ при $\kappa=0$: $z^s_0\equiv z^s,\ \ s=1,2,\ldots.$ В частности, такими важными частными случаями являются следующие (подробности см. в [15], [30]).

1. Для любого $\zeta\in\partial^P\beta(p)$ неравенство

$$\beta(p) - \langle\zeta,p\rangle \le \beta(p') - \langle\zeta,p'\rangle + l\|p'-p\|^2\ \ \forall\, p'\in H$$

выполняется с некоторым независящим от ζ штрафным коэффициентом $l=l(p)$. Последнее выполняется, например, тогда, когда множество $\partial^P\beta(p)$ является одноточечным, что, в силу включения $\partial^P\beta(p)\subset\hat{\partial}\beta(p)$, имеет место, в частности, тогда, когда $\partial^P\beta(p)\neq\emptyset$ и одноточечным является множество $\hat{\partial}\beta(p)$.

2. Проксимальный субградиент $\partial^P\beta(p)$ содержит минимальный по норме элемент. Это будет заведомо так, если, например, $\partial^P\beta(p)$ замкнутое множество.

3.1.1.4. Теорема сходимости процедуры двойственной регуляризация для решения исходной нелинейной задачи на условный экстремум в случае A). Подытоживая сказанное выше, можно утверждать, что для процесса построения ОМП в задаче (P_p^0) важнейшее значение имеет точность решения задачи минимизации МФЛ $L_{p,c}^{\delta}(z,\lambda_{p,c}^{\delta,\alpha(\delta)})$, $z\in\mathcal{D}$ при каждом $\delta>0$. Если при практическом решении задачи (P_p^0) такая минимизация может быть проведена с любой наперед заданной точностью, то в ней конструктивно указывается ОМП. Таким образом, следуя [13], [15], [30], можем сформулировать следующую теорему сходимости метода двойственной регуляризации в нелинейной задаче (P_p^0) (см. [15, теорема 3.1]). Обозначим с этой целью через $Z_p^{c,\varkappa,\gamma,\delta}[\lambda]\subset\mathcal{D}$, $\varkappa\geq 0$, множество всех элементов $z^\gamma\in\mathcal{D}$, удовлетворяющих неравенству $L_{p,c+\varkappa}^{\delta}(z^\gamma,\lambda)\leq L_{p,c+\varkappa}^{\delta}(z,\lambda)+\gamma\ \forall z\in\mathcal{D}$, т.е. множество всех γ-оптимальных элементов в задаче минимизации МФЛ

$$L_{p,c+\varkappa}^{\delta}(z,\lambda)\to\inf,\ z\in\mathcal{D}. \tag{3.8}$$

Примем также обозначение $Z_p^{c,0,\gamma,\delta}[\lambda]\equiv Z_p^{c,\gamma,\delta}[\lambda]$.

Теорема 3.1. *Пусть $\delta^s,\ s=1,2,\ldots,$ — произвольная сходящаяся к нулю последовательность положительных чисел. Тогда, если задача (P_p^0) обладает вектором Куна–Таккера в указанном выше обобщенном смысле (случай **A**)) или, другими словами, точка $p\in\operatorname{dom}\beta$ принадлежит плотному в $\operatorname{dom}\beta$ множеству всех точек, для которых эта параметрическая задача обладает вектором Куна–Таккера, то в предположении положительности хотя бы одного из двух штрафных коэффициентов l_1, l_2 найдется достаточно большое $c>0$ такое, что оператор $R_p(\cdot,\cdot,\delta^s)$, ставящий в соответствие каждому набору исходных данных $(f^{\delta^s},g^{\delta^s})$, удовлетворяющих оценкам (1.2) при $\delta=\delta^s$, множество*

$$R_p(f^{\delta^s},g^{\delta^s},\delta^s)\equiv Z_p^{c,\varkappa,\epsilon^s,\delta^s}[\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}]\subset\mathcal{D},$$

где $\epsilon^s\to 0,\ s\to\infty$ и выполняется условие согласования $\delta^s/\alpha(\delta^s)\to 0,\ s\to\infty$, является ОМП-образующим в задаче (P_p^0). Одновременно выполняются предельные соотношения

$$\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}\to\lambda_{p,c}^0,\quad V_{p,c}^0(\lambda_{p,c}^{\delta^s,\alpha(\delta^s)})\to\beta(p),\quad s\to\infty,$$

*где $\lambda_{p,c}^0$ — минимальный по норме во множестве Λ_c вектор Куна–Таккера задачи (P_p^0). Если же функция значений β обладает в точке p некоторыми дополнительными субдифференциальными свойствами в указанном выше смысле **1, 2** (см. п. 3.1.1.3), то величину $\varkappa$ можно считать равной нулю. В случае разрешимости задачи минимизации (3.8) можно считать, что $\epsilon^s=0,\ s=1,2,\ldots$.*

Замечание 3.1. Разрешимость задачи минимизации (3.8) имеет место во многих важных конкретных ситуациях. Опишем одну из таких ситуаций, возникающих в естественнонаучных приложениях, связанных с нелинейными некорректными задачами (см., например, [28], [32, гл. 1, §4]). Пусть Z — гильбертово пространство, функционал $f^\delta:\mathcal{D}\to R^1$ слабо полунепрерывен снизу (в качестве такого функционала можно взять, например, $f^\delta(\cdot)\equiv\|\cdot\|^2$), оператор $g^\delta:\mathcal{D}\to H$ переводит слабо сходящиеся к элементу z^0 последовательности $z^i\in\mathcal{D},\ i=1,2,\ldots$, в сходящиеся сильно к $g^\delta(z^0)$ элементы последовательности $g^\delta(z^i),\ i=1,2,\ldots$ (“полная непрерывность” оператора g^δ), $\mathcal{D}\subset Z$ — ограниченное, выпуклое и замкнутое множество. Тогда можно утверждать, что задача минимизации МФЛ (3.8) разрешима и, стало быть, в этом случае можно считать, что $\epsilon^s=0,\ s=1,2,\ldots$, в рамках теоремы 3.1. Можно заметить также, что при сделанных предположениях разрешимой является и сама исходная задача нелинейного программирования (P_p^0). Можно привести и другие группы условий на исходные данные задачи (P_p^δ), при которых имеет место разрешимость задачи (3.8).

Замечание 3.2. Важно заметить, что в зависимости от свойств исходных данных задачи (P_p^0) конструируемые ОМП обладают соответствующими свойствами сходимости. Проиллюстрируем сказанное на примере из замечания 3.1. Во-первых, в условиях этого примера все слабые предельные точки каждой ОМП в задаче (P_p^0) являются ее решениями. Это доказывается посредством стандартных рассуждений, основанных на свойствах слабой компактности ограниченного выпуклого замкнутого множества в гильбертовом пространстве, слабой полунепрерывности снизу функционала f^0 и “полной непрерывности” оператора g^0. Если же, во-вторых, в дополнение к описанным в примере из замечания 3.1 условиям функционал $f^0:\mathcal{D}\to R^1$ является еще сильно выпуклым и субдифференцируемым (в смысле выпуклого анализа), как, например, функционал $f^0(\cdot)\equiv\|\cdot\|^2$, то все указанные выше слабые предельные точки — решения задачи (P_p^0) — являются предельными точками в исходной сильной топологии гильбертова пространства Z: из слабой сходимости последовательности $z^i\in\mathcal{D},\ i=1,2,\ldots$, к элементу $z^0\in\mathcal{D}$ и числовой сходимости $f^0(z^i)\to f^0(z^0)=\beta(p),\ i\to\infty$, следует сходимость по норме $z^i\to z^0,\ i\to\infty$. Последнее можно утверждать в силу хорошо известного характеристического свойства субдифференцируемых сильно выпуклых функционалов в гильбертовом пространстве.

3.1.2. Случай Б). Пусть теперь задача (P_p^0) при условии $\beta(p) < +\infty$ не обладает вектором Куна–Таккера в указанном обобщенном смысле (случай **А)**) или, другими словами, функция $V_{p,c}^0$ ни при каком $c > 0$ не имеет точки максимума на H, в которой принимает значение $\beta(p)$. Тем не менее, предельное соотношение $\lambda_{p,c}^{\delta^s,\alpha(\delta^s)} \to \lambda_{p,c}^0, \ \ s \to \infty$, где $\lambda_{p,c}^0$ — минимальное по норме решение задачи (3.3), имеет место в силу утверждения 3.2 при каждом фиксированном $c > 0$ также и в этом случае (см. (3.6), когда $K_{p,c} \cap \Lambda_c = \emptyset$). Будем, однако, теперь стремить δ к нулю согласованно со стремлением штрафного коэффициента c к $+\infty$. В этом случае метод двойственной регуляризации сопрягается (объединяется), по сути дела, с методом штрафов, так как слагаемое $c\psi(\|g^\delta(z) - p\|)$ в МФЛ имеет вид именно штрафного слагаемого с коэффициентом штрафа c. В дальнейших построениях в случае **Б)** считаем, что (при $\beta(p) < +\infty$) оба весовых множителя l_1, l_2 в штрафной функции ψ положительны, т.е. $l_1 = l_2 = 1$, и выполняется условие согласования

$$c^s\delta^s \to 0, \ \ c^s \to \infty, \ \ \delta^s \to 0, \ \ s \to \infty. \tag{3.9}$$

Одновременно, мы будем далее предполагать, что в МФЛ с коэффициентом штрафа c^s двойственная переменная λ, как и в задаче (3.1), удовлетворяет неравенству $\|\lambda\| \le c^s$.

3.1.2.1. Конструирование ОМП в исходной нелинейной задаче условной минимизации. Итак, рассматриваем по аналогии с соотношениями (3.7) при $c = c^s$, $\kappa = 0$ последовательность $z^{c^s,\delta^s,i}, \ i = 1, 2, \ldots$, являющуюся минимизирующей для функционала $L_{p,c^s}^{\delta^s}(z, \lambda^s)$, $z \in \mathcal{D}$ с $\lambda^s \in \Lambda_{c^s}$. Тогда подобно (3.7) можем записать

$$f^{\delta^s}(z^{c^s,\delta^s,i(c^s,s)}) + \langle \lambda^s, g^{\delta^s}(z^{c^s,\delta^s,i(c^s,s)}) - p\rangle + c^s\psi(\|g^{\delta^s}(z^{c^s,\delta^s,i(c^s,s)}) - p\|) \le \tag{3.10}$$

$$\le V_{p,c^s}^{\delta^s}(\lambda^s) + \epsilon^{c^s,\delta^s,i(c^s,s)}, \ \ s \to \infty.$$

При этом номер $i(c^s, s)$ выбираем так, что $\epsilon^{c^s,\delta^s,i(c^s,s)} \le \gamma^s$, $s \to \infty$, где $\gamma^s, \ s = 1, 2, \ldots$, — некоторая произвольным образом заданная сходящаяся к нулю последовательность положительных чисел.

Подчеркнем, что в качестве последовательности $\lambda^s \in H, \ s = 1, 2, \ldots$, в (3.10) годится любая последовательность, элементы которой удовлетворяют неравенству $\|\lambda^s\| \le c^s$. Например, в целях согласования процессов построения ОМП в случаях **А)** и **Б)**, можно воспользоваться "двойственными точками" $\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}$ с достаточно большими c и в случае **Б)**. В качестве другого возможного варианта можно взять последовательность, все элементы которой совпадают с некоторым заданным фиксированным элементом $\lambda \in H$. В частности, ниже отдельно рассмотрим случай $\lambda^s = 0, \ s = 1, 2, \ldots$, при котором можно положить $l_1 = 0$.

Обозначим $z^s \equiv z^{c^s,\delta^s,i(c^s,s)}$. В силу ограниченности $\mathcal{D}$, оценки $\|\lambda^s\| \le c^s$, условия согласования (3.9) и оценок (1.2), (2.8) из (3.10) следует

$$f^0(z^s) + \langle \lambda^s, g^0(z^s) - p\rangle + c^s\psi(\|g^0(z^s) - p\|) \le V_{p,c^s}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s, \ \ s \to \infty, \tag{3.11}$$

где

$$\widetilde{\gamma}^s \equiv |f^{\delta^s}(z^s) - f^0(z^s| + |\langle \lambda^s, g^{\delta^s}(z^s) - g^0(z^s)\rangle| +$$

$$+c^s|\psi(\|g^{\delta^s}(z^s) - p\|) - \psi(\|g^0(z^s) - p\|)| + |V_{p,c^s}^{\delta^s}(\lambda^s) - V_{p,c^s}^0(\lambda^s)| \to 0, \ \ s \to \infty.$$

Тогда, во-первых, из неравенства (3.11) в силу неравенств ($\|\lambda^s\| \le c^s$)

$$\langle \lambda^s, g^0(z^s) - p\rangle + c^s\psi(\|g^0(z^s) - p\|) \ge 0, \ \ V_{p,c^s}^0(\lambda^s) \le \beta(p)$$

следует, что

$$f^0(z^s) \le \beta(p) + \gamma^s + \widetilde{\gamma}^s. \tag{3.12}$$

Во-вторых, из того же неравенства (3.11), так как $\|\lambda^s\| \le c^s$, одновременно следует, что

$$f^0(z^s) + c^s l_2\|g^0(z^s) - p\|^2 \le V_{p,c^s}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s \tag{3.13}$$

или ($l_2 = 1$)

$$l_2\|g^0(z^s) - p\|^2 \le (V_{p,c^s}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s - f^0(z^s))/c^s. \tag{3.14}$$

Поэтому из (3.12), (3.14) получаем для последовательности $z^s \equiv z^{c^s,\delta^s,i(c^s,s)}, \ s = 1, 2, \ldots$, соотношения

$$f^0(z^s) \to \beta(p), \quad z^s \in \mathcal{D}^{0,\zeta^s}, \ \ \zeta^s \to 0, \ \ s \to \infty \tag{3.15}$$

с

$$\zeta^s = \sqrt{(V_{p,c^s}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s - f^0(z^s))/c^s},$$

говорящие о том, что построенная выше последовательность $z^s,\ s = 1, 2, \ldots$, является ОМП в задаче (P_p^0). Одновременно, из (3.13) получаем, так как $V_{p,c^s}^0(\lambda) \le \beta(p)\ \forall \lambda \in H$ и $f^0(z^s) \ge \beta(p) - \chi^s$ (см. (3.15)),

$$c^s l_2 \|g^0(z^s) - p\|^2 \le V_{p,c^s}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s - f^0(z^s) \le \tag{3.16}$$

$$\le \beta(p) + \gamma^s + \widetilde{\gamma}^s - \beta(p) + \chi^s = \gamma^s + \widetilde{\gamma}^s + \chi^s \to 0,\ s \to \infty,$$

где $\chi^s,\ s = 1, 2, \ldots$, — некоторая сходящаяся к нулю последовательность положительных чисел. Отсюда, с учетом (3.15), следует, что $V_{p,c^s}^0(\lambda^s) \to \beta(p),\ \ s \to \infty$. И, наконец, последнее предельное соотношение, оценка (3.11), предельное соотношение в (3.15), а также оценка (3.16) приводят и к предельному соотношению

$$\langle \lambda^s, g^0(z^s) - p\rangle + c^s l_1 \|g^0(z^s) - p\| \to 0,\ \ s \to \infty.$$

3.1.2.2. Случай конечного коэффициента c, ненулевые "зазоры" по функции и "по ограничениям". Полученные предельные соотношения и оценки (3.12), (3.14), (3.15) говорят о том, что построенная последовательность $z^s,\ s = 1, 2, \ldots$, является ОМП в задаче (P_p^0). При этом $c^s \to \infty$, $s \to \infty$ и выполняется условие согласования (3.9). Можно однако не стремить c к бесконечности, взять произвольным достаточно большим и заменить предельные соотношения и оценки (3.12), (3.14), (3.15) соответствующими оценками, говорящими о том, что конструируемые элементы z^s "близки" по функции к $\beta(p)$ и удовлетворяют ограничению-равенству с некоторыми ненулевыми зазорами, которые стремятся к нулю при неограниченном увеличении коэффициента c. Действительно, пусть $c > 0$ фиксировано. В соответствии с утверждением 3.2 зафиксируем элемент $\lambda_{p,c}^0$ — минимальное по норме решение задачи (3.3), к которому, согласно этому утверждению, стремятся точки $\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}$ при $s \to \infty$. Теперь возьмем в качестве точек λ^s в проведенных выше рассуждениях точки $\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}$, $\|\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}\| \le c$, с достаточно большими номерами s. Тогда, во-первых, $V_{p,c}^0(\lambda_{p,c}^{\delta^s,\alpha(\delta^s)}) = V_{p,c}^0(\lambda^s) \to V_{p,c}^0(\lambda_{p,c}^0) \ge \beta(p) - \tilde{\chi}(c),\ s \to \infty$, с $\tilde{\chi}(c) \ge 0,\ \tilde{\chi}(c) \to 0,\ c \to \infty$. И, во-вторых, вместо (3.10) можем записать

$$f^{\delta^s}(z^{c,\delta^s,i(c,s)}) + \langle \lambda^s, g^{\delta^s}(z^{c,\delta^s,i(c,s)}) - p\rangle + c\psi(\|g^{\delta^s}(z^{c,\delta^s,i(c,s)}) - p\|) \le$$

$$\le V_{p,c}^{\delta^s}(\lambda^s) + \gamma^s,\ \ \gamma^s \to 0,\ s \to \infty.$$

Рассуждая далее точно так же, как при получении предельных соотношений и оценок (3.12), (3.14), (3.15), получаем, введя предварительно обозначение $\tilde{z}^s \equiv z^{c,\delta^s,i(c,s)}$

$$f^0(\tilde{z}^s) \le \beta(p) + \gamma^s + \widetilde{\gamma}^s,$$

$$l_2 \|g^0(\tilde{z}^s) - p\|^2 \le (V_{p,c}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s - f^0(\tilde{z}^s))/c$$

и, как следствие,

$$\tilde{z}^s \in \mathcal{D}^{0,\zeta^s},\ \ \zeta^s = \sqrt{(V_{p,c}^0(\lambda^s) + \gamma^s + \widetilde{\gamma}^s - f^0(\tilde{z}^s))/c},$$

где

$$\widetilde{\gamma}^s \equiv |f^{\delta^s}(\tilde{z}^s) - f^0(\tilde{z}^s)| + |\langle \lambda^s, g^{\delta^s}(\tilde{z}^s) - g^0(\tilde{z}^s)\rangle| +$$

$$+c|\psi(\|g^{\delta^s}(\tilde{z}^s) - p\|) - \psi(\|g^0(\tilde{z}^s) - p\|)| + |V_{p,c}^{\delta^s}(\lambda^s) - V_{p,c}^0(\lambda^s)| \to 0,\ s \to \infty.$$

Таким образом, выбирая коэффициент $c > 0$ сколь угодно большим, мы можем конструировать элементы $z \in \mathcal{D}$ которые "сколь угодно близки" по функции к $\beta(p)$ и "сколь угодно точно" удовлетворяют ограничению-равенству.

3.1.2.3. Связь с методом регуляризации Тихонова, характеристическое свойство основанной на двойственности регуляризации. Рассмотрим теперь отдельно случай $\lambda^s = 0,\ s = 1, 2, \ldots$, в котором хотя бы одно из чисел $l_1,\ l_2$ отлично от нуля. Таким образом, имеем дело с минимизацией МФЛ

$$f^{\delta^s}(z) + \langle \lambda^s, g^{\delta^s}(z) - p\rangle + c^s \psi(\|g^{\delta^s}(z) - p\|),\ \ z \in \mathcal{D}, \tag{3.17}$$

при $\lambda^s = 0$.

Рассуждая точно так же, как и при получении предельных соотношений и оценок (3.12), (3.14), (3.15), можем записать в силу (3.10) при $\lambda^s = 0$, ограниченности $\mathcal{D}$, условия согласования (3.9) и оценок (1.2), (2.8) с учетом обозначения $z^s \equiv z^{c^s,\delta^s,i(c^s,s)}$

$$f^0(z^s) + c^s \psi(\|g^0(z^s) - p\|) \le V_{p,c^s}^0(0) + \gamma^s + \widetilde{\gamma}^s,\ \ s \to \infty, \tag{3.18}$$

где

$$\widetilde{\gamma}^s \equiv |f^{\delta^s}(z^s) - f^0(z^s)| + c^s|\psi(\|g^{\delta^s}(z^s) - p\|) - \psi(\|g^0(z^s) - p\|)| + |V^{\delta^s}_{p,c^s}(0) - V^0_{p,c^s}(0)| \to 0,\ s \to \infty.$$

Тогда, во-первых, из неравенства (3.18) в силу неравенства $V^0_{p,c^s}(0) \le \beta(p)$ следует, что

$$f^0(z^s) \le \beta(p) + \gamma^s + \widetilde{\gamma}^s. \tag{3.19}$$

Во-вторых, из того же неравенства (3.18) одновременно следует, что ($(l_1, l_2) \ne 0$)

$$\psi(\|g^0(z^s) - p\|) \le (V^0_{p,c^s}(0) + \gamma^s + \widetilde{\gamma}^s - f^0(z^s))/c^s. \tag{3.20}$$

Поэтому из (3.19), (3.20) получаем для последовательности $z^s \equiv z^{c^s,\delta^s,i(c^s,s)},\ s = 1, 2, \ldots$, соотношения

$$f^0(z^s) \to \beta(p), \quad z^s \in \mathcal{D}^{0,\zeta^s},\ \zeta^s = \sqrt{\big(V^0_{p,c^s}(0) + \gamma^s + \widetilde{\gamma}^s - f^0(z^s)\big)/c^s} \to 0,\ s \to \infty,$$

говорящие о том, что построенная выше последовательность $z^s,\ s = 1, 2, \ldots$, также является ОМП в задаче (P^0_p). Попутно здесь заметим, что γ-оптимальные элементы в задаче минимизации при $\lambda^s = 0$ функционала (3.17) являются одновременно γ/c^s-оптимальными элементами в задаче минимизации функционала

$$\frac{1}{c^s} f^{\delta^s}(z) + \psi(\|g^{\delta^s}(z) - p\|),\ z \in \mathcal{D}. \tag{3.21}$$

Мы получили хорошо известную классическую конструкцию (см., например, [32, гл. 1, §4]) сглаживающего функционала (функционала Тихонова) в методе регуляризации Тихонова [5], [12], [32] с параметром регуляризации $1/c^s$, $\delta^s/(1/c^s) \to 0,\ s \to \infty$, в случае нелинейной задачи на условный экстремум (строго говоря, в [32, гл. 1, §4] рассматривается некорректная задача поиска так называемых Ω-нормальных решений нелинейного операторного уравнения).

Подчеркнем, наконец, принципиальное отличие процесса регуляризации в задаче (P^0_p) — процесса конструирования ОМП в этой задаче, основанного на минимизации МФЛ (3.17) ($\lambda^s \ne 0$) от аналогичного процесса минимизации функционала Тихонова (3.21) ($\lambda^s = 0$). В первом случае можно утверждать, что для плотного в $\operatorname{dom}\beta$ множества точек p как двойственная переменная λ, так и величина штрафного коэффициента c в указанном процессе остаются ограниченными сверху некоторой не зависящей от $s = 1, 2, \ldots$ положительной величиной (см. пп. 2.2, 2.3, 2.6, а также теорему 3.1). Такое характеристическое свойство регуляризации правила множителей является следствием опоры в этом случае на теорию возмущений, конструкции негладкого анализа и аппарат теории двойственности. Во втором же случае величину $c^s,\ s = 1, 2, \ldots$, необходимо всегда стремить к бесконечности. Теория решения нелинейных некорректных задач, в основе которой лежат различные конструкции сглаживающих функционалов (функционалов Тихонова) детально изложена в [32].

3.1.2.4. Теорема сходимости процедуры двойственной регуляризация для решения исходной нелинейной задачи условной минимизации в случае Б). Подытоживая сказанное, можно утверждать, что для процесса построения ОМП в задаче (P^0_p) наиважнейшее значение имеет "точность" решения задачи минимизации МФЛ $L^{\delta}_{p,c^s}(z, \lambda^s)$, $z \in \mathcal{D}$ при каждом $\delta > 0$. Если такая минимизация может быть проведена с любой наперед заданной точностью, то в задаче (P^0_p) конструктивно указывается ОМП. Таким образом, доказана следующая теорема сходимости по функции и "по ограничению" метода двойственной регуляризации в случае **Б)**, который в указанном случае приобретает вид варианта метода штрафов.

Теорема 3.2. *Пусть $\delta^s,\ s = 1, 2, \ldots$, — произвольная сходящаяся к нулю последовательность положительных чисел, $c^s,\ s = 1, 2, \ldots$, — произвольная фиксированная сходящаяся к $+\infty$ последовательность чисел такая, что $c^s\delta^s \to 0,\ s \to \infty$. Пусть также процесс минимизации МФЛ $L^{\delta^s}_{p,c}(z, \lambda)$, $z \in \mathcal{D}$, при каждом $\delta^s > 0$ и любых $c > 0,\ \lambda \in H$ может проходить с любой наперед заданной точностью. Тогда:*

1. *Если $\beta(p) < +\infty$, но обобщенного вектора Куна–Таккера в задаче (P^0_p) не существует, то найдется последовательность двойственной переменной λ^s, $\|\lambda^s\| \le c^s,\ s = 1, 2, \ldots$, такая, что для последовательности $z^s \in \mathcal{D},\ s = 1, 2, \ldots$, элементы которой удовлетворяют при $l_1 = l_2 = 1$ соотношениям*

$$z^s \in Z_p^{c^s,\epsilon^s,\delta^s}[\lambda^s] \subset \mathcal{D},\ \epsilon^s \to 0,\ s \to \infty, \tag{3.22}$$

справедливы предельные соотношения

$$g^{\delta^s}(z^s) - p \to 0,\ s \to \infty, \tag{3.23}$$

$$\langle \lambda^s, g^{\delta^s}(z^s) - p\rangle + c^s\psi(\|g^{\delta^s}(z^s) - p\|) \to 0,\ s \to \infty, \tag{3.24}$$

$$f^0(z^s) \to \beta(p),\ s \to \infty, \tag{3.25}$$

и, как следствие, предельное соотношение

$$V^0_{p,c^s}(\lambda^s) \to \beta(p),\quad s \to \infty. \tag{3.26}$$

Элементы z^s в (3.22) произвольным образом выбираются из множеств $Z_p^{c^s,\epsilon^s,\delta^s}[\lambda^s] \subset \mathcal{D}$, а МФЛ $L^{\delta^s}_{p,c^s}(z,\lambda^s)$, $z \in \mathcal{D}$ берется при $l_1 = l_2 = 1$ (определение множества $Z_p^{c,\epsilon,\delta}[\lambda] \subset \mathcal{D}$ см. в п. 3.1.1.4).

Другими словами, ОМП-образующим в этом случае является оператор $R_p(\cdot,\cdot,\delta^s)$, ставящий в соответствие каждому набору исходных данных $(f^{\delta^s}, g^{\delta^s})$, удовлетворяющих оценкам (1.2) при $\delta = \delta^s$, множество $Z_p^{c^s,\epsilon^s,\delta^s}[\lambda^s] \subset \subset \mathcal{D}$ с $\epsilon^s \to 0,\ s \to \infty$.

***2.** В качестве последовательности $\lambda^s,\ s = 1,2,\ldots,$ годится любая последовательность, элементы которой удовлетворяют неравенству $\|\lambda^s\| \le c^s$. Например, в качестве такой последовательности может быть взята последовательность элементов $\lambda^{\delta^s,\alpha(\delta^s)}_{p,c^k},\ s = 1,2,\ldots,$ максимизирующих на множестве Λ_{c^k} сильно вогнутый функционал $R^{\delta^s,\alpha(\delta^s)}_{p,c^k}$ (см. первое предложение утверждения 3.2) с достаточно большими номерами s при каждом фиксированном достаточно большом c^k, $c^k \to \infty,\ k \to \infty$.*

Взятые в совокупности, теоремы 3.1 и 3.2, представляют собою утверждение, которое можно трактовать как теорему сходимости метода двойственной регуляризации, учитывающую одновременно как случай **А)**, так и случай **Б)**.

3.2. Регуляризованные теорема Куна–Таккера и ПЛ в нелинейной задаче на условный экстремум

Теоремы 3.1, 3.2 сходимости процедуры двойственной регуляризации в случаях **А)** и **Б)** можно трактовать также, как необходимые условия существования ОМП в задаче (P^0_p) в этих случаях. Формулируемые ниже теоремы помимо указанных необходимых условий содержат так же и достаточные условия существования ОМП в рассматриваемой задаче. По этой причине ниже в данном разделе первую из этих теорем, "обслуживающую" регулярный случай **А)**, мы называем регуляризованной теоремой Куна–Таккера. Во второй же из них мы говорим о регуляризованном ПЛ в недифференциальной форме в нелинейной задаче на условный экстремум (P^0_p), имея ввиду, что формулировка теоремы "обслуживает" как случай **А)**, так и случай **Б)**.

3.2.1. Регуляризованная теорема Куна–Таккера в недифференциальной форме в нелинейной задаче на условный экстремум. Дополняя утверждение теоремы 3.1 "достаточной частью" мы приходим к теореме Куна-Таккера в секвенциальном регуляризованном недифференциальном модифицированном виде для нелинейной задачи (P^0_p). Ее характерным свойством является "устойчивость по отношению к ошибкам исходных данных". Это понимается в том смысле, что в утверждении теоремы указывается множество допустимых элементов в каждой возмущенной задаче, которое состоит из таких элементов, что при стремлении к нулю ошибки задания исходных данных $\delta^s,\ s = 1,2,\ldots,$ они, взятые произвольным образом из указанного множества и последовательно при $\delta^s \to 0,\ s \to \infty$, составляют ОМП в исходной задаче. Следуя [13], [15], [30], можем сформулировать следующую теорему (см. доказательство теоремы 4.1 в [15]).

Теорема 3.3. *Пусть задача (P^0_p) обладает вектором Куна–Таккера в указанном выше обобщенном смысле, $\delta^s,\ s = 1,2,\ldots,$ — произвольная сходящаяся к нулю последовательность положительных чисел, $(l_1,l_2) \ne 0$. Тогда:*

***1.** Найдутся достаточно большое $c > 0$ и ограниченная последовательность двойственной переменной $\lambda^s \in H$, $s = 1,2,\ldots,$ такие, что для последовательности $z^s,\ s = 1,2,\ldots,$ элементы которой при $\kappa > 0$ удовлетворяют соотношениям (определение множества $Z_p^{c,\kappa,\epsilon,\delta}[\lambda] \subset \mathcal{D}$ см. в п. 3.1.1.4)*

$$z^s \in Z_p^{c,\kappa,\epsilon^s,\delta^s}[\lambda^s] \subset \mathcal{D},\ \ \epsilon^s \to 0,\ s \to \infty, \tag{3.27}$$

справедливы предельные соотношения

$$f^0(z^s) \to \beta(p),\ \ V^0_{p,c}(\lambda^s) \to \beta(p),\quad s \to \infty, \tag{3.28}$$

$$g^0(z^s) - p \to 0,\ s \to \infty, \tag{3.29}$$

и, как следствие, предельное соотношение

$$V^0_{p,c+\kappa}(\lambda^s) \to \beta(p),\quad s \to \infty. \tag{3.30}$$

Элементы z^s в (3.27) произвольным образом выбираются из множеств $Z_p^{c^s,\kappa,\epsilon^s,\delta^s}[\lambda^s] \subset \mathcal{D}$, а МФЛ $L^{\delta^s}_{p,c+\kappa}(z,\lambda^s)$, $z \in \mathcal{D}$ берется при $(l_1,l_2) \ne 0$.

Другими словами, в этом случае оператор $R_p(\cdot,\cdot,\delta^s)$, ставящий в соответствие каждому набору исходных данных $(f^{\delta^s}, g^{\delta^s})$, удовлетворяющих оценкам (1.2) при $\delta = \delta^s$, множество

$$R_p(f^{\delta^s}, g^{\delta^s}, \delta^s) \equiv Z_p^{c,\varkappa,\epsilon^s,\delta^s}[\lambda^s] \subset \mathcal{D},$$

где $\epsilon^s \to 0,\ s \to \infty$, является ОМП-образующим в задаче (P_p^0).

*В качестве указанной выше последовательности $\lambda^s,\ s = 1,2,\ldots$, может быть взята последовательность $\lambda_{p,c}^{\delta^s,\alpha(\delta^s)},\ s = 1,2,\ldots$, из теоремы 3.1, элементы которой максимизируют на множестве H сильно вогнутый функционал $R_{p,c}^{\delta^s,\alpha(\delta^s)}$ при условии согласования $\delta^s/\alpha(\delta^s) \to 0,\ s \to \infty$. При этом $\lambda^s \to \lambda_{p,c}^0$, $s \to \infty$, где $\lambda_{p,c}^0$ — минимальный по норме во множестве $K_{p,c}$ обобщенный вектор Куна–Таккера задачи (P_p^0). Наконец, если функция значений β обладает в точке p дополнительными субдифференциальными свойствами **1, 2** (см. п. 3.1.1.3), то величину $\varkappa$ можно считать равной нулю.*

***2.** И наоборот, если при некотором достаточно большом $c > 0$ существует ограниченная последовательность двойственных переменных $\lambda^s \in H,\ s = 1,2,\ldots$, такая что элементы последовательности $z^s,\ s = 1,2,\ldots$, удовлетворяют при $\varkappa \geq 0$ соотношениям (3.27) и предельному соотношению (3.29), то выполняется и первое предельное соотношение (3.28), т.е. последовательность $z^s,\ s = 1,2,\ldots$, является ОМП в задаче (P_p^0). При этом одновременно выполняется и предельное соотношение (3.30).*

3.2.2. Регуляризованный ПЛ в недифференциальной форме в нелинейной задаче на условный экстремум. В свою очередь, сформулированные в теоремах 3.1–3.3 утверждения можно трансформировать в следующий ПЛ в секвенциальном регуляризованном недифференциальном модифицированном виде, учитывающий как случай **А)**, так и случай **Б)**, для нелинейной задачи (P_p^0). Как и в предыдущей ситуации, связанной с теоремой 3.3, его характерным свойством является "устойчивость по отношению к ошибкам исходных данных". Прежде чем формулировать указанный результат, напомним, что ввиду ограниченности множества $\mathcal{D}$ значение $\beta(p)$ задачи (P_p^0) конечно тогда и только тогда, когда в ней существует ОМП.

Теорема 3.4. *Пусть значение $\beta(p)$ задачи (P_p^0) конечно, $\delta^s,\ s = 1,2,\ldots$, — произвольная сходящаяся к нулю последовательность положительных чисел, $(l_1, l_2) \neq 0$. Тогда справедливы следующие два утверждения.*

***1.** Пусть задача (P_p^0) обладает вектором Куна–Таккера в указанном выше обобщенном смысле. Тогда справедливы все утверждения теоремы 3.3, которые в совокупности можно трактовать как регуляризованную теорему Куна–Таккера для нелинейной задачи на условный экстремум (P_p^0).*

***2.** Пусть в задаче (P_p^0) не существует обобщенного вектора Куна–Таккера в указанном выше обобщенном смысле, $c^s,\ s = 1,2,\ldots$, — произвольная фиксированная сходящаяся к $+\infty$ последовательности чисел такая, что $c^s\delta^s \to 0,\ s \to \infty$, оба штрафных коэффициента $l_1,\ l_2$ являются положительными. Тогда имеют место следующие два утверждения:*

***2.1.** Справедливы все утверждения теоремы 3.2.*

***2.2.** Если существует последовательность двойственной переменной $\lambda^s,\ s = 1,2,\ldots$, такая, что $\|\lambda^s\| \leq c^s$, $s = 1,2,\ldots$, и для последовательности $z^s,\ s = 1,2,\ldots$, элементы которой удовлетворяют включениям (3.22), выполняются предельные соотношения (3.23), (3.24), то выполняются и предельные соотношения (3.25), (3.26).*

Доказательство. Первое утверждение теоремы, представляющее собою ее "регулярную" часть, совпадает с утверждением теоремы 3.3. По этой причине доказываем второе утверждение, связанное с ее "нерегулярной" частью. При этом "необходимая" часть **2.1** "нерегулярной части" есть "непосредственное следствие" утверждений теоремы 3.2. Поэтому доказываем здесь лишь "достаточную" часть **2.2** второго утверждения теоремы.

Так как точка z^s удовлетворяет включениям (3.22), то можем записать

$$L_{p,c^s}^{\delta^s}(z^s, \lambda^s) \leq L_{p,c^s}^{\delta^s}(z, \lambda^s) + \epsilon^s \ \ \forall\, z \in \mathcal{D}.$$

Отсюда в силу предельных соотношений (3.23), (3.24), условий теоремы и ограниченности $\mathcal{D}$ получаем

$$f^0(z^s) \leq L_{p,c^s}^0(z, \lambda^s) + \widetilde{\epsilon}^s \ \forall\, z \in \mathcal{D},\ \ \widetilde{\epsilon}^s \to 0,\ \ s \to \infty. \tag{3.31}$$

Кроме того, так как выполняется предельное соотношение (3.23), то можно утверждать, что $\beta(p) < +\infty$ и в задаче (P_p^0) существуют ОМП. Так как мы можем подставить $z = \bar{z}^k,\ k = 1,2,\ldots$, в неравенство (3.31), где $\bar{z}^k,\ k = 1,2,\ldots$, любое ОМП в задаче (P_p^0), то получаем $f^0(z^s) \leq \beta(p) + \bar{\epsilon}^s$, $\bar{\epsilon}^s \to 0,\ s \to \infty$. Так как, к тому же, последовательность $z^s,\ s = 1,2,\ldots$, в силу предельного соотношения (3.23) удовлетворяет "в пределе" ограничению-равенству, то выполняется и предельное соотношение (3.25), т.е. последовательность $z^s,\ s = 1,2,\ldots$, является ОМП в задаче (P_p^0). Одновременное выполнение в этом случае и предельного соотношения (3.26) очевидно, так как наряду с предельным соотношением (3.24) в силу условий теоремы и ограниченности $\mathcal{D}$ выполняется и предельное соотношение

$$\langle \lambda^s, g^0(z^s) - p\rangle + c^s\psi(\|g^0(z^s) - p\|) \to 0,\ \ s \to \infty.$$

Замечание 3.3. Если в рамках теорем 3.3, 3.4 в задачах минимизации

$$L^{\delta^s}_{p,c+\kappa}(z,\lambda^s) \to \min,\ z \in \mathcal{D}, \quad L^{\delta^s}_{p,c^s}(z,\lambda^s) \to \min,\ z \in \mathcal{D}$$

минимальные значения достигаются, то в качестве элементов z^s, $s = 1,2,\ldots$, минимизирующих последовательностей в указанных теоремах могут быть естественно взяты непосредственно точки минимума в этих задачах минимизации МФЛ. Обсуждение условий, при которых указанные задачи разрешимы см. в замечании 3.1. В свою очередь, в замечании 3.2 можно найти обсуждение условий, при которых конструируемые в теоремах 3.3, 3.4 ОМП обладают свойствами слабой или сильной сходимости к решениям задачи (P^0_p) в случае гильбертова пространства Z.

СПИСОК ЛИТЕРАТУРЫ

1. *Лагранж Ж.Л.* Аналитическая механика. В 2-х томах. 2-е изд. М.–Л.: Гостехиздат, 1950. 603+442 с.
2. *Алексеев В.М., Тихомиров В.М., Фомин С.В.* Оптимальное управление. М.: Наука, 1979. 432 с.
3. *Тихомиров В.М.* Рассказы о максимумах и минимумах. М.: Наука, 1986. 192 с.
4. *Аваков Е.Р., Магарил-Ильяев Г.Г., Тихомиров В.М.* О принципе Лагранжа в задачах на экстремум при наличии ограничений // Успехи матем. наук. 2013. Т. 68. Вып. 3(411). С. 5–38.
5. *Васильев Ф.П.* Методы оптимизации: в 2-х кн. М.: МЦНМО, 2011. 1056 с.
6. *Сумин М.И.* Регуляризованная параметрическая теорема Куна-Таккера в гильбертовом пространстве // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. № 9. С. 1594–1615.
7. *Сумин М.И.* Устойчивое секвенциальное выпуклое программирование в гильбертовом пространстве и его приложение к решению неустойчивых задач // Ж. вычисл. матем. и матем. физ. 2014. Т. 54. № 1. С. 25–49.
8. *Сумин М.И.* Регуляризованные принцип Лагранжа и принцип максимума Понтрягина в оптимальном управлении и обратных задачах // Тр. Ин-та матем. и механ. УрО РАН. 2019. Т. 25. № 1. С. 279–296.
9. *Сумин М.И.* О некорректных задачах, экстремалях функционала Тихонова и регуляризованных принципах Лагранжа // Вестн. российских университетов. Математика. 2022. Т. 27. Вып. 137. С. 58–79.
10. *Треногин В.А.* Функциональный анализ. Москва: Наука, 1980. 496 с.
11. *Сумин М.И.* Недифференциальные теоремы Куна–Таккера в задачах на условный экстремум и субдифференциалы негладкого анализа // Вестн. российских университетов. Математика. 2020. Т. 25. Вып. 131. С. 307–330.
12. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М.: Наука, 1986. 288 с.
13. *Сумин М.И.* Регуляризация в линейно выпуклой задаче математического программирования на основе теории двойственности // Ж. вычисл. матем. и матем. физ. 2007. Т. 47. № 4. С. 602–625.
14. *Сумин М.И.* Регуляризованный двойственный метод решения нелинейной задачи математического программирования // Ж. вычисл. матем. и матем. физ. 2007. Т. 47. № 5. С. 796–816.
15. *Сумин М.И.* Устойчивая секвенциальная теорема Куна–Таккера в итерационной форме или регуляризованный алгоритм Удзавы в регулярной задаче нелинейного программирования // Ж. вычисл. матем. и матем. физ. 2015. Т. 55. № 6. С. 947–977. doi: 10.7868/S0044466915060137 .
16. *Гольштейн Е.Г.* Теория двойственности в математическом программировании и ее приложения. М.: Наука, 1971. 352 с.
17. *Варга Дж.* Оптимальное управление дифференциальными и функциональными уравнениями. М.: Наука, 1977. 624 с.
18. *Сумин М.И.* Метод возмущений и регуляризация правила множителей Лагранжа в выпуклых задачах на условный экстремум // Тр. Ин-та математики и механики УрО РАН. 2024. Т. 30. № 2. С. 203–221.
19. *Сумин М.И.* О регуляризации классических условий оптимальности в выпуклых задачах оптимального управления // Тр. Ин-та математики и механики УрО РАН. 2020. Т. 26. № 2. С. 252–269. doi: 10.21538/0134-4889-2020-26-2-252-269.
20. *Сумин М.И.* О регуляризации принципа Лагранжа и построении обобщенных минимизирующих последовательностей в выпуклых задачах условной оптимизации // Вестник Удмуртского университета. Математика. Механика. Компьютерные науки. 2020. Т. 30. Вып. 3. С. 410–428.
21. *Borwein J.M., Strojwas H.M.* Proximal Analysis and Boundaries of Closed Sets in Banach Space, Part I: Theory // Can. J. Math. 1986. V.38. No.2. P.431-452; Part II: Applications // Can. J. Math. 1987. V. 39. No 2. P. 428–472.

22. *Loewen P.D.* Optimal control via nonsmooth analysis. CRM Proceedings and Lecture Notes. Vol. 2. Providence, RI: AMS, 1993. 153 p. DOI: https://doi.org/10.1090/crmp/002
23. *Clarke F.H., Ledyaev Yu.S., Stern R.J., Wolenski P.R.* Nonsmooth analysis and control theory. Graduate texts in mathematics. V. 178. New York: Springer-Verlag, 1998. 278 p. DOI: https://doi.org/10.1007/b97650
24. *Mordukhovich B.S.* Variational analysis and generalized differentiation, I: Basic Theory. Berlin: Springer, 2006. 579 p. DOI: https://doi.org/10.1007/3-540-31247-1
25. *Бертсекас Д.* Условная оптимизация и методы множителей Лагранжа. М.: Радио и связь, 1987. 400 с.
26. *Мину М.* Математическое программирование. Теория и алгоритмы. М.: Наука, 1990. 488 с.
27. *Гольштейн Е.Г., Третьяков Н.В.* Модифицированные функции Лагранжа. Теория и методы оптимизации. М.: Наука, 1989. 399 с.
28. *Сумин М.И.* Метод возмущений, субдифференциалы негладкого анализа и регуляризация правила множителей Лагранжа в нелинейном оптимальном управлении // Тр. Ин-та математики и механики УрО РАН. 2022. Т. 28, № 3. С. 202–221.
29. *Сумин М.И.* О регуляризации недифференциальной теоремы Куна–Таккера в нелинейной задаче на условный экстремум // Вестн. российских университетов. Математика. 2022. Т. 27. Вып. 140. С. 351–374.
30. *Sumin M.I.* Parametric Dual Regularization in a Nonlinear Mathematical Programming // In book “Advances in Mathematics Research, Volume 11”. Chapter 5. New-York: Nova Sci. Publ. Inc. 2010. P. 103–134.
31. *Обен Ж.-П.* Нелинейный анализ и его экономические приложения. М.: Мир, 1988. 264 с.
32. *Тихонов А.Н., Леонов А.С., Ягола А.Г.* Нелинейные некорректные задачи. М.: Наука, 1995. 312 с.

THE PERTURBATION METHOD AND THE REGULARIZATION OF THE LAGRANGE PRINCIPLE IN NONLINEAR CONSTRAINED EXTREMUM PROBLEMS

M. I. Sumin*

392000 Tambov, Internationalnaya str., 33, G.R. Derzhavin TSU, Russia
**e-mail: m.sumin@mail.ru*

Received: 08.07.2024
Revised: 08.07.2024
Accepted: 23.08.2024

Abstract. The regularization of the Lagrange principle (LP) in a non-differential form in a nonlinear (non-convex) constrained extremum problem with an operator constraint-equality in Hilbert space is considered. The set of its permissible elements belongs to a complete metric space, the existence of a solution to the problem is not assumed a priori. The equality constraint contains an additively included parameter, which makes it possible to apply the “nonlinear variant” of the perturbation method to study the problem. The main purpose of the regularized LP is the stable generation of generalized minimizing sequences (GMSs) in the nonlinear problem under consideration. It can be interpreted as a GMS-generating (regularizing) operator that matches each set of initial data of the problem with a subminimal (minimal) of its regular augmented Lagrangian corresponding to this set, the dual variable in which is generated in accordance with the Tikhonov stabilization procedure of the dual problem. The structure of the augmented Lagrangian is completely determined by the type of “nonlinear” subdifferentials of a value function that is below semicontinuous and, generally speaking, non-convex as a function of the problem parameter. The Frechet proximal subgradient and the subdifferential, well-known in non-smooth (nonlinear) analysis, are used as such subdifferentials. The regularized LP “overcomes” the properties of the ill-posedness of the classical analogue and can be interpreted as a regularizing algorithm, thereby forming the theoretical basis for creating stable methods for the practical solving nonlinear constrained extremum problems.

Keywords: nonlinear onstrained extremum problem, operator constraint equality, Lagrange multiplier rule in non-differential form, regularization, perturbation method, value function, proximal subgradient, Frechet subdifferential, modified dual problem, generalized minimizing sequence, regularizing algorithm.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2332–2354

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

УДК 519.62

О МИНИМАЛЬНОСТИ КВАДРАТИЧНОЙ ПОГРЕШНОСТИ РЕШЕНИЯ ПРЕОБРАЗОВАННЫХ К НАИЛУЧШЕМУ ПАРАМЕТРУ СИСТЕМ УРАВНЕНИЙ ПРИ МАЛЫХ ОДНОРОДНЫХ ВОЗМУЩЕНИЯХ

© 2024 г. Е.Б. Кузнецов[1,*], С.С. Леонов[1,2,]**

[1] 125993 Москва, Волоколамское ш., 4, Московский авиационный институт (национальный исследовательский университет), Россия
[2] 117198 Москва, ул. Миклухо-Маклая, 6, Российский университет дружбы народов имени Патриса Лумумбы, Россия
**e-mail: kuznetsov@mai.ru*
***e-mail: powerandglory@yandex.ru*

Поступила в редакцию 05.05.2024 г.
Переработанный вариант 05.08.2024 г.
Принята к публикации 23.08.2024 г.

В статье рассматривается решение систем нелинейных уравнений с одним скалярным параметром. Множеством решений подобных систем является кривая в пространстве неизвестных системы уравнений и параметра. Ее построение проводится, как правило, при помощи численных методов и сопряжено с многочисленными трудностями, возникающими вследствие наличия на кривой множества решений предельных и существенно особых точек. Для нахождения таких кривых используется метод продолжения решения по параметру и наилучшей параметризации, позволяющий свести решение к начальной задаче для системы дифференциальных уравнений продолжения решения. В данной работе исследуется устойчивость решения системы продолжения решения на вносимые в нее возмущения. Впервые полностью доказано сформулированное ранее утверждение о минимальности квадратичной ошибки решения системы продолжения решения при однородных малых возмущениях ее матрицы. Теоретические результаты проиллюстрированы на примере численного построения лемнискаты Бернулли. Библ. 10. Фиг. 2. Табл. 1.

Ключевые слова: системы нелинейных уравнений, продолжение решения по параметру, наилучший параметр, система продолжения решения, малые возмущения, квадратичная погрешность.

DOI: 10.31857/S0044466924120087, **EDN**: KBTDTO

1. ВВЕДЕНИЕ

При описании задач механики деформируемого твердого тела, механики жидкости и газа, биологии, химии и экономики важную роль играют системы алгебраических и трансцендентных уравнений, которые в общем случае называются нелинейными. Но получить решение подобных уравнений и их систем весьма сложно. Начиная с XVI века разрабатываются общие методы, как аналитического, так и приближенного решения нелинейных уравнений и их систем. Поскольку далеко не каждое нелинейное уравнение или система могут быть решены аналитически, на первый план выходят приближенные методы. При этом и приближенные методы решения имеют существенные недостатки. Наиболее известны такие методы решения нелинейных уравнений и систем, как метод простой итерации и метод Ньютона. Оба этих метода обладают своими достоинствами и недостатками. Метод Ньютона по сравнению с методом простой итерации имеет квадратичную скорость сходимости, однако, в использовании требует выполнения значительно больших условий: подбор начального приближения, вычисление первой производной функции уравнения (матрицы Якоби для систем уравнений), значительные ограничения на функцию уравнения или системы, необходимость в анализе сходимости метода. Наличие подобных недостатков и значительная скорость сходимости метода Ньютона привели к появлению его модификаций и обобщений.

В 30-е годы прошлого века бельгийским математиком М. Лаэем был предложен метод решения нелинейных уравнений с параметром на основе метода Ньютона [1]. Метод Лаэя заключается в разбиении интервала изменения параметра уравнения и построении для каждой точки разбиения итерационной последовательности по

методу Ньютона, принимая в качестве начального приближения последнюю точку, полученную для предыдущего значения параметра. Этот метод нацелен на устранение сразу двух недостатков метода Ньютона: выбор начальной точки и обеспечение сходимости итерационной последовательности к корню уравнения с квадратичной скоростью. По всей видимости метод Лаэя является одним из первых примеров применения параметризации в вычислительной математике. Позднее М. Лаэй обобщил предложенный им метод на случай систем нелинейных уравнений [2].

Другую идею в своих работах [3, 4] использовал советский математик Д. Ф. Давиденко. Он рассматривал параметр системы нелинейных уравнений как аргумент, а кривые множества решений отождествлял с интегральными кривыми задачи Коши для системы дифференциальных уравнений, полученной из исходной системы нелинейных уравнений путем ее дифференцирования по параметру (полагая, что неизвестные системы являются функциями параметра). Выбирая начальное условие при начальном значении параметра можно свести решение системы нелинейных уравнений к решению соответствующей задачи Коши для системы дифференциальных уравнений.

Существенным недостатком, как метода Лаэя, так и метода Давиденко, является необходимость смены параметра продолжения решения, если кривая множества решений содержит предельные или существенно особые точки (например, в случае замкнутых или самопересекающихся кривых множества решений). Сам факт необходимости смены параметра говорит о неединственности его способа выбора, что ставит вопрос о наличии оптимального в некотором плане параметра продолжения решения.

Впервые гипотеза о том, что параметр, отсчитываемый по касательной к кривой множества решений рассматриваемой нелинейной системы, является в некотором смысле наилучшим высказал акад. И. И. Ворович [5]. Идея доказательства этой гипотезы принадлежит Э. Риксу [6], однако полное доказательство оптимальности параметра, отсчитываемого по касательной к кривой множества решений рассматриваемой нелинейной системы, было дано лишь в работе В. И. Шалашилина и Е. Б. Кузнецова [7]. В работе Э. Рикса под оптимальностью параметра продолжения решения понимается наилучшая обусловленность линеаризованной системы продолжения решения, но это предположение не доказывается. В работах В. И. Шалашилина и Е. Б. Кузнецова помимо наилучшей обусловленности доказывается еще и минимальность квадратичной погрешности решения системы продолжения решения, возникающей при возмущении матрицы системы и вектора правой части. Это дополнение является существенным, так как позволяет утверждать, что при использовании наилучшего параметра система продолжения решения не только обладает наилучшей обусловленностью, но и минимизируется влияние на решение этой системы возмущений различного рода: ошибок округления, вычислительных погрешностей, неустранимых погрешностей. Для ряда задач последнее может быть важнее, чем обусловленность.

К сожалению, доказательство минимальности квадратичной погрешности в работе [7] дано не полностью – строго проверены лишь необходимые условия локального условного экстремума. В данной статье авторы ставят целью устранить этот недостаток, строго доказав достаточные условия локального условного минимума квадратичной погрешности решения системы продолжения решения, возникающей при возмущении матрицы системы малыми однородными возмущениями, в случае использования наилучшего параметра.

2. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим систему n нелинейных уравнений относительно $n+1$ неизвестного

$$F_i\left(x_1,\dots,x_n,x_{n+1}\right)=0,\quad i=1,\dots,n. \tag{1}$$

Здесь и далее будем полагать, что функции F_i определены и непрерывны на всей рассматриваемой области изменения неизвестных $x_1,\dots,x_n,x_{n+1}$. Такими системами описываются многие задачи механики деформируемого твердого тела, физики, биологии, экономики. Тот факт, что количество неизвестных больше количества уравнений говорит о неединственности решения системы (1).

При определенных условиях можно построить множество решений системы (1). Примем одну из переменных системы (1) как параметр, для каждого значения которого (из некоторой области определения) система (1) является замкнутой. Без ограничения общности, в качестве параметра можно принять неизвестную x_{n+1}. Пусть известно какое-либо решение системы (1) для некоторого начального значения параметра $x_{(n+1)0}$, т. е. известна точка $M_0\left(x_{10},\dots,x_{n0},x_{(n+1)0}\right)$, удовлетворяющая всем уравнениям системы

$$F_i\left(x_{10},\dots,x_{n0},x_{(n+1)0}\right)=0,\quad i=1,\dots,n.$$

Если матрица Якоби относительно переменных $x_1,\dots,x_n$ невырожденна в окрестности точки M_0, то по теореме о неявной функции можно получить кривую множества решений в окрестности точки M_0, переходя по

параметру от значения параметра $x_{(n+1)0}$ к значению $x_{(n+1)1} = x_{(n+1)0} + h$, где h – заданная достаточно малая величина. При использовании метода Ньютона этот процесс определяется следующей итерационной процедурой:

$$\mathbf{x}_1^{(j)} = \mathbf{x}_1^{(j-1)} - J^{-1}\left(\mathbf{x}_1^{(j-1)}, x_{(n+1)1}\right) \cdot \mathbf{F}\left(\mathbf{x}_1^{(j-1)}, x_{(n+1)1}\right), \tag{2}$$

где $\mathbf{x} = (x_1, \ldots, x_n)^\top$, $\mathbf{x}_1^{(0)} = (x_{10}, \ldots, x_{n0})^\top$ – начальное приближение, $\mathbf{F} = (F_1, \ldots, F_n)^\top$, номер итерации задает индекс $j = 1, 2, \ldots$, а матрица Якоби относительно переменных $x_1, \ldots, x_n$ имеет вид

$$J(\mathbf{x}, x_{n+1}) = \left[\frac{\partial F_j}{\partial x_k}\right]_{j,k=1}^{n}.$$

Итерационный процесс (2) повторяется до момента выполнения условия $\left\|\mathbf{x}_1^{(j)} - \mathbf{x}_1^{(j-1)}\right\|_2 < \varepsilon$, где $\|\cdot\|_2$ – квадратичная векторная норма, которая для вектора $\mathbf{a} = (a_1, \ldots, a_n)^\top$ вычисляется по формуле $\|\mathbf{a}\|_2 = \sqrt{a_1^2 + \cdots + a_n^2}$, а ε задает величину требуемой погрешности. В результате будет получена точка $M_1\left(x_{11}, \ldots, x_{n1}, x_{(n+1)1}\right)$, также являющаяся решением системы (1). Продолжая этот процесс, можно продвигаться далее по кривой множества решений.

Описанный выше метод построения кривой множества решений системы (1) имеет значительный недостаток – он применим только в случае, когда матрица Якоби не вырождается. Если матрица Якоби вырождается в некоторой точке $M_k\left(x_{1k}, \ldots, x_{nk}, x_{(n+1)k}\right)$, то построение кривой множества решений из нее с использованием итерационной процедуры (2) невозможно. Используемый параметр x_{n+1} становится непригодным для продолжения решения и требует смены. Процедура смены параметра продолжения решения довольно трудоемка и слабо формализуема. Она состоит в следующем. На смену параметра x_{n+1} из переменных системы (1) выбирается новый параметр продолжения решения x_i, для которого в точке M_k будет невырожденной матрица Якоби для переменных $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, x_{n+1}$

$$\tilde{J}(\tilde{\mathbf{x}}, x_i) = \left[\frac{\partial F_j}{\partial x_k}\right]_{\substack{j,k=1 \\ j\neq n+1 \\ k\neq i}}^{n+1},$$

где $\tilde{\mathbf{x}} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, x_{n+1})^\top$. При наличии такого параметра x_i, итерационный процесс (2) преобразуется к виду

$$\tilde{\mathbf{x}}_{k+1}^{(j)} = \tilde{\mathbf{x}}_{k+1}^{(j-1)} - \tilde{J}^{-1}\left(\tilde{\mathbf{x}}_{k+1}^{(j-1)}, x_{i(k+1)}\right) \cdot \mathbf{F}\left(\tilde{\mathbf{x}}_{k+1}^{(j-1)}, x_{i(k+1)}\right),$$

где $\tilde{\mathbf{x}}_{k+1}^{(0)} = \left(x_{1k}, \ldots, x_{(i-1)k}, x_{(i+1)k}, \ldots, x_{nk}, x_{(n+1)k}\right)$, $x_{i(k+1)} = x_{ik} + h$, и процесс построения кривой множества решений продолжается. Описанная идея смены параметра продолжения решения впервые в численном анализе была отмечена в работах М. Лаэя.

Иную идею использовал в своих работах Д. Ф. Давиденко. Он полагал, что при выборе параметра, например, неизвестной x_{n+1}, остальные неизвестные $x_1, \ldots, x_n$ непрерывно зависят от него. Тогда дифференцируя уравнения системы (1) по переменной x_{n+1}, получим систему дифференциальных уравнений вида

$$\frac{\partial F_i}{\partial x_1} \cdot \frac{dx_1}{dx_{n+1}} + \cdots + \frac{\partial F_i}{\partial x_n} \cdot \frac{dx_n}{dx_{n+1}} + \frac{\partial F_i}{\partial x_{n+1}} = 0, \quad i = 1, \ldots, n, \tag{3}$$

или в векторно-матричной форме:

$$J\left(\mathbf{x}, x_{n+1}\right) \cdot \frac{d\mathbf{x}}{dx_{n+1}} + \frac{\partial \mathbf{F}}{\partial x_{n+1}} = 0. \tag{4}$$

Разрешая систему (4) относительно производной $\dfrac{d\mathbf{x}}{dx_{n+1}}$, перейдем к нормальной форме Коши

$$\frac{d\mathbf{x}}{dx_{n+1}} = -J^{-1}\left(\mathbf{x}, x_{n+1}\right) \cdot \frac{\partial \mathbf{F}}{\partial x_{n+1}}. \tag{5}$$

Дополняя систему уравнений (5) начальным условием

$$\mathbf{x}(x_{(n+1)0}) = \mathbf{x}_0, \quad \mathbf{x}_0 = (x_{10}, \ldots, x_{n0})^\top, \tag{6}$$

получим начальную задачу (5), (6).

В сравнении с методом Лаэя, метод Давиденко более привлекательный в вычислительном плане, поскольку не требует итерационного уточнения решения для каждого значения параметра. Однако метод Давиденко обладает тем же недостатком, что и метод Лаэя. Если матрица Якоби, входящая в правую часть системы (5), вырождается в некоторой точке $M_k\left(x_{1k},\ldots,x_{nk},x_{(n+1)k}\right)$, то дальнейшее движение вдоль кривой множества решений исходной системы (1) по параметру x_{n+1} невозможно. Для продолжения движения по кривой множества решений также необходима смена параметра продолжения. Эта идея впервые была применена Д. Ф. Давиденко. Для нового параметра продолжения решения x_i потребуем выполнения того же условия, что и в методе Лаэя: невырожденность матрицы Якоби $\tilde{J}(\tilde{\mathbf{x}},x_i)$. Преобразованная к новому параметру x_i задача (5), (6) примет вид системы

$$\frac{d\tilde{\mathbf{x}}}{dx_i}=-\tilde{J}^{-1}\left(\tilde{\mathbf{x}},x_i\right)\cdot\frac{\partial\mathbf{F}}{\partial x_i} \tag{7}$$

с начальным условием

$$\tilde{\mathbf{x}}(x_{ik})=\tilde{\mathbf{x}}_k,\quad \tilde{\mathbf{x}}_k=\left(x_{1k},\ldots,x_{(i-1)k},x_{(i+1)k}\ldots,x_{nk},x_{(n+1)k}\right)^{\top}. \tag{8}$$

Решая задачу (7), (8), мы продолжаем движение по кривой множества решений исходной системы (1) до момента, когда не возникнет новой точки, в которой вырождается определитель матрицы Якоби.

Метод Давиденко хоть и не преодолевает недостатков метода Лаэя, но дает идею построения кривой множества решений в случае, когда ни одна из неизвестных системы (1) не подходит на роль нового параметра. Эта идея состоит в использовании параметров продолжения иного вида.

Метод продолжения решения по параметру предполагает использование параметров продолжения более общего вида. Как правило они задаются локально, в каждой точке кривой множества решений. Определим параметр продолжения μ для системы (1) в форме

$$d\mu=\alpha_1\cdot dx_1+\cdots+\alpha_n\cdot dx_n+\alpha_{n+1}\cdot dx_{n+1}, \tag{9}$$

где $dx_1,\ldots,dx_{n+1}$ – дифференциалы неизвестных системы (1), $\alpha_1,\ldots,\alpha_{n+1}$ – заданные числовые коэффициенты. Стоит отметить, что при переходе от точки к точке множества решений системы (1), можно изменять параметр μ, т. е. придавать коэффициентам $\alpha_1,\ldots,\alpha_{n+1}$ другие значения, поэтому глобально $\alpha_1,\ldots,\alpha_{n+1}$ являются функциями неизвестных системы (1). Геометрически коэффициенты $\alpha_1,\ldots,\alpha_{n+1}$ можно интерпретировать как компоненты направляющего вектора $\bar{\alpha}$, вдоль которого отсчитывается параметр μ.

Полагая, что неизвестные системы (1) зависят от параметра μ, т. е.

$$x_1=x_1(\mu),\ldots,x_n=x_n(\mu),x_{n+1}=x_{n+1}(\mu),$$

можно продифференцировать уравнения системы (1) по параметру μ, получив систему дифференциальных уравнений

$$\frac{\partial F_i}{\partial x_1}\cdot\frac{dx_1}{d\mu}+\cdots+\frac{\partial F_i}{\partial x_n}\cdot\frac{dx_n}{d\mu}+\frac{\partial F_i}{\partial x_{n+1}}\cdot\frac{dx_{n+1}}{d\mu}=0,\quad i=1,\ldots,n. \tag{10}$$

Дополняя полученную систему уравнением, полученным из (9) делением на $d\mu$:

$$\alpha_1\cdot\frac{dx_1}{d\mu}+\cdots+\alpha_n\cdot\frac{dx_n}{d\mu}+\alpha_{n+1}\cdot\frac{dx_{n+1}}{d\mu}=1, \tag{11}$$

получим замкнутую систему обыкновенных уравнений с $n+1$ неизвестной $\dfrac{dx_1}{d\mu},\ldots,\dfrac{dx_{n+1}}{d\mu}$. Для удобства вычислений, наложим на коэффициенты $\alpha_1,\ldots,\alpha_{n+1}$ условие нормировки:

$$\alpha_1^2+\cdots+\alpha_{n+1}^2=1. \tag{12}$$

Последнее равенство отражает тот факт, что все направления отсчета параметра продолжения являются равноправными.

Можно видеть, что система (3) схожа по структуре с системой (10), поскольку использует ту же идею зависимости неизвестных системы (1) от параметра продолжения.

Запишем систему (10), (11) в векторно-матричном виде

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{n+1} \\ \dfrac{\partial F_1}{\partial x_1} & \dfrac{\partial F_1}{\partial x_2} & \cdots & \dfrac{\partial F_1}{\partial x_{n+1}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} & \dfrac{\partial F_n}{\partial x_2} & \cdots & \dfrac{\partial F_n}{\partial x_{n+1}} \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dx_1}{d\mu} \\ \dfrac{dx_2}{d\mu} \\ \vdots \\ \dfrac{dx_{n+1}}{d\mu} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{13}$$

Систему (13) будем называть системой *продолжения решения*.

Сам процесс решения системы (13) не отличается от используемого в методе Давиденко. Система (13) разрешается относительно производных неизвестных системы (1) и решается при начальных условиях

$$x_1(\mu_0) = x_{10}, \ldots, x_{n+1}(\mu_0) = x_{(n+1)0}.$$

Заметим, что система (13) не зависит явно от параметра μ, значит можно положить $\mu_0 = 0$.

Разрешение системы (13) и решение полученной начальной задачи проводится аналитически только в исключительных случаях. Эффективность же численного решения зависит существенно от обусловленности матрицы системы (13). Если выбрать параметр продолжения решения (т. е. коэффициенты $\alpha_1, \ldots, \alpha_{n+1}$) неудачно, то матрица системы (13) будет вырождаться, что приведет к необходимости смены параметра. Как видно из приводимых выше выкладок, эта процедура нежелательна. Стоит выбирать такой параметр продолжения решения, который не требуют смены в процессе решения. Очевидно, что подобный параметр будет не единственным, что ставит задачу выбора оптимального параметра продолжения решения. В работе [7] было доказано, что параметр продолжения решения, отсчитываемый по касательной к кривой множества решений, т. е. тот параметр, для которого коэффициенты

$$\alpha_1 = \frac{dx_1}{d\mu}, \ldots, \alpha_{n+1} = \frac{dx_{n+1}}{d\mu}, \tag{14}$$

будет доставлять матрице системы (13) наилучшую обусловленность. Такой параметр продолжения получил название наилучшего и обозначение λ.

Помимо наилучшей обусловленности параметр λ обладает тем свойством, что при его использовании становится минимальной квадратичная погрешность решения, возникающая при возмущении элементов матрицы системы (13). Это утверждение также было сформулировано в работе [7], однако полного его доказательства дано не было. Доказательство было ограничено лишь проверкой необходимых условий локального минимума, достаточные условия были лишь намечены. Целью данной работы является устранение этого недостатка.

3. ОСНОВНЫЕ ЛЕММЫ

Предварительно докажем две леммы.

Лемма 1. *Для любого натурального значения n, определитель n-го порядка*

$$\begin{vmatrix} a_1^2 + a_0 & -a_1 \cdot a_2 & a_1 \cdot a_3 & \cdots & (-1)^{n+1} a_1 \cdot a_n \\ -a_2 \cdot a_1 & a_2^2 + a_0 & -a_2 \cdot a_3 & \cdots & (-1)^{n+2} a_2 \cdot a_n \\ a_3 \cdot a_1 & -a_3 \cdot a_2 & a_3^2 + a_0 & \cdots & (-1)^{n+3} a_3 \cdot a_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (-1)^{n+1} a_n \cdot a_1 & (-1)^{n+2} a_n \cdot a_2 & (-1)^{n+3} a_n \cdot a_3 & \cdots & a_n^2 + a_0 \end{vmatrix}_n = \tag{15}$$

$$= a_0^{n-1} \cdot \left(a_0 + \sum_{i=1}^{n} a_i^2 \right),$$

где $a_0, a_1, \ldots, a_n$ – заданные вещественные числа.

Доказательство леммы 1. Для доказательства данной леммы используем метод математической индукции. Сформируем базу индукции.

Для значения $n = 1$ определитель вычисляется тривиально:

$$\left| a_1^2 + a_0 \right| = a_1^2 + a_0.$$

Для значения $n = 2$:

$$\begin{vmatrix} a_1^2 + a_0 & -a_1 \cdot a_2 \\ -a_2 \cdot a_1 & a_2^2 + a_0 \end{vmatrix} = a_0 \cdot \left(a_1^2 + a_2^2 + a_0\right).$$

Для значения $n = 3$, применяя разложение по последней строке, получим

$$\begin{vmatrix} a_1^2 + a_0 & -a_1 \cdot a_2 & a_1 \cdot a_3 \\ -a_2 \cdot a_1 & a_2^2 + a_0 & -a_2 \cdot a_3 \\ a_3 \cdot a_1 & -a_3 \cdot a_2 & a_3^2 + a_0 \end{vmatrix} = a_0^2 \cdot \left(a_1^2 + a_2^2 + a_3^2 + a_0\right).$$

Выполненные вычисления позволяют сформулировать гипотезу: для произвольного натурального n будет справедлива формула (15).

Проверим гипотезу, вычислив определитель размерности $n + 1$, разлагая его по последней строке. После применения формулы (15) и упрощения, получим

$$D_{n+1} = \begin{vmatrix} a_1^2 + a_0 & -a_1 \cdot a_2 & a_1 \cdot a_3 & \cdots & (-1)^{n+2} a_1 \cdot a_{n+1} \\ -a_2 \cdot a_1 & a_2^2 + a_0 & -a_2 \cdot a_3 & \cdots & (-1)^{n+3} a_2 \cdot a_{n+1} \\ a_3 \cdot a_1 & -a_3 \cdot a_2 & a_3^2 + a_0 & \cdots & (-1)^{n+4} a_3 \cdot a_{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (-1)^{n+2} a_{n+1} \cdot a_1 & (-1)^{n+3} a_{n+1} \cdot a_2 & (-1)^{n+4} a_{n+1} \cdot a_3 & \cdots & a_{n+1}^2 + a_0 \end{vmatrix}_{n+1} =$$

$$= \left(a_{n+1}^2 + a_0\right) \cdot a_0^{n-1} \cdot \left(a_0 + \sum_{i=1}^{n} a_i^2\right) +$$

$$+ a_{n+1}^2 \cdot a_1 \cdot \begin{vmatrix} -a_1 \cdot a_2 & a_1 \cdot a_3 & \cdots & (-1)^{n+1} a_1 \cdot a_n & (-1)^{n+2} a_1 \\ a_2^2 + a_0 & -a_2 \cdot a_3 & \cdots & (-1)^{n+2} a_2 \cdot a_n & (-1)^{n+3} a_2 \\ -a_3 \cdot a_2 & a_3^2 + a_0 & \cdots & (-1)^{n+3} a_3 \cdot a_n & (-1)^{n+4} a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (-1)^{n+2} a_n \cdot a_2 & (-1)^{n+3} a_n \cdot a_3 & \cdots & a_n^2 + a_0 & (-1)^{2n+1} a_n \end{vmatrix}_{n} + \cdots$$

$$\cdots + a_{n+1}^2 \cdot a_n \cdot \begin{vmatrix} a_1^2 + a_0 & -a_1 \cdot a_2 & \cdots & (-1)^{n} a_1 \cdot a_{n-1} & (-1)^{n+2} a_1 \\ -a_2 \cdot a_1 & a_2^2 + a_0 & \cdots & (-1)^{n+1} a_2 \cdot a_{n-1} & (-1)^{n+3} a_2 \\ a_3 \cdot a_1 & -a_3 \cdot a_2 & \cdots & (-1)^{n+2} a_3 \cdot a_{n-1} & (-1)^{n+4} a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (-1)^{n+1} a_n \cdot a_1 & (-1)^{n+2} a_n \cdot a_2 & \cdots & (-1)^{2n-1} a_n \cdot a_{n-1} & (-1)^{2n+1} a_n \end{vmatrix}_{n}.$$

Остается лишь вычислить полученные n определителей порядка n. Все они вычисляются одинаково. Продемонстрируем процедуру их вычисления на примере последнего определителя, обозначив его через D_n:

$$D_n = a_1 \cdot \begin{vmatrix} a_1 & -a_1 \cdot a_2 & \cdots & (-1)^{n} a_1 \cdot a_{n-1} & (-1)^{n+2} a_1 \\ -a_2 & a_2^2 + a_0 & \cdots & (-1)^{n+1} a_2 \cdot a_{n-1} & (-1)^{n+3} a_2 \\ a_3 & -a_3 \cdot a_2 & \cdots & (-1)^{n+2} a_3 \cdot a_{n-1} & (-1)^{n+4} a_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (-1)^{n+1} a_n & (-1)^{n+2} a_n \cdot a_2 & \cdots & (-1)^{2n-1} a_n \cdot a_{n-1} & (-1)^{2n+1} a_n \end{vmatrix}_{n} +$$

$$+ a_0 \cdot \begin{vmatrix} a_2^2 + a_0 & -a_2 \cdot a_3 & \cdots & (-1)^{n+1} a_2 \cdot a_{n-1} & (-1)^{n+3} a_2 \\ -a_3 \cdot a_2 & a_3^2 + a_0 & \cdots & (-1)^{n+2} a_3 \cdot a_{n-1} & (-1)^{n+4} a_3 \\ a_4 \cdot a_2 & -a_4 \cdot a_3 & \cdots & (-1)^{n+3} a_4 \cdot a_{n-1} & (-1)^{n+5} a_4 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (-1)^{n+2} a_n \cdot a_2 & (-1)^{n+3} a_n \cdot a_3 & \cdots & (-1)^{2n-1} a_n \cdot a_{n-1} & (-1)^{2n+1} a_n \end{vmatrix}_{n-1}.$$

Рассмотрим определитель, стоящий в первом слагаемом. Если из его последнего столбца вынести множитель $(-1)^n$, то у полученного определителя будут совпадать первый и последний столбцы, следовательно, он будет равняться нулю. Полученный в итоге определитель по структуре аналогичен исходному, поэтому проводя проделанные операции еще $n - 3$ раз, мы придем к виду

$$D_n = a_0^{n-2} \cdot \begin{vmatrix} a_0 & (-1)^{2n} a_{n-1} \\ 0 & (-1)^{2n+1} a_n \end{vmatrix} = -a_0^{n-1} a_n.$$

Таким же образом, можно вычислить и $n-1$ оставшихся определителей, которые все будут иметь отрицательный знак, как это видно при вычислении определителя для произвольного значения n. Таким образом, исходный определитель $n+1$ порядка может быть записан в виде

$$D_{n+1} = \left(a_{n+1}^2 + a_0\right) \cdot a_0^{n-1} \cdot \left(a_0 + \sum_{i=1}^{n} a_i^2\right) - a_0^{n-1} a_{n+1}^2 \sum_{i=1}^{n} a_i^2 = a_0^n \cdot \left(a_0 + \sum_{i=1}^{n+1} a_i^2\right).$$

По методу математической индукции формула (15) справедлива для всех натуральных значений n.

Доказательство леммы 1 завершено.

Используя лемму 1 можно доказать, что справедлива

Лемма 2. *Функция*

$$J(\alpha_1, \ldots, \alpha_{n+1}) = \frac{A}{\Delta^m} \tag{16}$$

достигает своего минимального значения при ограничении типа равенств (12) *в точке*

$$\alpha_k^* = \frac{(-1)^{k+1} \cdot \Delta_k}{\Delta}, \quad k = 1, \ldots, n+1, \tag{17}$$

если $A > 0$ и m принимает четные значения. Если к тому же $\Delta > 0$, то значения (17) *доставляют минимум функции* (16) *при $A > 0$ и любом натуральном m.*

В функции (16) *знаменатель задается формулой*

$$\Delta = \sum_{i=1}^{n+1} (-1)^{i+1} \alpha_i \Delta_i, \tag{18}$$

постоянные величины $A, \Delta_1, \ldots, \Delta_{n+1}$ не зависят от переменных $\alpha_1, \ldots, \alpha_{n+1}$. В выражениях (17) *значение знаменателя $\Delta = \pm\left(\Delta_1^2 + \cdots + \Delta_{n+1}^2\right)^{1/2}$.*

Доказательство леммы 2. Перепишем задачу минимизации функции (16) при ограничении типа равенств (12), прибегая к используемым в тензорной алгебре обозначениям:

$$\frac{A}{\Delta^m} \to \min_{\overline{\alpha} \in \mathbb{R}^{n+1}}, \quad \alpha_i \cdot \alpha_i = 1, \quad i = 1, \ldots, n+1, \tag{19}$$

где $\overline{\alpha} = (\alpha_1, \ldots, \alpha_{n+1})^\top$ – вектор переменных задачи. Здесь и далее будем предполагать суммирование по повторяющимся индексам в диапазоне их изменения, если не оговорено обратное.

Для решения задачи минимизации с ограничением типа равенств (19) воспользуемся методом множителей Лагранжа. Для этого составим функцию Лагранжа

$$L(\alpha_1, \ldots, \alpha_{n+1}, \gamma) = \frac{A}{\Delta^m} + \gamma \cdot (\alpha_i \cdot \alpha_i - 1), \quad i = 1, \ldots, n+1, \tag{20}$$

где γ – множитель Лагранжа.

Необходимое условие. Для проверки необходимых условий условного экстремума вычислим первые производные функции Лагранжа (20):

$$\frac{\partial L}{\partial \alpha_k} = -m \cdot \frac{A}{\Delta^{m+1}} \cdot (-1)^{k+1} \cdot \Delta_k + 2\gamma \cdot \alpha_k, \quad k = 1, \ldots, n+1. \tag{21}$$

Приравнивая производные (21) к нулю, получим систему из $n+1$ уравнения относительно $\alpha_1, \ldots, \alpha_{n+1}$ и γ:

$$-m \cdot \frac{A}{\Delta^{m+1}} \cdot (-1)^{k+1} \cdot \Delta_k + 2 \cdot \gamma \cdot \alpha_k = 0, \quad k = 1, \ldots, n+1. \tag{22}$$

Выразим α_k из системы (22):

$$\alpha_k = \frac{m \cdot A}{2 \cdot \gamma \cdot \Delta^{m+1}} \cdot (-1)^{k+1} \cdot \Delta_k, \quad k = 1, \ldots, n+1. \tag{23}$$

Пусть

$$\omega = \frac{m \cdot A}{2 \cdot \Delta^m}.$$

Тогда выражение (23) для α_k перепишется в виде

$$\alpha_k = \frac{\omega}{\gamma} \cdot \frac{(-1)^{k+1} \cdot \Delta_k}{\Delta}, \quad k = 1, \ldots, n+1. \tag{24}$$

Используем ограничение задачи (12):

$$\alpha_i \cdot \alpha_i = \frac{\omega^2}{\gamma^2} \cdot \frac{\Delta_i \cdot \Delta_i}{\Delta^2} = 1, \quad i = 1, \ldots, n+1. \tag{25}$$

Из равенства (25) получим

$$\gamma = \pm\omega \cdot \frac{(\Delta_i \cdot \Delta_i)^{1/2}}{\Delta}, \quad i = 1, \ldots, n+1. \tag{26}$$

Подставляя множитель Лагранжа (26) в соотношение (24), найдем

$$\alpha_k = \pm\frac{(-1)^{k+1} \cdot \Delta_k}{(\Delta_i \cdot \Delta_i)^{1/2}}, \quad i, k = 1, \ldots, n+1. \tag{27}$$

Упростим выражения (27). Подставим эти значения компонент α_k в формулу для определителя Δ (18):

$$\Delta = (-1)^{i+1} \cdot \alpha_i \cdot \Delta_i = \pm\frac{\Delta_i \cdot \Delta_i}{(\Delta_i \cdot \Delta_i)^{1/2}} = \pm(\Delta_i \cdot \Delta_i)^{1/2}, \quad i = 1, \ldots, n+1.$$

Подставляя полученное равенство в выражение для компонент α_k, найдем условно стационарную точку (17).

Достаточное условие. Как известно, характер экстремума задачи (19) связан со знакоопределенностью второго дифференциала функции Лагранжа (20)

$$d^2L = \sum_{i,j=1}^{n+1} \frac{\partial^2 L}{\partial\alpha_i \partial\alpha_j} d\alpha_i d\alpha_j. \tag{28}$$

В свою очередь знакоопределенность второго дифференциала функции Лагранжа зависит от знакоопределенности матрицы вторых производных функции Лагранжа. Вычислим эти производные, используя выражения для первых производных (21):

$$\frac{\partial^2 L}{\partial\alpha_k^2} = m \cdot (m+1) \cdot \frac{A}{\Delta^{m+2}} \cdot \Delta_k^2 + 2\gamma,$$
$$\frac{\partial^2 L}{\partial\alpha_k \partial\alpha_l} = m \cdot (m+1) \cdot \frac{A}{\Delta^{m+2}} \cdot (-1)^{k+l} \cdot \Delta_k \cdot \Delta_l, \quad i, k, l = 1, \ldots, n+1.$$

Учитывая, что, согласно достаточному условию условного экстремума второго рода, второй дифференциал функции Лагранжа (28) должен быть знакоопределенным в окрестности найденной условно стационарной точки, перепишем вторые производные в виде

$$\left.\frac{\partial^2 L}{\partial\alpha_k^2}\right|_{\alpha_k = \alpha_k^*} = \frac{A_1}{\Delta^{m+2}} \cdot \Delta_k^2 + 2\gamma^*,$$
$$\left.\frac{\partial^2 L}{\partial\alpha_k \partial\alpha_l}\right|_{\alpha_k = \alpha_k^*} = \frac{A_1}{\Delta^{m+2}} \cdot (-1)^{k+l} \cdot \Delta_k \cdot \Delta_l, \quad i, k, l = 1, \ldots, n+1,$$

где $\gamma^* = \omega = \dfrac{m \cdot A}{2 \cdot \Delta^m}$, $\Delta = \pm(\Delta_i \cdot \Delta_i)^{1/2}$, $A_1 = m \cdot (m+1) \cdot A$.

Таким образом, в окрестности полученной условно стационарной точки (17), матрица вторых производных функции Лагранжа (20) примет вид

$$\left[\frac{\partial^2 L}{\partial\alpha_i\partial\alpha_j}\right]_{i,j=1}^{n+1} =$$

$$= \begin{pmatrix} \frac{A_1}{\Delta^{m+2}}\cdot\Delta_1^2+2\gamma^* & -\frac{A_1}{\Delta^{m+2}}\cdot\Delta_1\cdot\Delta_2 & \cdots & \frac{(-1)^{n+2}\cdot A_1}{\Delta^{m+2}}\cdot\Delta_1\cdot\Delta_{n+1} \\ -\frac{A_1}{\Delta^{m+2}}\cdot\Delta_1\cdot\Delta_2 & \frac{A_1}{\Delta^{m+2}}\cdot\Delta_2^2+2\gamma^* & \cdots & \frac{(-1)^{n+3}\cdot A_1}{\Delta^{m+2}}\cdot\Delta_2\cdot\Delta_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(-1)^{n+2}\cdot A_1}{\Delta^{m+2}}\cdot\Delta_1\cdot\Delta_{n+1} & \frac{(-1)^{n+3}\cdot A_1}{\Delta^{m+2}}\cdot\Delta_2\cdot\Delta_{n+1} & \cdots & \frac{A_1}{\Delta^{m+2}}\cdot\Delta_{n+1}^2+2\gamma^* \end{pmatrix}. \tag{29}$$

Для анализа знакоопределенности матрицы (29) используем критерий Сильвестра. Рассмотрим M_k – угловой минор k-го порядка матрицы (29), равный

$$M_k = \left(\frac{A_1}{\Delta^{m+2}}\right)^k \begin{vmatrix} \Delta_1^2+\frac{\Delta^2}{m+1} & -\Delta_1\cdot\Delta_2 & \cdots & (-1)^{k+1}\cdot\Delta_1\cdot\Delta_k \\ -\Delta_1\cdot\Delta_2 & \Delta_2^2+\frac{\Delta^2}{m+1} & \cdots & (-1)^{k+2}\cdot\Delta_2\cdot\Delta_k \\ \vdots & \vdots & \ddots & \vdots \\ (-1)^{k+1}\cdot\Delta_1\cdot\Delta_k & (-1)^{k+2}\cdot\Delta_2\cdot\Delta_k & \cdots & \Delta_k^2+\frac{\Delta^2}{m+1} \end{vmatrix}_k.$$

Полученный определитель по структуре полностью удовлетворяет лемме 1 при $a_0 = \dfrac{\Delta^2}{m+1}$, $a_1 = \Delta_1, \ldots, a_k = \Delta_k$. Поскольку лемма 1 справедлива для любого натурального значения n, то все угловые миноры матрицы вторых производных (29) можно вычислить по формуле

$$M_k = \frac{m+1}{\Delta^2}\cdot\left(\frac{m\cdot A}{\Delta^m}\right)^k\cdot\left(\frac{\Delta^2}{m+1}+\sum_{i=1}^{k}\Delta_i^2\right), \quad k=1,\ldots,n+1. \tag{30}$$

Знак миноров (30) определяется множителем $\left(\dfrac{m\cdot A}{\Delta^m}\right)^k$. Этот множитель положителен, если $A>0$ и m принимает четные значения. Если же дополнительно $\Delta>0$, то $\left(\dfrac{m\cdot A}{\Delta^m}\right)^k>0$ при $A>0$ и любых натуральных m.

При указанных условиях все угловые миноры матрицы вторых производных функции Лагранжа (29) строго положительны и согласно критерию Сильвестра она будет положительно определена, а значит, и $d^2L>0$. Тогда, согласно достаточным условиям условного экстремума с ограничением типа равенств, найденная условно стационарная точка (17) является точкой условного минимума, доставляя функции (16) наименьшее значение при условии (12).

Доказательство леммы 2 завершено.

4. МИНИМАЛЬНОСТЬ КВАДРАТИЧНОЙ ПОГРЕШНОСТИ СИСТЕМЫ ПРОДОЛЖЕНИЯ РЕШЕНИЯ ПРИ МАЛЫХ ОДНОРОДНЫХ ВОЗМУЩЕНИЯХ ЕЕ МАТРИЦЫ

Структура системы продолжения решения (13) позволяет явно выписать формулы для компонент ее решения. Выпишем эти формулы, прежде чем перейти к возмущенному случаю.

Здесь и далее будем обозначать $x_1,\ldots,x_{n+1}$ невозмущенные неизвестные исходной системы (1), а производными $\dfrac{dx_1}{d\mu},\ldots,\dfrac{dx_{n+1}}{d\mu}$ будем обозначать компоненты решения невозмущенной системы продолжения решения (13).

Следуя монографии [8], используем метод Крамера для нахождения решения системы продолжения решения (13):

$$\frac{dx_i}{d\mu} = (-1)^{i+1}\frac{\Delta_i}{\Delta}, \quad i=1,\ldots,n+1. \tag{31}$$

Здесь Δ – определитель матрицы системы (13), который может быть вычислен по формуле (18), в которой $\Delta_1, \ldots, \Delta_{n+1}$ являются минорами, получаемыми из матрицы системы (13) вычеркиванием первой строки и столбца, номер которого соответствует индексу, т. е.

$$\Delta_i = \begin{vmatrix} \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_1}{\partial x_{i-1}} & \dfrac{\partial F_1}{\partial x_{i+1}} & \cdots & \dfrac{\partial F_1}{\partial x_n} & \dfrac{\partial F_1}{\partial x_{n+1}} \\ \dfrac{\partial F_2}{\partial x_1} & \cdots & \dfrac{\partial F_2}{\partial x_{i-1}} & \dfrac{\partial F_2}{\partial x_{i+1}} & \cdots & \dfrac{\partial F_2}{\partial x_n} & \dfrac{\partial F_2}{\partial x_{n+1}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} & \cdots & \dfrac{\partial F_n}{\partial x_{i-1}} & \dfrac{\partial F_n}{\partial x_{i+1}} & \cdots & \dfrac{\partial F_n}{\partial x_n} & \dfrac{\partial F_n}{\partial x_{n+1}} \end{vmatrix}. \tag{32}$$

Будем полагать далее, что $\Delta_1^2 + \cdots + \Delta_{n+1}^2 \neq 0$, т. е. множество решений системы (1) не содержит существенно особых точек.

Обозначим далее $y_1, \ldots, y_{n+1}$ возмущенные неизвестные исходной системы (1), а производными $\dfrac{dy_1}{d\mu}, \ldots, \dfrac{dy_{n+1}}{d\mu}$ будем обозначать компоненты решения системы продолжения решения (13) после наложения на ее матрицу возмущений.

В данной статье под возмущениями некоторой строки (или столбца) матрицы системы продолжения решения (13) вида

$$(a_1, a_2, \ldots, a_{n+1})$$

будет пониматься строка (или столбец) вида

$$(\varepsilon_1 \cdot a_1, \varepsilon_2 \cdot a_2, \ldots, \varepsilon_{n+1} \cdot a_{n+1}),$$

которая прибавляется к исходной строке (или столбцу) при заданных значениях $\varepsilon_1, \ldots, \varepsilon_{n+1}$. Будем говорить, что возмущения однородные, если выполняется равенство

$$\varepsilon_1 = \cdots = \varepsilon_{n+1} = \varepsilon.$$

Рассмотрим малые возмущения, т. е. будем полагать, что слагаемыми, содержащими квадраты и более высокие степени $\varepsilon_1, \ldots, \varepsilon_{n+1}$, можно пренебречь по сравнению со слагаемыми, содержащими меньшие степени $\varepsilon_1, \ldots, \varepsilon_{n+1}$.

Естественным ограничением на значения $\varepsilon_1, \ldots, \varepsilon_{n+1}$ является их неравенство нулю. В противном случае, исходная и возмущенная системы продолжения решения будут совпадать.

В прикладных задачах наличие возмущений в системе продолжения решения (13) неизбежно. Поэтому важно иметь возможность прогнозирования их влияния на полученное решение. Возможны случаи, когда возмущения системы (13) полностью искажают искомое решение. Проанализируем влияние возмущений на решение системы (13). Введем погрешность компоненты решения возмущенной системы (13) в форме разности

$$\delta_i = \frac{dy_i}{d\mu} - \frac{dx_i}{d\mu}, \quad i = 1, \ldots, n+1. \tag{33}$$

В качестве меры погрешности будем использовать квадратичную погрешность

$$\delta = \sum_{i=1}^{n+1} \delta_i^2 = \sum_{i=1}^{n+1} \left(\frac{dy_i}{d\mu} - \frac{dx_i}{d\mu} \right)^2. \tag{34}$$

Разумеется квадратичная погрешность будет зависеть от выбора коэффициентов $\alpha_1, \ldots, \alpha_{n+1}$. В статье [7] сформулирована теорема о минимальности квадратичной погрешности решения возмущенной системы продолжения решения при выборе в окрестности каждой точки кривой множества решений в качестве значений $\alpha_1, \ldots, \alpha_{n+1}$ компонент вектора, касательного к кривой множества решений системы (1).

Полного доказательства этой теоремы в работе [7] не дано. Приведем его для случая однородных возмущений.

Прежде, чем переходить к полному доказательству, рассмотрим три частных случая: возмущение первой строки матрицы системы (13), возмущение любой строки матрицы системы (13), отличной от первой, и возмущение любого столбца матрицы системы (13). Это приводит к формулировке и доказательству трех теорем.

Придадим первой строчке матрицы системы (13) возмущения, прибавив к ней строку $\varepsilon\,(\alpha_1,\ldots,\alpha_{n+1})$, где ε – заданная малая величина, квадратом которой можно пренебречь по сравнению с меньшими степенями. Тогда возмущенная система продолжения решения запишется в виде

$$\begin{pmatrix} \alpha_1+\varepsilon\cdot\alpha_1 & \alpha_2+\varepsilon\cdot\alpha_2 & \cdots & \alpha_{n+1}+\varepsilon\cdot\alpha_{n+1} \\ \dfrac{\partial F_1}{\partial x_1} & \dfrac{\partial F_1}{\partial x_2} & \cdots & \dfrac{\partial F_1}{\partial x_{n+1}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial x_n} & \dfrac{\partial F_n}{\partial x_2} & \cdots & \dfrac{\partial F_n}{\partial x_{n+1}} \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} \\ \dfrac{dy_2}{d\mu} \\ \vdots \\ \dfrac{dy_{n+1}}{d\mu} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{35}$$

Для возмущенной системы продолжения решения (35) справедлива

Теорема 1. *Квадратичная погрешность решения возмущенной системы продолжения решения* (35)*, полученной из системы продолжения решения* (13) *наложением малых однородных возмущений на первую строку ее матрицы системы, достигает наименьшего значения в том случае, когда направляющий вектор* $\overline{\alpha}=(\alpha_1,\ldots,\alpha_{n+1})^{\top}$ *направлен по касательной к кривой множества решений системы* (1) *в рассматриваемой точке, т. е. компоненты вектора* $\overline{\alpha}$ *задаются в виде* (14).

Доказательство теоремы 1. По аналогии с формулами (31), (32), можно дать решение и для возмущенной системы (35):

$$\frac{dy_i}{d\mu}=(-1)^{i+1}\frac{\Delta_{i\varepsilon}}{\Delta_\varepsilon},\quad i=1,\ldots,n+1. \tag{36}$$

Определитель Δ_ε может быть непосредственно вычислен вынесением общего множителя из первой строки:

$$\Delta_\varepsilon=(1+\varepsilon)\cdot\Delta, \tag{37}$$

а поскольку возмущению подвержена только первая строка матрицы системы (13), то справедливы равенства

$$\Delta_{i\varepsilon}=\Delta_i,\quad i=1,\ldots,n+1. \tag{38}$$

Подставляя равенства (37), (38) в формулы (36), получим

$$\frac{dy_i}{d\mu}=(-1)^{i+1}\frac{\Delta_{i\varepsilon}}{\Delta_\varepsilon}=(-1)^{i+1}\frac{\Delta_i}{(1+\varepsilon)\cdot\Delta},\quad i=1,\ldots,n+1. \tag{39}$$

Воспользуемся теперь малостью возмущений. Домножим числитель и знаменатель правой части формулы (39) на $(1-\varepsilon)$ и отбросим слагаемое, включающее ε^2, перейдя к виду

$$\frac{dy_i}{d\mu}=(-1)^{i+1}\frac{(1-\varepsilon)\cdot\Delta_i}{\Delta},\quad i=1,\ldots,n+1. \tag{40}$$

Используя формулы (31) и (40), вычислим погрешности компонент решения возмущенной системы (33):

$$\delta_i=\frac{dy_i}{d\mu}-\frac{dx_i}{d\mu}=-(-1)^{i+1}\frac{\varepsilon\cdot\Delta_i}{\Delta},\quad i=1,\ldots,n+1. \tag{41}$$

Тогда, согласно формуле (34), квадратичная погрешность будет иметь вид

$$\delta=\sum_{i=1}^{n+1}\delta_i^2=\frac{\varepsilon^2}{\Delta^2}\sum_{i=1}^{n+1}\Delta_i^2. \tag{42}$$

Поставим задачу минимизации квадратичной ошибки (42), учитывая ограничение (12), накладываемое на выбор направления отсчета параметра продолжения:

$$\frac{\varepsilon^2\cdot\Delta_i\cdot\Delta_i}{\Delta^2}\to\min_{\overline{\alpha}\in\mathbb{R}^{n+1}},\quad \alpha_i\cdot\alpha_i=1,\quad i=1,\ldots,n+1. \tag{43}$$

Полученная задача (43) поиска минимума при ограничении типа равенств удовлетворяет условиям леммы 2 при $A=\varepsilon^2\cdot\Delta_i\cdot\Delta_i>0$, $i=1,\ldots,n+1$, и $m=2$. Это означает, что эта задача имеет точку минимума

с компонентами (17), совпадающими с компонентами касательного вектора к кривой множества решений в рассматриваемой точке:

$$\alpha_k = \alpha_k^* = \frac{dx_k}{d\mu}, \quad k = 1, \ldots, n+1. \tag{44}$$

В этой точке квадратичная погрешность (42) принимает наименьшее значение.

Доказательство теоремы 1 завершено.

Замечание 1. В данной работе внесение возмущений в матрицу системы продолжения решения (13) происходит путем добавления к некоторой строке (или столбцу) матрицы системы вида

$$(a_1, a_2, \ldots, a_{n+1})$$

возмущающей строки (или столбца) вида

$$(\varepsilon_1 \cdot a_1, \varepsilon_2 \cdot a_2, \ldots, \varepsilon_{n+1} \cdot a_{n+1}),$$

т. е. возмущение пропорционально самим компонентам возмущаемых строк (или столбцов) с коэффициентами пропорциональности $\varepsilon_1, \ldots, \varepsilon_{n+1}$. Такой способ наложения возмущений очень схож с процессом накопления вычислительной погрешности при использовании итерационных процессов решения нелинейных систем или пошаговых процессов решения систем обыкновенных дифференциальных уравнений.

В отличие от указанного способа, можно проводить возмущение матрицы системы путем добавления строки (или столбца) вида

$$(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{n+1}).$$

В таком случае возмущения не зависят от возмущаемой строки (или столбца). Подобный вид возмущений можно сравнить с наложением неустранимой погрешности, возникающей на уровне построения модели, а не на уровне вычислений.

При всей схожести процессов наложения возмущений результаты их применения могут сильно отличаться, что является отдельной задачей для исследования.

Придадим $j+1$-й строке матрицы системы (13) возмущения, прибавив к ней строку $\varepsilon \cdot \left(\dfrac{\partial F_j}{\partial x_1}, \ldots, \dfrac{\partial F_j}{\partial x_{n+1}}\right)$, где ε – заданная малая величина, квадратом которой можно пренебречь по сравнению с меньшими степенями. Тогда возмущенная система продолжения решения запишется в виде

$$\begin{pmatrix} \alpha_1 & \cdots & \alpha_{n+1} \\ \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_1}{\partial x_{n+1}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_j}{\partial x_1} + \varepsilon \cdot \dfrac{\partial F_j}{\partial x_1} & \cdots & \dfrac{\partial F_j}{\partial x_{n+1}} + \varepsilon \cdot \dfrac{\partial F_j}{\partial x_{n+1}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} & \cdots & \dfrac{\partial F_n}{\partial x_{n+1}} \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} \\ \dfrac{dy_2}{d\mu} \\ \vdots \\ \dfrac{dy_{j+1}}{d\mu} \\ \vdots \\ \dfrac{dy_{n+1}}{d\mu} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{45}$$

Докажем вторую теорему.

Теорема 2. *Компоненты решения системы продолжения решения* (13) *при наложении малых однородных возмущений на любую строку ее матрицы системы, отличную от первой, не изменяются.*

Доказательство теоремы 2. По аналогии с доказательством теоремы 1, дадим решение для возмущенной системы (45):

$$\frac{dy_i}{d\mu} = (-1)^{i+1} \frac{\Delta_{i\varepsilon}}{\Delta_\varepsilon}, \quad i = 1, \ldots, n+1. \tag{46}$$

Определитель Δ_ε может быть непосредственно вычислен вынесением общего множителя из $j+1$-й строки:

$$\Delta_\varepsilon = (1+\varepsilon) \cdot \Delta, \tag{47}$$

а для определителей $\Delta_{i\varepsilon}$ будут справедливы равенства

$$\Delta_{i\varepsilon} = (1+\varepsilon) \cdot \Delta_i, \quad i = 1, \ldots, n+1. \tag{48}$$

Подставляя равенства (47), (48) в формулы (46), получим

$$\frac{dy_i}{d\mu} = (-1)^{i+1}\frac{\Delta_{i\varepsilon}}{\Delta_\varepsilon} = (-1)^{i+1}\frac{(1+\varepsilon)\cdot\Delta_i}{(1+\varepsilon)\cdot\Delta} = (-1)^{i+1}\frac{\Delta_i}{\Delta} = \frac{dx_i}{d\mu}, \quad i = 1,\ldots,n+1.$$

Таким образом, компоненты решения возмущенной системы (45) совпадают с компонентами решения исходной системы (13). При этом погрешности компонент решений $\delta_i \equiv 0,\ i = 1,\ldots,n+1$, и квадратичная погрешность $\delta \equiv 0$.

Доказательство теоремы 2 завершено.

Придадим j-му столбцу матрицы системы (13) возмущения, прибавив к нему столбец $\varepsilon\left(\alpha_j, \dfrac{\partial F_1}{\partial x_j}\ldots, \dfrac{\partial F_{n+1}}{\partial x_j}\right)^\top$, где ε – заданная малая величина, квадратом которой можно пренебречь по сравнению с меньшими степенями. Тогда возмущенная система продолжения решения запишется в виде

$$\begin{pmatrix} \alpha_1 & \cdots & \alpha_j + \varepsilon\cdot\alpha_j & \cdots & \alpha_{n+1} \\ \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_1}{\partial x_j} + \varepsilon\cdot\dfrac{\partial F_1}{\partial x_j} & \cdots & \dfrac{\partial F_1}{\partial x_{n+1}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} & \cdots & \dfrac{\partial F_n}{\partial x_j} + \varepsilon\cdot\dfrac{\partial F_n}{\partial x_j} & \cdots & \dfrac{\partial F_n}{\partial x_{n+1}} \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} \\ \dfrac{dy_2}{d\mu} \\ \vdots \\ \dfrac{dy_{n+1}}{d\mu} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{49}$$

Докажем третью теорему.

Теорема 3. *Квадратичная погрешность решения возмущенной системы продолжения решения* (49)*, полученной из системы продолжения решения* (13) *наложением малых однородных возмущений на любой j-й столбец ее матрицы системы при условии $\Delta_j \neq 0$, достигает наименьшего значения в том случае, когда вектор $\overline{\alpha} = (\alpha_1,\ldots,\alpha_{n+1})^\top$ направлен по касательной к кривой множества решений системы* (1) *в рассматриваемой точке, т. е. компоненты вектора $\overline{\alpha}$ задаются в виде* (14).

Доказательство теоремы 3. По аналогии с формулами (31), (32), найдем решение и для возмущенной системы (49) в виде (36). Определитель Δ_ε будет совпадать с выражением (37), а для миноров возмущенной матрицы системы (49) будут справедливы равенства

$$\Delta_{i\varepsilon} = (1+\varepsilon)\cdot\Delta_i, \quad i = 1,\ldots,n+1, \quad i \neq j, \quad \Delta_{j\varepsilon} = \Delta_j. \tag{50}$$

Подставляя равенства (37) и (50) в формулы (36), получим

$$\frac{dy_i}{d\mu} = (-1)^{i+1}\frac{\Delta_{i\varepsilon}}{\Delta_\varepsilon} = (-1)^{i+1}\frac{\Delta_i}{\Delta} = \frac{dx_i}{d\mu}, \quad i = 1,\ldots,n+1, \quad i \neq j,$$
$$\frac{dy_j}{d\mu} = (-1)^{j+1}\frac{\Delta_{j\varepsilon}}{\Delta_\varepsilon} = (-1)^{i+1}\frac{\Delta_j}{(1+\varepsilon)\cdot\Delta}.$$

Как и в теореме 1, воспользовавшись малостью возмущений, получим

$$\frac{dy_j}{d\mu} = (-1)^{j+1}\frac{(1-\varepsilon)\cdot\Delta_j}{\Delta}. \tag{51}$$

Учитывая равенства $\dfrac{dy_i}{d\mu} = \dfrac{dx_i}{d\mu}$ при $i \neq j$ и (51), вычислим погрешности компонент решения (33):

$$\delta_i = 0, \quad i = 1,\ldots,n+1, \quad i \neq j, \quad \delta_j = -(-1)^{j+1}\frac{\varepsilon\cdot\Delta_j}{\Delta}.$$

Тогда, согласно формуле (34), квадратичная погрешность будет иметь вид

$$\delta = \frac{\varepsilon^2\cdot\Delta_j^2}{\Delta^2}. \tag{52}$$

Поставим задачу минимизации квадратичной ошибки (52) при ограничении типа равенств (12):

$$\frac{\varepsilon^2\Delta_j^2}{\Delta^2} \to \min_{\overline{\alpha}\in\mathbb{R}^{n+1}}, \quad \alpha_i\cdot\alpha_i = 1, \quad i = 1,\ldots,n+1. \tag{53}$$

Полученная задача поиска минимума при ограничении типа равенств (53) удовлетворяет условиям леммы 2 при $A = \varepsilon^2\Delta_j^2 > 0$ и $m = 2$. Это означает, что эта задача имеет точку минимума с компонентами (17), совпадающими с компонентами касательного вектора к кривой множества решений в рассматриваемой точке, т.е. удовлетворяющими равенствам (44). В этой точке квадратичная погрешность (52) принимает наименьшее значение.

Доказательство теоремы 3 завершено.

Замечание 2. Условие $\Delta_j \neq 0$ в теореме 3 является существенным, поскольку в противном случае квадратичная погрешность $\delta \equiv 0$ и постановка задачи ее минимизации теряет смысл.

Можно обобщить теоремы 2 и 3 на случай, когда возмущаются не одна строка или столбец, а произвольное их количество.

Теорема 4. *Компоненты решения системы продолжения решения* (13) *при малых однородных возмущениях любых k строк (где $k = 1, \ldots, n$) ее матрицы системы, отличных от первой, не изменяются.*

Теорема 5. *Квадратичная погрешность решения возмущенной системы продолжения решения, полученной из системы продолжения решения* (13) *наложением малых однородных возмущениях на любые k столбцов (где $k = 1, \ldots, n+1$) ее матрицы системы с номерами $i_1, \ldots, i_k$ при условии $\Delta_{i_1}^2 + \cdots + \Delta_{i_k}^2 \neq 0$, достигает наименьшего значения в том случае, когда вектор $\overline{\alpha} = (\alpha_1, \ldots, \alpha_{n+1})^\top$ направлен по касательной к кривой множества решений системы* (1) *в рассматриваемой точке, т. е. компоненты вектора $\overline{\alpha}$ задаются в виде* (14).

При этом доказательства этих утверждений аналогичны приведенным.

Для полного доказательства теоремы о минимальности квадратичной погрешности, возникающей при возмущении элементов матрицы системы продолжения решения (13), при малых однородных возмущениях остается показать, что это свойство также будет справедливо и при возмущении произвольных k строк и m столбцов матрицы системы, где $k, m = 1, \ldots, n+1$.

Поскольку система продолжения решения (13) является линейной, то можно провести ее декомпозицию на случаи, рассмотренные в теоремах 1-5, что и доказывает теорему о минимальной квадратичной погрешности. Однако целесообразно дать и конструктивное доказательство. Докажем теорему для случая, когда к каждой строке и столбцу матрицы системы (13) прибавлены они же, домноженные на $\varepsilon/2$, где ε – заданная малая величина, квадратом которой можно пренебречь. Тогда возмущенная система продолжения решения запишется в виде

$$\begin{pmatrix} \alpha_1 + \varepsilon\cdot\alpha_1 & \alpha_2 + \varepsilon\cdot\alpha_2 & \cdots & \alpha_{n+1} + \varepsilon\cdot\alpha_{n+1} \\ \dfrac{\partial F_1}{\partial x_1} + \varepsilon\cdot\dfrac{\partial F_1}{\partial x_1} & \dfrac{\partial F_1}{\partial x_2} + \varepsilon\cdot\dfrac{\partial F_1}{\partial x_2} & \cdots & \dfrac{\partial F_1}{\partial x_{n+1}} + \varepsilon\cdot\dfrac{\partial F_1}{\partial x_{n+1}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial F_n}{\partial x_1} + \varepsilon\cdot\dfrac{\partial F_n}{\partial x_1} & \dfrac{\partial F_n}{\partial x_2} + \varepsilon\cdot\dfrac{\partial F_n}{\partial x_2} & \cdots & \dfrac{\partial F_n}{\partial x_{n+1}} + \varepsilon\cdot\dfrac{\partial F_n}{\partial x_{n+1}} \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} \\ \dfrac{dy_2}{d\mu} \\ \vdots \\ \dfrac{dy_{n+1}}{d\mu} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{54}$$

В других случаях доказательство не будет изменяться, за исключением вида квадратичной ошибки для конкретного рассматриваемого случая.

Докажем общую теорему.

Теорема 6. *Квадратичная погрешность решения возмущенной системы продолжения решения, полученной из системы продолжения решения* (13) *наложением малых однородных возмущений на произвольные k строк и m столбцов (где $k, m = 1, \ldots, n+1$) матрицы системы, достигает наименьшего значения в том случае, когда вектор $\overline{\alpha} = (\alpha_1, \ldots, \alpha_{n+1})^\top$ направлен по касательной к кривой множества решений системы* (1) *в рассматриваемой точке, т. е. компоненты вектора $\overline{\alpha}$ задаются в виде* (14).

Доказательство теоремы 6. По аналогии с формулами (31), (32), найдем решение и для возмущенной системы (54) в виде (36). Определитель Δ_ε будет вычисляться по формуле

$$\Delta_\varepsilon = (1+\varepsilon)^{n+1}\cdot\Delta, \tag{55}$$

а для миноров возмущенной матрицы системы (54) будут справедливы равенства

$$\Delta_{i\varepsilon} = (1+\varepsilon)^n\cdot\Delta_i, \quad i = 1, \ldots, n+1. \tag{56}$$

Учитывая малость возмущений, можно упростить выражения (55), (56):

$$\Delta_\varepsilon = [1 + (n+1)\cdot\varepsilon]\cdot\Delta, \quad \Delta_{i\varepsilon} = (1 + n\cdot\varepsilon)\cdot\Delta_i, \quad i = 1, \ldots, n+1. \tag{57}$$

Подставляя равенства (57) в формулы (36), получим

$$\frac{dy_i}{d\mu} = (-1)^{i+1}\frac{\Delta_{i\varepsilon}}{\Delta_\varepsilon} = (-1)^{i+1}\frac{(1+n\cdot\varepsilon)\cdot\Delta_i}{[1+(n+1)\cdot\varepsilon]\cdot\Delta}, \quad i=1,\ldots,n+1.$$

Используя соотношения (33), вычислим

$$\delta_i = \frac{dy_i}{d\mu} - \frac{dx_i}{d\mu} = (-1)^{i+1}\frac{(1+n\cdot\varepsilon)\cdot\Delta_i}{[1+(n+1)\cdot\varepsilon]\cdot\Delta} - (-1)^{i+1}\frac{\Delta_i}{\Delta}, \quad i=1,\ldots,n+1.$$

Приведя дроби к общему знаменателю, домножая числитель и знаменатель полученной дроби на $1-(n+1)\cdot\varepsilon$ и отбрасывая все слагаемые, содержащие ε^2, получим $\delta_i, i=1,\ldots,n+1$, вида (41). Тогда, согласно формуле (34), квадратичная погрешность будет иметь вид (42), аналогичный полученному в теореме 1. Задача ее минимизации при ограничении типа равенств (12) будет совпадать с задачей (43). Она имеет точку минимума с компонентами (17), совпадающими с компонентами касательного вектора к кривой множества решений в рассматриваемой точке, т. е. удовлетворяющими равенствам (44). В этой точке квадратичная погрешность (42) принимает наименьшее значение.

Доказательство теоремы 6 завершено.

5. ЧИСЛЕННЫЙ ПРИМЕР

В качестве примера, иллюстрирующего полученные теоретические результаты, рассмотрим задачу построения лемнискаты Бернулли, задаваемой в декартовой системе координат уравнением

$$F(x_1,x_2) = \left(x_1^2+x_2^2\right)^2 - 2a^2\left(x_1^2-x_2^2\right) = 0, \tag{58}$$

где a – заданный вещественный параметр.

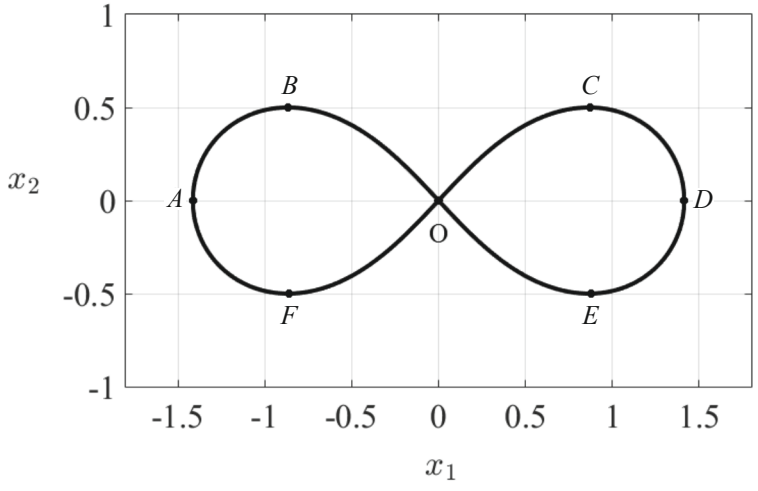
Для построения графика лемнискаты Бернулли удобнее воспользоваться ее уравнением в полярных координатах

$$r^2 = 2a^2\cos 2\varphi,$$

используя которое можно дать выражения для координат x_1 и x_2:

$$x_1 = r\cos\varphi = \pm\sqrt{2}a\sqrt{\cos 2\varphi}\cos\varphi,$$
$$x_2 = r\sin\varphi = \pm\sqrt{2}a\sqrt{\cos 2\varphi}\sin\varphi.$$

Здесь угол $\varphi\in[-\pi/4;\pi/4]$. Положительному знаку соответствует часть лемнискаты, расположенная в правой полуплоскости, а отрицательному – в левой. При $a=1$ график лемнискаты Бернулли изображен на фиг. 1.



Фиг. 1. Лемниската Бернулли.

Если рассматривать построение лемнискаты Бернулли как задачу численного решения уравнения (58), зависящего от двух неизвестных, то процесс решения при помощи алгоритмов, описанных в разд. 2, будет сопряжен со значительными трудностями. Это связано с наличием на лемнискате двух предельных особых точек

для параметра продолжения x_1 (точки A и D), четырех предельных особых точек для параметра продолжения x_2 (точки B, C, E и F) и одной существенно особой точки (точку ветвления), расположенной в начале координат O.

Построение лемнискаты Бернулли методом продолжения решения предполагает зависимость неизвестных уравнения (58) от параметра продолжения μ:

$$x_1 = x_1(\mu), \quad x_2 = x_2(\mu). \tag{59}$$

Параметр продолжения решения для данного уравнения задается локально, в окрестности каждой точки кривой множества решений, в форме

$$d\mu = \alpha_1 dx_1 + \alpha_2 dx_2.$$

Последнее соотношение можно переписать в виде

$$\alpha_1 \frac{dx_1}{d\mu} + \alpha_2 \frac{dx_2}{d\mu} = 1. \tag{60}$$

Если теперь продифференцировать уравнение (58) по параметру μ, учитывая зависимости (59), и дополнить его уравнением (60), то будет получена замкнутая система продолжения решения из двух дифференциальных уравнений относительно производных $\frac{dx_1}{d\mu}$ и $\frac{dx_2}{d\mu}$, которую в векторно-матричном виде можно записать в форме

$$\begin{pmatrix} \alpha_1 & \alpha_2 \\ F_1 & F_2 \end{pmatrix} \cdot \begin{pmatrix} \frac{dx_1}{d\mu} & \frac{dx_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{61}$$

где

$$F_1 = \frac{\partial F}{\partial x_1} = 2x_1\left(x_1^2 + x_2^2\right) - 4a^2 x_1, \quad F_2 = \frac{\partial F}{\partial x_2} = 2x_2\left(x_1^2 + x_2^2\right) + 4a^2 x_2.$$

Таким образом, построение кривой множества решений уравнения (58) из начальной точки $M_0(x_{10}, x_{20})$ по выбранному параметру продолжения решения (с заданными значениями α_1 и α_2), сводится к решению начальной задачи для системы уравнений (61) с начальными условиями

$$x_1(0) = x_{10},\ x_2(0) = x_{20}. \tag{62}$$

При построении лемнискаты Бернулли методом Давиденко используем два параметра продолжения: x_1, которому соответствуют значения $\alpha_1 = 1$ и $\alpha_2 = 0$; x_2, которому соответствуют значения $\alpha_1 = 0$ и $\alpha_2 = 1$. Для параметра продолжения x_1 начальная задача (61), (62) перепишется в виде

$$\frac{dx_2}{dx_1} = -\frac{F_1}{F_2}, \quad x_2(x_{10}) = x_{20}, \tag{63}$$

а для параметра продолжения x_2:

$$\frac{dx_1}{dx_2} = -\frac{F_2}{F_1}, \quad x_1(x_{20}) = x_{10}. \tag{64}$$

Обе задачи (63) и (64) решались для значения $a = 1$ численно явным методом Эйлера с постоянным шагом интегрирования $h = 10^{-4}$. В процессе решения приходилось проводить смену параметра продолжения решения четыре раза при специальном выборе начальной точки $x_{10} = 1,\ x_{20} = \sqrt{\sqrt{5} - 2}$. Из начальной точки задача (63) решалась до момента достижения параметром x_1 значения -1. Затем производилась смена параметра продолжения на x_2 и из последней полученной точки решалась задача (64) до достижения параметром x_2 значения $-\sqrt{\sqrt{5} - 2}$. После этого производился возврат к параметру x_1 и задача (63) решалась из последней полученной точки до достижения параметром x_1 значения 1. Последняя смена параметра производилась на x_2, после этого задача (64) решалась до возврата в начальную точку.

График лемнискаты, полученный методом Давиденко при $a = 1$, изображен на фиг. 2а.

Для оценки погрешности вычислений в точке будем использовать абсолютную погрешность

$$\epsilon(x_1, x_2) = \left| \left(x_1^2 + x_2^2\right)^2 - 2a^2\left(x_1^2 - x_2^2\right) \right|. \tag{65}$$

В табл. 1 даны значения погрешностей решения уравнения (58), полученного методом Давиденко, где ϵ_{av} – среднее значение погрешности в точке, ϵ_{med} – медиана значений погрешности, ϵ_{std} – среднее квадратическое отклонение значений погрешности от среднего значения.

Таблица 1. Погрешность ϵ вида (65), возникающая при численном построении лемнискаты Бернулли

	Метод Давиденко			Наилучшая парамеризация		
Невозмущенная система продолжения решения						
	ϵ_{av}	ϵ_{med}	ϵ_{std}	ϵ_{av}	ϵ_{med}	ϵ_{std}
	0.0011	0.0014	$8.0898 \cdot 10^{-4}$	$8.4093 \cdot 10^{-4}$	0.0010	$5.7732 \cdot 10^{-4}$
Возмущение первой строки матрицы системы продолжения решения						
$\varepsilon = 0.01$	0.0011	0.0014	$8.0125 \cdot 10^{-4}$	$8.3265 \cdot 10^{-4}$	$9.9890 \cdot 10^{-4}$	$5.7163 \cdot 10^{-4}$
$\varepsilon = 0.05$	0.0011	0.0014	$7.7066 \cdot 10^{-4}$	$8.0108 \cdot 10^{-4}$	$9.6099 \cdot 10^{-4}$	$5.4994 \cdot 10^{-4}$
$\varepsilon = 0.1$	0.0010	0.0013	$7.3585 \cdot 10^{-4}$	$7.6483 \cdot 10^{-4}$	$9.1747 \cdot 10^{-4}$	$5.2504 \cdot 10^{-4}$
Возмущение второго столбца матрицы системы продолжения решения						
$\varepsilon = 0.01$	0.0055	0.0049	0.0043	0.0051	0.0046	0.0040
$\varepsilon = 0.05$	0.0231	0.0200	0.0203	0.0221	0.0191	0.0192
$\varepsilon = 0.1$	0.0454	0.0382	0.0407	0.0434	0.0362	0.0388

Более эффективным способом решения уравнения (58) является применение наилучшего параметра, дифференциал которого для уравнения (58) удовлетворяет соотношению

$$(d\lambda)^2 = (dx_1)^2 + (dx_2)^2. \tag{66}$$

Наилучшему параметру соответствуют значения $\alpha_1 = \dfrac{dx_1}{d\mu}$ и $\alpha_2 = \dfrac{dx_2}{d\mu}$, входящие в систему продолжения решения (61) с заменой обозначения μ на λ.

Решение уравнения (58) с использованием наилучшего параметра при выборе положительного направления обхода кривой множества решений сводится к решению задачи Коши

$$\frac{dx_1}{d\lambda} = -\frac{F_2}{\sqrt{F_1^2 + F_2^2}},\ \frac{dx_2}{d\lambda} = \frac{F_1}{\sqrt{F_1^2 + F_2^2}},\quad x_1(0) = x_{10},\ x_2(0) = x_{20}. \tag{67}$$

Решение задачи (67) также проводилось для значения $a = 1$ явным методом Эйлера с постоянным шагом $l = 10^{-4}$. Можно видеть, что при решении начальной задачи (67) исчезает проблема прохождения предельных особых точек, что делает ненужным смену параметра продолжения решения. Однако остается проблема в прохождении существенно особой точки, лежащей в начале координат.

График лемнискаты, полученный методом наилучшей параметризации при $a = 1$ имеет вид, аналогичный изображенному на фиг. 2а.

В табл. 1 даны значения погрешностей решения уравнения (58), полученного методом наилучшей параметризации.

Проанализируем полученные решения. Основным неудобством при решении задач (63) и (64) является необходимость смены параметра продолжения. Но это не единственный недостаток. Из табл. 1 видно, что даже используя специально выбранные начальную точку и точки смены параметра, лежащие вне малых окрестностей предельных и существенно особых точек, погрешность решения задач (63) и (64) превосходит погрешность решения задачи (67). При специальном выборе начальной точки и точек смены параметра погрешности, конечно, различаются незначительно. Но при произвольном выборе начальной точки и момента смены параметра погрешность решения задач (63) и (64) может значительно вырасти, как в окрестности начальной точки, так и в окрестности точек смены параметра. Последнее происходит из-за трудности выявления предельных и существенно особых точек при численном счете. В прикладных задачах аналитическое решение возможно лишь в исключительных случаях, а методы оценки локальной погрешности могут давать искаженные результаты в окрестности предельных и существенно особых точек.

5.1. Возмущение первой строки матрицы продолжения решения

Рассмотрим наложение малых возмущений на первую строку матрицы системы продолжения решения. Тогда возмущенная система (61) запишется в виде

$$\begin{pmatrix} \alpha_1 + \varepsilon \cdot \alpha_1 & \alpha_2 + \varepsilon \cdot \alpha_2 \\ F_1 & F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{68}$$

где ε – малое заданное значение, квадратом которого можно пренебречь по сравнению с меньшими степенями. Для визуального разделения будем компоненты возмущенного решения обозначать y_1 и y_2.

Для трех рассмотренных ранее начальных задач (63), (64) и (67) можно получить возмущенные аналоги. Для возмущенного параметра $\mu = x_1$ система продолжения решения (68) перепишется в виде

$$\begin{pmatrix} 1 + \varepsilon & 0 \\ F_1 & F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

и сведется к начальной задаче

$$\frac{dy_1}{d\mu} = \frac{1}{1+\varepsilon}, \quad \frac{dy_2}{d\mu} = -\frac{1}{1+\varepsilon} \cdot \frac{F_1}{F_2}, \quad y_1(0) = x_{10},\ y_2(0) = x_{20}. \tag{69}$$

Для возмущенного параметра продолжения $\mu = x_2$ система продолжения решения (68) перепишется в виде

$$\begin{pmatrix} 0 & 1 + \varepsilon \\ F_1 & F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

и сведется к начальной задаче

$$\frac{dy_1}{d\mu} = -\frac{1}{1+\varepsilon} \cdot \frac{F_2}{F_1}, \quad \frac{dy_2}{d\mu} = \frac{1}{1+\varepsilon}, \quad y_1(0) = x_{10},\ y_2(0) = x_{20}. \tag{70}$$

При использовании возмущенного наилучшего параметра (66) система продолжения решения (68) перепишется в виде

$$\begin{pmatrix} (1+\varepsilon) \cdot \dfrac{dx_1}{d\lambda} & (1+\varepsilon) \cdot \dfrac{dx_2}{d\lambda} \\ F_1 & F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

и сведется к начальной задаче

$$\begin{aligned} &\frac{dy_1}{d\mu} = -\frac{1}{1+\varepsilon} \cdot \frac{F_2}{\sqrt{F_1^2 + F_2^2}}, \ \frac{dy_2}{d\mu} = \frac{1}{1+\varepsilon} \cdot \frac{F_1}{\sqrt{F_1^2 + F_2^2}}, \\ &y_1(0) = x_{10},\ y_2(0) = x_{20}. \end{aligned} \tag{71}$$

Отметим, что в функциях F_1 и F_2, входящих в начальные задачи (69), (70) и (71), переменными являются уже не x_1 и x_2, а y_1 и y_2.

Вычислим погрешности решения системы продолжения решения для случая выбора в качестве параметра продолжения возмущенной переменной $\mu = x_2$:

$$\delta_{21} = \frac{dx_1}{dx_2} - \frac{dy_1}{d\mu} = -\frac{\varepsilon}{1+\varepsilon} \cdot \frac{F_2}{F_1}, \quad \delta_{22} = \frac{dx_2}{dx_2} - \frac{dy_2}{d\mu} = \frac{\varepsilon}{1+\varepsilon}.$$

Тогда квадратичная погрешность для данного случая будет равна

$$\delta_2 = \delta_{21}^2 + \delta_{22}^2 = \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \left[1 + \left(\frac{F_2}{F_1}\right)^2\right].$$

Если выбрать в качестве параметра продолжения возмущенную переменную $\mu = x_1$, то

$$\delta_{11} = \frac{dx_1}{dx_1} - \frac{dy_1}{d\mu} = \frac{\varepsilon}{1+\varepsilon}, \quad \delta_{12} = \frac{dx_2}{dx_1} - \frac{dy_2}{d\mu} = -\frac{\varepsilon}{1+\varepsilon} \cdot \frac{F_1}{F_2}.$$

Тогда квадратичная погрешность для данного случая будет равна

$$\delta_1 = \delta_{11}^2 + \delta_{12}^2 = \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \left[1 + \left(\frac{F_1}{F_2}\right)^2\right].$$

Для возмущенного наилучшего параметра продолжения решения погрешности будут вычислены в виде

$$\delta_{\lambda 1} = \frac{dx_1}{d\lambda} - \frac{dy_1}{d\lambda} = -\frac{\varepsilon \cdot F_2}{(1+\varepsilon) \cdot \sqrt{F_1^2 + F_2^2}},$$
$$\delta_{\lambda 2} = \frac{dx_2}{d\lambda} - \frac{dy_2}{d\lambda} = -\frac{\varepsilon \cdot F_1}{(1+\varepsilon) \cdot \sqrt{F_1^2 + F_2^2}}.$$

Тогда квадратичная погрешность для данного случая будет равна

$$\delta_\lambda = \delta_{\lambda 1}^2 + \delta_{\lambda 2}^2 = \frac{\varepsilon^2}{(1+\varepsilon)^2}. \tag{72}$$

Используя оценки сверху для квадратичной погрешности (72)

$$\frac{\varepsilon^2}{(1+\varepsilon)^2} \le \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \left[1 + \left(\frac{F_2}{F_1}\right)^2\right], \quad \frac{\varepsilon^2}{(1+\varepsilon)^2} \le \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \left[1 + \left(\frac{F_1}{F_2}\right)^2\right],$$

можно получить неравенства

$$\delta_\lambda \le \delta_1, \quad \delta_\lambda \le \delta_2,$$

т. е. квадратичная погрешность возмущенной системы продолжения решения (68) при выборе наилучшего параметра не превосходит значений, полученных с использованием параметров продолжения x_1 и x_2. Согласно полученным теоретическим результатам, эти неравенства будут справедливы и для других параметров продолжения решения, отличных от x_1 и x_2.

Эти результаты позволяют утверждать, что и для уравнения (58) воздействие возмущений (вычислительных погрешностей) будет наименьшим при использовании для его решения наилучшего параметра. Продемонстрируем это на примере численного решения возмущенных задач.

Возмущенные задачи (69), (70) и (71) решались для значения $a = 1$ явным методом Эйлера с постоянным шагом интегрирования, таким же как и для невозмущенных задач. Параметр возмущения $\varepsilon = 0.01; 0.05; 0.1$.

Графики лемнискаты для возмущенных задач при $a = 1$ и различных значениях ε, полученные методами Давиденко и наилучшей параметризации, имеют вид аналогичный изображенному на фиг. 1. Графики решений возмущенных задач близки к графику точного решения, за исключением существенной особой точки, лежащей в начале координат. В ее окрестности происходит малое отклонение от точного решения.

В табл. 1 даны значения погрешностей решений возмущенных задач для всех рассматриваемых случаев.

При однородных малых возмущениях первой строки матрицы системы продолжения решения, погрешность решения задач (69), (70) незначительно больше по сравнению с погрешностью решения задачи (71). Это связано, как и в невозмущенном случае, с удачным выбором начальной точки и точек смены параметра.

Отличительной особенностью этого случая возмущения является уменьшение погрешности по сравнению с невозмущенным случаем. Это можно объяснить тем, что возмущениям подвержена только первая строка матрицы системы продолжения решения, отвечающая за задание параметра μ. Поэтому возмущенный параметр продолжения решения μ можно рассматривать как удлиненный в $1 + \varepsilon$ раз параметр продолжения x_1 и x_2 для задач (69) и (70) соответственно, и удлиненный в $\sqrt{1+\varepsilon}$ раз наилучший параметр для задачи (71). Поскольку происходит удлинение параметра продолжения, то участки кривой множества решений большой кривизны удается преодолеть с меньшей погрешностью. Этим и объясняется уменьшение погрешности решения для данного класса возмущенных задач.

5.2. Возмущение столбца матрицы продолжения решения

Рассмотрим наложение малых возмущений на второй столбец матрицы системы продолжения решения. Тогда возмущенная система (61) запишется в виде

$$\begin{pmatrix} \alpha_1 & \alpha_2 + \varepsilon \cdot \alpha_2 \\ F_1 & F_2 + \varepsilon \cdot F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^\top = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{73}$$

Как и в п. 5.1, для трех рассмотренных ранее начальных задач (63), (64) и (67) получим возмущенные аналоги. Для параметра продолжения $\mu = x_1$ система продолжения решения (73) перепишется в виде

$$\begin{pmatrix} 1 & 0 \\ F_1 & F_2 + \varepsilon \cdot F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

и сведется к начальной задаче

$$\frac{dy_1}{d\mu} = 1, \quad \frac{dy_2}{d\mu} = -\frac{1}{1+\varepsilon} \cdot \frac{F_1}{F_2}, \quad y_1(0) = x_{10},\ y_2(0) = x_{20}, \tag{74}$$

эквивалентной начальной задаче

$$\frac{dy_2}{dx_1} = -\frac{1}{1+\varepsilon} \cdot \frac{F_1}{F_2}, \quad y_2(x_{10}) = x_{20}.$$

Для возмущенного параметра продолжения $\mu = x_2$ система продолжения решения (73) перепишется в виде

$$\begin{pmatrix} 0 & 1+\varepsilon \\ F_1 & F_2 + \varepsilon \cdot F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

и сведется к начальной задаче

$$\frac{dy_1}{d\mu} = -\frac{F_2}{F_1}, \quad \frac{dy_2}{d\mu} = \frac{1}{1+\varepsilon}, \quad y_1(0) = x_{10},\ y_2(0) = x_{20}. \tag{75}$$

При использовании возмущенного наилучшего параметра (66) система продолжения решения (73) перепишется в виде

$$\begin{pmatrix} \dfrac{dx_1}{d\lambda} & (1+\varepsilon) \cdot \dfrac{dx_2}{d\lambda} \\ F_1 & (1+\varepsilon) \cdot F_2 \end{pmatrix} \cdot \begin{pmatrix} \dfrac{dy_1}{d\mu} & \dfrac{dy_2}{d\mu} \end{pmatrix}^{\top} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

и сведется к начальной задаче

$$\frac{dy_1}{d\mu} = -\frac{F_2}{\sqrt{F_1^2 + F_2^2}}, \ \frac{dy_2}{d\mu} = \frac{F_1}{(1+\varepsilon) \cdot \sqrt{F_1^2 + F_2^2}}, \quad y_1(0) = x_{10},\ y_2(0) = x_{20}. \tag{76}$$

Вычислим погрешности решения системы продолжения решения для случая выбора в качестве параметра продолжения возмущенной переменной $\mu = x_2$:

$$\delta_{21} = \frac{dx_1}{dx_2} - \frac{dy_1}{d\mu} = 0, \quad \delta_{22} = \frac{dx_2}{dx_2} - \frac{dy_2}{d\mu} = \frac{\varepsilon}{1+\varepsilon}.$$

Тогда квадратичная погрешность для данного случая будет равна

$$\delta_2 = \delta_{21}^2 + \delta_{22}^2 = \frac{\varepsilon^2}{(1+\varepsilon)^2}.$$

Если выбрать в качестве параметра продолжения переменную $\mu = x_1$, то

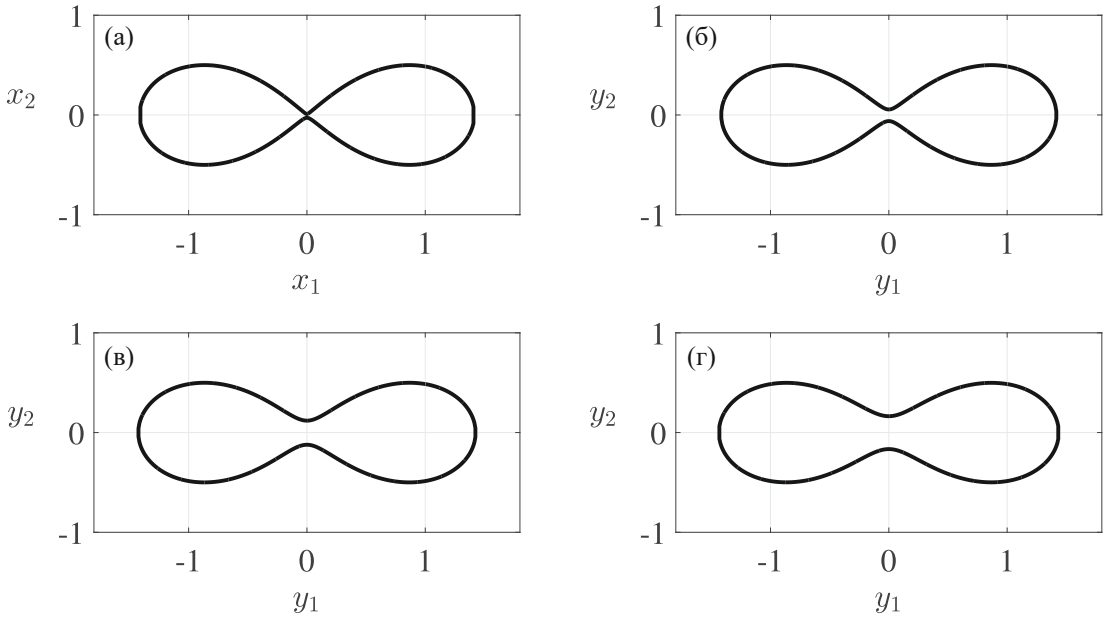
$$\delta_{11} = \frac{dx_1}{dx_1} - \frac{dy_1}{dy_1} = 0, \quad \delta_{12} = \frac{dx_2}{dx_1} - \frac{dy_2}{dy_1} = -\frac{\varepsilon}{1+\varepsilon} \cdot \frac{F_1}{F_2}.$$

Тогда квадратичная погрешность для данного случая будет равна

$$\delta_1 = \delta_{11}^2 + \delta_{12}^2 = \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \left(\frac{F_1}{F_2}\right)^2.$$

Для наилучшего параметра продолжения решения погрешности будут вычислены в виде

$$\delta_{\lambda 1} = \frac{dx_1}{d\lambda} - \frac{dy_1}{d\mu} = 0, \quad \delta_{\lambda 2} = \frac{dx_2}{d\lambda} - \frac{dy_2}{d\mu} = \frac{\varepsilon \cdot F_1}{(1+\varepsilon) \cdot \sqrt{F_1^2 + F_2^2}}.$$



Фиг. 2. Лемниската Бернулли, метод Давиденко, система продолжения решения с возмущенным вторым столбцом: (а) — для $\varepsilon = 0$, (б) — для $\varepsilon = 0.01$, (в) — для $\varepsilon = 0.05$, (г) — для $\varepsilon = 0.1$.

Тогда квадратичная погрешность для данного случая будет равна

$$\delta_\lambda = \delta_{\lambda 1}^2 + \delta_{\lambda 2}^2 = \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \frac{F_1^2}{F_1^2 + F_2^2}. \tag{77}$$

Используя оценки сверху для квадратичной погрешности (77)

$$\frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \frac{F_1^2}{F_1^2 + F_2^2} \le \frac{\varepsilon^2}{(1+\varepsilon)^2}, \quad \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \frac{F_1^2}{F_1^2 + F_2^2} \le \frac{\varepsilon^2}{(1+\varepsilon)^2} \cdot \frac{F_1^2}{F_2^2},$$

можно получить неравенства

$$\delta_\lambda \le \delta_1, \quad \delta_\lambda \le \delta_2.$$

Как и в предыдущем рассмотренном случае, квадратичная погрешность решения возмущенной системы продолжения решения (73) меньше при выборе наилучшего параметра, по сравнению с использованием параметров продолжения x_1 и x_2. Согласно полученным теоретическим результатам, эти неравенства будут справедливы и для других параметров продолжения решения, отличных от x_1 и x_2.

Возмущенные задачи (75), (74) и (76) решались для значения $a = 1$ явным методом Эйлера с постоянным шагом интегрирования, таким же как и для невозмущенных задач. Параметр возмущения $\varepsilon = 0.01; 0.05; 0.1$.

Графики лемнискаты, полученные для возмущенных задач при $a = 1$ для различных значений ε, изображены на фиг. 2б–2г. Так как погрешности решений задач (75), (74) и (76) мало отличаются, то на фиг. 2 приводится только графики, полученные методом Давиденко. Кривые множества решений, полученные с использованием наилучшего параметра имеют аналогичный вид.

В табл. 1 даны значения погрешностей решений возмущенных задач для всех рассматриваемых случаев. Все полученные расчетные данные полностью согласуются с теоретическими результатами.

6. ЗАКЛЮЧЕНИЕ

В данной работе исследуется одно из свойств наилучшего параметра, которое имеет важное значение в прикладных расчетах. Это свойство минимальности квадратичной погрешности, возникающей при возмущении элементов матрицы системы продолжения решения (13).

Целью данной работы было строгое доказательство минимальности квадратичной погрешности решения системы продолжения решения (13) с возмущенной матрицей системы при использовании наилучшего параметра. Это удалось доказать для случая малых однородных возмущений, значения которых одинаковы, а их квадратами можно пренебречь.

Все теоретические результаты были полностью подтверждены на примере численного построения лемнискаты Бернулли (численного решения нелинейного уравнения с двумя неизвестными). Построение лемнискаты Бернулли показывает, что при использовании в качестве параметров продолжения решения x_1 и x_2 даже в случае удачного выбора начальной точки и точек смены параметра (которые не должны попадать в окрестность предельных и существенно особых точек), полученная квадратичная погрешность превосходит значение, полученное при использовании наилучшего параметра. Это оказывается справедливым даже при наличии у лемнискаты существенной особой точки в начале координат, хотя ее наличия и не предполагалось при доказательстве утверждений статьи. При прохождении существенно особой точки эффективность использования наилучшего параметра падает. Более эффективные способы прохождения существенно особых точек с использованием наилучшего параметра рассмотрены в работах [9, 10].

При доказательстве теорем в статье предполагалось, что возмущения накладываются на элемент матрицы системы продолжения решения пропорционально его значению (с коэффициентом пропорциональности ε), т. е. возмущения являются зависимыми, что соответствует возникновению вычислительной погрешности при использовании приближенных методов решения. Таким образом, преобразование задачи к наилучшему параметру позволяет минимизировать воздействие на решение зависимых возмущений, в том числе и погрешность вычислений.

СПИСОК ЛИТЕРАТУРЫ

1. *Lahaye M. E.* Une metode de resolution d'une categorie d'equations transcendentes // Comptes Rendus hebdomadaires des seances de L'Academie des sciences. 1934. Vol. 198. No. 21. P. 1840–1842.
2. *Lahaye M. E.* Solution of system of transcendental equations // Acad. Roy. Belg. Bull. Cl. Sci. 1948. Vol. 5. P. 805–822.
3. *Давиденко Д. Ф.* Об одном новом методе численного решения систем нелинейных уравнений // Докл. АН СССР. 1953 Т. 88. № 4. С. 601–602.
4. *Давиденко Д. Ф.* О приближенном решении систем нелинейных уравнений // Украинский матем. ж. 1953 Т. 5. № 2. С. 196–206.
5. *Ворович И. И., Зипалова В. Ф.* К решению нелинейных краевых задач теории упругости методом перехода к задаче Коши // Прикл. матем. и механ. 1965. Т. 29. Вып. 5. С. 894–901.
6. *Рикс Э.* Применение метода Ньютона к задаче упругой устойчивости // Прикл. механ. 1972. № 5. С. 204–210.
7. *Кузнецов Е. Б., Шалашилин В. И.* Задача Коши как задача продолжения по наилучшему параметру // Дифференц. ур-ния. 1994. Т. 30. № 6. С. 964–971.
8. *Шалашилин В. И., Кузнецов Е. Б.* Метод продолжения решения по параметру и наилучшая параметризация в прикладной математике и механике. М.: Эдиториал УРСС, 1999.
9. *Красников С.Д., Кузнецов Е.Б.* Численное продолжение решения в особых точках коразмерности единица // Ж. вычисл. матем. и матем. физ. 2015. Т. 55. № 11. С. 1835–1856.
10. *Красников С.Д., Кузнецов Е.Б.* Численное продолжение решения в особых точках высокой коразмерности для систем нелинейных алгебраических или трансцендентных уравнений // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 9. С. 1571–1585.

ON THE MINIMALITY OF SQUARED ERROR OF SOLUTIONS TO SYSTEMS OF EQUATIONS TRANSFORMED TO THE BEST PARAMETER UNDER SMALL HOMOGENEOUS PERTURBATIONS

E. B. Kuznetsov[a,*], **S. S. Leonov**[a,b,**]

[a] *125993 Moscow, Volokolamsk Highway 4, Moscow Aviation Institute (National Research University), Russia*
[b] *117198 6 Miklukho-Maklay str., Patrice Lumumba Peoples' Friendship University of Russia, Moscow, Russia*
**e-mail: kuznetsov@mai.ru*
***e-mail: powerandglory@yandex.ru*

Received: 05.05.2024
Revised: 05.08.2024
Accepted: 23.08.2024

Abstract. Solving of systems of nonlinear equations with a scalar parameter is studied. The set of solutions to such systems is a curve in the space of variables of the equation system and the parameter. Its construction is usually carried out using numerical methods and is associated with numerous difficulties arising due to the presence of limiting and essentially singular points on the curve of the set of solutions. To find such curves, the method of solution continuation with respect to a parameter and the best parameterization is used, which allows us to reduce the solution to the Cauchy problem for a system of differential equations of solution continuation. Stability of the solution to perturbations introduced into the continuation system is investigated. For the first time, the previously formulated proposition about the minimality of the squared error of the solution to the continuation system under homogeneous small perturbations of its matrix is completely proved. The theoretical results are illustrated by the example of the numerical construction of Bernoulli's lemniscate.

Keywords: systems of nonlinear equations, solution continuation with respect to a parameter, best parameterization, system of solution continuation, small perturbations, squared error.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2355–2370

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.9

МЕТОД УСРЕДНЕНИЯ В ЗАДАЧЕ ПОСТРОЕНИЯ АВТОКОЛЕБАТЕЛЬНЫХ РЕШЕНИЙ РАСПРЕДЕЛЕННЫХ КИНЕТИЧЕСКИХ СИСТЕМ[1)]

© 2024 г. Е. П. Кубышкин[1,*]

[1] 150000 Ярославль, ул. Советская, 14, ЯрГУ им. П.Г. Демидова, матем. ф-т, Россия
**e-mail: kubysh.e@yandex.ru*

Поступила в редакцию 28.11.2023 г.
Переработанный вариант 29.07.2024 г.
Принята к публикации 23.08.2024 г.

Построен метод усреднения для двухкомпонентных распределенных кинетических систем с малой диффузией в ограниченной одномерной области с условиями непроницаемости на границе. Построены преобразования рассматриваемой распределенной системы, позволяющие выделить одну “быструю” и счетное число “медленных” переменных. Доказаны теоремы о соответствии стационарных и периодических решений, а также инвариантных торов усредненных уравнений “медленных” переменных соответственно пространственно неоднородным периодическим решениям и инвариантным торам исходных уравнений аналогичного характера устойчивости. Предложены алгоритмы построения периодических решений (циклов) и инвариантных торов исходных уравнений в виде разложения по степеням малого параметра, обеспечивающих построение асимптотических формул указанных автоколебательных объектов. Сформулированы условия сходимости соответствующих разложений. Библ. 20.

Ключевые слова: метод усреднения, распределенные кинетические системы, системы уравнений реакция–диффузия, пространственно неоднородные решения, теория бифуркаций.

DOI: 10.31857/S0044466924120094, **EDN**: KBTCFG

1. ВВЕДЕНИЕ

Распределенные кинетические системы (системы уравнений реакция–диффузия) находят широкое применение в химической кинетике, синергетике, астрофизике, биологии, нелинейной оптике, теории фазовых переходов и многих других областях естествознания. Исследованию таких систем посвящена обширная литература. Основной вопрос, который интересует авторов – это изучение условий и механизмов возникновения пространственно неоднородных решений и построение для них аналитических и асимптотических формул. Интересны как стационарные решения, особенно имеющие узкие области быстрого изменения и получившие название контрастные структуры, так и изменяющиеся по времени – движущиеся фронты, периодические, квазипериодические и хаотические решения. Важным вопросом является задача исследования устойчивости таких решений. Отметим кратко некоторые работы. Теории контрастных структур посвящены работы [1]–[4], в которых сформулированы условия существования контрастных структур и разработана совокупность методов асимптотического построения таких решений для различных классов обыкновенных уравнений и уравнений в частных производных. В работах имеется обширная библиография по указанной проблематике. Начало исследованию условий и механизмов возникновения автоколебательных пространственно неоднородных решений в системах с малой диффузией было положено в работе [5]. В дальнейшем идеи этой работы получили развитие в работах [6]–[10]. Основная идея этих работ связана с построением специальных начально-краевых задач для уравнений в частных производных, не содержащих малых параметров, автоколебательные решения которых в определенном смысле описывают автоколебательные решения исходной задачи. Такие начально-краевые задачи получили название квазинормальных форм, а соответствующий подход – метод квазинормальных форм. Этот подход подробно изложен в монографии [11]. Иной подход к построению пространственно неоднородных

[1)] Работа выполнена в рамках реализации программы развития регионального научно-образовательного математического центра (ЯрГУ) при финансовой поддержке Минобрнауки РФ (Соглашение о предоставлении из федерального бюджета субсидии № 075-02-2024-1442).

периодических решений предложен в работе [12]. Мощным аппаратом исследования колебательных решений нелинейных обыкновенных дифференциальных уравнений является метод усреднения Крылова–Боголюбова (см. [13]). Этот метод получил развитие для других классов уравнений, в том числе, и для некоторых классов уравнений в частных производных. Метод позволяет строить колебательные решения, бифурцирующие из стационарных решений при изменении параметров уравнений. При этом удается в явном виде получать аналитические или асимптотические формулы решений. Основная сложность этого метода – выделение "быстрых" и "медленных" переменных. Это часто связано с физической постановкой задачи, математической моделью которой являются исследуемые уравнения. Другой сложностью метода усреднения является доказательство теорем соответствия между стационарными решениями усредненных уравнений медленных переменных и исследуемых уравнений. Эта проблема для обыкновенных дифференциальных уравнений в полном объеме была решена в монографии [14].

В настоящей работе построен метод усреднения для двухкомпонентных распределенных кинетических систем с малой диффузией в ограниченной одномерной области с условиями непроницаемости на границе. Построены преобразования, позволяющие выделить одну "быструю" и счетное число "медленных" переменных. Доказаны теоремы о соответствии стационарных и периодических решений, а также инвариантных торов усредненных уравнений "медленных" переменных соответственно пространственно неоднородным периодическим решениям и инвариантным торам исходных уравнений аналогичного характера устойчивости. Предложены алгоритмы построения периодических решений (циклов) и инвариантных торов исходных уравнений в виде разложения по степеням малого параметра, обеспечивающих построение асимптотических формул указанных автоколебательных объектов. Сформулированы условия сходимости соответствующих разложений. Предложенный метод позволяет исследовать автоколебательные решения, бифуцирующие при изменении параметров из однородных состояний равновесия распределенных кинетических систем с малой диффузией. Используемые при построении метода усреднения идеи были успешно апробированы при исследовании автоколебательных решений дифференциально-разностных уравнений с малым параметром при производной (см. [15]–[17]).

2. МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ

Рассматривается начально-краевая задача

$$u_t = \varepsilon D u_{xx} + A(\varepsilon)u + F(u;\varepsilon), \tag{2.1}$$

$$u_x(0,t) = u_x(1,t) = 0, \;\; u(x,0) = u_0(x), \tag{2.2}$$

где $u(x,t) = (u_1(x,t), u_2(x,t))$, $0 \leqslant x \leqslant 0$, $t \geqslant 0$, $0 < \varepsilon \leqslant \varepsilon_0$ – малый параметр, $D = \operatorname{diag}\{d_1, d_2\}$, матрица второго порядка $A(\varepsilon)$ гладко зависит от ε при $|\varepsilon| < \varepsilon_0$ и имеет собственные значения вида

$$\lambda(\varepsilon) = \gamma(\varepsilon) \pm i\sigma(\varepsilon),\; i = \sqrt{-1},\; \gamma(0) = 0,\; \gamma'(0) = \gamma_1 > 0,\; \sigma(0) = \sigma_0 > 0, \tag{2.3}$$

нелинейная вектор-функция $F(u;\varepsilon), ||F(u;\varepsilon)|| = o(||u||)$ гладко зависит от u и ε при $||u|| < R_0$, $|\varepsilon| < \varepsilon_0$ (принадлежит классу $C^k, k \geqslant 5$). Здесь $||.||$ норма в $\mathbb{R}^2$.

В дальнейшем $H = L_2(0,1)$ – пространство комплекснозначных определенных на $[0,1]$ функций $u(x)$, для которых $||u(x)||_H = (u(x), u(x))_H^{1/2} < \infty$, $(u(x), v(x))_H = \int_0^1 u(x)\bar{v}(x)dx$, H^2 – пространство определенных на $[0,1]$ комплекснозначных функций $u(x)$, полученное замыканием множества функций $\{u(x) : u(x) \in C^2(0,1)\bigcap C^1[0,1], u'(0) = u'(1) = 0)\}$ в норме $||u(x)||_{H^2} = (u(x), u(x))_{H^2}^{1/2} < \infty$, $(u(x), v(x))_{H^2} = (u(x), v(x))_H + (u''(x), v''(x))_H$, где $C^1[0,1]$ и $C^2(0,1)$ – пространства комплекснозначных непрерывно дифференцируемых на $[0,1]$ и дважды непрерывно дифференцируемых на $(0,1)$ функций, нормы в которых определены стандартным способом.

Начально-краевая задача (2.1), (2.2) для любой вещественнозначной начальной функции $u_0(x) = (u_{01}(x), u_{02}(x)) \in S(R_0) = \{u(x) = (u_1(x), u_2(x)) : u_1(x), u_2(x) \in H^2, (||u_1(x)||_{H^2}^2 + ||u_2(x)||_{H^2}^2)^{1/2} < R_0\}$ при $t > 0$ однозначно разрешима, пока $u(x,t) = (u_1(x,t), u_2(x,t)) \in S(R_0)$. При этом решение $u(x,t)$ непрерывно по совокупности переменных, дифференцируемо по t и при каждом t $u_1(x,t), u_2(x,t) \in H^2$ (см., например, [18]).

Ниже исследуются условия бифуркации из нулевого состояния равновесия начально-краевой задачи (2.1), (2.2) при изменении ее параметров пространственно-неоднородных периодических решений и инвариантных торов, принадлежащих $S(R_0)$, исследуется их устойчивость. Предложена методика построения асимптотических и аналитических формул для указанных автоколебательных решений.

3. НОРМАЛИЗУЮЩЕЕ ПРЕОБРАЗОВАНИЕ НАЧАЛЬНО-КРАЕВОЙ ЗАДАЧИ (2.1), (2.2)

Рассмотрим систему обыкновенных дифференциальных уравнений

$$\dot{v} = A(\varepsilon)v + F(v;\varepsilon) \tag{3.1}$$

относительно вектор-функции $v(t) = (v_1(t), v_2(t))$, в которой матрица $A(\varepsilon)$ и нелинейная вектор-функция $F(v;\varepsilon)$ определены в (2.1). Считая $||v(t)|| < R_0$, выполним в (3.1) нормализующую замену

$$v(t) = v(w,\bar{w};\varepsilon) = e(\varepsilon)w + \bar{e}(\varepsilon)\bar{w} + v_{20}(\varepsilon)w^2 + v_{11}(\varepsilon)w\bar{w} + v_{02}(\varepsilon)\bar{w}^2 +$$

$$+ v_{30}(\varepsilon)w^3 + v_{21}(\varepsilon)w^2\bar{w} + v_{21}(\varepsilon)w\bar{w}^2 + v_{03}(\varepsilon)\bar{w}^3, \;\; v_{pq}(\varepsilon) = \bar{v}_{qp}(\varepsilon), \tag{3.2}$$

приводящую (3.1) к системе дифференциальных уравнений

$$\dot{w} = \lambda(\varepsilon)w + a(\varepsilon)w^2\bar{w} + W(w,\bar{w};\varepsilon) \equiv W^*(w,\bar{w};\varepsilon), \; |W(w,\bar{w};\varepsilon)| = o(|w|^3), \tag{3.3}$$

относительно комплекснозначных функций $w(t)$ и $\bar{w}(t)$. Уравнение для $\bar{w}(t)$ является комплексно сопряженным с (3.3). В (3.2) $A(\varepsilon)e(\varepsilon) = \lambda(\varepsilon)e(\varepsilon), ||e(\varepsilon)|| = 1$, гладко зависящие от ε двумерные векторы $v_{pq}(\varepsilon)$, а также $a(\varepsilon)$ и функция $W(w,\bar{w};\varepsilon)$ подлежат определению. Здесь и в дальнейшем $||.||$ – норма в $\mathbb{C}^2$. Условие принадлежности траекторий (3.3) в силу замены (3.2) системе уравнений (3.1) определяет тождество

$$\frac{dv}{dt} = \frac{\partial v(w,\bar{w};\varepsilon)}{\partial w}W^*(w,\bar{w};\varepsilon) + \frac{\partial v(w,\bar{w};\varepsilon)}{\partial \bar{w}}\bar{W}^*(w,\bar{w};\varepsilon) \equiv A(\varepsilon)v(w,\bar{w};\varepsilon) + F(v(w,\bar{w};\varepsilon);\varepsilon), \tag{3.4}$$

которое должно выполнятся при всех w и $\bar{w}$. При первых степенях w и $\bar{w}$ тождество (3.4) выполняется в силу выбора $e(\varepsilon)$ и $\bar{e}(\varepsilon)$. Записав $F(v;\varepsilon)$ в окрестности $v = 0$ в виде $F((v;\varepsilon) = F_2(v;\varepsilon) + F_3(v;\varepsilon) + F_*(v;\varepsilon)$, где $F_2(.), F_3(.), F_*(.)$ – соответственно квадратичные, кубические слагаемые и остаточный в форме Лагранжа, приравняем в тождестве (3.4) слева и справа коэффициенты при одинаковых квадратичных и кубических степенях w и $\bar{w}$. В результате получим рекуррентную последовательность линейных неоднородных алгебраических уравнений для определения $v_{pq}(\varepsilon)$, которые однозначно разрешимы всегда, за исключением уравнения для определения $v_{21}(\varepsilon)$ ($v_{12}(\varepsilon) = \bar{v}_{21}(\varepsilon)$), которое имеет вид

$$B(\varepsilon)v_{21}(\varepsilon) \equiv ((2\lambda(\varepsilon) + \bar{\lambda}(\varepsilon))I - A(\varepsilon))v_{21}(\varepsilon) = -e(\varepsilon)a(\varepsilon) + f_{21}(\varepsilon) = f_{21}^*(\varepsilon), \tag{3.5}$$

где I – единичная матрица, $f_{21}(\varepsilon)$ – известный вектор. Матрица $B(0)$, как легко заметить, имеет однократное нулевое собственное значение. Обозначим через $\gamma_1(\varepsilon)$ и $\gamma_2(\varepsilon)$ $(\gamma_1(0) = 0, \gamma_2(0) = -i2\sigma_0)$ собственные значения матрицы $B(\varepsilon)$, а соответствующие собственные векторы – $e_1(\varepsilon)$ и $e_2(\varepsilon)$ $(||e_j(\varepsilon)|| = 1, j = 1,2)$. Обозначим $h(\varepsilon)$ и $h_1(\varepsilon)$, $h_2(\varepsilon)$ соответственно собственные векторы сопряженных матриц

$$A^*(\varepsilon)h(\varepsilon) = \bar{\lambda}(\varepsilon)h(\varepsilon)\;(e(\varepsilon),h(\varepsilon)) = 1, (\bar{e}(\varepsilon),h(\varepsilon)) = 0, \tag{3.6}$$

$$B^*(\varepsilon)h_j(\varepsilon) = \bar{\gamma}_j(\varepsilon)h_j(\varepsilon)\;(e_j(\varepsilon),h_j(\varepsilon)) = 1\;(j = 1,2), \;\; (e_1(\varepsilon),h_2(\varepsilon)) = 0, (e_2(\varepsilon),h_1(\varepsilon)) = 0.$$

Здесь (.,.) скалярное произведение в $\mathbb{C}^2$. Отметим, что $e(0) = e_1(0), h(0) = h_1(0)$. С учетом этого из условия разрешимости системы уравнений (3.5)

$$(f_{21}^*(\varepsilon), h_1(\varepsilon)) = (-e(\varepsilon)a(\varepsilon) + f_{21}(\varepsilon), h_1(\varepsilon)) = 0$$

однозначно определяется $a(\varepsilon)$, а из (3.5) – единственное решение, которые соответственно имеют вид

$$a(\varepsilon) = (f_{21}(\varepsilon), h_1(\varepsilon))/((e(\varepsilon), h_1(\varepsilon)), \; v_{21}(\varepsilon) = (f_{21}^*(\varepsilon), h_2(\varepsilon))/\gamma_2(\varepsilon)e_2(\varepsilon).$$

Обозначим $G(w,\bar{w};\varepsilon)$ совокупность слагаемых в правой части тождества (3.4), имеющих порядок малости по $w,\bar{w}$ (при $|w| \to 0$) выше третьего. Представим

$$G(w,\bar{w};\varepsilon) = W(w,\bar{w};\varepsilon)e(\varepsilon) + \bar{W}(w,\bar{w};\varepsilon)\bar{e}(\varepsilon), \; W(w,\bar{w};\varepsilon) = (G(w,\bar{w};\varepsilon), h(\varepsilon)).$$

Таким образом, нелинейное преобразование (3.2) и система дифференциальных уравнений (3.3) определяются однозначно.

Считая теперь преобразование (3.2) построенным, положим в нем $w = w(x,t) \in H^2$, $||w(x,t)||_{H^2} < R_0$ при каждом $t \geqslant 0$ и непрерывно диференцируемой по t в норме H^2, выполним в уравнении (2.1) замену переменных $u(x,t) = v(w(x,t), \bar{w}(x,t); \varepsilon)$. В результате получим тождество

$$u_t = \frac{\partial v(w,\bar{w};\varepsilon)}{\partial w} W^{**}(w,\bar{w},w_x,\bar{w}_x,w_{xx},\bar{w}_{xx};\varepsilon) + \frac{\partial v(w,\bar{w};\varepsilon)}{\partial \bar{w}} \bar{W}^{**}(w,\bar{w},w_x,\bar{w}_x,w_{xx},\bar{w}_{xx};\varepsilon) \equiv$$

$$\equiv A(\varepsilon)v(w,\bar{w};\varepsilon) + F(v(w,\bar{w};\varepsilon);\varepsilon)$$

относительно $w(x,t)$ и $\bar{w}(x,t)$ при $||w(x,t)||_{H^2} < R_0$, если $w(x,t), \bar{w}(x,t)$ являются решениями системы уравнений

$$w_t = \varepsilon(D(v_w(w,\bar{w};\varepsilon)w_{xx} + v_{\bar{w}}(w,\bar{w};\varepsilon)\bar{w}_{xx} + v_{ww}(w,\bar{w};\varepsilon)w_x^2 + 2v_{w\bar{w}}(w,\bar{w};\varepsilon)w_x\bar{w}_x +$$
$$+ v_{\bar{w}\bar{w}}(w,\bar{w};\varepsilon)\bar{w}_x^2), h(\varepsilon)) + \lambda(\varepsilon)w + a(\varepsilon)w^2\bar{w} + W(w,\bar{w};\varepsilon) \equiv W^{**}(w,\bar{w},w_x,\bar{w}_x,w_{xx},\bar{w}_{xx};\varepsilon), \tag{3.7}$$

в которой уравнение для $\bar{w}(x,t)$ будет комплексно сопряженным с (3.7). Фазовым пространством (3.7) является пространство $H \times H$, областью определения правой части – пространство $H^2 \times H^2$.

4. АНАЛИЗ АВТОКОЛЕБАТЕЛЬНЫХ РЕШЕНИЙ НАЧАЛЬНО-КРАЕВОЙ ЗАДАЧИ (2.1), (2.2)

Изучение возможных бифуркаций периодических решений и инвариантных торов из нулевого решения начально-краевой задачи (2.1), (2.2), соглано результатам предыдущего раздела, достаточно рассмотреть для системы уравнений (3.7). Нормируем в (3.7) $w(x,t) \to \varepsilon^{1/2}w(x,t)$ и выделим в явном виде слагаемые, перед которыми стоит множитель нулевого и первого порядков по ε, объединив все остальные в единую функцию. В результате будем иметь квазилинейное операторное уравнение

$$w_t = \lambda(\varepsilon)w + \varepsilon((De(\varepsilon),h(\varepsilon))w_{xx} + (D\bar{e}(\varepsilon),h(\varepsilon))\bar{w}_{xx} + a(\varepsilon)w^2\bar{w}) +$$
$$+\varepsilon^{3/2}((D(v_0(w,\bar{w};\varepsilon^{1/2})w_{xx} + \bar{v}_0(w,\bar{w};\varepsilon^{1/2})\bar{w}_{xx} + v_1(w,\bar{w};\varepsilon^{1/2})w_x^2 + v_{11}(w,\bar{w};\varepsilon^{1/2})w_x\bar{w}_x +$$
$$+ \bar{v}_1(w,\bar{w};\varepsilon^{1/2})\bar{w}_x^2), h(\varepsilon)) + W(w,\bar{w};\varepsilon^{1/2})) \ \ (W(0,0;\varepsilon^{1/2}) \equiv 0) \tag{4.1}$$

в пространстве $H \times H$ для определения $w(x,t)$ с областью определения правой части $H^2 \times H^2$. Функции $v_*(w,\bar{w};\varepsilon^{1/2}), W(w,\bar{w};\varepsilon^{1/2})$ легко строятся по (3.2) и (3.7).

Функции $g_0(x) = 1, g_k(x) = 2^{1/2}\cos(\pi kx), k = 1,2,\dots$, образуют ортонормированный базис в H и ортогональный в H^2. Введем пространства l_2 и $l_2^p, p = 2,4,\dots$, последовательностей вида $z = (z_0, z_1, \dots)$, $z_n \in \mathbb{C}$, удовлетворяющих соответственно условиям $||z||_{l_2} = (z,z)_{l_2}^{1/2} < \infty$, $(z_1,z_2)_{l_2} = \sum_{k=0}^{\infty} z_{1k}\bar{z}_{2k}$ и $||z||_{l_2^p} = ((z,z)_{l_2} + \sum_{k=1}^{\infty} k^{2p}|z_k|^2)^{1/2} < \infty$. Если $||z||_{l_2^p} < \infty$ при всех $p = 1,2,\dots$, соответствующее пространство обозначим l_2^∞. Обозначим через $s(r_0) = \{z \in l_2 : ||z||_{l_2} < r_0\}$, $s_1(r_0) = \{z \in l_2^2 : ||z||_{l_2^2} < r_0\}$. Представим $w(x,t) \in H^2$ в виде:

$$w(x,t) = \sum_{k=0}^{\infty} z_k(t)g_k(x), \ \ z(t) = (z_0(t), z_1(t), \dots) \in l_2^2, \tag{4.2}$$

подставим в (4.1) и спроектируем на $g_k(x), k = 0,1,\dots$. В результате получим систему обыкновенных дифференциальных уравнений в l_2

$$\dot{z}_k = \lambda(\varepsilon)z_k - \varepsilon(\pi^2k^2((De(\varepsilon),h(\varepsilon))z_k + (D\bar{e}(\varepsilon),h(\varepsilon))\bar{z}_k) - a(\varepsilon)\sum_{(k_1,k_2,k_3)\in\Omega_k} \alpha_{k_1k_2k_3}z_{k_1}z_{k_2}\bar{z}_{k_3}) +$$
$$+ \varepsilon^{3/2}Z_j(z,\bar{z};\varepsilon^{1/2}) \ (z_{-k_j} = \bar{z}_{k_j}, Z_j(0,0;\varepsilon^{1/2}) \equiv 0), \tag{4.3}$$

где $\Omega_k = \{(k_1,k_2,k_3) : k_j \geqslant 0, j = 1,2,3, \ \pm k_1 \pm k_2 \pm k_3 = \pm k\}$; $\alpha_{k_1k_2k_3} = 1$, если $k_1 = k_2 = k_3 = 0$, либо один $k_j > 0$, $\alpha_{k_1k_2k_3} = 2^{-1/2}$, если один $k_j = 0$, $\alpha_{k_1k_2k_3} = 2^{-1}$, если $k_1,k_2,k_3 > 0$; $Z(z,\bar{z};\varepsilon^{1/2}) = (Z_0(z,\bar{z};\varepsilon^{1/2}), Z_1(z,\bar{z};\varepsilon^{1/2}),\dots)$ нелинейный оператор, действующий из $s_1(r_0) \times s_1(r_0) \times [0,\varepsilon_0^{1/2}]$ в $s(r_0)$ и гладко зависящий от своих переменных.

Представим

$$(De(\varepsilon),h(\varepsilon)) = \alpha_1(\varepsilon) + i\beta_1(\varepsilon), (D\bar{e}(\varepsilon),h(\varepsilon)) = \alpha_2(\varepsilon) + i\beta_2(\varepsilon), \gamma(\varepsilon) = \varepsilon\gamma_1(\varepsilon), \sigma(\varepsilon) = \sigma_0 + \varepsilon\sigma_1(\varepsilon). \tag{4.4}$$

При этом

$$\alpha_1(0) + i\beta_1(0) = \frac{d_1+d_2}{2} + i\frac{(d_2-d_1)a_{11}(0)}{2\sigma_0}, \ \alpha_2(0) + i\beta_2(0), = \frac{d_1-d_2}{2} + i\frac{(d_2-d_1)a_{11}(0)}{2\sigma_0},$$

в чем несложно убедиться непосредственно ($a_{11}(0)$ элемент матрицы $A(0)$). Введем в рассмотрение полярные координаты, положив $z_k = \rho_k e^{i\tau_k}$ ($\rho_k \geqslant 0, -\infty < \tau_k < \infty, \rho = (\rho_0, \rho_1, \dots), \rho \in l_2^2, \tau = (\tau_0, \tau_1, \dots)$. Система уравнений (4.3) в силу структуры правой части позволяет ввести одну "быструю" переменную и счетное число "медленных" переменных. Выберем в качестве "быстрой" переменной τ_0, а в качестве "медленных" переменных $\rho_0, q_k = z_k e^{-i\tau_0} = \rho_k e^{i(\tau_k - \tau_0)} = \rho_k e^{i\theta_k}, k = 1, 2, \dots$. Обозначим $q = (q_0, q_1, \dots), q_0 = \rho_0$. Из первого уравнения (4.3) выразим

$$\dot\tau_0 = \sigma(\varepsilon) - (\varepsilon \sum_{(k_1,k_2,k_3)\in\Omega_0} \alpha_{k_1k_2k_3} i(a(\varepsilon) q_{k_1} q_{k_2} \bar q_{k_3} - \bar a(\varepsilon) \bar q_{k_1} \bar q_{k_2} q_{k_3})/2 +$$

$$+ \varepsilon^{3/2} i(Z_0(e^{i\tau_0} q, e^{-i\tau_0}\bar q; \varepsilon^{1/2}) e^{-i\tau_0} - \bar Z_0(e^{-i\tau_0}\bar q, e^{i\tau_0} q; \varepsilon^{1/2}) e^{i\tau_0})/2)/q_0, \tag{4.5}$$

$$\dot q_0 = \varepsilon(\gamma(\varepsilon) q_0 + \sum_{(k_1,k_2,k_3)\in\Omega_0} \alpha_{k_1k_2k_3} (a(\varepsilon) q_{k_1} q_{k_2} \bar q_{k_3} + \bar a(\varepsilon) \bar q_{k_1} \bar q_{k_2} q_{k_3})/2) +$$

$$+ \varepsilon^{3/2} (Z_0(e^{i\tau_0} q, e^{-i\tau_0}\bar q; \varepsilon^{1/2}) e^{-i\tau_0} + \bar Z_0(e^{-i\tau_0}\bar q, e^{i\tau_0} q; \varepsilon^{1/2}) e^{i\tau_0})/2, \tag{4.6}$$

$$\dot q_k = \varepsilon(\gamma(\varepsilon) q_k - \pi^2 k^2((\alpha_1(\varepsilon) + i\beta_1(\varepsilon)) q_k + (\alpha_2(\varepsilon) + i\beta_2(\varepsilon)) e^{-i2\tau_0}\bar q_k) +$$

$$+ a(\varepsilon) \sum_{(k_1,k_2,k_3)\in\Omega_k} \alpha_{k_1k_2k_3} q_{k_1} q_{k_2} \bar q_{k_3} - q_k/q_0 \sum_{(k_1,k_2,k_3)\in\Omega_0} \alpha_{k_1k_2k_3} (a(\varepsilon) q_{k_1} q_{k_2} \bar q_{k_3} - \bar a(\varepsilon) \bar q_{k_1} \bar q_{k_2} q_{k_3})/2) +$$

$$+ \varepsilon^{3/2} (Z_k(e^{i\tau_0} q, e^{-i\tau_0}\bar q; \varepsilon^{1/2}) e^{-i\tau_0} - q_k/q_0 (Z_0(e^{i\tau_0} q, e^{-i\tau_0}\bar q; \varepsilon^{1/2}) e^{-i\tau_0} - \bar Z_0(e^{-i\tau_0}\bar q, e^{i\tau_0} q; \varepsilon^{1/2}) e^{i\tau_0})/2), \tag{4.7}$$

$k = 1, 2, \dots$. Правые части уравнений (4.6), (4.7) пропорциональны ε, т.е. это уравнения "медленных" переменных. Отметим также, что переменная q_0 всегда положительна, она может быть равна нулю лишь при условии, что $q_k = 0, k = 1, 2, \dots$.

Перейдем в системе уравнений (4.6)-(4.7) к "новому времени" $\tau = \tau_0$ согласно (4.5) и запишем ее в операторной форме в виде

$$\frac{d}{d\tau} q = \varepsilon Q(\tau, q; \varepsilon^{1/2}) = \varepsilon Q^\circ(\tau, q; \varepsilon) + \varepsilon^{3/2} Q^*(\tau, q; \varepsilon^{1/2}), \tag{4.8}$$

где гладко зависящие от своих переменных операторы $Q^\circ(.) = (Q_0^\circ(.), Q_1^\circ(.), \dots)$ и $Q^*(.) = (Q_0^*(.), Q_1^*(.), Q_2^*(.), \dots)$ 2π-периодические по τ, действуют из $s_1(r_0) \in s(r_0)$ и определены правыми частями (4.6), (4.7). При этом первый определяется слагаемыми, стоящими в скобках при ε, деленными согласно (4.5) на $\sigma(\varepsilon)$, а второй содержит все остальные слагаемые. Отметим, что множество $q^\circ = (q_0, 0, 0, \dots), q_0 > 0$, инвариантно для решений (4.8) и для всех $k > 0$ $Q_k^\circ(\tau, q^\circ; \varepsilon) \equiv Q_k^*(\tau, q^\circ; \varepsilon) \equiv 0$. Поведение решений на q° определяется уравнением

$$\dot q_0 = \varepsilon(\gamma(\varepsilon) + Re\, a(\varepsilon) q_0^2 + \varepsilon^{1/2} R^*(q_0; \varepsilon^{1/2})) q_0/\sigma(\varepsilon) \equiv \varepsilon Q(\tau, q^\circ; \varepsilon^{1/2})\ (R^*(q_0; \varepsilon^{1/2}) = o(q_0^2)).$$

Представим

$$Q(\tau, q; \varepsilon^{1/2}) = Q_0(q; \varepsilon^{1/2}) + Q_1(\tau, q; \varepsilon^{1/2}),\ Q_0(q; \varepsilon^{1/2}) = M(Q(\tau, q; \varepsilon^{1/2})) = (2\pi)^{-1} \int_0^{2\pi} Q(\tau, q; \varepsilon^{1/2}) d\tau \tag{4.9}$$

и рассмотрим систему "усредненных" уравнений

$$\frac{dq}{d\tau} = \varepsilon Q_0(q; \varepsilon^{1/2}) = \varepsilon Q_0^\circ(q; \varepsilon) + \varepsilon^{3/2} Q_0^*(q; \varepsilon^{1/2}). \tag{4.10}$$

Пусть $q_*(\varepsilon^{1/2}) = q_{*0} + \varepsilon^{1/2} q_{*1}(\varepsilon^{1/2})$ – решение следующего операторного уравнения в l_2:

$$Q_0(q; \varepsilon^{1/2}) = Q_0^\circ(q; \varepsilon) + \varepsilon^{1/2} Q_0^*(q; \varepsilon^{1/2}) = 0, \tag{4.11}$$

которое согласно структуре оператора $Q_0(q; \varepsilon^{1/2})$, определяемой (4.7), принадлежит l_2^4. Представим $Q_0(q; \varepsilon^{1/2})$ в окрестности точки $q_*(\varepsilon^{1/2}) \in l_2^2$ рядом Тейлора (см. [19])

$$Q_0(q_*(\varepsilon^{1/2}) + q; \varepsilon^{1/2}) = L(\varepsilon^{1/2}) q + Q_{02}(q; \varepsilon^{1/2})\ (||Q_{02}(q; \varepsilon^{1/2})||_{l_2} = O(||q||_{l_2^2})^2), \tag{4.12}$$

где линейный оператор $L(\varepsilon^{1/2}) : l_2^2 \to l_2$ может быть определен через бесконечную матрицу

$$L(\varepsilon^{1/2}) = \{\frac{\partial Q_{0k}(q; \varepsilon^{1/2})}{\partial q_j}\}\ (k, j = 0, 1, \dots), \tag{4.13}$$

вычисленную в точке $q = q_*(\varepsilon^{1/2})$.

Пусть $\lambda_j(\varepsilon^{1/2})$ ($L(\varepsilon^{1/2})p_k(\varepsilon^{1/2}) = \lambda_j(\varepsilon^{1/2})p_k(\varepsilon^{1/2})$) – собственные значения линейной части оператора (4.12). Оператор $L(\varepsilon^{1/2})$ имеет счетное число собственных значений, которые могут быть пронумерованы в порядке возрастания их модулей, и в любой ограниченной области комплексной плоскости может быть лишь конечное число собственных значений конечной кратности. Предельной может быть лишь точка бесконечность и при этом $\lim_{j\to\infty} \operatorname{Re} \lambda_j(\varepsilon^{1/2}) = -\infty$. Это следует из свойств дифференциального уравнения

$$\frac{dq}{d\tau} = L(\varepsilon^{1/2})q \tag{4.14}$$

в l_2, которое согласно (4.6), (4.7) и (4.13) является абстрактно параболическим (см. [20]).

Предположим, что оператор $L(0)$ имеет $m \geqslant 0$ собственных значений, принадлежащих правой комплексной полуплоскости, и не имеет собственных значений, лежащих на мнимой оси комплексной плоскости. В силу сказанного выше, это свойство собственных значений будет выполнено также при малых $\varepsilon^{1/2}$. Покажем, что в этом случае уравнение (4.8) имеет в окрестности состояния равновесия $q_*(\varepsilon^{1/2})$ 2π-периодическое решение, характер устойчивости которого определяется устойчивостью состояния равновесия.

Введем в рассмотрение банаховы пространства $C(2\pi), C^1(2\pi), C^2(2\pi)$ соответственно непрерывных, непрерывно дифференцируемых и дважды непрерывно дифференцируемых 2π-периодических функций $f(\tau)(-\infty < \tau < \infty)$ со значениями в l_2, определив нормы $||f||_{C(2\pi)} = \sup_\tau ||f(\tau)||_{l_2}$, $||f||_{C^1(2\pi)} = ||f||_{C(2\pi)} + ||f'||_{C(2\pi)}$, $||f||_{C^2(2\pi)} = ||f||_{C(2\pi)} + ||f''||_{C(2\pi)}$. Через $C^\circ(2\pi), C^{\circ 1}(2\pi), C^{\circ 2}(2\pi)$ соответственно обозначим подпространства $C(2\pi), C^1(2\pi), C^2(2\pi)$ функций $f(\tau)$, удовлетворяющих условию $M(f(\tau)) = 0$.

2π-Периодическое решение уравнения (4.8) будем строить в виде

$$q_\circ(\tau, \varepsilon^{1/2}) = q_*(\varepsilon^{1/2}) + \varepsilon q_1(\tau, \varepsilon^{1/2}) + \varepsilon^2 q_2(\tau), \tag{4.15}$$

в котором $q_1(\tau, \varepsilon^{1/2})$ – решение уравнения

$$\frac{dq_1}{d\tau} = Q_1(\tau, q_*(\varepsilon^{1/2}); \varepsilon^{1/2}) = Q_1^\circ(\tau, q_*(\varepsilon^{1/2}); \varepsilon) + \varepsilon^{1/2} Q_1^*(\tau, q_*(\varepsilon^{1/2}); \varepsilon^{1/2}),$$

имеющее согласно (4.6)–(4.7) вид

$$q_1(\tau, \varepsilon^{1/2}) = (i\alpha_2(\varepsilon) - \beta_2(\varepsilon))e^{-2i\tau} p(\varepsilon^{1/2})/(2\sigma(\varepsilon)) + \varepsilon^{1/2} p_1^*(\tau; \varepsilon^{1/2}) + c,$$

$$p(\varepsilon^{1/2}) = (p_0(\varepsilon^{1/2}), p_1(\varepsilon^{1/2}), \dots),\ p_k(\varepsilon^{1/2}) = \pi^2 k^2 \bar{q}_{*k}(\varepsilon^{1/2}),$$

где функция $p_1^*(\tau; \varepsilon^{1/2}) \in C^{\circ 2}(2\pi)$ получена интегрированием $Q_1^*(\tau, q_*(\varepsilon^{1/2}); \varepsilon^{1/2})$, $c \in l_2^2$ – произвольный вектор, а функция $q_2(\tau) \in C^2(2\pi)$ подлежит определению.

Подставим (4.15) в уравнение (4.8) и приравняем слагаемые, имеющие порядок по ε второй и выше. В результате с учетом (4.11), (4.12) получим уравнение вида

$$\dot{q}_2(\tau) = L(\varepsilon^{1/2})c + P(\tau, c, q_2(\tau); \varepsilon^{1/2}), \tag{4.16}$$

где 2π-периодический по τ и гладкий по совокупности переменным нелинейный оператор $P(\tau, c, q_2(\tau); \varepsilon^{1/2})$ удовлетворяет условию $M(P(\tau, c, 0; 0)) = 0$.

Уравнение

$$\dot{q}_2(\tau) = f(\tau),\ f(\tau) \in C^{\circ 1}(2\pi), q_2(\tau) \in C^{\circ 2}(2\pi) \tag{4.17}$$

имеет единственное решение $q_2(\tau) = Jf(\tau)$, где J – линейный вполне непрерывный оператор. Из нелинейного уравнения в пространстве l_2^2

$$c + L^{-1}(\varepsilon^{1/2}) M(P(\tau, c, q_2(\tau); \varepsilon^{1/2})) = 0$$

с учетом условия $M(P(\tau, c, 0; 0)) = 0$ на основании теоремы о неявной функции в банаховом пространстве (см. [19]) находим решение (оператор-функцию) $c = c(\tau, q_2(\tau); \varepsilon^{1/2})$, которое подставим в правую часть (4.16). В результате получим операторное уравнение

$$q_2(\tau) = J(L(\varepsilon^{1/2})c(\tau, q_2(\tau); \varepsilon^{1/2}) + P(\tau, c(\tau, q_2(\tau); \varepsilon^{1/2}), q_2(\tau); \varepsilon^{1/2})) \tag{4.18}$$

в пространстве $C^{\circ 2}(2\pi)$ для определения $q_2(\tau)$. Осталось вновь применить к (4.18) теорему о неявной функции (см. [19]), на основании которой находим $q_2(\tau; \varepsilon^{1/2})$. Подставив теперь $q_2(\tau; \varepsilon^{1/2})$ в (4.15), получим 2π-периодическое решение $q_\circ(\tau; \varepsilon^{1/2})$ уравнения (4.8).

Линеаризуем теперь на $q_\circ(\tau;\varepsilon^{1/2})$ уравнение (4.8). С учетом вида (4.8) и (4.12) будем иметь линейное уравнение

$$\frac{dq}{d\tau} = \varepsilon(L(\varepsilon^{1/2}) + \varepsilon^{1/2}L_1(\tau;\varepsilon^{1/2}))q \tag{4.19}$$

в пространстве l_2, где $L_1(\tau;\varepsilon^{1/2}) : l_2^2 \to l_2$ есть 2π-периодический по τ, гладко зависящий от $\varepsilon^{1/2}$ линейный оператор. В силу сделанных предположений относительно собственных значений оператора $L(\varepsilon^{1/2})$, поведение решений уравнения (4.19) при $t \to \infty$ определяется поведением решений уравнения (4.14). Отсюда фазовое пространство уравнения (4.19) имеет m-мерное подпространство решений, нормы которых неограниченно возрастают при $t \to \infty$. Таким образом, в случае $m = 0$ периодическое решение $q_\circ(\tau;\varepsilon^{1/2})$ уравнения (4.8) асимптотически орбитально устойчиво.

Подставим $q_\circ(\tau;\varepsilon^{1/2})$ в (4.5). В результате получим уравнение

$$\dot\tau = \sigma(\varepsilon) + \varepsilon\sigma_1(\varepsilon) + \varepsilon^{3/2}\sigma_*(\tau;\varepsilon^{1/2}), \tag{4.20}$$

для определения $\tau(t)$, в котором $\sigma_1(\varepsilon)$ и $\sigma_*(\tau;\varepsilon^{1/2})$ – гладко зависящие от $\varepsilon^{1/2}$ функции, вторая – 2π-периодическая функция по τ. Общее решение уравнения (4.20) имеет вид

$$\tau(\sigma_\circ(\varepsilon^{1/2})t + c;\varepsilon^{1/2}) = \sigma_\circ(\varepsilon^{1/2})t + c + \varepsilon^{3/2}\tau_*(\sigma_\circ(\varepsilon)t + c;\varepsilon^{1/2}),$$

$$\tau_*(t + 2\pi;\varepsilon^{1/2}) \equiv \tau_*(t;\varepsilon^{1/2}),\ M(\tau_*(t;\varepsilon^{1/2})) = 0,\ \sigma_\circ(\varepsilon^{1/2}) = \sigma(\varepsilon) + \varepsilon\sigma_1(\varepsilon) + O(\varepsilon^{3/2}),$$

c – произвольная постоянная.

Подставим теперь $z_\circ(\tau;\varepsilon^{1/2}) = q_\circ(\tau;\varepsilon^{1/2})e^{i\tau}$ (оставив обозначение τ для переменной τ_0) в (4.2), а также полученное $w_\circ(x,\tau;\varepsilon^{1/2})$ с учетом выполненной нормировки в (3.2). В результате получим периодическое решение уравнения (2.1), представимое в виде

$$u_\circ(x,\tau;\varepsilon^{1/2}) = v(\varepsilon^{1/2}w_\circ(x,\tau;\varepsilon^{1/2}), \varepsilon^{1/2}\bar w_\circ(x,\tau;\varepsilon^{1/2});\varepsilon) =$$

$$= \varepsilon^{1/2}(e(\varepsilon)w_\circ(x,\tau;\varepsilon^{1/2}) + \bar e(\varepsilon)\bar w_\circ(x,\tau;\varepsilon^{1/2})) + \varepsilon u_*(x,\tau;\varepsilon^{1/2}), \tag{4.21}$$

в котором $u_*(x,\tau;\varepsilon^{1/2})$ – гладко зависящая от своих переменных и 2π-периодическая функция по τ вектор-функция, $\tau = \tau(\sigma_\circ(\varepsilon^{1/2})t + c;\varepsilon^{1/2})$ – общее решение уравнения (4.20).

Изложенное сформулируем в виде следующей теоремы.

Теорема 1. *Пусть при* $0 < \varepsilon < \varepsilon_0$ *система уравнений* (4.10) *имеет состояние равновесия* $q_*(\varepsilon^{1/2})$*, а линейный оператор* $L(0)$*, построенный согласно* (4.13)*, на этом решении имеет* $m \geqslant 0$ *собственных значений, принадлежащих правой комплексной полуплоскости, и не имеет собственных значений, лежащих на ее мнимой оси. Тогда уравнение* (4.8) *имеет* 2π*-периодическое решение* $q_\circ(\tau;\varepsilon^{1/2})$ *вида* (4.15)*, принадлежащее окрестности состояния равновесия, устойчивость которого определяется устойчивостью состояния равновесия. Этому периодическому решению в начально-краевой задаче* (2.1)*,* (2.2) *соответствует* $2\pi/\sigma_\circ(\varepsilon^{1/2})$*-периодическое решение вида* (4.21)*, устойчивость которого определяется устойчивостью состояния равновесия* $q_*(\varepsilon^{1/2})$*. Это периодическое решение (цикл) имеет в своей окрестности фазового пространства* m*-мерное многообразие решений, нормы которых неограниченно возрастают. В случае* $m = 0$ *периодическое решение* (4.21) *асимптотически орбитально устойчиво.*

Пусть уравнение (4.10) имеет периодическое решение $q_*(\varepsilon\tau;\varepsilon^{1/2}), \varepsilon\tau = \omega, q_*(\omega;\varepsilon^{1/2}) = q_*(\omega + T;\varepsilon^{1/2})$, которое в силу структуры $Q_0(q;\varepsilon^{1/2})$ принадлежит l_2^4. Представим $Q_0(q;\varepsilon^{1/2})$ в окрестности точки $q_*(\omega,\varepsilon^{1/2})$ рядом Тейлора (см. [19])

$$Q_0(q_*(\omega;\varepsilon^{1/2}) + q;\varepsilon^{1/2}) = L(\omega;\varepsilon^{1/2})q + Q_{02}(\omega,q;\varepsilon^{1/2})\ (||Q_{02}(\omega,q;\varepsilon^{1/2})||_{l_2} = O(||q||_{l_2^2})^2), \tag{4.22}$$

где линейный оператор $L(\omega;\varepsilon^{1/2})$ $(L(\omega;\varepsilon^{1/2}) \equiv L(\omega + T;\varepsilon^{1/2})) : l_2^2 \to l_2$ может быть определен через бесконечную матрицу (4.13), вычисленную в точке $q = q_*(\omega;\varepsilon^{1/2})$.

Рассмотрим в пространстве l_2 линейное уравнение

$$\frac{dq}{d\omega} = L(\omega;\varepsilon^{1/2})q. \tag{4.23}$$

В силу структуры уравнения (4.8) ((4.5)–(4.7)), уравнение (4.23) является абстрактно параболическим (см. [20]). Обозначим через $U(T;\varepsilon^{1/2}) : l_2^2 \to l_2^2$ оператор монодромии уравнения (4.23), который является вполне непрерывным. Функция $\dot q_*(\omega;\varepsilon^{1/2})$ является, как следует из (4.13), (4.22), T-периодическим решением (4.23), поэтому одно собственное значение оператора монодромии (мультипликатор) равно единице.

Предположим, что оператор $U(T;0)$ имеет только однократный единичный мультипликатор, модуль которого равен единице. Заметим, что в силу отмеченных свойств уравнения (4.23) число мультипликаторов оператора $U(T;\varepsilon^{1/2})$, модуль которых больше единицы, конечно. Покажем, что в этом случае уравнение (4.8) имеет в окрестности T-периодического решения $q_*(\omega;\varepsilon^{1/2})$ инвариантный тор $\mathbb{T}^2$, характер устойчивости которого определяется устойчивостью решений уравнения (4.23).

Введем в рассмотрение банаховы пространства $C(2\pi,T), C^1(2\pi,T), C^2(2\pi,T)$ соответственно непрерывных, непрерывно дифференцируемых и дважды непрерывно дифференцируемых функций $f(\tau,\omega)(-\infty < \tau,\omega < \infty)$ со значениями в l_2, 2π и T-периодических по τ и ω соответственно, определив нормы $||f||_{C(2\pi,T)} = \sup_{\tau,\omega}||f(\tau,\omega)||_{l_2}$, $||f||_{C^1(2\pi,T)} = ||f||_{C(2\pi,T)} + ||f_\tau||_{C(2\pi,T)} + ||f_\omega||_{C(2\pi,T)}$, $||f||_{C^2(2\pi,T)} = ||f||_{C(2\pi,T)} + ||f_{\tau\tau}||_{C(2\pi,T)} + ||f_{\tau\omega}||_{C(2\pi,T)} + ||f_{\omega\omega}||_{C(2\pi,T)}$. Через $C^\circ(2\pi,T), C^{\circ 1}(2\pi,T), C^{\circ 2}(2\pi,T)$ соответственно обозначим подпространства $C(2\pi,T)$, $C^1(2\pi,T), C^2(2\pi,T)$ функций $f(\tau,\omega)$, удовлетворяющих условию $M(f(\tau,\omega)) = 0$ (здесь и в дальнейшем среднее (4.9) берется по переменной τ).

Инвариантный тор $\mathbb{T}^2$ уравнения (4.8) будем строить в виде разложения

$$q_{\circ\circ}(\tau,\omega,\varepsilon^{1/2}) = q_*(\omega,\varepsilon^{1/2}) + \varepsilon q_1(\tau,\omega,\varepsilon^{1/2}) + \varepsilon^2 q_2(\tau,\omega) \tag{4.24}$$

одновременно с уравнением траекторий на нем

$$\frac{d\omega}{d\tau} = \varepsilon(1+\varepsilon\Omega(\varepsilon^{1/2})). \tag{4.25}$$

При этом функции $q_2(\tau,\omega) \in C^{\circ 2}(2\pi,T)$ и $\Omega(\varepsilon^{1/2})$ подлежат определению, а $q_1(\tau,\omega,\varepsilon^{1/2})$ есть общее решение уравнения

$$\frac{dq_1}{d\tau} = Q_1(\tau,q_*(\omega;\varepsilon^{1/2});\varepsilon^{1/2}) = Q_1^\circ(\tau,q_*(\omega;\varepsilon^{1/2});\varepsilon) + \varepsilon^{1/2}Q_1^*(\tau,q_*(\omega;\varepsilon^{1/2});\varepsilon^{1/2}),$$

в котором ω рассматривается как параметр. Согласно (4.6), (4.7) имеем

$$q_1(\tau,\omega;\varepsilon^{1/2}) = (i\alpha_2(\varepsilon) - \beta_2(\varepsilon))e^{-2i\tau}p(\omega;\varepsilon^{1/2})/(2\sigma(\varepsilon)) + \varepsilon^{1/2}p_1^*(\tau,\omega;\varepsilon^{1/2}) + c(\omega),$$

$$p(\omega;\varepsilon^{1/2}) = (p_0(\omega;\varepsilon^{1/2}), p_1(\omega;\varepsilon^{1/2}),\dots),\ p_k(\omega;\varepsilon^{1/2}) = \pi^2k^2\bar{q}_{*k}(\omega;\varepsilon^{1/2}),$$

где функция $p_1^*(\tau,\omega;\varepsilon^{1/2})$ $(M(p_1^*(\tau,\omega;\varepsilon^{1/2})) = 0)$ получена интегрированием по τ функции $Q_1^*(\tau,q_*(\omega;\varepsilon^{1/2});\varepsilon^{1/2})$, $c(\omega) \in l_2^2$ – произвольная непрерывно дифференцируемая T-периодическая функция, которая будет определена далее.

Подставим (4.24) в уравнение (4.8) и приравняем слагаемые, имеющие порядок по ε второй и выше (слагаемые первого порядка сократятся). В результате с учетом (4.17), (4.25) получим уравнение вида

$$q_{2\tau}(\tau,\omega) = -\dot{q}_*(\omega,\varepsilon^{1/2})\Omega(\varepsilon^{1/2}) - \dot{c}(\omega) - \varepsilon q_{2\omega}(\tau,\omega) + L(\omega;\varepsilon^{1/2})(c(\omega)+\varepsilon q_2(\tau,\omega)) +$$
$$+ F(\tau,\omega,q_2(\tau,\omega),\Omega(\varepsilon^{1/2}),c(\omega);\varepsilon^{1/2}), \tag{4.26}$$

где нелинейный оператор

$$F(\tau,\omega,q_2(\tau,\omega),\Omega(\varepsilon^{1/2}),c(\omega);\varepsilon^{1/2}),$$
$$F(\tau,\omega,q_2(\tau,\omega),\Omega(\varepsilon^{1/2}),c(\omega);0) \equiv F(\tau,\omega,0,0,0;\varepsilon^{1/2}) \equiv f(\tau,\omega;\varepsilon^{1/2}) \tag{4.27}$$

вобрал в себя все слагаемые, не приведенные в явном виде.

Для однозначной разрешимости (4.26) относительно $q_2(\tau,\omega) \in C^{\circ 2}(2\pi,T)$ (ω рассматриваем как параметр) необходимо и достаточно согласно (4.17), чтобы

$$-\dot{q}_*(\omega,\varepsilon^{1/2})\Omega(\varepsilon^{1/2}) - \dot{c}(\omega) - \varepsilon q_{2\omega}(\tau,\omega) + L(\omega;\varepsilon^{1/2})(c(\omega)+\varepsilon q_2(\tau,\omega)) +$$
$$+ F(\tau,\omega,q_2(\tau,\omega),\Omega(\varepsilon^{1/2}),c(\omega);\varepsilon^{1/2}) \in C^{\circ 1}(2\pi,T). \tag{4.28}$$

Рассмотрим в пространстве l_2 линейное уравнение

$$\frac{dq}{d\omega} = L(\omega;\varepsilon^{1/2})q + f(\omega), \tag{4.29}$$

где $f(\omega) \in C^1(T)$ – пространство непрерывно дифференцируемых T-периодических на $-\infty < \omega < \infty$ функций со значениями в l_2, норма в котором определена стандартным способом. Отметим некоторые свойства его решений (см. [20]). Введем в рассмотрение сопряженный с $L(\omega;\varepsilon^{1/2})$ оператор $L^*(\omega;\varepsilon^{1/2})$: $(L(\omega;\varepsilon^{1/2})q_1,q_2)_{l_2} = (q_1,L^*(\omega;\varepsilon^{1/2})q_2)_{l_2}$, $q_1,q_2 \in l_2^2, -\infty < \omega < \infty$, и сопряженное с (4.23) уравнение

$$\frac{dp}{d\omega} = -L^*(\omega;\varepsilon^{1/2})p. \tag{4.30}$$

Отметим, что $L^*(\omega;\varepsilon^{1/2})$ может быть определен через сопряженную с (4.13) матрицу. Уравнение (4.30) имеет единственное T-периодическое решение $p_*(\omega;\varepsilon^{1/2})$, которое может быть выбрано, удовлетворяющим условию

$$(\dot q_*(\omega;\varepsilon^{1/2}), p_*(\omega;\varepsilon^{1/2}))_{l_2} = 1,\ -\infty < \omega < \infty. \tag{4.31}$$

Условие

$$\int_0^T (f(\omega), p_*(\omega;\varepsilon^{1/2}))_{l_2} d\omega = 0 \tag{4.32}$$

является необходимым и достаточным для существования T-периодического решения уравнения (4.29) $q_p(\omega;\varepsilon^{1/2}) \in C^2(T)$ – пространству дважды непрерывно дифференцируемых T-периодических на $-\infty < \omega < \infty$ функций со значениями в l_2. При этом условие

$$\int_0^T (q_p(\omega;\varepsilon^{1/2}), p_*(\omega;\varepsilon^{1/2}))_{l_2} d\omega = 0$$

обеспечивает его единственственность и представление в виде

$$q_p(\omega;\varepsilon^{1/2}) = G(\varepsilon^{1/2}) f(\omega), \tag{4.33}$$

где $G(\varepsilon^{1/2}) : C^{\circ 1}(T) \to C^{\circ 2}(T)$ – гладко зависящий от $\varepsilon^{1/2}$ линейный вполне непрерывный оператор, определенный в $C^1(T)$ на функциях, удовлетворяющих (4.32), и действующий в $C^2(T)$ в подпространство функций, также удовлетворяющих (4.32).

Рассмотрим определяемое выражением (4.28) условие

$$-\dot q_*(\omega,\varepsilon^{1/2})\Omega(\varepsilon^{1/2}) - \dot c(\omega) - \varepsilon q_{2\omega}(\tau,\omega) + L(\omega;\varepsilon^{1/2})(c(\omega) + \varepsilon q_2(\tau,\omega)) +$$
$$+ M(F(\tau,\omega,q_2(\tau,\omega),\Omega(\varepsilon^{1/2}),c(\omega);\varepsilon^{1/2})) = 0 \tag{4.34}$$

как дифференциальное уравнение вида (4.29) относительно $\dot c(\omega) + \varepsilon q_{2\omega}(\tau,\omega)$ (τ рассматриваем как параметр). Условие (4.32) в этом случае примет вид

$$\Omega(\varepsilon^{1/2}) - T^{-1}\int_0^T (M(F(\tau,\omega,q_2(\tau,\omega),\Omega(\varepsilon^{1/2}),c(\omega);\varepsilon^{1/2})), p_*(\omega;\varepsilon^{1/2}))_{l_2}) d\omega = 0, \tag{4.35}$$

из которого с учетом условия (4.27) на основании теоремы о неявной функции в банаховом пространстве (см. [19]) однозначно определим гладко зависящую от своих переменных оператор-функцию $\Omega(\varepsilon^{1/2}) = \Omega(q_2(\tau,\omega),c(\omega);\varepsilon^{1/2})$ $(\Omega(q_2(\tau,\omega),c(\omega);0) \equiv \Omega_0(\omega))$. Выбрав таким образом $\Omega(\varepsilon^{1/2})$, обеспечим выполнение условия (4.32), что на основании (4.33) дает нелинейное операторное уравнение

$$c(\omega) + \varepsilon q_2(\tau,\omega) = G(\varepsilon^{1/2})(-\dot q_*(\omega,\varepsilon^{1/2})\Omega(q_2(\tau,\omega),c(\omega);\varepsilon^{1/2})) +$$
$$+ M(F(\tau,\omega,q_2(\tau,\omega),\Omega(q_2(\tau,\omega),c(\omega);\varepsilon^{1/2}),c(\omega);\varepsilon^{1/2})) \tag{4.36}$$

для определения $c(\omega)$, из которого на основании теоремы о неявной функции (см. [19]) в банаховом пространстве $C^1(T)$ с учетом условия (4.27) находим однозначное решение (оператор-функцию) $c(\omega) = c(\omega,q_2(\tau,\omega);\varepsilon^{1/2})$, $(c(\omega,q_2(\tau,\omega);0) \equiv c(\omega,0;\varepsilon^{1/2}) \equiv c_0(\omega))$, гладко зависящее от $q_2(.)$ при $||q_2(.)||_{C(2\pi,T)} < q_0$ и малых $\varepsilon^{1/2}$.

Отсюда на основании (4.17), (4.26), (4.34) получим нелинейное операторное уравнение

$$q_2(\tau,\omega) = J(F(\tau,\omega,q_2(\tau,\omega),\Omega(q_2(\tau,\omega),c(\omega,q_2(\tau,\omega);\varepsilon^{1/2});\varepsilon^{1/2}),c(\omega,q_2(\tau,\omega);\varepsilon^{1/2});\varepsilon^{1/2}) -$$
$$- M(F(\tau,\omega,q_2(\tau,\omega),\Omega(q_2(\tau,\omega),c(\omega,q_2(\tau,\omega);\varepsilon^{1/2});\varepsilon^{1/2}),c(\omega,q_2(\tau,\omega));\varepsilon^{1/2});\varepsilon^{1/2})), \tag{4.37}$$

в пространстве $C^{\circ 1}(2\pi,T)$ для определения $q_2(\tau,\omega)$. Используя теорему о неявной функции (см. [19]) в банаховом пространстве $C^{\circ 1}(2\pi,T)$ с учетом свойств (4.27), находим однозначное решение $q_2(\tau,\omega;\varepsilon^{1/2}) \in C^{\circ 2}(2\pi,T)$, гладко зависящее $\varepsilon^{1/2}$.

Подставив теперь $q_2(\tau,\omega;\varepsilon^{1/2})$ в (4.24), а также в $\Omega(\varepsilon^{1/2}) = \Omega(q_2(\tau,\omega),c(\omega,q_2(\tau,\omega));\varepsilon^{1/2})$ и $c(\omega,q_2(\tau,\omega))$ получим выражение $q_{\circ\circ}(\tau,\omega,\varepsilon^{1/2})$ для инвариантного тора $\mathbb{T}^2$ уравнения (4.8) и уравнение (4.25) траекторий на нем.

Линеаризуем теперь на $q_{\circ\circ}(\tau,\omega;\varepsilon^{1/2})$ уравнение (4.8) с учетом (4.25). В результате будем иметь линейное уравнение

$$\frac{dq}{d\tau} = \varepsilon(L(\varepsilon\tau;\varepsilon^{1/2}) + \varepsilon^{1/2}L_1(\tau,\omega;\varepsilon^{1/2}))q, \tag{4.38}$$

в пространстве l_2, где $L_1(\tau,\omega;\varepsilon^{1/2}) : l_2^2 \to l_2$ – гладко зависящий от своих переменных 2π-периодический по τ и T-периодический по ω оператор. В силу сделанных предположений относительно собственных значений оператора монодромии $U(T;\varepsilon^{1/2})$ уравнения (4.23) поведение решений уравнения (4.38) при $t \to \infty$ определяется поведением решений уравнения (4.23). Отметим, что один характеристический показатель уравнения (4.38) равен нулю. При этом фазовое пространство уравнения (4.38) имеет m-мерное подпространство решений, нормы которых неограниченно возрастают при $t \to \infty$. Таким образом, в случае $m = 0$ инвариантный тор $q_{\circ\circ}(\tau,\omega;\varepsilon^{1/2})$ асимптотически орбитально устойчив.

Подставим теперь $q_{\circ\circ}(\tau,\omega;\varepsilon^{1/2})$ и $z_\circ(\tau,\omega;\varepsilon^{1/2}) = q_{\circ\circ}(\tau,\omega;\varepsilon^{1/2})e^{i\tau}$ соответственно в (4.5) (оставив обозначение τ за переменной τ_0) и в (4.2), а также полученную функцию $w_\circ(x,\tau;\varepsilon^{1/2})$ с учетом выполненной нормировки, в выражение (3.2). В результате получим уравнение

$$\dot\tau = \sigma(\varepsilon) + \varepsilon\sigma_*(\tau,\omega;\varepsilon^{1/2}), \tag{4.39}$$

где $\sigma_*(\tau,\omega;\varepsilon^{1/2})$ – гладко зависящая от $\tau,\omega,\varepsilon^{1/2}$, 2π-периодическая по τ и T-периодическая по ω функция, а также инвариантный тор $\mathbb{T}^2$ начально-краевой задачи (2.1), представимый в виде

$$u_{\circ\circ}(x,\tau,\omega;\varepsilon^{1/2}) = v(\varepsilon^{1/2}w_\circ(x,\tau,\omega;\varepsilon^{1/2}),\varepsilon^{1/2}\bar w_\circ(x,\tau,\omega;\varepsilon^{1/2});\varepsilon) =$$

$$= \varepsilon^{1/2}(e(\varepsilon)w_\circ(x,\tau,\omega;\varepsilon^{1/2}) + \bar e(\varepsilon)\bar w_\circ(x,\tau,\omega;\varepsilon^{1/2})) + \varepsilon u_*(x,\tau,\omega;\varepsilon^{1/2}), \tag{4.40}$$

в котором $u_*(x,\tau,\omega;\varepsilon^{1/2})$ – гладко зависящая от своих переменных 2π-периодическая функция по τ и T-периодическая по ω функция.

Перейдем теперь в уравнении (4.25) к времени t в соответствии с (4.39). В результате имеем уравнение

$$\dot\omega = \varepsilon(\sigma(\varepsilon) + \varepsilon\Omega_*(\tau,\omega;\varepsilon^{1/2})), \tag{4.41}$$

где $\Omega_*(\tau,\omega;\varepsilon^{1/2})$ – гладко зависящая от $\varepsilon^{1/2}$ и 2π-периодическая по τ функции и T-периодическая по ω функция. Уравнения (4.39), (4.41) являются уравнениями траекторий на инвариантном торе (4.40).

Сказанное сформулируем в виде следующей теоремы.

Теорема 2. *Пусть при* $0 < \varepsilon < \varepsilon_0$ *система уравнений* (4.10) *имеет периодическое решение* $q_*(\varepsilon\tau;\varepsilon^{1/2})$ *периода* T ($q_*(\omega;\varepsilon^{1/2}) \equiv q_*(\omega + T;\varepsilon^{1/2})$)*, а оператор монодромии* $U(T;0)$ ($U(T;\varepsilon^{1/2})$) *уравнения* (4.23) *имеет однократный мультипликатор равный единице, не имеет других мультипликаторов по модулю равных единице и имеет* $m \geqslant 0$ *мультипликаторов по модулю больше единицы. Тогда уравнение* (4.8) *имеет инвариантный тор* $\mathbb{T}^2$ $q_{\circ\circ}(\tau,\omega;\varepsilon^{1/2})$ *вида* (4.24)*, устойчивость которого определяется устойчивостью периодического решения* $q_*(\omega;\varepsilon^{1/2})$*. Этому инвариантному тору в начально-краевой задаче* (2.1), (2.2) *соответствует инвариантный тор* $\mathbb{T}^2$ *вида* (4.40)*, устойчивость которого также определяется устойчивостью периодического решения* $q_*(\omega;\varepsilon^{1/2})$*. Уравнения траекторий на инвариантном торе* (4.40) *имеют вид* (4.39), (4.41)*. Инвариантный тор* (4.40) *имеет в своей окрестности фазового пространства* m*-мерное многообразие решений, нормы которых неограниченно возрастают. В случае* $m = 0$ *инвариантный тор* (4.40) *асимптотически орбитально устойчив.*

Предположим теперь, что система уравнений (4.10) имеет инвариантный тор $\mathbb{T}^k$ $(k > 1)$ вида

$$q_*(\omega_1,\dots,\omega_k;\varepsilon^{1/2}) \equiv q_*(\omega_1 + 2\pi,\dots,\omega_k + 2\pi;\varepsilon^{1/2}),\ \ \dot\omega_j = \varepsilon\Omega_j(\varepsilon^{1/2}),\ \Omega_j(0) \neq 0,\ j = 1,\dots,k.$$

который в силу структуры $Q_0(q;\varepsilon^{1/2})$ принадлежит l_2^4. Представим оператор $Q_0(q;\varepsilon^{1/2})$ в окрестности точки $q_*(\omega_1,\dots,\omega_k;\varepsilon^{1/2})$ рядом Тейлора [4]

$$Q_0(q_*(\omega_1,\dots,\omega_k;\varepsilon^{1/2}) + q;\varepsilon^{1/2}) = L(\omega_1,\dots,\omega_k;\varepsilon^{1/2})q + Q_{02}(\omega_1,\dots,\omega_k,q;\varepsilon^{1/2}),$$

$$(||Q_{02}(\omega_1,\dots,\omega_k,q;\varepsilon^{1/2})||_{l_2} = O(||q||_{l_2^2})^2),$$

где линейный оператор $L(\omega_1,\dots,\omega_k;\varepsilon^{1/2}) : l_2^2 \to l_2$ 2π-периодический по ω_j и может быть определен через бесконечную матрицу (4.13), вычисленную в точке $q = q_*(\omega_1,\dots,\omega_k;\varepsilon^{1/2})$.

Рассмотрим в пространстве l_2 уравнение

$$\frac{dq}{d\tau} = L(\omega_1,\dots,\omega_k;0)q,\ \ \frac{d\omega_j}{d\tau} = \Omega_j(0). \tag{4.42}$$

Уравнение (4.42) является абстрактно параболическим (см. [20]) и имеет k решений вида $q_{*\omega_j}(\omega_1,\dots,\omega_k;0)$, $j = 1,\dots,k$, характкеристические показатели которых равны нулю.

Справедлива следующая теорема, доказательство которой аналогично доказательсву теоремы 2.

Теорема 3. *Предположим, что уравнение* (4.42) *имеет ровно* k *нулевых и* m *положительных характеристических показателей. Остальные характеристические показатели отрицательные. Тогда уравнение* (4.8) *имеет инвариантный тор* $\mathbb{T}^{k+1}$ $q_{\circ k}(\tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})$*, устойчивость которого определяется устойчивостью решений уравнения* (4.42)*. Этому инвариантному тору в начально-краевой задаче* (2.1)*,* (2.2) *соответствует инвариантный тор* $\mathbb{T}^{k+1}$ *вида*

$$u_{\circ k}(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}) = v(\varepsilon^{1/2} w_\circ(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}), \varepsilon^{1/2} \bar{w}_\circ(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}); \varepsilon) =$$
$$= \varepsilon^{1/2}(e(\varepsilon) w_\circ(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}) + \bar{e}(\varepsilon)\bar{w}_\circ(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})) + \varepsilon u_*(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}), \tag{4.43}$$

в котором $w_\circ(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}), u_*(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})$ *— гладко зависящие от своих переменных* 2π*-периодические по* τ *и* ω_j *функции. Уравнения траекторий на инвариантных торах имеют вид*

$$\dot{\tau} = \sigma(\varepsilon) + \varepsilon \sigma_*(\tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2}), \tag{4.44}$$

$$\dot{\omega}_j = \varepsilon(\Omega_j(\varepsilon^{1/2})\sigma(\varepsilon) + \varepsilon \Omega_{*j}(\tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})),\ j = 1, \dots, k, \tag{4.45}$$

где $\sigma_*(\tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})$ *и* $\Omega_{*j}(\tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})$ *— гладко зависящие от своих переменных функции,* 2π*-периодические по* τ *и* ω_j*. Инвариантный тор* (4.43) *имеет в своей окрестности фазового пространства* m*-мерное многообразие решений, нормы которых неограниченно возрастают. В случае* $m = 0$ *инвариантный тор* (4.43) *асимптотически орбитально устойчив.*

Отметим следующее. Предположим, что матрица $A(\varepsilon)$ и функция $F(u; \varepsilon)$ в (2.1) аналитически зависят от ε и u при $|\varepsilon| < \varepsilon_0$, и $||u|| < R_0$. В связи с этим будут аналитически зависить от своих переменных и параметра ε преобразование (3.2) и правая часть уравнения (3.3). Следствием этого является аналитическая зависимость правых частей уравнений (4.8) от фазовых переменных и параметра $\mu = \varepsilon^{1/2}$ при $|\mu| < \mu_0$, а также операторных уравнений (4.18) и (4.37). Теорема о неявной функции для аналитических операторных уравнений с аналитической зависимостью от входящих параметров (см. [19]) позволяет утверждать, что периодическое решение (4.21), инвариантный торы (4.40), (4.43), правые части уравнений (4.39), (4.41) и (4.44), (4.45) траекторий на торах разлагаются в сходящиеся ряды по $\varepsilon^{1/2}$.

5. АЛГОРИТМ ПОСТРОЕНИЯ АВТОКОЛЕБАТЕЛЬНЫХ РЕШЕНИЙ

Предположим, что матрица $A(\varepsilon)$ и функция $F(u; \varepsilon)$ в (2.1) являются бесконечно дифференцируемыми соответственно по ε и u при $|\varepsilon| < \varepsilon_0$, и $||u|| < R_0$. В этом случае периодическое решение (4.21), инвариантный тор (4.40) и правые части уравнений траекторий на инвариантном торе (4.39) и (4.41) являются бесконечно дифференцируемыми функциями $x, \tau, \omega, \mu = \varepsilon^{1/2}$ (см. [18]). При этом

$$\frac{\partial^k u_\circ(x, \tau; \varepsilon^{1/2})}{\partial x^k}\Big|_{x=0, x=1} = \frac{\partial^k u_{\circ\circ}(x, \tau; \varepsilon^{1/2})}{\partial x^k}\Big|_{x=0, x=1} = 0,\ k = 1, 3, \dots. \tag{5.1}$$

Рассмотрим сначала алгоритм построения периодического решения $u_\circ(x, \tau; \varepsilon^{1/2})$, определяемого выражениями (4.20), (4.21). Пусть $q_*(0) = q_* = (q_{*0}, q_{*1}, \dots)$ главная часть состояния равновесия уравнения (4.10), т.е.

$$Q_0^\circ(q_*; 0) = 0, \tag{5.2}$$

для которого оператор (4.12) удовлетворяет условиям теоремы 1. Обозначим

$$u_0(x) = \sum_{k=0}^{\infty} q_{*k} g_k(x). \tag{5.3}$$

Функция (5.3) является бесконечно дифференцируемой и удовлетворяет условиям (5.1) (принадлежит классу $C_\circ^\infty$). Запишем уравнение (2.1) в виде

$$u_t = \varepsilon D u_{xx} + (A(0) + \varepsilon A_1 + \varepsilon^2 A_2 + \cdots)u + F_2(u, u; \varepsilon) + F_3(u, u, u; \varepsilon) + \cdots, \tag{5.4}$$

где $A_1 = A'(0), A_2 = A''(0)$, $F_2(.)F_3(.)$ соответственно квадратичные и кубические слагаемые, записанные в симметризованном виде, здесь и в дальнейшем точками обозначены слагаемые, имеющие более высокий порядок малости по соответствующим переменным.

Будем строить периодическое решение $u_\circ(.)$ в виде разложения по $\varepsilon^{1/2}$:

$$u_\circ(x, \tau; \varepsilon^{1/2}) = \varepsilon^{1/2}(e_0(u_0(x) + \varepsilon u_3(x) + \cdots)e^{i\tau} + \bar{e}_0(\bar{u}_0(x) + \varepsilon \bar{u}_3(x) + \cdots)e^{-i\tau}) +$$

$$+\varepsilon(e_{20}u_{20}(x)e^{2i\tau}+e_{11}u_{11}(x)+\bar{e}_{20}\bar{u}_{20}(x)e^{-2i\tau})+$$
$$+\varepsilon^{3/2}(e_{30}u_{30}(x)e^{3i\tau}+e_{21}u_{21}(x)e^{i\tau}+\bar{e}_{21}\bar{u}_{21}(x)e^{-i\tau}+\bar{e}_{30}\bar{u}_{30}(x)e^{-3i\tau})+\cdots, \tag{5.5}$$
$$\dot{\tau}=\sigma_0+\varepsilon\sigma_{*1}+\varepsilon^2\sigma_{*2}+\cdots, \tag{5.6}$$

в котором $e_0=e(0)$ ($e(\varepsilon)$ определен в (3.2)), векторы $e_{jk}\in\mathbb{C}^2$, функции $u_j(x),u_{jk}(x)\in C_\circ^\infty$ и σ_{*j} подлежат определению.

Подставим (5.5), (5.6) в (5.4) и приравняем слева и справа коэффициенты при одинаковых степенях ε с учетом соответствующих гармоник по τ. В результате получим рекуррентную последовательность операторных уравнений в пространстве $H\times H$ определенного вида. При $\varepsilon^{1/2}$ равенства будут выполнены в силу (3.2). При ε будем иметь уравнения

$$(2i\sigma_0I-A(0))e_{20}u_{20}(x)=F_2(e_0,e_0;0)u_0(x)^2,\ \ -A(0)e_{11}u_{11}(x)=2F_2(e_0,\bar{e}_0;0)|u_0(x)|^2,$$

из которых однозначно определяем $e_{20}u_{20}(x)=e_{20}u_0(x)^2, e_{11}u_{11}(x)=e_{11}|u_0(x)|^2$. Здесь I – единичная матрица. При $\varepsilon^{3/2}$ получим уравнение для определения $e_{30}u_{30}(x)$

$$(3i\sigma_0I-A(0))e_{30}u_{30}(x)=(2F_2(e_{20},e_0;0)+F_3(e_0,e_0,e_0))u_0(x)^3,$$

из которого однозначно определим $e_{30}u_{30}(x)=e_{30}u_0(x)^3$, и уравнение для определения $e_{21}u_{21}(x)$

$$(i\sigma_0I-A(0))e_{21}u_{21}(x)=(A_1-i\sigma_{*1}I)e_0u_0(x)+De_0u_{0xx}(x)+(2F_2(e_{20},\bar{e}_0;0)+$$
$$+2F_2(e_{11},e_0;0)+3F_3(e_0,e_0,\bar{e}_0;0))|u_0(x)|^2u_0(x)=\Phi_3(x)=e_0\phi_{31}(x)+\bar{e}_0\phi_{32}(x), \tag{5.7}$$

правая часть которого представлена в виде разложения по векторам $e_0,\bar{e}_0$. При этом $\phi_{31}(x)=(\Phi_3(x),h_0),\phi_{32}(x)=(\Phi_3(x),\bar{h}_0)$, где вектор $h_0=h(0)$, $h(\varepsilon)$ определен в (3.6). Так как $(i\sigma_0I-A(0))e_0g_k(x)=0,k=0,1,\ldots$, то уравнение (5.7) будет разрешимо в $H\times H$ лишь при условии

$$(\phi_{31}(x),g_k(x))_H=((A_1-i\sigma_{*1}I)e_0u_0(x)+De_0u_{0xx}(x)+(2F_2(e_{20},\bar{e}_0;0)+$$
$$+2F_2(e_{11},e_0;0)+3F_3(e_0,e_0,\bar{e}_0;0))|u_0(x)|^2u_0(x),h_0g_k(x))_H=0,\ k=0,1\ldots. \tag{5.8}$$

В первом уравнении (5.8) (при $k=0$) приравняем нулю мнимую часть и выразим из полученного равенства σ_{*1}. Подставим полученное выражение σ_{*1} в оставшиеся уравнения системы (5.8). В результате получим, что несложно заметить, операторное уравнение (5.2) в пространстве l_2 относительно q_*. Таким образом, уравнение (5.7) разрешимо, а его общее решение с учетом (5.5) имеет вид

$$e_{21}u_{21}(x)+e_0u_3(x)=\bar{e}_0\phi_{32}(x)/(2i\sigma_0)+e_0\sum_{k=0}^{\infty}q_{3k}g_k(x),$$

где $q_3=(q_{30},q_{31},\ldots)\in l_2^\infty$ – произвольный вектор, который будет определен в дальнейшем.

Отметим, что изначально вектор q_* в (5.3) можно рассматривать как неопределенный параметр и, дойдя до уравнения (5.8), его определить как решение уравнения $Q_0^\circ(q;0)=0$. Затем уже определить функции $u_{20}(x),u_{11}(x),u_{30}(x)$.

Приравняв теперь коэффициенты при ε^2, получим уравнения для определения соответствующих слагаемых, решения которых находятся однозначно. Дополняющие (5.5) слагаемые будут иметь вид

$$\varepsilon^2(e_{40}u_{40}(x)e^{4i\tau}+(e_{31}u_{31}(x)+2e_{20}u_0(x)u_3(x))e^{2i\tau}+e_{22}u_{22}(x)+e_{11}(u_0(x)\bar{u}_3(x)+\bar{u}_0(x)u_3(x))+\cdots),$$

в которых векторы e_{40},e_{31},e_{22} и функции $u_{40}(x),u_{31}(x),u_{22}(x)\in C_\circ^\infty$ однозначно определяются, а точками обозначены комплексно сопряженные к первым двум слагаемые. Здесь выделены слагаемые, содержащие функцию $u_3(x)$. Приравняв теперь коэффициенты при $\varepsilon^{5/2}$, однозначно определим $e_{50}u_{50}(x)$ при $e^{5i\tau}$, $e_{41}u_{41}(x)$ при $e^{3i\tau}(u_{50}(x),u_{41}(x)\in C_\circ^\infty)$. Сложность возникнет при определении вектора $e_{32}u_{32}(x)$ при $e^{i\tau}$, для определения которого получим уравнение

$$(i\sigma_0I-A(0))e_{32}u_{32}(x)=-i\sigma_{*2}e_0u_0(x)+(A_1-i\sigma_{*1}I)e_0u_3(x)+De_0u_{3xx}(x)+$$
$$+(2F_2(e_{20},\bar{e}_0;0)+2F_2(e_{11},e_0;0)+3F_3(e_0,e_0,\bar{e}_0;0))(u_0(x)^2\bar{u}_3(x)+$$
$$+2|u_0(x)|^2u_3(x))+f_{32}(x)=\Phi_5(x)=e_0\phi_{51}(x)+\bar{e}_0\phi_{52}(x), \tag{5.9}$$

где $f_{32}(x)$ – определенная вектор-функция, координаты которой принадлежат $C_\circ^\infty$, $\phi_{51}(x) = (\Phi_5(x), h_0), \phi_{52}(x) = (\Phi_5(x), \bar{h}_0)$, выражение для σ_{*1} определено на предыдущем шаге. Так как $(i\sigma_0 I - A(0))e(0)g_k(x) = 0, k = 0, 1, \dots$, уравнение (5.9) будет разрешимо в $H \times H$ лишь при условии

$$(\phi_{51}(x), g_k(x))_H = (-i\sigma_{*2}e_0u_0(x) + (A_1 - i\sigma_{*1}I)e_0u_3(x) + De_0u_{3xx}(x) + (2F_2(e_{20}, \bar{e}_0; 0) +$$

$$+ 2F_2(e_{11}, e_0; 0) + 3F_3(e_0, e_0, \bar{e}_0; 0))(u_0(x)^2\bar{u}_3(x) + 2|u_0(x)|^2u_3(x) + f_{32}(x), h_0g_k(x))_H = 0, \tag{5.10}$$

$k = 0, 1 \dots$. В первом уравнении системы (5.10) (при $k = 0$) приравняем нулю мнимую часть, выразим из полученного равенства σ_{*2} и подставим полученное выражение σ_{*2} в оставшиеся уравнения системы (5.10). В результате получим линейное операторное уравнение в пространстве l_2 вида

$$L(0)q_3 = f_{32}^*, \tag{5.11}$$

в котором координаты вектора $f_{32}^* \in l_2$ определяются величинами $(f_{32}(x), h_0g_k(x))_H, k = 0, 1, \dots$, с учетом выражения для σ_{*2}. При сделанных предположениях относительно свойств оператора $L(0)$ (теорема 1) уравнение (5.11) имеет единственное решение $q_3 = L^{-1}(0)f_{32}^* \in l_2^\infty$. Выбрав таким образом q_3, получим разрешимость уравнения (5.9), а его общее решение с учетом (5.5) будет иметь вид

$$e_{32}u_{32}(x) + e_0u_5(x) = \bar{e}_0\phi_{52}(x)/(2i\sigma_0) + e_0\sum_{k=0}^{\infty} q_{5k}g_k(x), \tag{5.12}$$

где $q_5 = (q_{50}, q_{51}, \dots) \in l_2^\infty$ – произвольный вектор, который будет определен из уравнений, полученных при $\varepsilon^{7/2}$. Этот процесс можно продолжать до бесконечности, на каждом последующем шаге будем иметь уравнение вида (5.9), а соответствующее общее решение вида (5.12).

Рассмотрим теперь алгоритм построения инвариантного тора $\mathbb{T}^2$ $u_{\circ\circ}(x, \tau, \omega; \varepsilon^{1/2})$ вида (4.40) и уравнений траекторий на нем (4.39), (4.41). Пусть $q_*(\omega; 0) = q_*(\omega) = (q_{*0}(\omega), q_{*1}(\omega), \dots)$ – главная часть T-периодического решения $q_*(\omega; \varepsilon^{1/2})$ уравнения (4.10), т.е.

$$\frac{dq_*}{d\omega} = Q_0(q_*(\omega); 0), \tag{5.13}$$

для которого оператор монодромии $U(T; 0)$ $(U(T; \varepsilon^{1/2}))$ уравнения (4.23), построенного на этом решении, удовлетворяет условиям теоремы 2. Обозначим

$$u_0(x, \omega) = \sum_{k=0}^{\infty} q_{*k}(\omega)g_k(x). \tag{5.14}$$

Функция (5.3) является бесконечно дифференцируемой по совокупности переменных и удовлетворяет условиям (5.1) (принадлежит классу $C_{\circ\circ}^\infty$).

Будем строить инвариантный тор $u_{\circ\circ}(x, \tau, \omega; \varepsilon^{1/2})$ и уравнения траекторий на нем в виде разложений по $\varepsilon^{1/2}$ в следующей форме:

$$u_{\circ\circ}(x, \tau, \omega; \varepsilon^{1/2}) = \varepsilon^{1/2}(e_0(u_0(x, \omega) + \varepsilon u_3(x, \omega) + \cdots)e^{i\tau} + \bar{e}_0(\bar{u}_0(x, \omega) + \varepsilon\bar{u}_3(x, \omega) + \cdots)e^{-i\tau}) +$$

$$+\varepsilon(e_{20}u_{20}(x, \omega)e^{2i\tau} + e_{11}u_{11}(x, \omega) + \bar{e}_{20}\bar{u}_{20}(x, \omega)e^{-2i\tau}) +$$

$$+ \varepsilon^{3/2}(e_{30}u_{30}(x, \omega)e^{3i\tau} + e_{21}u_{21}(x, \omega)e^{i\tau} + \bar{e}_{21}\bar{u}_{21}(x, \omega)e^{-i\tau} + \bar{e}_{30}\bar{u}_{30}(x, \omega)e^{-3i\tau}) + \cdots, \tag{5.15}$$

$$\dot{\tau} = \sigma_0 + \varepsilon\sigma_{*1}(\omega) + \varepsilon^2\sigma_{*2}(\omega) + \cdots,$$

$$\frac{d\omega}{d\tau} = \varepsilon(1 + \varepsilon\Omega_1 + \varepsilon^2\Omega_2 + \cdots), \tag{5.16}$$

гле $e_0 = e(0)$ ($e(\varepsilon)$ определен в (3.2)), а векторы $e_{jk} \in \mathbb{C}^2$, функции $u_j(x, \omega), u_{jk}(x, \omega) \in C_{\circ\circ}^\infty$, $\sigma_{*j}(\omega)$ и величины Ω_j подлежат определению.

Подставим (5.15) с учетом (5.16) в (5.4) и приравняем слева и справа коэффициенты при одинаковых степенях ε. В результате получим рекуррентную последовательность операторных уравнений в пространстве $H \times H$ определенного вида. При $\varepsilon^{1/2}$ равенства будут выполнены в силу (3.2). При ε будем иметь уравнения в пространстве $H \times H$

$$(2i\sigma_0 I - A(0))e_{20}u_{20}(x, \omega) = F_2(e_0, e_0; 0)u_0(x, \omega)^2, \quad -A(0)e_{11}u_{11}(x, \omega) = 2F_2(e_0, \bar{e}_0; 0)|u_0(x, \omega)|^2,$$

из которых однозначно определяем $e_{20}u_{20}(x,\omega) = e_{20}u_0(x,\omega)^2$, $e_{11}u_{11}(x,\omega) = e_{11}|u_0(x,\omega)|^2$. При $\varepsilon^{3/2}$ получим уравнение для определения $e_{30}u_{30}(x,\omega)$

$$(3i\sigma_0 I - A(0))e_{30}u_{30}(x,\omega) = (2F_2(e_{20},e_0;0) + F_3(e_0,e_0,e_0))u_0(x,\omega)^3,$$

из которого однозначно определим $e_{30}u_{30}(x,\omega) = e_{30}u_0(x,\omega)^3$, и уравнение для определения $e_{21}u_{21}(x)$

$$(i\sigma_0 I - A(0))e_{21}u_{21}(x,\omega) = -e_0u_{0\omega}(x,\omega)\sigma_0 + (A_1 - i\sigma_{*1}(\omega)I)e_0u_0(x,\omega) + De_0u_{0xx}(x,\omega) +$$

$$+(2F_2(e_{20},\bar e_0;0) + 2F_2(e_{11},e_0;0) + 3F_3(e_0,e_0,\bar e_0;0))|u_0(x,\omega)|^2u_0(x,\omega) =$$

$$= \Phi_3(x,\omega) = e_0\phi_{31}(x,\omega) + \bar e_0\phi_{32}(x,\omega), \tag{5.17}$$

правая часть которого представлена в виде разложения по векторам $e_0, \bar e_0$. При этом $\phi_{31}(x,\omega) = (\Phi_3(x,\omega),h_0), \phi_{32}(x,\omega) = (\Phi_3(x,\omega),\bar h_0)$. Так как $(i\sigma_0 I - A(0))e_0g_k(x) = 0, k = 0,1,\cdots$, то уравнение (5.17) будет разрешимо в $H\times H$ (здесь и в дальнейшем ω рассматривается как параметр) лишь при условии

$$(\phi_{31}(x,\omega),g_k(x))_H = (-e_0u_{0\omega}(x,\omega)\sigma_0 + (A_1 - i\sigma_{*1}(\omega)I)e_0u_0(x,\omega) + De_0u_{0xx}(x,\omega) +$$

$$+ (2F_2(e_{20},\bar e_0;0) + 2F_2(e_{11},e_0;0) + 3F_3(e_0,e_0,\bar e_0;0))|u_0(x,\omega)|^2u_0(x,\omega),h_0g_k(x))_H = 0,\ k = 0,1\cdots. \tag{5.18}$$

В первом уравнении (5.18) (при $k=0$) приравняем нулю мнимую часть и выразим из полученного равенства $\sigma_{*1}(\omega)$. Так как $q_{*0}(\omega)\in\mathbb{R}$, то $\sigma_{*1}(\omega)$ не зависит от $\dot q_*(\omega)$. Подставим полученное выражение $\sigma_{*1}(\omega)$ в оставшиеся уравнения системы (5.18). В результате получим уравнение (5.13) в пространстве l_2. Таким образом, уравнение (5.17) в силу (5.14) разрешимо, а его общее решение с учетом (5.15) имеет вид

$$e_{21}u_{21}(x,\omega) + e_0u_3(x,\omega) = \bar e_0\phi_{32}(x,\omega)/(2i\sigma_0) + e_0\sum_{k=0}^{\infty}q_{3k}(\omega)g_k(x),$$

где $q_3(\omega) = (q_{30}(\omega),q_{31}(\omega),\cdots)\in l_2^\infty$ – произвольный бесконечно дифференцируемый по ω T- периодический вектор, который будет определен в дальнейшем.

Отметим, что здесь также изначально можно рассматривать вектор $q_*(\omega)$ в (5.14) как неопределенный параметр и, дойдя до уравнения (5.18), его определить как решение уравнения $dq_*(\omega)/d\omega = Q_0(q_*(\omega);0)$, получаемого из (5.18) выделением выражения $\sigma_{*1}(\omega)$, а затем уже определить функции $u_{20}(x,\omega), u_{11}(x,\omega), u_{30}(x,\omega)$.

Приравняв теперь коэффициенты при ε^2, получим уравнения для определения соответствующих слагаемых, решения которых находятся однозначно. В результате дополняющее (5.15) слагаемое будет иметь вид

$$\varepsilon^2(e_{40}u_{40}(x,\omega)e^{4i\tau} + (e_{31}u_{31}(x,\omega) + 2e_{20}u_0(x,\omega)u_3(x,\omega))e^{2i\tau} + e_{22}u_{22}(x,\omega) +$$
$$+e_{11}(u_0(x,\omega)\bar u_3(x,\omega) + \bar u_0(x,\omega)u_3(x,\omega)) + \cdots),$$

в котором векторы e_{40},e_{31},e_{22} и функции $u_{40}(x,\omega),u_{31}(x,\omega),u_{22}(x,\omega)\in C_{\circ\circ}^\infty$ однозначно определяются, точками обозначены комплексно сопряженные к первым двум слагаемые. Здесь выделены слагаемые, содержащие функцию $u_3(x,\omega)$. Приравняв теперь коэффициенты при $\varepsilon^{5/2}$, однозначно определим $e_{50}u_{50}(x,\omega)$ при $e^{5i\tau}$, $e_{41}u_{41}(x,\omega)$ при $e^{3i\tau}(u_{50}(x,\omega),u_{41}(x,\omega)\in C_{\circ\circ}^\infty)$. Сложность возникнет при определении вектора $e_{32}u_{32}(x,\omega)$ при $e^{i\tau}$, для которого получим уравнение

$$(i\sigma_0 I - A(0))e_{32}u_{32}(x,\omega) = -e_0u_{0\omega}(x,\omega)\sigma_0\Omega_1 - e_0u_{3\omega}(x,\omega)\sigma_0 + (A_2 - i\sigma_{*2}(\omega)I)e_0u_0(x,\omega) +$$

$$+(A_1 - i\sigma_{*1}(\omega)I)e_0u_3(x,\omega) + De_0u_{3xx}(x,\omega) + (2F_2(e_{20},\bar e_0;0) + 2F_2(e_{11},e_0;0) + 3F_3(e_0,e_0,\bar e_0;0))\times$$
$$\times(u_0(x,\omega)^2\bar u_3(x,\omega) + 2|u_0(x,\omega)|^2u_3(x,\omega)) + f_{32}(x,\omega) = \Phi_5(x,\omega) = e_0\phi_{51}(x,\omega) + \bar e_0\phi_{52}(x,\omega), \tag{5.19}$$

где $f_{32}(x,\omega)$ – определенная вектор-функция, координаты которой принадлежат $\in C_{\circ\circ}^\infty$, $\phi_{51}(x,\omega) = (\Phi_5(x,\omega),h_0), \phi_{52}(x,\omega) = (\Phi_5(x,\omega),\bar h_0)$, выражение для $\sigma_{*1}(\omega)$ определено на предыдущем шаге. Так как $(i\sigma_0 I - A(0))e(0)g_k(x) = 0, k = 0,1,\ldots$, то уравнение (5.19) будет разрешимо в $H\times H$ лишь при условии

$$(\phi_{51}(x,\omega),g_k(x))_H = (-e_0u_{0\omega}(x,\omega)\sigma_0\Omega_1 - e_0u_{3\omega}(x,\omega)\sigma_0 + (A_2 - i\sigma_{*2}(\omega)I)e_0u_0(x,\omega) +$$

$$+(A_1 - i\sigma_{*1}(\omega)I)e_0u_3(x,\omega) + De_0u_{3xx}(x,\omega) + (2F_2(e_{20},\bar e_0;0) + 2F_2(e_{11},e_0;0) +$$
$$+ 3F_3(e_0,e_0,\bar e_0;0))(u_0(x,\omega)^2\bar u_3(x,\omega) + 2|u_0(x,\omega)|^2u_3(x,\omega)) + f_{32}(x,\omega),h_0g_k(x))_H = 0, \tag{5.20}$$

$k = 0, 1 \dots$. В первом уравнении системы (5.20) (при $k = 0$) приравняем нулю мнимую часть, выразим из полученного равенства $\sigma_{*2}(\omega)$ и подставим полученное выражение $\sigma_{*2}(\omega)$ в оставшиеся уравнения системы (5.20). Отметим, что $\sigma_{*2}(\omega)$ не зависит от $\dot{q}_*(\omega)$. В результате получим линейное дифференциальное уравнение вида (4.29) в пространстве l_2

$$\frac{dq_3(\omega)}{d\omega} = L(\omega;0)q_3(\omega) - \dot{q}_*(\omega)\sigma_0\Omega_1 + f^*_{32}(\omega), \tag{5.21}$$

в котором координаты вектора $f^*_{32}(\omega) \in l^\infty_2$ определяются величинами $(f_{32}(x,\omega), h_0 g_k(x))_H, k = 0, 1, \dots$, с учетом выражения для $\sigma_{*2}(\omega)$. При сделанных предположениях относительно свойств уравнения (4.23) (оператора $L(\omega;0)$) уравнение (5.21) имеет T-периодическое решение только при условии (4.32), которое в нашем случае с учетом равенства (4.31) примет вид

$$\Omega_1 = (\sigma_0 T)^{-1} \int_0^T (f^*_{32}(\omega), p_*(\omega;0))_{l_2} d\omega. \tag{5.22}$$

Выбрав Ω_1 в соответствии с (5.22), определим T-периодическое решение $q_3(\omega)$ уравнения (5.21), удовлетворяющее условию

$$\int_0^T (q_3(\omega), p_*(\omega;0))_{l_2} d\omega = 0,$$

которое обеспечивает его единственственность и согласно (4.33) представление в виде

$$q_3(\omega) = G(0)(-\dot{q}_*(\omega)\Omega_1 + f^*_{32}(\omega)), \tag{5.23}$$

где $G(0) : C^{\circ 1}(T) \to C^{\circ 2}(T)$ – линейный вполне непрерывный оператор. Выбрав таким образом $q_3(\omega)$, получим разрешимость уравнения (5.19), а его общее решение с учетом (5.15) будет иметь вид

$$e_{32}u_{32}(x,\omega) + e_0 u_5(x,\omega) = \bar{e}_0 \phi_{52}(x,\omega)/(2i\sigma_0) + e_0 \sum_{k=0}^{\infty} q_{5k}(\omega) g_k(x), \tag{5.24}$$

где $q_5(\omega) = (q_{50}(\omega), q_{51}(\omega), \dots) \in l^\infty_2$ – произвольный бесконечно дифференцируемый по ω T- периодический вектор, который будет определен из уравнений, полученных при $\varepsilon^{7/2}$. Этот процесс можно продолжать до бесконечности, на каждом последующем шаге будем иметь уравнения вида (5.20), (5.21), а соответствующие их решения будут иметь вид (5.23), (5.24).

Алгоритм построения инвариантного тора $\mathbb{T}^{k+1}$ $u_{\circ k}(x, \tau, \omega_1, \dots, \omega_k; \varepsilon^{1/2})$ вида (4.43) и уравнений траекторий на нем (4.44), (4.45) аналогичен приведенному. Для него также выполнены условия (5.1).

В случае аналитической зависимости матрицы $A(\varepsilon)$ и функции $F(u;\varepsilon)$ в (2.1) от ε и u при $|\varepsilon| < \varepsilon_0$ и $||u|| < R_0$ ряды в представлениях (5.5), (5.6) и (5.15), (5.16), а также в соответствующих разложениях для (4.43) и (4.44), (4.45) будут сходящимися при $0 < \varepsilon < \varepsilon_0$.

СПИСОК ЛИТЕРАТУРЫ

1. *Бутузов В.Ф., Васильева А.Б, Нефедов Н.Н.* Асимптотическая теория контрастных структур (обзор) // Автоматика и телемеханика. 1997. № 7. С. 4–31.
2. *Васильева А.Б, Бутузов В.Ф., Нефедов Н.Н.* Сингулярно возмущенные задачи с пограничными и внутренними слоями // Тр. матем. ин.та им. В.А. Стеклова. 2010. Т. 268. С. 268–283.
3. *Бутузов В.Ф., Нефедов Н.Н., Шнайдер К.Р.* Сингулярно возмущенные задачи в случае смены устойчивости // Итоги науки и техн. Сер. Соврем. матем. и ее прилож. Тематические обзоры. 2002. Т. 109. С. 5–242.
4. *Нефедов Н.Н.* Развитие методов асимптотического анализа переходных слоев в уравнениях реакции–диффузии–адвекции: теория и применение // Ж. вычисл. матем. и матем. физ. 2021. Т. 61. № 2074–2094.
5. *Васильева А.Б., Кащенко С.А., Колесов Ю.С., Розов Н.Х.* Бифуркация автоколебаний нелинейных параболических уравнений с малой диффузией // Матем. сб. 1986. Т. 130. № 4. С. 172.
6. *Кащенко С.А.* О квазинормальных формах для параболических уравнений с малой диффузией // Докл. АН СССР. 1988. Т. 299. № 5. С. 1049–1052.
7. *Кащенко С.А.* Пространственные особенности высокомодовых бифуркаций двухкомпонентных систем с малой диффузией // Дифференц. уравнения. 1989. Т. 5. № 2. С. 262–270.
8. *Кащенко С.А.* Простейшие критические случаи в динамике нелинейных систем с малой диффузией // Тр. ММО. 2018. Т. 79. № 1. С. 97–115.

9. *Колесов Ю.С.* Бифуркация инвариантных торов параболических систем с малой диффузией // Матем. сб. 1993. Т. 184. № 3. С. 121–136.
10. *Колесов А.Ю., Розов Н.Х., Садовничий В.А.* О проблеме возникновения автоволн в параболических системах с малой диффузией // Матем. сб. 2007. Т. 198. № 11. С. 67–106.
11. *Мищенко Е.Ф., Садовничий В.А., Колесов А.Ю., Розов Н.Х.* Автоволновые процессы в нелинейных средах с диффузией. М.: Физматлит, 2005. 432 с.
12. *Нефедов Н.Н.* Периодические контрастные структуры в задаче реакция–диффузия с быстрой реакцией и малой диффузией // Матем. заметки. 2022. Т. 112. № 4. С. 601–612.
13. *Крылов Н.М., Боголюбов Н.Н.* Введение в нелинейную механику. Киев: Из-во АН УССР, 1937. 352 с.
14. *Хейл Дж.* Колебания в нелинейных системах. М.: Мир, 1966. 230 с.
15. *Kubyshkin E.P., Moriakova A.R.* Features of Bifurcations of Periodic Solutions of the Ikeda Equation // Rus. J. Nonlin. Dyn. 201. V. 14. № 3. P. 301–324.
16. *Кубышкин Е.П., Морякова А.Р.* Особенности бифуркаций периодических решений уравнения Мэкки–Гласса // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 8. С. 1340–1357.
17. *Kubyshkin E.P., Moriakova A.R.* Analysis of special cases in the study of bifurcations of periodic solutions of the ikeda equation // Rus. J. Nonlin. Dyn. 2020. V. 16. № 3. P. 437–451.
18. *Соболевский П.Е.* Об уравнениях параболического типа в банаховом пространстве // Тр. ММО. 1961. Т. 10. С. 297–350.
19. *Красносельский М.А., Вайникко Г.М., Забрейко П.П. и др.* Приближенные методы решения операторных уравнений. М.: Наука, 1969.
20. *Крейн С.Г.* Линейные дифференциальные уравнения в банаховом пространстве. М.: Наука, 1967. 464 с.

THE AVERAGING METHOD IN THE PROBLEM OF CONSTRUCTING SELF-OSCILLATORY SOLUTIONS OF DISTRIBUTED KINETIC SYSTEMS

E. P. Kubyshkin*

150000 Yaroslavl, Sovetskaya str., 14, P.G.Demidov Yaroslavl State University, Faculty of Mathematics, Russia
**e-mail: kubysh.e@yandex.ru*

Received: 28.11.2023
Revised: 29.07.2024
Accepted: 23.08.2024

Abstract. An averaging method is constructed for two-component distributed kinetic systems with low diffusion in a limited one-dimensional region with impermeability conditions at the boundary. Transformations of the considered distributed system are constructed, which make it possible to allocate one “fast” and a countable number of “slow” variables. Theorems on the correspondence of stationary and periodic solutions, as well as invariant tori of averaged equations of “slow” variables, respectively, to spatially inhomogeneous periodic solutions and invariant tori of initial equations of a similar stability character are proved. Algorithms for constructing periodic solutions (cycles) and invariant tori of the initial equations in the form of a power expansion of a small parameter are proposed, providing the construction of asymptotic formulas for these self-oscillating objects. The conditions for convergence of the corresponding expansions are formulated.

Keywords: averaging method, distributed kinetic systems, reaction-diffusion equation systems, spatially inhomogeneous solutions, bifurcation theory.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2371–2377

УДК 517.956

МЕТОД МАЛОГО ПАРАМЕТРА В ТЕОРИИ УРАВНЕНИЙ ТИПА БЮРГЕРСА[1)]

© 2024 г. В. И. Качалов[1,*], Д. А. Маслов[1,]**

[1] 111250 Москва, ул. Красноказарменная, 14, НИУ «МЭИ», Россия
**e-mail: vikachalov@rambler.ru*
***e-mail: maslovdma@mpei.ru*

Поступила в редакцию 04.06.2024 г.
Переработанный вариант 30.07.2024 г.
Принята к публикации 23.08.2024 г.

Введенное Г. Бейтманом в 1915 г. и изученное Й. М. Бюргерсом в 1948 г. уравнение Бюргерса нашло широкое применение в механике жидкости, нелинейной акустике и других областях прикладной математики. Подходы к его решению были самые разнообразные: асимптотические, численные, аналитические. В данной работе развивается аналитический метод решения уравнения типа Бюргерса в банаховом пространстве. А именно, после искусственного введения в уравнение малого параметра доказывается существование аналитического по этому параметру решения. При этом, рассматривается также и многомерный вариант уравнения Бюргерса. Библ. 16.

Ключевые слова: уравнение Бюргерса, ε-регулярное решение, сильно непрерывная полугруппа, функция Грина.

DOI: 10.31857/S0044466924120104, **EDN**: KBRHHS

1. ВВЕДЕНИЕ

Известное уравнение вязкого Бюргерса в одном пространственном измерении для поля скоростей $u(x,t)$ и коэффициента диффузии ν

$$\partial_t u + u\partial_x u = \nu\partial_x^2 u, \quad u(x,0) = \varphi(x) \tag{1}$$

является одним из эталонных нелинейных уравнений математической физики. Оно является модельным для описания одномерной турбулентности, а в нелинейной акустике уравнение Бюргерса описывает распространение одномерных акустических волн конечной амплитуды в условиях проявления диссипации. В связи с тем, что это уравнение описывает довольно большое число различных по своей природе физических явлений, появилось много работ, посвященных анализу и методам его решений [1]–[3].

Уникальность одномерного уравнения Бюргерса прежде всего состоит в том, что с помощью замены Хопфа-Коула

$$u(x,t) = -2\nu\partial_x[\ln v(x,t)] \tag{2}$$

уравнение (1) сводится к уравнению теплопроводности

$$\begin{aligned} &\partial_t v = \nu\partial_x^2 v, \ \ t > 0, \\ &v(x,0) = \exp\left(-\tfrac{1}{2\nu}\int_0^x \varphi(\xi)d\xi\right). \end{aligned} \tag{3}$$

Тогда решение начальной задачи (1) следующее:

$$u(x,t) = \frac{\int_{-\infty}^{+\infty} \frac{x-y}{t}\exp\left(-\frac{G(y,x,t)}{2\nu}\right)dy}{\int_{-\infty}^{+\infty}\exp\left(-\frac{G(y,x,t)}{2\nu}\right)dy}, \tag{4}$$

где

$$G(y,x,t) = \frac{(x-y)^2}{2t} + \int\limits_0^y \varphi(\xi)d\xi.$$

[1)] Работа выполнена при финансовой поддержке РНФ (проект № 23-21-00496).

В работе [4] для начальной задачи (1) при $\nu = 1$ при условии, что $\psi(x) = v(x,0)$ является целой функцией порядка $0 < \rho < 2$, построено мероморфное по x решение

$$u(x,t) = -2\frac{\psi'(x) + t\psi''(x) + \ldots + \frac{t^n}{n!}\psi^{(2n+1)}(x) + \ldots}{\psi(x) + t\psi'(x) + \ldots + \frac{t^n}{n!}\psi^{(2n)}(x) + \ldots}. \tag{5}$$

Чтобы получить решение (5) методом малого параметра была решена задача

$$\begin{aligned} &\partial_t u + u\partial_x u = \nu\partial_x^2 u, \ \ t > 0, \\ &u(0,x) = \nu\varphi(x), \end{aligned} \tag{6}$$

а затем положили $\nu = 1$.

Следует также отметить, что развитие получил и метод решения краевых задач для уравнения Бюргерса [5, 6]. В представленной работе будет изучаться уравнение типа Бюргерса в банаховом пространстве.

2. ПОСТАНОВКА ЗАДАЧИ И ε-РЕГУЛЯРНЫЕ РЕШЕНИЯ

Пусть в банаховом пространстве E задана начальная задача

$$\begin{aligned} &\partial_t u + \varepsilon B(u, Hu) = Au, \ \ t \in (0,T], \\ &u\big|_{t=0} = u^0, \end{aligned} \tag{7}$$

в которой A и H — неограниченные линейные операторы; $B(u,v)$ — ограниченный билинейный оператор.

Уравнение (7) называется уравнением типа Бюргерса. Если в нем отсутствует малый параметр, и $u^0 = \varepsilon v^0$ (см. начальное условие в задаче (6)), то его можно ввести, пользуясь билинейностью оператора B с помощью замены $u = \varepsilon v$.

Сформулируем условия на данные задачи (7).

1°. Оператор $A : E \to E$ — неограниченный замкнутый и имеет непрерывный обратный A^{-1}.

2°. Оператор A — инфинитезимальный генератор сильно непрерывной полугруппы $\mathcal{U}(t)$.

3°. H — неограниченный замкнутый оператор с областью определения $D_H \supset D_A$.

Будем искать решение уравнения (7) в виде формального ряда по степеням малого параметра — это следствие того, что, благодаря условию 3°, задача является регулярно возмущенной:

$$u^f(t,\varepsilon) = u_0(t) + \varepsilon u_1(t) + \ldots + \varepsilon^n u_n(t) + \ldots . \tag{8}$$

Если ряд (8) подставить в уравнение (7) и приравнять коэффициенты при одинаковых степенях ε, то получится следующая серия задач:

$$\begin{aligned} &\partial_t u_0 = Au_0, \ \ u_0\big|_{t=0} = u^0, \\ &\partial_t u_1 = Au_1 - B(u_0, Hu_0), \ \ u_1\big|_{t=0} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\partial_t u_n = Au_n - \sum_{k=0}^{n-1} B(u_k, Hu_{n-k-1}), \ \ u_n\big|_{t=0} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{9}$$

Определение. Если ряд (8) с коэффициентами, определяемыми из серии (9), сходится равномерно на отрезке $[0,T]$ в некоторой окрестности точки $\varepsilon = 0$, то его сумма $u^r(t,\varepsilon)$ называется ε-регулярным решением задачи Коши (7).

В работе [7] приведены достаточные условия существования ε-регулярного решения, а в работе [8] указан способ построения приближений к точному решению $u(t,\varepsilon)$ ε-регулярными. Заметим, что $u^r(t,\varepsilon)$ совпадает с точным решением $u(t,\varepsilon)$ лишь тогда, когда $u^r(t,\varepsilon)$ будет иметь сильную производную по $t \in (0,T]$ и $u^r(t,\varepsilon) \in D_A$.

Для дальнейшего нам понадобится следующая

Лемма. *Пусть последовательность $\{a_n\}_{n=0}^{\infty}$ задана рекуррентно: $a_0 = 1$, $a_1 = 1, \ldots, a_n = a_0a_{n-1} + a_1a_{n-2} + \ldots + a_{n-2}a_1 + a_{n-1}a_0, \ldots$. Тогда производящая функция этой последовательности голоморфна в круге радиуса $1/4$.*

Доказательство. Пусть

$$\varphi(z) = a_0 + a_1 z + \ldots + a_n z^n + \ldots \tag{10}$$

есть производящая функция данной последовательности. Имеем

$$\varphi^2(z) = a_0^2 + (a_0a_1 + a_1a_0)z + \ldots + (a_0a_{n-1} + a_1a_{n-2} + \ldots + a_{n-2}a_1 + a_{n-1}a_0)z^{n-1} + \ldots ,$$

т.е.

$$\varphi^2(z) = a_1 + a_2 z + a_3 z^3 + \ldots + a_n z^n + \ldots . \tag{11}$$

Из (10) и (11) вытекает уравнение для $\varphi(z)$:

$$z\varphi^2(z) - \varphi(z) + 1 = 0,$$

из которого следует, что

$$\varphi(z) = (2z)^{-1}(1 - \sqrt{1-4z}).$$

Голоморфность производящей функции очевидна. Итак,

$$\varphi(z) = 1 + z + 2z^2 + 5z^3 + 14z^4 + 42z^5 + 132z^6 + 429z^7 + \ldots ,$$

и лемма доказана.

Сформулируем и докажем основной результат работы.

Теорема. *При выполнении условий* $1^\circ - 3^\circ$ ε*-регулярное решение* $u^r(t,\varepsilon)$ *существует, единственно и совпадает с точным решением* $u(t,\varepsilon)$.

Доказательство. В соответствии с условием 2°, все задачи серии (9) корректно разрешимы и

$$\begin{aligned}
&u_0(t) = \mathcal{U}(t)u^0, \;\; v_0(t) = A[u_0(t)] = A\mathcal{U}(t)[u^0];\\
&u_1(t) = \int\limits_0^t \mathcal{U}(t-\tau)B(u_0, Hu_0)d\tau, \;\; v_1(t) = A[u_1(t)] = A\left[\int\limits_0^t \mathcal{U}(t-\tau)B(u_0, Hu_0)d\tau\right];\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&u_n(t) = \int\limits_0^t \mathcal{U}(t-\tau)\left(\sum_{k=0}^{n-1} B(u_k, Hu_{n-k-1})\right)d\tau,\\
&\qquad\qquad v_n(t) = A[u_n(t)] = A\left[\int\limits_0^t \mathcal{U}(t-\tau)\left(\sum_{k=0}^{n-1} B(u_k, Hu_{n-k-1})\right)d\tau\right].\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{12}$$

Здесь, помимо $u_0(t), u_1(t), \ldots, u_n(t), \ldots$, введены функции $v_0(t), v_1(t), \ldots, v_n(t), \ldots$, которые понадобятся в дальнейшем.

Введем операторы $F(s) = A\mathcal{U}(s)$, $G = HA^{-1}$. Оба оператора являются ограниченными по теореме Банаха о замкнутом графике [9, гл. III, п. 15.4]. Действительно, $\mathcal{U}(s)$ является непрерывным в E и $\operatorname{Im}\mathcal{U}(s) \subset D_A$ при каждом $s \geqslant 0$, а значит, оператор $F(s)$ является замкнутым, как произведение замкнутого оператора на ограниченный оператор, и определен на всем пространстве E. То же самое можно сказать об операторе G: замкнутый, как произведение замкнутого оператора H на непрерывный оператор A^{-1}, и определенный на всем пространстве E ($\operatorname{Im} A^{-1} = D_A$, $D_A \subset D_H$). Представим функции $v_n(t)$, $n = 1, 2, \ldots$, из серии (12) с помощью операторов F и G:

$$\begin{aligned}
&v_1(t) = \int\limits_0^t F(t-\tau)B(A^{-1}v_0, Gv_0)d\tau,\\
&v_2(t) = \int\limits_0^t F(t-\tau)(B(A^{-1}v_0, Gv_1) + B(A^{-1}v_1, Gv_0))d\tau,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&v_n(t) = \int\limits_0^t F(t-\tau)\left(\sum_{k=0}^{n-1} B(A^{-1}v_k, Hv_{n-k-1})\right)d\tau,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{13}$$

Пусть $\|F(s)\| = f(s)$, $\|G\| = g$, $\|B\| = b$, $\|A^{-1}\| = \alpha$. Докажем методом математической индукции, что

$$\|v_n(t)\| \leqslant a_n t^n b^n f^n(s)\alpha^n g^n \|v_0\|^{n+1}. \tag{14}$$

При $n = 1$ неравенство (14), очевидно, выполняется. Пусть оно справедливо при $n = m$. Получим оценку для $v_{m+1}(t)$. Имеем

$$\|v_{m+1}(t)\| \leqslant tf(s)b\alpha g \sum_{k=0}^{m}(a_k t^k b^k f^k(s)\alpha^k g^k \|v_0(t)\|^{k+1} a_{m-k} t^{m-k} b^{m-k} f^{m-k}(s)\alpha^{m-k} g^{m-k}\|v_0\|^{m-k+1}) =$$

$$= \left(\sum_{k=0}^{m} a_k a_{m-k}\right) t^{m+1} b^{m+1} f^{m+1}(s) \alpha^{m+1} g^{m+1} \|v_0(t)\|^{m+2} = a_{m+1} t^{m+1} b^{m+1} f^{m+1}(s) \alpha^{m+1} g^{m+1} \|v_0(t)\|^{m+2}.$$

Неравенство (14) выполняется и для $n = m+1$ и, значит, доказано. Следовательно, ряд $\sum_{n=0}^{\infty} \varepsilon^n v_n(t)$ сходится в окрестности радиуса

$$r = \frac{1}{4T\alpha g b (\max_{s\in[0,T]} f(s))(\max_{t\in[0,T]} \|v_0(t)\|)}$$

точки $\varepsilon = 0$.

Таким образом, ряд

$$\sum_{n=0}^{\infty} \varepsilon^n A[u_n(t)] \tag{15}$$

сходится равномерно на отрезке $[0, T]$ при $|\varepsilon| < r$. Поскольку A — замкнутый оператор, то ряд (15) совпадает с рядом

$$A \sum_{n=0}^{\infty} \varepsilon^n u_n(t),$$

см. [9, гл. III, § 15], а так как A^{-1} — ограниченный оператор, то и ряд (8) сходится и его сумма $u(t, \varepsilon) \in D_A$, т.е. является решением уравнения (7). Теорема доказана.

Пример 1. В пространстве $L_2(\Omega)$, где $\Omega \subset \mathbb{R}^3$ — ограниченная область с кусочно-гладкой границей $\partial\Omega$, задана смешанная задача

$$\partial_t u + \varepsilon \iiint\limits_{\Omega} \mathcal{K}(x, y, z, x', y', z') u(x', y', z') (\partial_{x'} u + \partial_{y'} u + \partial_{z'} u) dx' dy' dz' = \partial_x^2 u + \partial_y^2 u + \partial_z^2 u, \tag{16}$$

$u\big|_{t=0} = \varphi(x, y, z)$, $u\big|_{\partial\Omega} = 0$.

Здесь оператор H является замыканием дифференциальной операции $\partial_x + \partial_y + \partial_z$, определенной на множестве $C^1(\Omega) \cap C(\overline{\Omega})$; A — замыкание лапласиана $\Delta = \partial_x^2 + \partial_y^2 + \partial_z^2$, заданного на множестве функций из $C^2(\Omega) \cap C(\overline{\Omega})$, обращающихся в ноль на границе $\partial\Omega$; ядро $\mathcal{K}(x, y, z, x', y', z') \in C^1(\overline{\Omega} \times \overline{\Omega})$ и положительно на компакте $\overline{\Omega} \times \overline{\Omega}$; билинейный оператор

$$B(v, w) = \iiint\limits_{\Omega} \mathcal{K}(x, y, z, x', y', z') v(x', y', z') w(x', y', z') dx' dy' dz',$$

действующий из $L_2(\Omega) \times L_2(\Omega)$ в $L_2(\Omega)$ является ограниченным, так как

$$\|B(v, w)\| \leqslant \max_{\overline{\Omega}\times\overline{\Omega}} \mathcal{K}(x, y, z, x', y', z') \iiint\limits_{\Omega} |v(x', y', z') w(x', y', z')| dx' dy' dz' \leqslant \max_{\overline{\Omega}\times\overline{\Omega}} \mathcal{K}(x, y, z, x', y', z') \|v\| \, \|w\|.$$

Пусть область Ω такова, что спектральная задача

$$\Delta w + \lambda w = 0, \quad w\big|_{\partial G} = 0$$

имеет собственные функции $\{w_m(x, y, z)\}_{m=1}^{\infty}$ из класса $C^2(\Omega) \cap C(\overline{\Omega})$ и образуют в $L_2(\Omega)$ ортонормированную систему; соответствующие им собственные значения $\lambda_m \to +\infty$, $m \to \infty$. В этом случае смешанная задача

$$\begin{aligned} &\partial_t v = \Delta v + f(x, y, z, t), \\ &v\big|_{t=0} = \varphi(x, y, z), \\ &v\big|_{\partial\Omega} = 0 \end{aligned}$$

имеет решение

$$v(x, y, z, t) = \sum_{m=1}^{\infty} e^{-\lambda_m t} \left[(\varphi(x, y, z), w_m(x, y, z)) + \int_0^t e^{\lambda_k \tau} (f(x, y, z, \tau), w_m(x, y, z)) d\tau \right] w_m(x, y, z) \tag{17}$$

и указанный ряд сходится регулярно в $\overline{\Omega}$. Непрерывная обратимость лапласиана в пространстве $L_2(\Omega)$ известна и для решения уравнения $\Delta w = g(x,y,z)$ имеет место формула

$$w(x,y,z) = \sum_{m=1}^{\infty} \lambda_m^{-1}(g, w_m)w_m. \tag{18}$$

Серия (9) для задачи (16) будет выглядеть следующим образом:

$$\begin{aligned}
&\partial_t u_0 = \Delta u_0, \;\; u_0\big|_{t=0} = \varphi(x,y,z);\\
&\partial_t u_1 = \Delta u_1 - \iiint\limits_{\Omega} \mathcal{K}(x,y,z,x_1,y_1,z_1)u_0(x_1,y_1,z_1,t)(\partial_{x_1} + \partial_{y_1} + \partial_{z_1})\cdot\\
&\qquad\qquad \cdot u_0(x_1,y_1,z_1,t)dx_1dy_1dz_1, \;\; u_1\big|_{\partial\Omega} = 0;\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\partial_t u_n = \Delta u_n - \sum_{k=0}^{n-1} \iiint\limits_{\Omega} \mathcal{K}(x,y,z,x_1,y_1,z_1)u_k(x_1,y_1,z_1,t)(\partial_{x_1} + \partial_{y_1} + \partial_{z_1})\cdot\\
&\qquad\qquad \cdot u_{n-k-1}(x_1,y_1,z_1,t)dx_1dy_1dz_1, \;\; u_n\big|_{\partial\Omega} = 0;\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

В соответствии с формулой (17) имеем:

$$u_0(x,y,z,t) = \sum_{m=1}^{\infty} e^{-\lambda_m t}(\varphi(x,y,z), w_m(x,y,z))w_m(x,y,z),$$

$$u_1(x,y,z,t) = -\sum_{m=1}^{\infty} \int\limits_0^t e^{-\lambda_m(t-\tau)} \times$$

$$\times \left(\iiint\limits_{\Omega} \mathcal{K}(x,y,z,x_1,y_1,z_1)u_0(x_1,y_1,z_1,\tau)(\partial_{x_1} + \partial_{y_1} + \partial_{z_1})u_0(x_1,y_1,z_1)dx_1dy_1dz_1, w_m(x,y,z) \right) w_m(x,y,z)$$

и т.д. В итоге будет построено решение задачи (16)

$$u(x,y,z,t,\varepsilon) = u_0(x,y,z,t) + \varepsilon u_1(x,y,z,t) + \ldots,$$

голоморфное в точке $\varepsilon = 0$.

Пример 2. Построим приближение к решению двумерного уравнения Бюргерса с малым параметром

$$\partial_t u + \varepsilon u(\partial_x u + \partial_y u) = \partial_x^2 u + \partial_y^2 u, \tag{19}$$

с начальным условием

$$u\big|_{t=0} = \psi(x,y), \;\; (x,y) \in \overline{\Pi}, \tag{20}$$

и краевым условием

$$u\big|_{\partial\Pi} = 0. \tag{21}$$

Здесь,

$$\Pi = \{(x,y) \in \mathbb{R}^2 : \;\; 0 < x < 1, \;\; 0 < y < 1\}$$

есть прямоугольник с границей $\partial\Pi$. Будем рассматривать задачу (18)–(20) в банаховом пространстве $E = C(\overline{\Pi})$, при этом оператор $H = \partial_x + \partial_y$ имеет область определения $D_H = \{v(x,y) \in C^1(\Pi) \cap C(\overline{\Pi}), \;\; v(0,y) = 0\}$, а оператор $A = \partial_x^2 + \partial_y^2$ имеет область определения $D_A = \{w(x,y) \in C^2(\Pi) \cap C(\overline{\Pi}), \;\; w\big|_{\partial\Pi} = 0\}$.

Замкнутость обоих операторов очевидна. Как известно [10, гл. 5, § 26], двумерная задача на собственные значения

$$\Delta u + \lambda u = 0, \quad u\big|_{\partial\Pi} = 0$$

имеет следующее решение: $\lambda_{kj} = \pi^2(k^2 + j^2)$ — собственные значения, $X_{kj} = 2\sin\pi kx \sin\pi jy$ — собственные функции.

Функция Грина первой краевой задачи

$$\begin{aligned}
&\partial_t w = \Delta w + g(x,y,t), \;\; (x,y) \in \Pi, \;\; t > 0,\\
&w\big|_{t=0} = \theta(x,y), \;\; w\big|_{\partial\Pi} = 0
\end{aligned} \tag{22}$$

строится в виде регулярно сходящегося при $t > 0$ ряда

$$G(x, y; \xi, \eta; t) = \sum_{k,j=1}^{\infty} e^{-\lambda_{kj} t} X_{kj}(x, y) X_{kj}(\xi, \eta),$$

при этом решение задачи (22) имеет следующий вид:

$$w(x, y, t) = \iint\limits_{\Pi} G(x, y; \xi, \eta; t)\theta(\xi, \eta) d\xi d\eta + \int\limits_0^t d\tau \iint\limits_{\Pi} G(x, y; \xi, \eta; t - \tau) g(\xi, \eta, \tau) d\xi d\eta. \tag{23}$$

Далее, фигурирующий в уравнении (19) билинейный оператор $B(u, v) = uv$ в пространстве E, очевидно, является ограниченным (в отличие от пространства $L_2(\Pi)$). Наряду с уравнением (19) рассмотрим уравнение

$$\partial_t v + \varepsilon v(H_m v) = \Delta v, \ \ m \in \mathbb{N}, \tag{24}$$

с теми же краевыми условиями, с ограниченными операторами [11, гл. I, § 2.2]

$$H_m = -mI - m^2 R_H(m),$$

сильно сходящимися к оператору H при $m \to +\infty$. Здесь

$$R_H(\lambda)[f] = \int\limits_0^x e^{\lambda(x-\xi)} f(\xi, \xi + y - x) d\xi$$

есть резольвента оператора H.

В работе [7] было доказано, что при каждом натуральном m уравнение (24) имеет ε-регулярное решение $v_m(x, y, t, \varepsilon)$, которое совпадает с его точным решением и, если $u = u(x, y, t, \varepsilon)$ — точное решение задачи (19)–(21), то

$$\|u(x, y, t, \varepsilon) - v_m(x, y, t, \varepsilon)\| \leqslant C_\varepsilon \|(H - H_m)u\|, \tag{25}$$

причем $C_\varepsilon \to 0$ при $\varepsilon \to 0$. Если мы выберем m столь большим, что $\|(H - H_m)u\| < \delta$ для заданного малого δ, то из неравенства (25) будет следовать, что решение v_m хорошо аппроксимирует решение исходной задачи при достаточно малом ε, порядка $1/\|H_m\|$.

Замечание. Если же $D_H \not\supset D_A$, т.е. возмущение, создаваемое билинейной частью уравнения (7), является сингулярным, то здесь уже применяются методы решения сингулярно возмущенных задач [12]–[14]. Для такого типа задач введено понятие ε-псевдорегулярного решения и разрабатываются как аналитические [15], так и асимптотические подходы [16].

СПИСОК ЛИТЕРАТУРЫ

1. *Burgers J.M.* A mathematical model illustraing the theory of turbulance // Advances in Applied Mechanics, 1, eds. R. von Mises. T. von Karman, New York: Acad. Press. 1948. P. 171–199.
2. *Pilant M.S., Rundell W.* An inverse problem for a nonlinear parabolic equation // Commun. Part. Differ. Equ. 1986. V. 11. № 4. P. 445–457.
3. *Henkin G.M.* Asimptotic structure for solutions of the Cauchy problem for Burgers type equations // J. Fixed Point Theory Appl. 2007. V. 1. № 2. P. 239–291.
4. *Качалов В.И., Федоров Ю.С.* О методе малого параметра в нелинейной математической физике // Сиб. электрон. матем. изв. 2018. Т. 15. С. 1680–1686.
5. *Нефедов Н.Н., Руденко О.В.* О движении, усилении и разрушении фронтов в уравнениях типа Бюргерса с квадратичной и модульной нелинейностью // Докл. АН. Матем., информ., проц. упр. 2020. Т. 493. С. 26–31.
6. *Волков В.Т., Нефедов Н.Н.* Асимптотическое решение коэффициентных обратных задач для уравнений типа Бюргерса // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. № 6. С. 975–984.
7. *Качалов В.И., Маслов Д.А.* Аналитичность и псевдоаналитичность в методе малого параметра // Ж. вычисл. матем. и матем. физ. 2023. Т. 63. № 11. С. 1806–1814.

8. *Качалов В.И.* Об ε-регулярных решениях дифференциальных уравнений с малым параметром // Сиб. матем. журнал. 2023. Т. 64. № 1. С. 113–122.

9. *Треногин В.А.* Функциональный анализ. М.: Наука, 1980.

10. *Владимиров В.С.* Уравнения математической физики. М.: Наука, 1988.

11. *Крейн С.Г.* Линейные дифференциальные уравнения в банаховом пространстве. М.: Наука, 1967.

12. *Ломов С.А., Ломов И.С.* Основы математической теории пограничного слоя. М.: Изд-во МГУ, 2011.

13. *Васильева А.Б., Бутузов В.Ф.* Асимптотические разложения решений сингулярно возмущенных задач. М.: Наука, 1973.

14. *Нефедов Н.Н.* Периодические контрастные структуры в задаче реакция-диффузия с быстрой реакцией и малой диффузией // Матем. заметки. 2022. Т. 112. № 4. С. 601–612.

15. *Качалов В.И.* Псевдоголоморфные и ε-псевдорегулярные решения сингулярно возмущенных задач // Дифференц. ур-ния. 2022. Т. 58. № 3. С. 361–370.

16. *Bobodzhanov A.A., Safonov V.F., Kachalov V.I.* Asymptotic and pseudoholomorphic solutions of singularly perturbed differential and integral equations in the Lomov's regularization method // Axioms. 2019. V. 8. № 27. https://doi.org/10.3390/axioms8010027

THE SMALL PARAMETER METHOD IN THE THEORY OF BURGERS-TYPE EQUATIONS

V. I. Kachalov*, D. A. Maslov*

111250 Moscow, Krasnokazarmennaya str., 14, NRU Moscow Energy Institute
**e-mail: vikachalov@rambler.ru*
***e-mail: maslovdma@mpei.ru*

Received: 04.06.2024
Revised: 30.07.2024
Accepted: 23.08.2024

Abstract. Introduced by G. Bateman in 1915 and studied by J. M. Burgers in 1948, the Burgers equation has found wide application in fluid mechanics, nonlinear acoustics and other fields of applied mathematics. The approaches to its solution were very diverse: asymptotic, numerical, and analytical. In this paper, an analytical method for solving a Burgers-type equation in a Banach space is developed. Namely, after artificially introducing a small parameter into the equation, the existence of an analytical solution for this parameter is proved. At the same time, a multidimensional version of the equation is also considered.

Keywords: Burgers equation, ε-regular solution, strongly continuous semigroup, Green's function.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2378–2389

УДК 519.642

ОБ ОДНОЗНАЧНОСТИ ОПРЕДЕЛЕНИЯ СЕТОЧНОГО ФУНДАМЕНТАЛЬНОГО РЕШЕНИЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ И ВОЛНОВОГО УРАВНЕНИЯ В РАМКАХ ТЕОРИИ ДИСКРЕТНОГО ПОТЕНЦИАЛА[1)]

© 2024 г. И.Э. Степанова[1,*], И.И. Колотов[2], А.В. Щепетилов[2], А.Г. Ягола[2], А.Н. Левашов[2]

[1] 123242 Москва, ул. Б.Грузинская, 10, стр. 1, Институт физики Земли им. О.Ю. Шмидта РАН, Россия
[2] 119992 Москва, Ленинские горы, МГУ им. М.В. Ломоносова, Физический факультет, Россия
**e-mail: tet@ifz.ru*

Поступила в редакцию 10.07.2024 г.
Переработанный вариант 10.07.2024 г.
Принята к публикации 23.08.2024 г.

В работе рассматривается проблема однозначного определения фундаментального решения сеточного аналога волнового уравнения, а также уравнения теплопроводности в рамках теории дискретного потенциала. Сеточные фундаментальные решения конечно-разностных аналогов уравнений в частных производных позволяют решать прямые и обратные задачи по восстановлению источников волн и тепла в различных средах по разнородной и разноточной информации о соответствующих физических полях. В статье рассматриваются постановки с условиями Дирихле в трехмерном и четырехмерном декартовых пространствах. Библ. 16.

Ключевые слова: однозначное определение, фундаментальное решение, дискретный тепловой потенциал, дискретный волновой потенциал.

DOI: 10.31857/S0044466924120114, **EDN**: KBPMXT

ВВЕДЕНИЕ

Спутниковое зондирование Земли и объектов в пределах Солнечной системы бросило целый ряд вызовов исследователям. В первую очередь, возникла проблема адекватной интерпретации больших и сверхбольших массивов меняющихся со временем (в том числе – “осциллирующих”) данных. Во-вторых, актуальным остается вопрос о целесообразности записи и хранения тех или иных выборок: объемы памяти на соответствующих устройствах обширны, но не безграничны. Поэтому требуется определить еще на этапе получения информации, какая ее часть будет использоваться для решения задач интерпретационного и иного характера, а какая будет отброшена, поскольку фактически не влияет на качество решения поставленной задачи.

При применении аппроксимационного подхода к решению обратных линейных и нелинейных задач геофизики, геодезии и геоморфологии [1–4] практически все постановки по определению параметров геологической среды требуют решения больших и очень больших систем линейных (в некоторых случаях и нелинейных) алгебраических уравнений. Свойства матриц таких систем (СЛАУ) зависят от типа оператора, фигурирующего в постановке прямой задачи, заключающейся в вычислении компонент некоторого векторного поля при известных зависимостях от координат и времени плотностей распределения источников этого поля в среде. Для решения обратной задачи важное значение приобретает резольвента оператора, описывающего создаваемое источниками физическое поле (поле скоростей, давлений, температур и т.п.) – фактически, матрица, обратная к матрице оператора прямой задачи за вычетом единичной матрицы, умноженной на некоторое комплексное, в общем случае, число.

При интерпретации геофизических данных необходимо учитывать дискретный характер передаваемой информации, даже в случае спутниковых систем высокого разрешения, как это было отмечено уже в [2]. Гладкость функций, представляющих свойства реальных физических полей, имеет решающее значение при постановке задач (как прямых, так и обратных) в пространствах с непрерывно меняющимися координатами и временем. Однако a priori бывает, как правило, не известно, каков характер зависимости плотности источников от пространственных координат и насколько резко эта плотность может варьироваться с течением времени.

[1)] Работа выполнена в рамках госзадания ИФЗ РАН.

Поэтому основные принципы теории дискретного гравитационного и магнитного потенциалов, предложенные В.Н. Страховым в начале 1990-х годов и развитые далее в работах [1, 3], могут, на наш взгляд, быть распространены на постановки прямых и обратных задач для уравнений с частными производными гиперболического и параболического типов. Вместе с теорией дискретного потенциала можно применять следующий метод линейных интегральных представлений (см. [4–6]); такой подход позволит обеспечить лучшую корреляцию между математическими моделями и геофизическими реалиями.

1. ТЕОРИЯ ДИСКРЕТНОГО ПОТЕНЦИАЛА В СЛУЧАЕ ВОЛНОВОГО УРАВНЕНИЯ И УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

В [1] описываются основные принципы теории теории дискретного гравитационного потенциала. Роль вектора с координатами в декартовом пространстве $\mathbb{R}^n : x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, “играет” его сеточный аналог. Общее число сеток равно 2^n. Из этих сеток одна считается основной, и в узлах этой сетки определены значения сеточного гравитационного потенциала $V_S(x^{(S)})$, где $x^{(S)}$ имеет следующий вид:

$$x^{(S)} = \left(x_1^{(S)}, x_2^{(S)}, \ldots, x_n^{(S)}\right)^{\mathrm{T}}, \quad x_k^{(S)} = mh_k, \quad m = 0, \pm 1, \pm 2, \ldots; \quad h_k = \mathrm{const},$$
$$h = (h_1, h_2, \ldots, h_n)^{\mathrm{T}}.$$

В узлах вспомогательных, или “неосновных”, сеток вычисляются значения высших производных потенциала. Если $\mathbf{e}_k$ — единичный орт вдоль k-й оси, то первая производная гравитационного потенциала запишется в виде

$$\frac{\partial V_S(x^{(S)})}{\partial x_k^{(S)}} \approx \frac{V_S(x^{(S)} + h_k e_k) - V_S(x^{(S)})}{h_k}. \tag{1}$$

Как показано в [1], значения производной по k-й оси относятся к узлам $x^{(S,k)}$, $k = 1, 2, \ldots, n$, вспомогательной сетки:

$$x^{(S,k)} = x^{(S)} + \frac{h_k}{2} e_k, \quad k = 1, 2, \ldots, n.$$

Аналогично, можно записать следующие выражения для вторых производных сеточного потенциала:

$$\frac{\partial^2 V_S(x^{(S)})}{\partial x_p^{(S)} \partial x_q^{(S)}} = \frac{1}{h_q}\left[\frac{\partial V_S(x^{(S)} + h_p e_p)}{\partial x_p^{(S)}} - \frac{\partial V_S(x^{(S)})}{\partial x_p^{(S)}}\right], \tag{2}$$

Значения $\frac{\partial^2 V_S(x^{(S)})}{\partial x_p^{(S)} \partial x_q^{(S)}}$ при $p \neq q$ относятся к узлам третьей дополнительной сетки: $x^{(S,3)} = x^{(S)} + \frac{h_p e_p}{2} + \frac{h_q e_q}{2}$; если же $p = q$, то разностные аналоги вторых производных вычисляются в узлах основной сети.

В настоящей работе, в отличие от [1], будем полагать, что одна из декартовых сеточных координат представляет собой время. Для удобства записи положим, что мы работаем в $\mathbb{R}^{n+1} = \mathbb{R}_x^n \times \mathbb{R}_t$. Первый сомножитель в декартовом произведении пространств соответствует пространственным координатам, а второй – времени. Тогда вид сеточных дифференциальных операторов практически не изменится, но нужно будет учитывать, что одна из координат $n+1$-мерного пространства – выделенная.

Аналоги операторов для уравнения теплопроводности и волнового уравнения в сеточном пространстве могут быть записаны следующим образом:

$$\begin{aligned} &\Delta_x \left\{V_S(x^{(S)}\right\} a^2 = \nabla_t \left\{V_S(x^{(S)}\right\}; \quad x^{(S)} \in R^{n+1}; \\ &\Delta_x \left\{V_S(x^{(S)}\right\} a^2 = \Delta_t \left\{V_S(x^{(S)}\right\}; \quad x^{(S)} \in R^{n+1}. \end{aligned} \tag{3}$$

В (3) Δ_x — выбранный конечно-разностный аналог оператора Лапласа по пространственным переменным, ∇_t — конечно-разностная аппроксимация первой производной по времени, Δ_t -конечно-разностный аналог оператора второй производной по времени. В дальнейшем будем обозначать дифференциальные операторы в частных производных через

$$\begin{aligned} L^{(h)}\left\{V_S(x^{(S)})\right\} &\equiv a^2 \Delta_x \left\{V_S(x^{(S)})\right\} - \nabla_t \left\{V_S(x^{(S)})\right\}; \\ L^{(W)}\left\{V_S(x^{(S)})\right\} &\equiv a^2 \Delta_x \left\{V_S(x^{(S)})\right\} - \Delta_t \left\{V_S(x^{(S)})\right\}. \end{aligned} \tag{4}$$

В (4) верхняя строка соответствует уравнению теплопроводности, а вторая – волновому уравнению.

Примем, что шаги по пространственным координатам будут обозначаться как h_p, $p = 1, \ldots, N$, а шаг по времени обозначим через τ, пусть a^2 — некоторый параметр, фигурирующий в континуальных уравнениях теплопроводности или волновом – он имеет смысл коэффициента теплопроводности, диффузии или квадрата скорости распространения сигнала.

Как мы уже отмечали в [2], в теории дискретного потенциала важнейшей методологической установкой считается формулировка вариационной задачи, в которой наряду с условиями (3) рассматривается дополнительное требование:

$$\left| L_2^{(w,h)} \left\{ V_S(x^{(S)}) \right\} \right|^2_{E(D_S)} + \alpha \Phi \left(V_S(x^{(S)}), u_S(x^{(S)}) \right) = \min_{V_S(x^{(S)}), x^{(S)} \in \mathrm{Int} D_S}, \tag{5}$$

где $\alpha \geqslant 0$ — некоторый параметр (параметр регуляризации [7]); D_S — заданная сеточная область (в четырехмерном пространстве !); $x^{(S)} \in D_S$; $E(D_S)$ — евклидова норма в сеточной области D_S; $\Phi(\mathbf{t}, \mathbf{z})$ — неотрицательный функционал на векторах одинаковой размерности $\mathbf{t}$ и $\mathbf{z}$. Через $u_S(x^{(S)})$ в (5) обозначен известный благодаря наблюдениям сигнал (меняющийся по гармоническому закону или температурное нестационарное поле); $L_2^{(w,h)}$ — это аппроксимация дифференциального оператора (в [2] — оператора Лапласа) на другом шаблоне, отличная от $L^{(w,h)} = L_1^{(w,h)}$. Вариационную задачу (3)–(5) можно решать одним из методов, описанных, например, в [8]. Но в наших предыдущих работах (см. [4–6]) мы показали, что при решении больших плохо обусловленных систем линейных алгебраических уравнений (СЛАУ) целесообразно применять специальные методы, которые позволяют эффективно устранять влияние помехи в правой части и матрице.

Система уравнений (3) не является, очевидно, замкнутой: она не позволяет определять сеточные потенциалы однозначно во всем пространстве (даже в некоторой ограниченной четырехмерной сеточной области ее решение будет неединственным). Вопросы однозначной разрешимости СЛАУ имеют принципиальное значение для адекватной реальности интерпретации геофизических данных [9–11]. Чем больше неопределенность в выборе эквивалентного по внешнему полю распределения масс, тем больше несоответствие между выбранной исследователем математической моделью геологической среды и ее “прототипом”, если можно так выразиться. Если априорная информация об объекте, порождающем поле, отсутствует, то выводы о локализации и геометрии источника можно назвать лишь умозрительными: источник сигнала будет представлять собой в таком случае некий “черный ящик”. Поэтому, прежде чем переходить к формулировке краевых задач для выделения единственного сеточного аналога волнового или теплового потенциала, необходимо рассмотреть принципиально важный для дальнейшего вопрос нахождения фундаментального решения разностного аналога оператора $L^{(w)}$ или $L^{(h)}$.

Дискретными аналогами волнового и теплового потенциалов, а также фундаментальных решений волнового уравнения и уравнения теплопроводности служат следующие выражения:

$$V_i^{(S)}(x^{(s)}) = \sum_{\xi^{(S)} \in J^{(S)}} m_S \left(\xi^{(S)} \right) \Omega_{n,i}^{(S)}(\xi^{(S)} - x^{(s)}), \quad i = 1, 2; \tag{6}$$

$$\begin{aligned} a^2 \Delta_x \left(\Omega_{n,1}^{(S)}(x^{(s)}) \right) - \nabla_t \left(\Omega_{n,1}^{(S)}(x^{(s)}) \right) &= -\frac{C_n}{H^{n-2}} e \left(x^{(S)} \right), \\ a^2 \Delta_x \left(\Omega_{n,2}^{(S)}(x^{(s)}) \right) - \Delta_{tt} \left(\Omega_{n,2}^{(S)}(x^{(s)}) \right) &= -\frac{C_n}{H^{n-2}} e \left(x^{(S)} \right); \end{aligned} \tag{7}$$

$$e(x^{(S)}) = \begin{cases} 1, & x^{(S)} = 0, \\ 0, & |x^{(S)}| > 0. \end{cases} \tag{8}$$

В (6) J_S есть совокупность векторов координат $\xi^{(S)}$ тех узлов сетки, в которых имеются ненулевые сеточные источники волн или тепла, обозначенные условно через $m_S \left(\xi^{(S)} \right)$; иначе говоря, J_S есть сеточный носитель источников поля. Функция $\Omega_{n,1}^{(S)}(x^{(s)})$ есть сеточный аналог фундаментального решения волнового уравнения в континуальной теории, $\Omega_{n,2}^{(S)}(x^{(s)})$ — сеточный аналог фундаментального решения уравнения теплопроводности, Δ_x — выбранный, как и в (3), конечно-разностный аналог оператора Лапласа. Ясно, что $\Omega_{n,i}^{(S)}(x^{(s)})$, $i = 1, 2$, играют роль сеточных фундаментальных решений в $\mathbb{R}^{n+1} = \mathbb{R}_x^n \times \mathbb{R}_t$. Сам индекс n у фундаментального решения волнового уравнения или уравнения теплопроводности соответствует размерности координатного пространства. Определенную в (8) правую часть уравнения, которому удовлетворяет сеточное фундаментальное решение, можно трактовать либо как тепло, выделившееся в начале координат в начальный момент времени (в случае уравнения теплопроводности), либо источник волн, также испустивший импульс при $t = 0$.

Условиями (6)—(8) сеточное фундаментальное решение уравнений (волнового и теплопроводности) ни в одной из размерностей однозначно не определяется. В [1] В.Н. Страхов выдвинул гипотезу о том, что имеет место единственность решения задачи по определению сеточного фундаментального решения уравнения Лапласа в $\mathbb{R}^3$, близкого по евклидовой норме к потенциалу призмы, заданной в области

$$|\bar{r}| \leqslant \frac{h}{2},\ \bar{r} = (x_1, x_2, x_3)^{\mathrm{T}}, \quad |\bar{r}| \equiv r = \sqrt{x_1^2 + x_2^2 + x_3^2}.$$

Можно отметить, что уравнениям (7), (8) удовлетворяют, например, следующие сеточные функции трех переменных и времени:

$$\begin{aligned}
\Omega_{3,1}^{(0)}(x^{(S)}) &= \frac{2\pi}{3H} \cdot \left\{ \left(x_1^{(S)}\right)^2 + \left(x_2^{(S)}\right)^2 + \left(x_3^{(S)}\right)^2 \right\} + Ax_1^{(S)} + Bx_2^{(S)} + Cx_3^{(S)} + D + \frac{8\pi t}{Ha^2}, \\
\Omega_{3,2}^{(0)}(x^{(S)}) &= \frac{2\pi}{3H} \cdot \left\{ \left(x_1^{(S)}\right)^2 + \left(x_2^{(S)}\right)^2 + \left(x_3^{(S)}\right)^2 \right\} + Ax_1^{(S)} + Bx_2^{(S)} + Cx_3^{(S)} + D + Et + \frac{8\pi t^2}{Ha^2},
\end{aligned} \tag{9}$$

где A, B, C, D, E – произвольные вещественные постоянные. Вместо линейной функции сеточных переменных можно взять любой гармонический полином, который удовлетворяет сеточному аналогу уравнения Лапласа (к примеру, $2(x_1^{(S)})^2 + 2(x_2^{(S)})^2 - 4(x_3^{(S)})^2$). Таким образом, об однозначной разрешимости задачи по определению сеточного фундаментального волнового уравнения или уравнения теплопроводности речь может идти только при наложении дополнительных ограничений на функцию $\Omega_{3,i}^{(S)}(x^{(S)})$, $i = 1, 2$. Далее, сеточное фундаментальное решение мы будем искать в некоторой ограниченной области сеточного трехмерного пространства. В качестве такой области может выступать как прямоугольный параллелепипед (в случае декартовой системы координат), так и шар (если рассматривать сеточные аналоги уравнений в частных производных в сферической системе координат).

2. СЕТОЧНОЕ ФУНДАМЕНТАЛЬНОЕ РЕШЕНИЕ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ В ДЕКАРТОВОЙ СИСТЕМЕ КООРДИНАТ

Оператор Лапласа в (7) будем считать заданным на шаблоне “крест”(см. [13]). Тогда для трехмерного сеточного пространства будем иметь (не ограничивая общности, будем считать, что a^2=1) :

$$\begin{aligned}
&\Lambda^+(\Omega_{3,1}^{(S)}(x^{(S),j+1})) - \frac{\Omega_{3,1}^{(S)}(x^{(S),j+1}) - \Omega_{3,1}^{(S)}(x^{(S),j})}{\tau} = -\frac{4\pi}{H} e_S(x^{(S),j}); \\
&x^{(S),j} = \left(x_1^{(S)}, x_2^{(S)}, \ldots, x_n^{(S)}; x_{n+1} = j\tau\right), \quad x^{(S),j+1} = \left(x_1^{(S)}, x_2^{(S)}, \ldots, x_n^{(S)}; x_{n+1} = (j+1)\tau\right), \quad j = 0, \ldots, J-1.
\end{aligned}$$

$$\begin{aligned}
&\Lambda^+(\Omega_3^{(S)}(x^{(S),j})) = -3\left(\sum_{p=1}^{3} C_p\right)\Omega_3^{(S)}\left(x^{(S),j}\right) + \sum_{p=1}^{3}\left\{\Omega_3^{(S)}(x^{(S),j} - h_p e_p) + \Omega_3^{(S)}(x^{(S),j} + h_p e_p)\right\}, \quad C_p = \frac{H^3}{h_1 h_2 h_3}, \\
&H = \sqrt[3]{h_1 h_2 h_3}, \quad x^{(S),j} = \left(x_1^{(s)}, x_2^{(S)}, x_3^{(S)}, t = j\tau\right), \quad x_p^{(S)} = kh_p, \quad -K \leqslant k \leqslant +K; \quad h_p = \mathrm{const}, \quad p = 1, 2, 3; \\
&\left|x_p^{(S)}\right| \leqslant d_p = Kh_p, \quad p = 1, 2, 3; \quad j = 0, \ldots, J-1; \quad J\tau = T.
\end{aligned} \tag{10}$$

Индекс j в (10) соответствует дискретному времени. Мы полагаем, что сеточное фундаментальное решение уравнения теплопроводности существует на интервале: $x_4 \doteq t \in (0, J\tau) = (0, T)$. Конечно-разностную производную по времени в (10) можно также задавать различными способами. В настоящей статье мы не будем акцентировать внимание на порядке точности аппроксимации производных по координатам и времени при переходе от уравнения теплопроводности и волнового уравнения к их дифференциально-разностным аналогам. Отметим следующий важный момент: конечно-разностные аналоги дифференциальных операторов должны быть такими, чтобы имела место устойчивость по начальным данным [13]. Ниже мы покажем на примере волнового уравнения, каким соотношениям должно удовлетворять сеточное фундаментальное решение волнового уравнения, чтобы одновременно учитывались критерии применимости метода матричной прогонки и устойчивости по начальным данным.

Как утверждается в [1], введение условия

$$\lim_{|x^{(S)})| \to \infty} \left|\Omega_n^{(S)}(x^{(S)})\right| = 0 \quad \text{при} \quad n \geq 3$$

не устраняет проблему неединственности решения сеточного аналога фундаментального решения уравнения Лапласа. Такой же вывод можно сделать и относительно сеточных фундаментальных решений для уравнения теплопроводности и волнового уравнения.

Действительно, соображения симметрии показывают, что имеют место равенства :

$$\Omega^{(S)}_{3,k}(x^{(S)}) \equiv \Omega^{(S)}_{3,k}\left(x^{(S)}_1, x^{(S)}_2, x^{(S)}_3, x^{(S)}_4\right) = \Omega^{(S)}_{3,k}\left(\left|x^{(S)}_1\right|, \left|x^{(S)}_2\right|, \left|x^{(S)}_3\right|, x^{(S)}_4\right) = \Omega^{(S)}_{3,k}\left(\left|x^{(S)}_{i_1}\right|, \left|x^{(S)}_{i_2}\right|, \left|x^{(S)}_{i_3}\right|, x^{(S)}_4\right),$$
$$(i_1, i_2, i_3) \in S_3, \quad k = 1, 2,$$

где S_3 — группа перестановок трех индексов.

Следовательно, достаточно найти значения $\Omega^{(S)}_{3,k}(x^{(S)})$, $k = 1, 2$, в области

$$S^4_{+++} = \left\{x^{(S)} : x^{(S)}_1 \geqslant 0, x^{(S)}_2 \geqslant 0, x^{(S)}_3 \geqslant 0, x^{(S)}_1 \leqslant x^{(S)}_2 \leqslant x^{(S)}_3, 0 \leqslant x^{(S)}_4 \leqslant T\right\},$$

т.е. в части первого координатного октанта. Кроме того, для волнового уравнения фундаментальное решение симметрично и по четвертой координате (времени): от знака этой координаты решение не зависит, но мы будем рассматривать решения только при положительных временах.

При определении значений сеточного фундаментального решения в узлах области S^4_{+++} можно поставить условие минимальности отклонения сеточного решения от его континуального аналога:

$$F\left[\Omega^{(S)}_{3,k}\left(x^{(S)}\right)\right] = \sum_{x^{(S)} \in S^4_{+++}} \left(\frac{\Omega^{(S)}_{3,k}\left(x^{(S)}\right) - \Omega_{3,kc}\left(x^{(S)}\right)}{\Omega_{3,kc}\left(x^{(S)}\right)}\right)^2 = \min, \quad k = 1, 2, \tag{11}$$

где $\Omega_{3,kc}\left(x^{(S)}\right)$ — значения континуального фундаментального решения уравнения теплопроводности (значение индекса $k = 1$) и волнового уравнения (значение индекса $k = 2$) в точках сеточной области.

Таким образом, одним из вариантов “устранения” неопределенности при нахождении сеточного фундаментального решения уравнений в частных производных параболического и гиперболического типов является минимизация функционала (11) при условиях, аналогичных (10); такого рода вариационные постановки исследовались нами неоднократно при применении метода линейных интегральных представлений к решению обратных линейных задач геофизики [4–6].

В настоящей работе мы будем рассматривать постановки задач для определения сеточных фундаментальных решений уравнения теплопроводности и волнового уравнения на семействе расширяющихся компактов: четырехмерное сеточное пространство ,т.е. неограниченное множество точек:

$$x^{(S)} = \left(x^{(S)}_1, x^{(S)}_2, \ldots, x^{(S)}_n, x^{(S)}_{n+1}\right)^{\mathrm{T}}, \quad x^{(S)}_k = mh_k; \quad m = 0, \pm 1, \pm 2, \ldots;$$
$$h_k = \mathrm{const}, \quad x^{(S)}_{n+1} \in (0, T),$$

заменяется ансамблем, или семейством, расширяющихся компактов, представляющих собой прямоугольные параллелепипеды $\left|x^{(S)}_p\right| \leqslant d_p = K_n h_p$, $p = 1, 2, 3$; здесь d_p, K_n и h_p, $p = 1, 2, 3$, — некоторые положительные константы, причем $K_n \to \infty$ при $n \to \infty$. Четвертая координата (время), по-прежнему, принимает значения из фиксированного интервала $(0, T)$.

Рассмотрим уравнение (10) в прямом произведении некоторого прямоугольного параллелепипеда на интервал $(0, T)$: $K_{n_q,t} = K_{n_q} \times (0, T)$. Для того чтобы корректно поставить задачу по определению $\Omega(x^{(S)})$ в указанной области, нужно задать граничные и начальные условия. Будем считать, что на гранях параллелепипеда K_{n_q} с номером n_q, принадлежащего описанному выше семейству расширяющихся компактов, значения фундаментального решения равны нулю (континуальным аналогом дискретной постановки такого рода является задача Дирихле для уравнения теплопроводности):

$$\begin{aligned} &\Omega^{(S)}_{3,1}(x^{(S)}) = 0, \quad x^{(S)} \in \Gamma K_{n_q}; \\ &\Omega^{(S)}_{3,1}(x^{(S)}) = \Omega^{(S)}_{3,10}(\bar{x}^{(S)}), \quad x^{(S)}_4 = 0, \quad \bar{x}^{(S)} = (x^{(S)}_1, x^{(S)}_2, x^{(S)}_3). \end{aligned} \tag{12}$$

В качестве $\Omega^{(S)}_{3,10}(\bar{x}^{(S)})$ можно взять сеточное фундаментальное решение уравнения Лапласа в соответствующем трехмерном параллелепипеде, как это было описано в [2].

Вариационная задача (10)–(12) может быть решена с помощью метода матричной прогонки, если для конечно-разностного аналога уравнения теплопроводности рассмотреть неявную схему. Матрицы систем линейных алгебраических уравнений имеют при этом в каждом слое по времени (т.е. при каждом значении

$x_4^{(S)} = j\tau$, $0 = 1, \ldots, J-1$) трехдиагональный или блочно-трехдиагональный вид (если постановка рассматривается в трехмерном сеточном пространстве), а векторы неизвестных $\mathbf{Y}_k^{(j)}$ и правых частей $\mathbf{\Phi}_k^{(j)}$ – блочный вид:

$$
\begin{aligned}
&-\mathbf{B}_{j,k+1}\mathbf{Y}_{k+1}^{(j)} + \mathbf{A}_{j,k}\mathbf{Y}_k^{(j)} - \mathbf{B}_{j,k-1}\mathbf{Y}_{k-1}^{(j)} = -\mathbf{F}_k^{(j)}, j = 1, \ldots, J-1; k = 1, \ldots, N_3 - 1;\\
&\mathbf{Y}_0^{(j)} = \mathbf{F}_0^{(j)}; \mathbf{Y}_{N_3}^{(j)} = \mathbf{F}_{N_3}^{(j)};\\
&\mathbf{A}_{j,k} = \begin{bmatrix} A_{11} & B_{12} & \ldots 0 \\ B_{21} & \ldots & \ldots \\ 0 & \ldots & A_{N_3N_3} \end{bmatrix},\ A_{ll} = \begin{bmatrix} a & b\ldots & 0 \\ b & \ldots & 0 \\ 0 & \ldots & a \end{bmatrix}, \quad \mathbf{Y}_k^{(j)} = \begin{pmatrix} \mathbf{y}_1^{(k,j)} \\ \mathbf{y}_2^{(k,j)} \\ \ldots \\ \mathbf{y}_{N_2}^{(k,j)} \end{pmatrix},\\
&a = 2\cdot(1+\alpha_1+\alpha_2) + \frac{1}{\tau},\ \alpha_1 = \frac{h_3^2}{h_1^2}; \alpha_2 = \frac{h_3^2}{h_2^2}; b = -\alpha_1; c = -\alpha_2; \mathbf{y}_l^{(k,j)} = \left(y_{1,l}^{(k,j)}, y_{2,l}^{(k,j)} \ldots, y_{N_1,l}^{(k,j)}\right)^{\mathrm{T}},\\
&1 \leqslant k \leqslant N_3 - 1;\ \mathbf{N} = (N_1, N_2, N_3);\\
&\mathbf{F}_k^{(j)} = -\frac{\mathbf{Y}_k^{(j-1)}}{\tau} + \mathbf{\Phi}_k^{(j)}, j = 1, \ldots, J-1;\\
&B_{12} = B_{21}, B_{il} = \begin{bmatrix} c & 0\ldots & 0 \\ 0 & \ldots & 0 \\ 0 & \ldots & c \end{bmatrix}, \quad l = i \pm 1; \quad i = 1, \ldots, N_1 - 1, \quad l = 1, \ldots, N_2 - 1.
\end{aligned}
\tag{13}
$$

В (13) блочные матрицы не зависят от дискретного времени в нашем случае, поэтому индекс "j" у них можно опустить. Но сами векторы неизвестных $\mathbf{Y}_k^{(j)}$ от времени, конечно, зависят. Индекс "j" в (13) соответствует значению времени $x_4^{(S)} = j\tau$, $j = 1, \ldots, J-1$.

Необходимо сделать важное замечание: соотношение (10) – это конечно-разностный аналог уравнения теплопроводности, заданный в неограниченной сеточной области, фактически – в прямом произведении всего трехмерного сеточного пространства на конечный интервал времени. Для однозначной разрешимости такого рода уравнений необходимо задать граничные и начальные условия. Поэтому нужно более подробно описать свойства сеточных областей, в которых мы в дальнейшем будем рассматривать краевые задачи для конечно-разностных аналогов уравнения теплопроводности и волнового уравнения. Введем следующие обозначения:

$$
\begin{aligned}
&\omega_3 = (ih_1, lh_2, kh_3), \quad 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant l \leqslant N_2 - 1, \quad 1 \leqslant k \leqslant N_3 - 1;\\
&\omega = (ih_1, lh_2), \quad 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant l \leqslant N_2 - 1;\\
&\gamma_3 = (0, lh_2, kh_3) \cup (N_1, lh_2, kh_3) \cup (ih_1, 0, kh_3) \cup (ih_1, N_2, kh_3) \cup (ih_1, lh_2, 0) \cup (ih_1, lh_2, N_3),\\
&1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant l \leqslant N_2 - 1, \quad 1 \leqslant k \leqslant N_3 - 1;\\
&\gamma = (0, lh_2) \cup (N_1, lh_2) \cup (ih_1, 0) \cup (ih_1, N_2), \quad 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant l \leqslant N_2 - 1.
\end{aligned}
\tag{14}
$$

Таким образом, мы задали равномерные сетки в трехмерном и двумерном пространствах (ω_3, ω), а также обозначили границы соответствующих областей. В (14) мы намеренно не использовали индекс j при нумерации узлов по второй декартовой координате: этот индекс у нас "зарезервирован" для времени.

Дифференциально-разностному оператору в (10) соответствует при решении краевых задач схема матричной прогонки [13] следующего вида (чтобы не усложнять запись формул, мы опустили индекс "j", соответствующий конкретному моменту дискретного времени):

$$
\begin{aligned}
&-\mathbf{Y}_{k-1} + \mathbf{C}_k\mathbf{Y}_k + \mathbf{Y}_{k+1} = \mathbf{F}_k, \quad 1 \leqslant k \leqslant N_3 - 1,\\
&\mathbf{CY} = (\Lambda\mathbf{y}_1, \Lambda\mathbf{y}_2, \ldots, \Lambda\mathbf{y}_{N_2-1}), \quad \mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_2-1}),\\
&\mathbf{y}_j = (y_{1,l}, y_{2,l}, \ldots, y_{N_1-1,l}), \quad 1 \leqslant l \leqslant N_2 - 1,\\
&\Lambda\mathbf{y}_l = (\Lambda_1 y_{1,l}, \Lambda_1 y_{2,l}, \ldots, \Lambda_1 y_{N_1-1,l}),\\
&\Lambda_1 y_{i,l} = \left(2 + \frac{1}{\tau}\right) y_{i,l} - C_3^2 y_{\bar{x}_1 x_1; i,l} - C_3^2 y_{\bar{x}_2 x_2; i,l}, \quad 1 \leqslant i \leqslant N_1 - 1; \quad 1 \leqslant l \leqslant N_2 - 1.
\end{aligned}
\tag{15}
$$

В (15) через $\mathbf{Y}_k$ обозначен блочный вектор значений сеточной функции в слое, соответствующем значению третьей координаты $x_3^{(S)} = kh_3$, $1 \leqslant k \leqslant N_3 - 1$. Правые части $\mathbf{F}_k$ при изменении индекса $1 \leqslant k \leqslant N_3 - 1$ – это сумма значений сеточной функции $\frac{4\pi}{H} e_S(x^{(S)})$ в соответствующих точках и значений деленной на шаг по времени вектор-функции решения, вычисленной на предыдущем по времени шаге (см. (13)); константа C_3^2 определяется в (10) (точнее говоря, это – квадрат C_3). К уравнениям (15) необходимо добавить граничные условия

с тем, чтобы возникла постановка краевой задачи. Подчеркнем еще раз, что мы используем неявную схему для конечно-разностного аналога уравнения теплопроводности: в этом случае значения сеточного аналога фундаментального решения уравнения теплопроводности вычисляются последовательно для каждого значения времени с помощью метода матричной прогонки в трехмерном сеточном пространстве координат. Как показано в [14], неявная схема для уравнения теплопроводности устойчива по начальным данным при любом соотношении $\gamma = \tau/H$.

Рассмотрим граничные условия I рода:

$$\mathbf{Y}_0 = \mathbf{F}_0, \quad \mathbf{Y}_{N_3} = \mathbf{F}_{N_3}. \tag{16}$$

Постановка (15), (16) – это частный случай более общей задачи: найти векторы $\mathbf{Y}_k$, $0 \leqslant k \leqslant N_3$, удовлетворяющие следующей системе:

$$\begin{aligned} &-\mathbf{A}_k\mathbf{Y}_{k-1} + \mathbf{C}_k\mathbf{Y}_k + \mathbf{B}_k\mathbf{Y}_{k+1} = \mathbf{F}_k, \quad 1 \leqslant k \leqslant N_3 - 1, \\ &C_0\mathbf{Y}_0 - \mathbf{B_0}\mathbf{Y_1} = \mathbf{F}_0, \quad k = 0, \\ &-\mathbf{A}_{\mathbf{N_3}}\mathbf{Y}_{N_3-1} + \mathbf{C}_{\mathbf{N_3}}\mathbf{Y}_{\mathbf{N_3}} = \mathbf{F}_{N_3}, \quad k = N_3, \end{aligned} \tag{17}$$

где $\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k$ — блочные матрицы; $\mathbf{C}_k$ — блочная квадратная матрица $N_2 \times N_2$, каждый блок которых имеет размер $N_1 \times N_1$; $\mathbf{A}_k$ — блочная прямоугольная матрица размера $N_2 - 1 \times N_2$, блоки которой имеют размер $N_1 \times N_1$; $\mathbf{B}_k$ — блочная прямоугольная матрица $N_2 + 1 \times N_2$, с блоками размера $N_1 \times N_1$; $\mathbf{Y}_k$ — блочный вектор размера N_2 с блоками длины N_1.

Формулы матричной прогонки имеют вид:

$$\begin{aligned} &\alpha_{k+1} = (\mathbf{C}_k - \mathbf{A}_k\alpha_{k-1})^{-1}\mathbf{B}_k, \quad k = 1, 2, \ldots, N_3 - 1, \quad \alpha_1 = \mathbf{C}_0^{-1}\mathbf{B}_0; \\ &\beta_{k+1} = (\mathbf{C}_k - \mathbf{A}_k\alpha_{k-1})^{-1}(\mathbf{F}_k + \mathbf{A}_k\beta_k), \quad k = 1, 2, \ldots, N_3, \quad \beta_1 = \mathbf{C}_0^{-1}\mathbf{F}_0; \\ &\mathbf{Y}_k = \alpha_{k+1}\mathbf{Y}_{k+1} + \beta_{k+1}, \quad k = N_3 - 1, N_3 - 2, \ldots, 0, \quad \mathbf{Y}_{N_3} = \beta_{N_3+1}, \end{aligned} \tag{18}$$

причем алгоритм (18) корректен, если матрицы $(\mathbf{C}_k - \mathbf{A}_k\alpha_{k-1})^{-1}$ не вырождены для $k = 1, 2, \ldots, N_3$. В [13] показано, что если $\mathbf{C}_k = \mathbf{C}, A_{N_3} = B_0 = 0, C_0 = C_{N_3} = E, B_k = A_k = E$, $1 \leqslant k \leqslant N_3$, а квадратная матрица $\mathbf{C}_k$ задана в (13), то условия корректности алгоритма (18) принимают вид

$$\left\|\mathbf{C}^{-1}\right\| = \max_m \left|\lambda_m(\mathbf{C}^{-1})\right| = \frac{1}{\min\limits_m |\lambda_m(\mathbf{C})|} \leqslant 0.5. \tag{19}$$

В (19) максимум и минимум берется по собственным значениям матрицы $\mathbf{C}^{-1}$ и $\mathbf{C}$ соответственно. Из определения матрицы $\mathbf{C}$ можно получить, что при замене $\lambda_m = 2 + \frac{1}{\tau} + C_3^2\lambda_{m_1}^{(1)} + C_3^2\lambda_{m_2}^{(2)}$, $m = (m_1, m_2)$, задача по поиску собственных значений λ_l для оператора

$$\begin{aligned} &\Lambda_1 y_{i,l} = \left(2 + \frac{1}{\tau}\right) y_{i,l} - C_3^2 y_{\bar{x}_1 x_1; i,l} - C_3^2 y_{\bar{x}_2 x_2; i,l} = \lambda_m y_{i,l}, \quad 1 \leqslant i \leqslant N_1 - 1; \quad 1 \leqslant l \leqslant N_2 - 1, \\ &y_{k,0} = y_{k,N_2} = y_{0,l} = y_{N_1,l} = 0, \end{aligned} \tag{20}$$

сводится к следующей задаче на собственные значения [13]:

$$\begin{gathered} \Lambda y(x) + \lambda_l y(x) = 0,\ x = (x_1, x_2) \in \omega, \quad \Lambda = \Lambda_x + \Lambda_y, \\ y(x) = 0, \quad x = (x_1, x_2) \in \gamma; \\ y_{il} = y(x_{1,i}, x_{2,l}); \quad x_{1,i} = ih_1, x_{2,l} = lh_2; \\ \Lambda_x\mu_{m_1}^{(1)} + \lambda_{m_1}^{(1)}\mu_{l_1}^{(1)} = 0, \quad 1 \leqslant i \leqslant N_1 - 1, \\ \Lambda_y\mu_{m_2}^{(2)} + \lambda_{l_2}^{(2)}\mu_{m_2}^{(2)} = 0, \quad 1 \leqslant l \leqslant N_2 - 1, \\ \mu_{m_1}^{(1)}(0) = \mu_{m_1}^{(1)}(N_1) = 0 = \mu_{m_2}^{(2)}(0) = \mu_{m_2}^{(2)}(N_2), \\ \lambda_{m_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2}\sin^2\left(\frac{m_\alpha\pi}{2N_\alpha}\right) = \frac{4}{h_\alpha^2}\sin^2\left(\frac{m_\alpha\pi h_\alpha}{2d_\alpha}\right), \quad m_\alpha = 1, 2, \ldots, N_\alpha - 1, \\ \mu_{m_1}^{(1)}(i) = \sqrt{\frac{2}{d_1}}\sin\left(\frac{m_1\pi i}{N_1}\right), \quad m_1 = 1, \ldots, N_1 - 1; \end{gathered} \tag{21}$$

$$\mu^{(2)}_{m_2}(j) = \sqrt{\frac{2}{d_2}}\sin\left(\frac{m_2\pi j}{N_2}\right), \quad m_2 = 1,\ldots,N_2-1;$$

$$y_{il} = \mu_m(i,l) = \mu^{(1)}_{m_1}(i)\cdot\mu^{(2)}_{m_2}(l),$$

$$\Lambda_x y_{il} = \frac{1}{h_1^2}\left[y_{i+1,l} - 2y_{il} + y_{i-1,l}\right], \quad \Lambda_y y_{il} = \frac{1}{h_2^2}\left[y_{i,l+1} - 2y_{il} + y_{i,l-1}\right].$$

Таким образом, собственное значение равно

$$\lambda_m = 2 + \frac{1}{\tau} + C_3^2\cdot\sum_{\alpha=1}^{2}\lambda^{(\alpha)}_{m_\alpha}. \tag{22}$$

В (19) через $y(x)$ обозначена сеточная функция двух аргументов, определенная на двумерной сетке $x = (x_1,x_2)\in\omega$ (см. (14)). При этом в каждом слое $x_3 = kh_3$, $1\leqslant k\leqslant N_3-1$, и при каждом значении времени $x_4 = j\tau$, $1\leqslant j\leqslant J-1$, рассматривается "своя" задача на собственные значения конечно-разностного оператора вида (20). Из выражения (22) для собственных значений краевой задачи (20) становится ясно, что $\lambda_l > 2$, поэтому алгоритм матричной прогонки для задачи (10)–(12) определен корректно, и мы можем констатировать, что верна следующая

Теорема 1. *Фундаментальное решение сеточного аналога уравнения теплопроводности в трехмерном случае определяется условиями* (10)–(12) *однозначно.*

Важно отметить, что корректность алгоритма матричной прогонки означает, что существует только одно решение задачи (10)–(12), которое определяется с помощью указанного алгоритма.

3. СЕТОЧНОЕ ФУНДАМЕНТАЛЬНОЕ РЕШЕНИЕ ВОЛНОВОГО УРАВНЕНИЯ В ДЕКАРТОВОЙ СИСТЕМЕ КООРДИНАТ

Перейдем теперь ко второму уравнению в (7), описывающему сеточное фундаментальное решение волнового уравнения. Запишем конечно-разностный аналог для оператора $L^{(w)}$ на шаблоне "крест" следующим образом:

$$\Delta_{t\bar{t}}\left(\Omega_{3,2}\left(x^{(S)}\right)\right) - \Delta_{x\bar{x}}\left(\Omega_{3,2}\left(x^{(S)}\right)\right) = -\frac{4\pi e_S(x^{(S)})}{H^{n-2}};$$

$$H = \left(\prod_{k=1}^{3}h_k\right)^{\frac{1}{3}}, \quad e_S(x^{(S)}) = \begin{cases}1, & x^{(S)} = 0,\\ 0, & \left|x^{(S)}\right| > 0,\end{cases}$$

$$\Delta_{t\bar{t}}\Omega^{(S)}_{3,2}(x^{(S),j}) = \frac{\Omega^{(S)}_{3,2}(x^{(S),j+1}) - 2\Omega^{(S)}_{3,2}(x^{(S),j}) + \Omega^{(S)}_{3,2}(x^{(S),j-1})}{\tau^2};$$

$$\Delta_{x\bar{x}}\Omega^{(S)}_{3,2}(x^{(S),j}) = \sum_{k=1}^{3}\frac{\Omega^{(S)}_{3,2}\left(x^{(S),j}_{k+1}\right) - 2\Omega^{(S)}_{3,2}\left(x^{(S),j}_{k}\right) + \Omega^{(S)}_{3,2}\left(x^{(S),j}_{k-1}\right)}{h_k^2};$$

$$x^{(S),j} = \left(x^{(S)}_1, x^{(S)}_2,\ldots,x^{(S)}_n; x_{n+1} = j\tau\right),$$

$$x^{(S),j+1} = \left(x^{(S)}_1, x^{(S)}_2,\ldots,x^{(S)}_n; x_{n+1} = (j+1)\tau\right), \quad j = 1,\ldots,J-1;$$

$$H = \sqrt[3]{h_1h_2h_3}, \quad x^{(S),j} = \left(x^{(s)}_1, x^{(S)}_2, x^{(S)}_3, t = j\tau\right), \quad x^{(S)}_p = kh_p, \quad -K\leqslant k\leqslant +K; \quad h_p = \text{const}, \quad p = 1,2,3;$$

$$\left|x^{(S)}_p\right| \leqslant d_p = Kh_p, \quad p = 1,2,3; \quad j = 1,\ldots,J-1; \quad J\tau = T. \tag{23}$$

Если применить неявную схему к волновому уравнению, то мы получим, как и в предыдущем разделе, некоторую краевую задачу для разностного аналога оператора Лапласа в трехмерном координатном пространстве. Вместо

$$\mathbf{F}^{(j)}_k = -\frac{\mathbf{Y}^{(j-1)}_k}{\tau} + \mathbf{\Phi}^{(j)}_k, \quad j = 1,\ldots,J-1,$$

т.е. правых частей в методе матричной прогонки (13) – теперь будут фигурировать следующие выражения:

$$\mathbf{F}^{(j)}_k = \frac{\mathbf{Y}^{(j-1)}_k}{\tau^2} + \mathbf{\Phi}^{(j)}_k - \frac{2\mathbf{Y}^{(j)}_k}{\tau^2}, \quad j = 1,\ldots,J-1. \tag{24}$$

Помимо значений сеточного аналога фундаментального решения волнового уравнения в начальный момент времени, нам нужно будет знать также значения его производной по времени при $t = 0$. Тогда выражения (12) запишутся следующим образом:

$$\begin{aligned}&\Omega^{(S)}_{3,2}(x^{(S)}) = 0, \quad x^{(S)} \in \Gamma K_{n_q};\\ &\Omega^{(S)}_{3,2}(x^{(S)}) = \Omega^{(S)}_{3,20}(\bar{x}^{(S)}), \quad x^{(S)}_4 = 0, \quad \bar{x}^{(S)} = (x^{(S)}_1, x^{(S)}_2, x^{(S)}_3);\\ &\nabla_t \left(\Omega^{(S)}_{3,2}(x^{(S)})\right) = \Omega^{(S)}_{3,21}(\bar{x}^{(S)}), \quad x^{(S)}_4 = 0, \quad \bar{x}^{(S)} = (x^{(S)}_1, x^{(S)}_2, x^{(S)}_3).\end{aligned} \tag{25}$$

Как и в предыдущем разделе, начальные данные определим как значения сеточного фундаментального решения уравнения Лапласа в некотором параллелепипеде, а в качестве производной по времени в начальный момент времени можно взять градиент сеточного фундаментального решения уравнения Лапласа (поскольку можно считать, что дифференцирование по времени эквивалентно дифференцированию по координатам в пространстве, если двигаться вдоль характеристик волнового уравнения; однако данный вопрос требует более глубокого изучения). Формулы (20) примут вид:

$$\begin{aligned}&\Lambda_1 y_{i,l} = \left(2 + \frac{h_3^2}{\tau^2}\right) y_{i,l} - h_3^2 y_{\bar{x}_1 x_1; i,l} - h_3^2 y_{\bar{x}_2 x_2; i,l} = \lambda_m y_{i,l}, \quad 1 \leqslant i \leqslant N_1 - 1; \quad 1 \leqslant l \leqslant N_2 - 1,\\ &y_{k,0} = y_{k,N_2} = y_{0,l} = y_{N_1,l} = 0.\end{aligned} \tag{26}$$

Оценить собственные значения оператора в (25) можно так же, как это было нами описано ранее. А именно, мы придем к выводу, что собственные значения оператора Λ_1 будут равны:

$$\lambda_m = 2 + \frac{h_3^2}{\tau^2} + h_3^2 \cdot \sum_{\alpha=1}^{2} \lambda^{(\alpha)}_{m_\alpha}. \tag{27}$$

Если еще учесть, что неявная схема для волнового уравнения устойчива при условии [14]:

$$\tau^2 \leqslant \frac{3}{\sum\limits_{p=1}^{3} \frac{1}{h_p^2}}, \tag{28}$$

то мы получим, что метод матричной прогонки позволяет найти единственное решение задачи (23)–(25), устойчивое по начальным данным в случае, когда (28) имеет место.

Таким образом, нами доказана

Теорема 2. *Сеточное фундаментальное решение волнового уравнения в трехмерном координатном сеточном пространстве определяется однозначно с помощью метода матричной прогонки при условии, что применяется неявная схема по времени.*

Если теперь рассмотреть двухпараметрическое (σ_1, σ_2 — некоторые вещественные числа) семейство конечно-разностных аппроксимаций волнового уравнения

$$\Delta_{t\bar{t}}\Omega^{(S)}_{3,2}(x^{(S),j}) = \sigma_1 \Delta_{x\bar{x}}\Omega^{(S)}_{3,2}(x^{(S),j+1}) + (1 - \sigma_1 - \sigma_2)\Delta_{x\bar{x}}\Omega^{(S)}_{3,2}(x^{(S),j}) + \sigma_2 \Delta_{x\bar{x}}\Omega^{(S)}_{3,2}(x^{(S),j-1}) - \frac{4\pi e_S(x^{(S)})}{H^{n-2}}; \tag{29}$$

$$\sigma_1 \geqslant \sigma_2, \ \frac{\tau^2}{\sum\limits_{p=1}^{3} \frac{1}{h_p^2}} \leqslant 1, \tag{30}$$

то мы придем к выводу, что справедлива

Теорема 3. *Сеточное фундаментальное решение волнового уравнения в трехмерном координатном сеточном пространстве определяется однозначно с помощью метода матричной прогонки при условии, что применяется схема* (29) *с параметрами* σ_1, σ_2, *удовлетворяющими соотношениям* (30).

4. ВАРИАЦИОННАЯ ПОСТАНОВКА ПО ОПРЕДЕЛЕНИЮ СЕТОЧНЫХ ФУНДАМЕНТАЛЬНЫХ РЕШЕНИЙ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ И ВОЛНОВОГО НА СЕМЕЙСТВЕ РАСШИРЯЮЩИХСЯ КОМПАКТОВ

Из свойств конечно-разностных аналогов операторов в уравнении теплопроводности и волновом уравнении следует, что функционал (11) выпуклый и непрерывный. Мы видим также, что дискретные аналоги дифференциальных операторов, определенные в (10) и (23), а также в (29) – это линейные операторы, действующие на пространстве сеточных функций в декартовой системе координат.

В [15] доказывается, что *для выпуклого, замкнутого и ограниченного множества G множество $G^* = \{\varphi^* \in G : J(\varphi^*) = \mu\} \neq \varnothing$ выпукло, замкнуто и ограничено.* Здесь $J(\varphi^*)$ — выпуклый и непрерывный функционал на некотором гильбертовом пространстве, φ^* — элемент этого пространства. В качестве гильбертова пространства мы выберем пространство сеточных функций Ξ со скалярным произведением, определенным по формуле:

$$
\begin{aligned}
(u,v) = {} & \sum_{j=1}^{J-1}\sum_{k=1}^{N_1-1}\sum_{l=1}^{N_2-1}\sum_{m=1}^{N_3-1} u(k,l,m;j)v(k,l,m;j)h_1h_2h_3\tau + \\
& + 0.5h_1h_2\tau\left[\sum_{j=1}^{J-1}\sum_{k=1}^{N_1-1}\sum_{l=1}^{N_2-1}\{u(k,l,0;j)v(k,l,0;j) + u(k,l,N_3;j)v(k,l,N_3;j)\}\right] + \\
& + 0.5h_1h_3\tau\left[\sum_{j=1}^{J-1}\sum_{k=1}^{N_1-1}\sum_{m=1}^{N_3-1}\{u(k,0,m;j)v(k,0,m;j) + u(k,N_2,m;j)v(k,N_2,m;j)\}\right] + \\
& + 0.5h_2h_3\tau\left[\sum_{j=1}^{J-1}\sum_{m=1}^{N_3-1}\sum_{l=1}^{N_2-1}\{u(0,l,m;j)v(0,l,m;j) + u(N_1,l,m;j)v(N_1,l,m;j)\}\right] + \\
& + 0.5h_1h_2h_3\left[\sum_{k=1}^{N_1-1}\sum_{l=1}^{N_2-1}\sum_{m=1}^{N_3-1}\{u(k,l,m;0)v(k,l,m;0) + u(k,l,m;J)v(k,l,m;J)\}\right].
\end{aligned}
\tag{31}
$$

При рассмотрении всех постановок задач в предыдущих разделах предполагалось, что сеточные функции $u(k,l,m;j), v(k,l,m;j) \in \Xi$.

Замечание. Введенное в (31) скалярное произведение соответствует сеточным краевым задачам с условиями Дирихле.

Определим некоторые вспомогательные множества

$$
\begin{gathered}
G_R = G \cap K_R, \quad F_R = \left\{ f \in \Phi, f = L^{(h,w)}\varphi,\ \varphi \in G_R \right\}, \\
\text{где} \quad K_R = \{\varphi \in \Phi :\ \|\varphi\| \leqslant R\}, \quad S_R = \{\varphi \in \Phi :\ \|\varphi\| = R\}, \quad \varphi \in \Xi.
\end{gathered}
\tag{32}
$$

В (32) через $L^{(h,w)}$обозначены конечно-разностные аналоги дифференциальных операторов в уравнении теплопроводности и волновом уравнении соответственно (см. (3), (7) или (10), (23), (29)). Как было показано в работе [15], из того, что $K_{R'} \subseteq K_{R''}$ при $\bar{R} \leqslant R' \leqslant R'' \leqslant \bar{\bar{R}}$ следует справедливость включения $G_{R'} \subseteq G_{R''}$, $F_{R'} \subseteq F_{R''}$.

Теорема 2, доказанная в [15], утверждает, что при $\forall R \quad F_R$ выпукло, слабо замкнуто и ограничено. Это значит, что при $\forall R \geqslant \bar{R}$ существует единственный $f_R^* \in F_R$ такой, что

$$
\|f_R^* - f\|_{\Xi'} = \inf_{f \in F_R} \left\|f - \bar{f}\right\|_{\Xi'}.
\tag{33}
$$

Через Ξ' мы обозначили сеточное пространство функций, определенных во внутренних узлах исходной сетки (поскольку шаблон “крест” для оператора Лапласа предполагает вычисление значений функций именно в таких узлах). Сами сеточные функции зависят от четырех сеточных координат. Назовем f^* и f_R^* *метрическими проекциями* f соответственно на множества F и F_R. Очевидно,

$$
\begin{aligned}
\mu &= \inf_{\varphi \in G} J(\varphi) = \inf_{f \in F} \left\|f - \bar{f}\right\|_{\Xi'}^2 = \left\|f^* - \bar{f}\right\|_{\Xi'}^2, \\
\mu(R) &= \inf_{\varphi \in G_R} J(\varphi) = \inf_{f \in F_R} \left\|f - \bar{f}\right\|_{\Xi'}^2 = \left\|f_R^* - \bar{f}\right\|_{\Xi'}^2.
\end{aligned}
\tag{34}
$$

Так как $J(\varphi)$ — непрерывный и выпуклый функционал, G_R — выпукло, замкнуто и ограничено, F_R — выпукло, слабо замкнуто и ограничено, G и F – подмножества гильбертовых пространств, (см. соответствующие теоремы и леммы в [15]), то $\mu(R)$ — непрерывная функция при $R \in [\bar{R}, \bar{\bar{R}}]$.

$J(\varphi)$ достигает своей точной нижней грани на G_R в единственной точке, такой что $\|\varphi_R^*\| = R$. Через R мы обозначили расстояние от начала координат до точки в пространстве сеточных фундаментальных решений дифференциальных уравнений в частных производных.

Если через $J(\varphi)$ обозначить функционал качества (11), отражающий близость сеточного фундаментального решения к классическому, континуальному, то можно сделать вывод о том, что справедлива

Теорема 4. *На семействе расширяющихся компактов $K_{R_n}, R_n \to \infty$, существует единственное решение задачи* (11), (12) *как в случае сеточного фундаментального решения уравнения теплопроводности, так и в случае волнового уравнения.*

Устойчивое приближенное решение задачи (11), (12) может быть получено только при помощи регуляризирующих алгоритмов, обеспечивающих высокую точность искомых сеточных фундаментальных решений уравнения теплопроводности и волнового уравнения [16]. Отметим также, что к условиям (11), (12) в рамках теории дискретного потенциала следует добавить требование достижения минимума функционала (5).

ЗАКЛЮЧЕНИЕ

1. В статье исследованы условия однозначной разрешимости системы линейных алгебраических уравнений, к которой редуцируется проблема поиска сеточных аналогов фундаментальных решений уравнения теплопроводности и волнового уравнения. Рассмотрены различные варианты представления дискретных фундаментальных решений в некоторых ограниченных областях четырехмерного сеточного пространства.

2. В работе описывается алгоритм построения сеточных фундаментальных решений на семействе расширяющихся компактов в сеточном четырехмерном пространстве, который позволяет находить значения дискретных теплового и волнового потенциалов и их производных в любой точке неограниченного сеточного пространства, что важно для приложений в области математической физики и математической геофизики [16].

Авторы выражают глубокую благодарность А.С. Леонову за полезные замечания и внимание к работе.

СПИСОК ЛИТЕРАТУРЫ

1. *Страхов В.Н., Степанова И.Э., Гричук Л.В.* Теория дискретного гравитационного потенциала и ее использование в гравиметрии // В сб.: “Вопросы теории и практики геологической интерпретации гравитационных, магнитных и электрических полей”. Труды междунар. конференции. Воронеж: Воронежский государственный университет. 1996. С. 49–71.

2. *Stepanova I.E., Kolotov I.I.* Solution of the interpretation tomography problem in geophysics under the linear integral representation method and theories of discrete gravity and magnetic potentials // Doklady. Earth Sciences. 2024. № 1. P. 1–9.

3. *Арсанукаев З.З.* Вычисление пространственных элементов аномальных полей с использованием методов теории дискретных гравитационных полей// Физ. Земли. 2004. № 11. С. 47–69.

4. *Страхов В.Н., Степанова И.Э.* Метод S-аппроксимаций и его использование при решении задач гравиметрии (локальный вариант) // Физ. Земли. 2002. № 2. С. 3–19.

5. *Страхов В.Н., Степанова И.Э.* Метод S-аппроксимаций и его использование при решении задач гравиметрии (региональный вариант) // Физ. Земли. 2002. № 7. С. 3–12.

6. *Stepanova I.E., Kerimov I.A., Yagola A.G.* Approximation approach in various modifications of the method of linear integral representations // Izvestiya. Physics of the Solid Earth. 2019. Vol. 55. No. 2. P. 218–231.

7. *Раевский Д.Н., Степанова И.Э.* О решении обратных задач гравиметрии с помощью модифицированного метода S-аппроксимаций // Физ. Земли. 2015. № 2. С. 44–54. гравиметрической и магнитометрической съемок // Физ. Земли. 2009. № 4. С. 17–30.

8. *Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г.* Численные методы решения некорректных задач. М.: Наука, 1990. 230 с.

9. *Лаврентьев М.А., Люстерник Л.А.* Курс вариационного исчисления. М.Л.: Гостоптехиздат. 1950. 296 с.

10. *Kolotov I.I., Lukyanenko D.V., Stepanova I.E., Shchepetilov A.V., Yagola A.G.* On the uniqueness of solution to systems of linear algebraic equations to which the inverse problems of gravimetry and magnetometry are reduced: a regional variant // Comput. Math. and Math. Phys. 2023. V. 63. № 9. P. 1588–1599.

11. *Kolotov I.I., Lukyanenko D.V., Stepanova I.E., Yagola A.G.* On the uniqueness of solutions to systems of linear algebraic equations resulting from the reduction of linear inverse problems of gravimetry and magnetometry: a local case // Comput. Math. and Math. Phys. 2023. V. 63. № 8. P. 1452–1465.

12. *Леонов А.С.* Метод минимальной псевдообратной матрицы// Ж. вычисл. матем. и матем. физ. 1987. Т. 27. № 8. С. 1123–1138.

13. *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. М.: Наука,1978. 592 с.

14. *Самарский А.А., Гулин А. В.* Численные методы. М.: Наука, 1989. 431 с.

15. *Гавурин М.К., Фабровская Ю.Б.* Об одном итеративном методе разыскания суммы квадратов // Ж. вычисл. матем. и матем. физ. 1966. Т. 6. № 6. С. 1094–1097.

16. *Leonov A.S.* Extraoptimal A Posteriori Estimates of the Solution Accuracy in the Ill-Posed Problems of the Continuation of Potential Geophysical Fields, Izvestiya, Physics of the Solid Earth. 2011. Vol. 47. No. 6. P. 531–540.

ON THE UNAMBIGUITY OF DETERMINING THE GRID FUNDAMENTAL SOLUTION OF THE THERMAL CONDUCTIVITY EQUATION AND THE WAVE EQUATION WITHIN THE FRAMEWORK OF THE THEORY OF DISCRETE POTENTIAL

I. E. Stepanova[a,*], I. I. Kolotov[b], A. V. Shchepetilov[b], A. G. Yagola [b], A. N. Levashov [b]

[a] 123242 Moscow, B.Gruzinskaya str.,10, bldg. 1, O.Yu. Schmidt Institute of Earth Physics of the Russian Academy of Sciences, Russia
[b] 119992 Moscow, Leninskie Gory, Lomonosov Moscow State University, Faculty of Physics, Russia
**e-mail: tet@ifz.ru*

Received: 10.07.2024
Revised: 10.07.2024
Accepted: 23.08.2024

Abstract. The paper considers the problem of unambiguously determining the fundamental solution of the grid analogue of the wave equation, as well as the equation of thermal conductivity within the framework of the theory of discrete potential. Grid-based fundamental solutions of finite-difference analogues of equations in partial derivatives allow solving direct and inverse problems of restoring wave and heat sources in various media from heterogeneous and different-flow information about the corresponding physical fields. The article considers statements with Dirichlet conditions in three-dimensional and four-dimensional Cartesian spaces.

Keywords: unambiguous definition, fundamental solution, discrete thermal potential, discrete wave potential.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2390–2400

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.635

РАЗВИТИЕ МЕТОДА АДАПТИВНОЙ ИСКУССТВЕННОЙ ВЯЗКОСТИ ДЛЯ ГАЗОДИНАМИЧЕСКИХ РАСЧЕТОВ НА НЕРАВНОМЕРНЫХ РАЗНОСТНЫХ СЕТКАХ[1)]

© 2024 г. А. Ю. Круковский[1], И. В. Попов[1,*], В. А. Гасилов[1]

[1] 125047 Москва, Миусская пл., 4, ИПМ им. М.В. Келдыша РАН, Россия
**e-mail: piv2964@mail.ru*

Поступила в редакцию 05.07.2024 г.
Переработанный вариант 27.07.2024 г.
Принята к публикации 23.08.2024 г.

Метод адаптивной искусственной вязкости в настоящей работе обобщается с целью построения разностных схем газовой динамики, обеспечивающих высокое разрешение структуры течений как на равномерных, так и на неравномерных сетках. Рассматриваются разностные схемы, аппроксимирующие одномерную систему уравнений газовой динамики. Полученные в работе оценки величины адаптивной вязкости учитывают неоднородность распределения газодинамических величин в расчетной области и неравномерность разностной сетки. Построенные схемы с адаптивной искусственной вязкостью обладают свойствами однородности и консервативности. Апробация предложенных схем выполнена на модельных задачах, решения которых описывают различные гладкие газодинамические структуры, а также сильные и слабые разрывы. Продемонстрирована возможность получения высокоточных решений на расчетных сетках с существенной разницей геометрических размеров соседних разностных ячеек. Библ. 13. Фиг. 5.

Ключевые слова: газовая динамика, адаптивная искусственная вязкость, однородная консервативная разностная схема, вычислительный эксперимент.

DOI: 10.31857/S0044466924120126, **EDN**: KBNHIO

ВВЕДЕНИЕ

Исходный вариант метода адаптивной искусственной вязкости (АИВ) (см. [1]) на равномерных сетках приводит к разностной схеме второго порядка точности, обладающей свойством монотонности. Целью разработки модифицированного метода было сохранение, по возможности, свойств исходного метода АИВ – монотонности и высокой точности решений, в случае его применения к решению газодинамических задач на неравномерных разностных сетках. Публикации, посвященные разработкам методик расчетов на неравномерных сетках, в основном содержат результаты, полученные в предположении о гладком изменении размеров сеточных ячеек. Под этим обычно понимается правило распределения сеточных элементов, квазиравномерное согласно определению А.А. Самарского: шаги такой сетки должны удовлетворять условию $h_{i+1} = h_i + O(h_i^2)$ (см., например, [2], [3]). В настоящей работе построена разностная схема с высоким разрешением структуры газовых течений на сетках, которые могут не удовлетворять условию квазиравномерности. Таковы, например, сетки иерархической (древовидной) либо нерегулярной структуры (см. [4]–[7]).

При построении модификации АИВ для определения величины искусственной вязкости использован часто применяемый в подобных случаях анализ по принципу “замороженных” коэффициентов (см. [8], [9]). Наряду с этим потребовалось провести более полное исследование, в ходе которого были учтены как неоднородности распределения вычисляемых искомых значений функций (плотности, скорости, давления, полной энергии и др.), так и неравномерность геометрических размеров разностной сетки. При этом использовались общие принципы построения однородных разностных схем (см. [10]). В итоге в настоящей работе построена однородная разностная схема, аппроксимирующая исходную систему газодинамики как на равномерной, так и на неравномерной расчетных сетках. Апробация модифицированного метода АИВ выполнена на известных тестовых задачах (см. [11], [12]). Численная методика на основе предложенной однородной схемы показала

[1)] Работа выполнена с использованием оборудования Центра коллективного пользования ИПМ им. М.В. Келдыша РАН (суперкомпьютеры К-60, К-100).

возможность расчета газодинамических структур с высоким разрешением даже в случае существенного, многократного различия размеров соседних расчетных ячеек на разностной сетке, построенной со сгущением в областях локализации больших градиентов газодинамических величин.

1. ПОСТАНОВКА ЗАДАЧИ

Разностная схема с адаптивной искусственной вязкостью рассматривалась для одномерных задач газовой динамики в декартовой системе координат в переменных Эйлера. Систему дифференциальных уравнений, описывающих одномерные течения сжимаемой среды, запишем в виде законов сохранения (см. [8]):

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) &= 0, \\ \frac{\partial I}{\partial t} + \frac{\partial}{\partial x}(Iu + p) &= 0, \\ \frac{\partial E}{\partial t} + \frac{\partial}{\partial x}((E + p)u) &= 0. \end{aligned} \tag{1}$$

Здесь использованы следующие обозначения: t – время, x – пространственная переменная, ρ – плотность, u – скорость, $I = \rho u$ – механический импульс, ε – внутренняя энергия, $E = \rho\varepsilon + \rho\frac{u^2}{2}$ – полная энергия. Система уравнений замыкается уравнением состояния $p = p(\rho, \varepsilon)$, например, уравнением состояния идеального газа $p = (\gamma - 1)\rho\varepsilon$, где γ – показатель адиабаты Пуассона.

Начально-краевая задача для системы дифференциальных уравнений (1) рассматривается в области $D(x, t)$ ($t \geqslant T_0$, $L_1 \leqslant x \leqslant L_2$). Начальные распределения газодинамических величин, например, ρ, u, I, E, в некоторых точках могут претерпевать существенные разрывы. Учитывая, что основное внимание в настоящей работе уделяется свойствам численного решения внутри расчетной области, особенности реализации граничных условий не рассматриваются. Соответственно, расчеты проводятся либо до заданного момента времени, либо до выхода возмущений на границу расчетной области.

2. ПОСТРОЕНИЕ РАЗНОСТНОЙ СХЕМЫ ВТОРОГО ПОРЯДКА АППРОКСИМАЦИИ

Методику аппроксимации системы дифференциальных уравнений (1) рассмотрим на примере уравнения неразрывности. Введем сетку по времени с переменным шагом τ_n $\omega_\tau = \{t_0 = T_0, t_{n+1} = t_n + \tau_n, n = 0,\ 1,\ \ldots\}$. Функции, которые относятся к моментам времени t_n и t_{n+1}, обозначим верхними индексами n и $n + 1$ соответственно: $r^n = r(t_n)$, $r^{n+1} = r(t_{n+1})$, где $r = \rho, u, I, E, p$. Разностную производную по времени от сеточной функции будем вычислять следующим образом:

$$r_t^n = \frac{r^{n+1} - r^n}{\tau_n}.$$

Построение явной разностной схемы второго порядка аппроксимации в областях гладкости решения основано, как и в исходном варианте метода АИВ (см. [1]), на использовании поправок Лакса–Вендроффа. Получим уравнение, являющееся дифференциально-разностным приближением к уравнению неразрывности. Для этого используем приближенное выражение, связывающее значение плотности газа ρ на моменты времени $t = t_n$, $t = t_{n+1}$, по формуле многочлена Тейлора, ограничиваясь слагаемыми до $O\left({\tau_n}^2\right)$:

$$\rho^{n+1} = \rho^n + \tau_n \frac{\partial \rho^n}{\partial t} + \frac{\tau_n^2}{2}\frac{\partial^2 \rho^n}{\partial t^2} + \cdots.$$

Производную по времени $\frac{\partial \rho^n}{\partial t} = \left.\frac{\partial \rho}{\partial t}\right|_{t=t_n}$ заменим с помощью уравнения неразрывности, записанного при $t = t^n$:

$$\frac{\partial \rho^n}{\partial t} = -\frac{\partial}{\partial x}(\rho u)^n.$$

Вторую производную от плотности $\frac{\partial^2(\rho)}{\partial t^2}$ найдем дифференцированием по времени уравнения неразрывности. Производная по времени от механического импульса ρu выражается через плотность потока импульса $\rho u^2 + p$ из уравнения баланса импульса:

$$\frac{\partial^2 \rho}{\partial t^2} = \frac{\partial^2}{\partial x^2}\left(\rho u^2 + p\right).$$

В результате имеем дифференциально-разностное уравнение баланса массы

$$\frac{\rho^{n+1}-\rho^n}{\tau_n}+\frac{\partial W_\rho^n}{\partial x}=0,$$

в котором поток вещества вычисляется по формуле

$$W_\rho^n=(\rho u)^n-\frac{\tau_n}{2}LW_\rho^n.$$

Выражение $\frac{\tau_n}{2}LW_\rho^n$ является поправкой Лакса–Вендроффа

$$LW_\rho^n=\frac{\partial}{\partial x}\left(\rho u^2+p\right)^n. \tag{2}$$

Аналогичным способом получаем поправки Лакса–Вендроффа для уравнений импульса и полной энергии:

$$W_I^n=(Iu+p)^n-\frac{\tau_n}{2}LW_I^n, W_E^n=\left((E+p)\,u\right)^n-\frac{\tau_n}{2}LW_E^n. \tag{3}$$

Поправки Лакса–Вендроффа при аппроксимации уравнений балансов импульса и энергии соответственно имеют следующий вид:

$$LW_I^n=\left(\frac{\partial}{\partial x}\left(Iu^2\right)+3u\frac{\partial p}{\partial x}+\gamma p\frac{\partial u}{\partial x}\right)^n, \tag{4}$$

$$LW_E^n=\left(\frac{\partial}{\partial x}\left((E+p)\,u^2\right)+\left(\frac{c^2}{\gamma-1}+\frac{3}{2}u^2\right)\frac{\partial p}{\partial x}+\gamma pu\frac{\partial u}{\partial x}\right)^n. \tag{5}$$

Также дифференциально-разностные уравнения можно записать единообразно в форме законов сохранения:

$$\frac{\mathbf{r}^{n+1}-\mathbf{r}^n}{\tau_n}+\frac{\partial\mathbf{W}^n(\mathbf{r})}{\partial x}=0, \tag{6}$$

где $\mathbf{r}=(\rho, I,\ E)$.

Далее в поправках Лакса–Вендроффа выделим те слагаемые, в которые входят производные от плотности, импульса и полной энергии газа:

$$LW_\rho=\left(u^2+c^2\right)\frac{\partial\rho}{\partial x}+\rho\frac{\partial u^2}{\partial x}+\varphi, \tag{7}$$

$$LW_I=\left(u^2+3c^2\right)\frac{\partial I}{\partial x}+2\rho\left(u^2-c^2\right)\frac{\partial u}{\partial x}+3u\varphi, \tag{8}$$

$$LW_E=\left(u^2+\varsigma_m c^2\right)\frac{\partial E}{\partial x}+\frac{1}{\gamma-1}\left[1+\frac{\gamma-1}{2}m^2-\frac{(\gamma-1)^2m}{1+\frac{\gamma-1}{2}m^2}\right]\rho\frac{\partial u^2}{\partial x}+\varsigma_m\frac{u^2}{2}\varphi, \tag{9}$$

Здесь переменная φ с учетом зависимости $p=p(\rho,S)$ выражается через производную от энтропии S следующим образом:

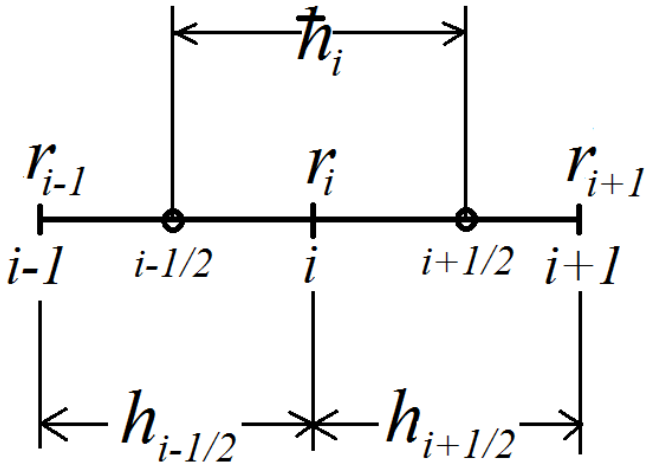
$$\varphi=\frac{\partial p}{\partial x}-c^2\frac{\partial\rho}{\partial x}=\left(\frac{\partial p}{\partial S}\right)_\rho\frac{\partial S}{\partial x},$$

где $c=\sqrt{\gamma\frac{p}{\rho}}$ – скорость звука,

$$\varsigma_m=\frac{1+\frac{5}{2}(\gamma-1)M^2}{1+\frac{1}{2}(\gamma-1)M^2},\quad M=\frac{u}{c},\quad 1\leqslant\varsigma_m\leqslant 5.$$

3. РАЗНОСТНАЯ СЕТКА ПО ПРОСТРАНСТВЕННОЙ ПЕРЕМЕННОЙ. АППРОКСИМАЦИЯ

В расчетной области $L_1\leqslant x\leqslant L_2$ введем неравномерную, вообще говоря, разностную сетку $\bar\Omega_h=\Omega_h\cup\Gamma_h$ узлов по пространственной переменной x: $x_1=L_1<x_2<\ldots<x_N=L_2$ и сетку $\bar\Omega_h'=\Omega_h'\cup\Gamma_h'$ потоковых узлов $\bar x_{1/2}=L_1$, $\bar x_{i+1/2}=0.5\,(x_i+x_{i+1})$, $i=1,2,\ldots,N-1$, $\bar x_{N+1/2}=L_2$. Граничными узлами являются точки $x_1,x_N\in\Gamma_h$, $\bar x_{1/2},\bar x_{N+1/2}\in\Gamma_h'$.



Фиг. 1. Шаблон аппроксимации потоков в разностных уравнениях балансов массы, импульса и энергии.

Обозначим шаги расчетных сеток $\bar{\Omega}_h$ и $\bar{\Omega}'_h$ соответственно $\{h_{i+1/2} = x_{i+1} - x_i,\ i = \overline{1, N-1}$ и $h_i = 0.5\left(h_{i+1/2} + h_{i-1/2}\right),\ i = \overline{2, N-1},\ h_1 = 0.5h_{3/2},\ h_N = 0.5h_{N-1/2}\}$.

К узлам исходной разностной сетки x_i будем относить искомые функции ρ_i, u_i, $I_i = \rho_i u_i$, p_i, E_i. К узлам потоковой сетки $\bar{x}_{i+1/2}$ отнесем разностные аппроксимации производных, например,

$$\rho_{x_{i+1/2}} = \frac{\rho_{i+1} - \rho_i}{h_{i+1/2}}, \quad u_{x_{i+1/2}} = \frac{u_{i+1} - u_i}{h_{i+1/2}} \text{ и т.д.},$$

а также средние величины

$$\bar{\rho}_{i+1/2} = \frac{1}{2}\left(\rho_{i+1} + \rho_i\right), \quad \bar{u}_{i+1/2} = \frac{1}{2}\left(u_{i+1} + u_i\right) \text{ и т.д.}$$

На границе области $\bar{\rho}_{1/2} \equiv \rho_1$, $\bar{u}_{1/2} \equiv u_1$, $\bar{\rho}_{N+1/2} \equiv \rho_N$, $\bar{u}_{N+1/2} \equiv u_N$ и т.д.

К потоковым узлам будем относить сеточные аналоги потоков $W^r_{i+1/2}$, $i = \overline{1, N}$, $r = \rho, I, E$.

Шаблон аппроксимации с расположением узлов основной и потоковой сеток приводится на фиг. 1.

Разностные производные величин, заданных в потоковых узлах, определяются следующим образом:

$$W^r_{\bar{x}_i} = \frac{W^r_{i+1/2} - W^r_{i-1/2}}{h_i}, \quad i = \overline{1, N},$$

где $r = \rho, I, E$.

Конструкция искусственной вязкости строится с использованием поправок Лакса–Вендроффа (2), (4), (5). Для наших целей достаточно взять только слагаемые поправок, в которые входят сомножителями производные от плотности, импульса и энергии, т.е.

$$LW_\rho = \left(u^2 + c^2\right)\frac{\partial \rho}{\partial x}, \quad LW_I = \left(u^2 + 3c^2\right)\frac{\partial I}{\partial x}, \quad LW_E = \left(u^2 + \varsigma_m c^2\right)\frac{\partial E}{\partial x}. \tag{10}$$

Поправки в этом “усеченном” виде позволяют получить выражение для искусственной вязкости в предельном случае малых возмущений, распространяющихся на фоне однородного течения газа. В акустическом приближении можно ограничиться поправками (10), при этом пренебрегая другими слагаемыми, входящими в уравнения (7)–(9), как величинами второго порядка малости.

Потоки в уравнениях баланса массы, импульса и энергии аппроксимируем в форме, записанной с использованием следующих индексных обозначений:

$$W_{\rho i+1/2} = (\overline{\rho u})_{i+1/2} - \frac{\tau}{2}(LW_\rho)_{i+1/2}, \quad (LW_\rho)_{i+1/2} = (\rho u^2 + p)_{xi+1/2},$$

$$W_{Ii+1/2} = (\overline{\rho u})_{i+1/2}\bar{u}_{i+1/2}\bar{p}_{i+1/2} - \frac{\tau}{2}(LW_I)_{i+1/2}, \quad (LW_I)_{i+1/2} = (Iu^2)_{xi+1/2} + 3\bar{u}_{i+1/2}p_{xi+1/2} + \gamma\bar{p}_{i+1/2}\bar{u}_{xi+1/2},$$

$$(LW_E)_{i+1/2} = \left((E+p)u^2\right)_{xi+1/2} + \left(\frac{\bar{c}^2_{i+1/2}}{\gamma - 1} + \frac{3}{2}(\bar{u}_{i+1/2})^2\right)p_{xi+1/2} + \gamma\bar{p}_{i+1/2}\bar{u}_{i+1/2}u_{xi+1/2}, \tag{11}$$

$$W_{Ei+1/2} = (\overline{\rho u})_{i+1/2}\left\langle \frac{u^2}{2}\right\rangle_{i+1/2} + \frac{\gamma}{\gamma-1}\bar{p}_{i+1/2}\bar{u}_{i+1/2} - \frac{\tau}{2}(LW_E)_{i+1/2},$$

$$\text{где } \left\langle \frac{u^2}{2}\right\rangle_{i+1/2} = \frac{u_{i+1/2}u_i}{2}, \quad \bar{c}^2_{i+1/2} \equiv \gamma\frac{\bar{p}_{i+1/2}}{\bar{\rho}_{i+1/2}}.$$

В граничных узлах $\bar{x}_{1/2} = L_1$ и $\bar{x}_{N+1/2} = L_2$ поправки Лакса–Вендроффа будем считать равными нулю, т.е. $LW^r_{1/2} = LW^r_{N-1/2} = 0$, где $r = \rho, I, E$.

В итоге разностная схема, аппроксимирующая систему уравнений (6), может быть записана в известной балансной форме (см. [8], [10]), в которой потоки вычисляются согласно (11):

$$\frac{\mathbf{r}^{n+1}_i - \mathbf{r}^n_i}{\tau_n} + \frac{\mathbf{W}^n_{i+1/2} - \mathbf{W}^n_{i-1/2}}{h_i} = 0.$$

4. АПРИОРНЫЕ ОЦЕНКИ ВЕЛИЧИНЫ АДАПТИВНОЙ ИСКУССТВЕННОЙ ВЯЗКОСТИ

Вводимая в разностные уравнения искусственная вязкость, очевидно, должна быть по возможности минимальной, необходимой для подавления нефизичных осцилляций решения. Известно, что осцилляции сеточной природы не возникают, если разностная схема удовлетворяет условиям принципа максимума (см. [10]). Найдем ограничения на коэффициенты разностной схемы, при которых данные условия выполнены. Из них будут получены ограничения для коэффициента искусственной вязкости $\left(Q_{i+1/2}\right)^{(\rho)}_{\min}$ и $\left(Q_{i+1/2}\right)^{(\rho)}_{\max}$.

Проведем необходимые выкладки для уравнения неразрывности. Разностный аналог уравнения неразрывности с учетом искусственной диссипации представим в следующем виде:

$$\begin{gathered}\frac{\rho^{n+1}_i - \rho^n_i}{\tau} + \frac{1}{h_i}\left(\bar{u}_{i+1/2}\frac{\rho^n_{i+1}+\rho^n_i}{2} - \bar{u}_{i-1/2}\frac{\rho^n_i+\rho^n_{i-1}}{2}\right) - \\ -\frac{\tau}{2h_i}\left[\frac{\rho^n_{i+1}(u^2_{i+1}+c^2_{i+1}) - \rho^n_i(u^2_i+c^2_i)}{h_{i+1/2}} - \frac{\rho^n_i(u^2_i+c^2_i) - \rho^n_{i-1}(u^2_{i-1}+c^2_{i-1})}{h_{i-1/2}}\right] = \\ = \frac{1}{h_i}\left(q_{i+1/2}\frac{\rho^n_{i+1}-\rho^n_i}{h_{i+1/2}} - q_{i-1/2}\frac{\rho^n_i-\rho^n_{i-1}}{h_{i-1/2}}\right).\end{gathered} \tag{12}$$

Здесь $q_{i+1/2}$ — коэффициент искусственная вязкость, $c^2_i = \frac{p_i}{\rho_i}$.

Преобразуем формулу (12) следующим образом:

$$\begin{gathered}\rho^{n+1}_i = \rho^n_i - \frac{\tau}{2h_i}\left(\bar{u}_{i+1/2}\rho^n_{i+1} + (\bar{u}_{i+1/2} - \bar{u}_{i-1/2})\rho^n_i - \bar{u}_{i-1/2}\rho^n_{i-1}\right) + \\ -\frac{\tau}{2h_i}\left[\frac{u^2_{i+1}+c^2_{i+1}}{h_{i+1/2}}\rho^n_{i+1} - (u^2_i+c^2_i)\left(\frac{1}{h_{i+1/2}} + \frac{1}{h_{i-1/2}}\right)\rho^n_i\right] + \\ +\frac{u^2_{i-1}+c^2_{i-1}}{h_{i-1/2}}\rho^n_{i-1} + \frac{\tau}{h_i}\left(\frac{q_{i+1/2}}{h_{i+1/2}}\rho^n_{i+1} - \left(\frac{q_{i+1/2}}{h_{i+1/2}} - \frac{q_{i-1/2}}{h_{i-1/2}}\right)\rho^n_i + \frac{q_{i-1/2}}{h_{i-1/2}}\rho^n_{i-1}\right).\end{gathered} \tag{13}$$

В итоге расчетная формула для ρ^{n+1}_i приобретает вид

$$\rho^{n+1}_i = A_i\,\rho^n_{i+1} + B_i\,\rho^n_i + C_i\,\rho^n_{i-1}, \tag{14}$$

где

$$\begin{gathered}A_i = \frac{\tau}{2h_i}\left(-\bar{u}_{i+1/2} + \frac{\tau}{h_{i+1/2}}(u^2_{i+1}+c^2_{i+1}) + 2\frac{q_{i+1/2}}{h_{i+1/2}}\right), \\ B_i = 1 - \frac{\tau}{h_i}\left[\frac{1}{2}(\bar{u}_{i+1/2} - \bar{c}_{i-1/2}) + \frac{\tau}{2}(u^2_i+c^2_i)\left(\frac{1}{h_{i+1/2}} + \frac{1}{h_{i-1/2}}\right) + \left(\frac{q_{i+1/2}}{h_{i+1/2}} + \frac{q_{i-1/2}}{h_{i-1/2}}\right)\right], \\ C_i = \frac{\tau}{2h_i}\left(\bar{u}_{i-1/2} + \frac{\tau}{h_{i-1/2}}(u^2_{i-1}+c^2_{i-1}) + 2\frac{q_{i-1/2}}{h_{i-1/2}}\right).\end{gathered} \tag{15}$$

Решение системы уравнений (12) будет подчиняться принципу максимума в случае выполнения следующих условий (см. [10]): $A_i \geqslant 0$, $B_i \geqslant 0$, $C_i \geqslant 0$ и $A_i + B_i + C_i \leqslant 1$.

Условия $A_i \geqslant 0$ приводят, в соответствии с выражениями (15), к неравенствам

$$q_{i+1/2} \geqslant \frac{h_{i+1/2}}{2}\left(\bar{u}_{i+1/2} - \frac{\tau}{h_{i+1/2}}(u_{i+1}^2 + c_{i+1}^2)\right),$$

откуда следует оценка коэффициента вязкости

$$q_{i+1/2} \geqslant \left(Q_{i+1/2}\right)_{\min}^{(\rho)} = \frac{h_{i+1/2}}{2}\left[\left|u_{i+1/2}\right| - \frac{\Delta t}{h_{i+1/2}}\left(u_{i+1}^2 + \frac{p_{i+1}}{\rho_{i+1}}\right)\right]. \tag{16}$$

Аналогично из условия $C_i \geqslant 0$ следует оценка

$$q_{i-1/2} \geqslant \left(Q_{i-1/2}\right)_{\min}^{(\rho)} = \frac{h_{i-1/2}}{2}\left[\left|u_{i-1/2}\right| - \frac{\Delta t}{h_{i-1/2}}\left(u_{i-1}^2 + \frac{p_{i-1}}{\rho_{i-1}}\right)\right]. \tag{17}$$

Заменим в уравнении (16) u_{i+1}^2 и в уравнении (17) u_{i-1}^2 на меньшие (или такие же) величины $\bar{u}_{i+1/2}^2 = \min(u_{i+1}^2, u_i^2)$; $\bar{u}_{i-1/2}^2 = \min(u_i^2, u_{i-1}^2)$.

Заменим также в уравнениях (16) и (17) отношения $\frac{p_{i+1}}{\rho_{i+1}}, \frac{p_{i-1}}{\rho_{i-1}}$ на меньшие (или равные им) величины

$$\bar{c}_{i+1/2}^2 = \min\left(\frac{p_{i+1}}{\rho_{i+1}}, \frac{p_i}{\rho_i}\right), \quad \bar{c}_{i-1/2}^2 = \min\left(\frac{p_i}{\rho_i}, \frac{p_{i-1}}{\rho_{i-1}}\right).$$

Потребуем, чтобы во всех ячейках расчетной сетки выполнялись более сильные условия, чем в уравнениях (16)—(17):

$$q_{i+1/2} \geqslant \left(Q_{i+1/2}\right)_{\min}^{(\rho)} = \frac{h_{i+1/2}}{2}\left[\left|u_{i+1/2}\right| - \frac{\Delta t}{h_{i+1/2}}\left(\bar{u}_{i+1/2}^2 + \bar{c}_{i+1/2}^2\right)\right]. \tag{18}$$

Очевидно, при выполнении условия

$$\Delta t < \frac{h_{i+1/2}\,\left|u_{i+1/2}\right|}{\bar{u}_{i+1/2}^2 + \bar{c}_{i+1/2}^2} \tag{19}$$

величина $\left(Q_{i+1/2}\right)_{\min}^{(\rho)}$ будет положительной. Найдем условия для выполнения неравенств $B_i \geqslant 0$. Для этого перепишем второе из равенств (15) в виде

$$\frac{\Delta t}{h_i}\left(\frac{q_{i+1/2}}{h_{i+1/2}} + \frac{q_{i-1/2}}{h_{i-1/2}}\right) = 1 - \frac{\Delta t}{h_i}\,\frac{u_{i+1/2}}{2} + \frac{\Delta t}{h_i}\,\frac{u_{i-1/2}}{2} - \frac{(\Delta t)^2}{2\,h_i\;h_{i+1/2}}\left(u_i^2 + \frac{p_i}{\rho_i}\right) - \frac{(\Delta t)^2}{2\,h_i\;h_{i-1/2}}\left(u_i^2 + \frac{p_i}{\rho_i}\right).$$

Отсюда следует, что условия $B_i \geqslant 0$ выполнены при выполнении неравенства

$$\begin{aligned}\frac{q_{i+1/2}}{h_{i+1/2}} + \frac{q_{i-1/2}}{h_{i-1/2}} \leqslant \frac{\left(Q_{i+1/2}\right)_{\max}^{(\rho)}}{h_{i+1/2}} + \frac{\left(Q_{i-1/2}\right)_{\max}^{(\rho)}}{h_{-1/2}} = \frac{h_i}{\Delta t} - \frac{u_{i+1/2}}{2} + \frac{u_{i-1/2}}{2} - \frac{\Delta t}{2h_{i+1/2}}\left(u_i^2 + \frac{p_i}{\rho_i}\right) - \\ - \frac{\Delta t}{2h_{i-1/2}}\left(u_i^2 + \frac{p_i}{\rho_i}\right) = \frac{h_{i+1/2}}{2\,\Delta t} + \frac{h_{i-1/2}}{2\,\Delta t} - \frac{u_{i+1/2}}{2} + \frac{u_{i-1/2}}{2} - \frac{\Delta t}{2h_{i+1/2}}\left(u_i^2 + \frac{p_i}{\rho_i}\right) - \frac{\Delta t}{2h_{i-1/2}}\left(u_i^2 + \frac{p_i}{\rho_i}\right).\end{aligned} \tag{20}$$

Здесь учтено, что $h_i = \frac{1}{2}\left(h_{i+1/2} + h_{i-1/2}\right)$.

Будем требовать выполнение более сильного неравенства, чем (20). Заменим

$$\frac{u_i^2 + \frac{p_i}{\rho_i}}{h_{i+1/2}} \quad \text{и} \quad \frac{u_i^2 + \frac{p_i}{\rho_i}}{h_{i-1/2}}$$

соответственно на большие (или равные) величины

$$\frac{\bar{u}_{i+1/2}^2 + \bar{c}_{i+1/2}^2}{h_{i+1/2}} \quad \text{и} \quad \frac{\bar{u}_{i-1/2}^2 + \bar{c}_{i-1/2}^2}{h_{i-1/2}},$$

где

$$\bar{u}_{i+1/2}^2 = \max\left(u_i^2, u_{i+1}^2\right), \quad \bar{c}_{i+1/2}^2 = \max\left(\frac{p_i}{\rho_i}, \frac{p_{i+1}}{\rho_{i+1}}\right), \quad \bar{u}_{i-1/2}^2 = \max\left(u_i^2, u_{i-1}^2\right), \quad \bar{c}_{i-1/2}^2 = \max\left(\frac{p_{i-1}}{\rho_{i-1}}, \frac{p_i}{\rho_i}\right).$$

Отсюда вытекает ограничение на величину искусственной диссипации в следующем виде:

$$q_{i+1/2} \leqslant \left(Q_{i+1/2}\right)^{(\rho)}_{\max} = \frac{0.5\ h^2_{i+1/2}}{\Delta t} - \frac{\left|u_{i+1/2}\right|}{2}\, h_{i+1/2} - \frac{\Delta t}{2}\left(\bar{u}^2_{i+1/2} + \bar{c}^2_{i+1/2}\right). \tag{21}$$

Условие (21) должно выполняться для всех ячеек расчетной сетки.

Данное условие означает, что

$$\frac{h^2_{i+1/2}}{\Delta t} - \left|u_{i+1/2}\right|\ h_{i+1/2} - \Delta t\left(u^2_{i+1/2} + \bar{c}^2_{i+1/2}\right) > 0, \quad \text{или} \quad \Delta t < \frac{h_{i+1/2}\ \left|u_{i+1/2}\right|}{\bar{u}^2_{i+1/2} + \bar{c}^2_{i+1/2}}. \tag{22}$$

Ограничим шаг по времени Δt сильнее, чем следует из (22):

$$\frac{h^2_{i+1/2}}{\Delta t} - 2\ \left|u_{i+1/2}\right|\ h_{i+1/2} > 0,$$

или

$$\Delta t < \frac{h_{i+1/2}}{2\ \left|u_{i+1/2}\right|}. \tag{23}$$

Условие (22) гарантирует выполнение неравенства $B_i \geqslant 0$.

Если $\bar{c}^2_{i+1/2} > \bar{u}^2_{i+1/2}$, то для шага Δt необходимо выполнение ограничения (22), в противном случае Δt ограничивается неравенством (23).

При проведении расчетов величина искусственной вязкости должна определяться с учетом неравенств (18) и (21):

$$\left(Q_{i+1/2}\right)^{(\rho)}_{\min} \leqslant q_{i+1/2} \leqslant \left(Q_{i+1/2}\right)^{(\rho)}_{\max}. \tag{24}$$

5. РАСЧЕТЫ МОДЕЛЬНЫХ ЗАДАЧ

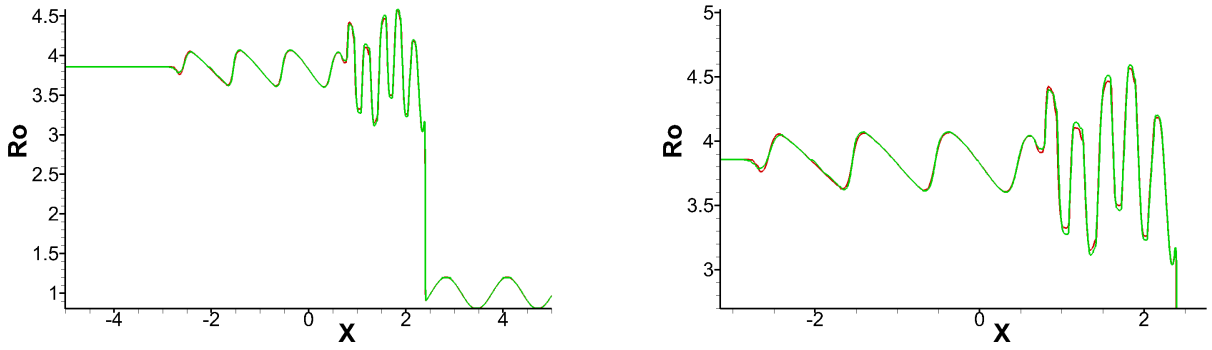
Для практического выяснения возможностей предложенного модифицированного метода АИВ были выполнены расчеты трех тестовых задач. Данные расчеты продемонстрировали возможность использования неравномерных сеток с достаточно большой разницей размеров соседних ячеек в расчетах течений, как с сильными, так и со слабыми разрывами, а также в случаях, когда начальные распределения газодинамических величин немонотонны.

В проведенных расчетах на обеих границах расчетной области задавались постоянные значения искомых функций, равные значениям в начальный момент времени. Тем самым расчеты были ограничены во времени выходом на границы области моделируемых волновых процессов.

В качестве первого теста была выбрана задача Шу–Ошера о взаимодействии сильного разрыва с гладкой, волнообразной маломасштабной структурой. Постановка задачи приведена в [11]. Ударная волна налетает на область покоящегося газа, в которой плотность газа возмущена по синусоиде. Течение газа рассчитывается в области $-5 \leqslant x \leqslant 5$, содержащей точку разрыва с координатой $x = -4$. Начальные значения плотности, скорости и давления на участке $-5 \leqslant x \leqslant -4$ постоянны и имеют следующие значения: $\rho = 3.857143$, $u = 2.629369$, $P = 10.33333$. В остальной расчетной области начальные распределения имеют вид $\rho = 1 + 0.2\sin(5x)$, $u = 0$, $P = 1$. Расчет выполнялся в интервале времени $0 < T \leqslant 1.8$. При прохождении ударной волны с числом Маха $M = 3$ через участок изначально возмущенного распределения плотности возникают осцилляции плотности в виде набора волн, длины которых отличаются от длины волны первоначального возмущения.

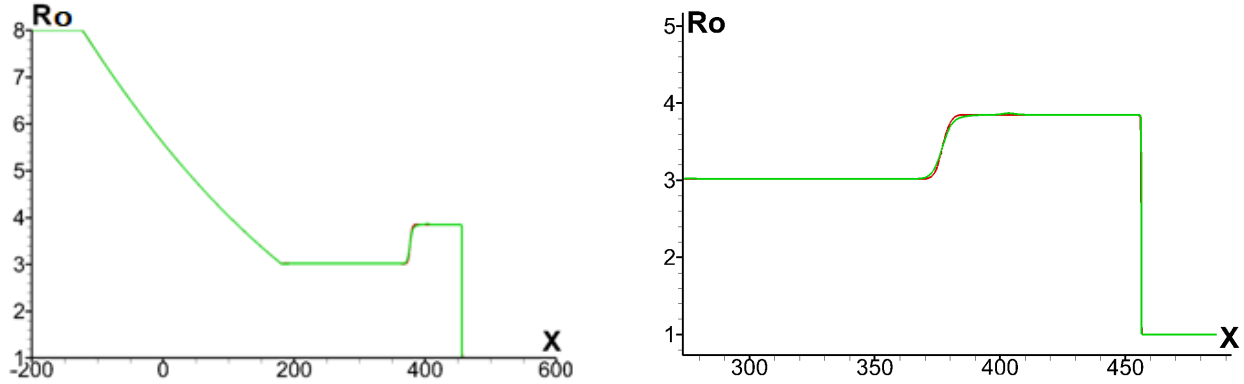
При тестировании метода АИВ с постоянным шагом по пространству расчеты проводились на сетке с числом ячеек 2000. Для модифицированного метода АИВ с переменным шагом расчетная сетка имела то же количеством узлов, при этом со сгущением в области формирования осцилляций за ударной волной с перепадом величины ячеек в 2 раза.

Проведенные расчеты показали повышение точности решения на неравномерной сетке. При этом увеличивается число шагов по времени, что связано с уменьшением шага по времени, так как для обеспечения устойчивости разностной схемы требуется выполнение условия Куранта. Заметим, что для получения той же точности в исследуемой области на равномерной сетке необходимо количество узлов почти в 2 раза больше, соответственно, время расчета увеличится практически в 2 раза. Результаты решения задачи (распределение плотности) на равномерной и неравномерной сетках представлены на фиг. 2. Расчеты на равномерной сетке изображены красным цветом, а на неравномерной сетке – зеленым цветом. На неравномерной сетке сгущение осуществлялось в области $-3 < x < 2.5$, измельчение шага было в 2 раза.



(а) (б)

Фиг. 2. Задача 1: (а) – решение во всей расчетной области; (б) – увеличенное изображение решения в области изменения частоты возмущения плотности.



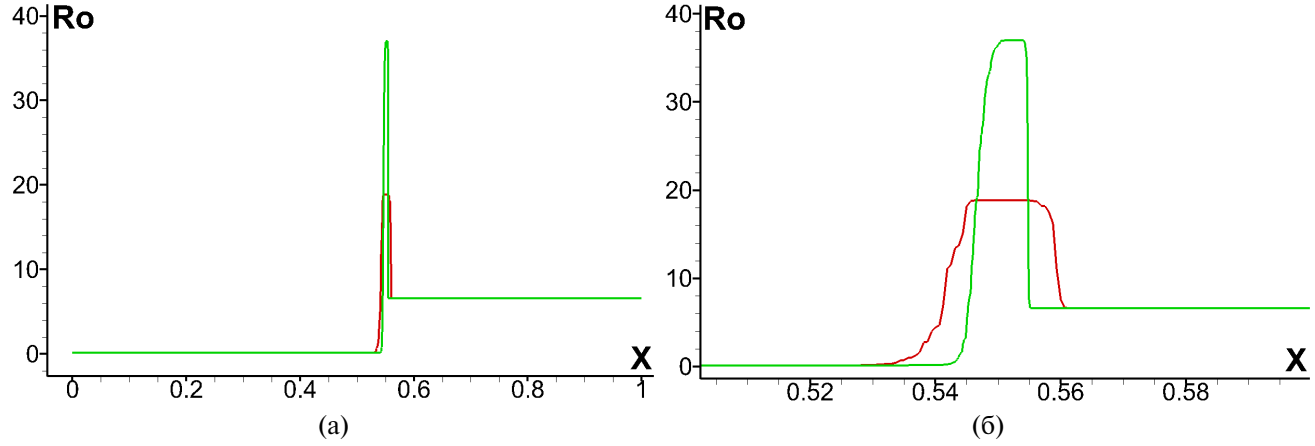
(а) (б)

Фиг. 3. Задача 2: (а) – решение во всей расчетной области; (б) – увеличенное изображение решения в области контактного разрыва и ударной волны.

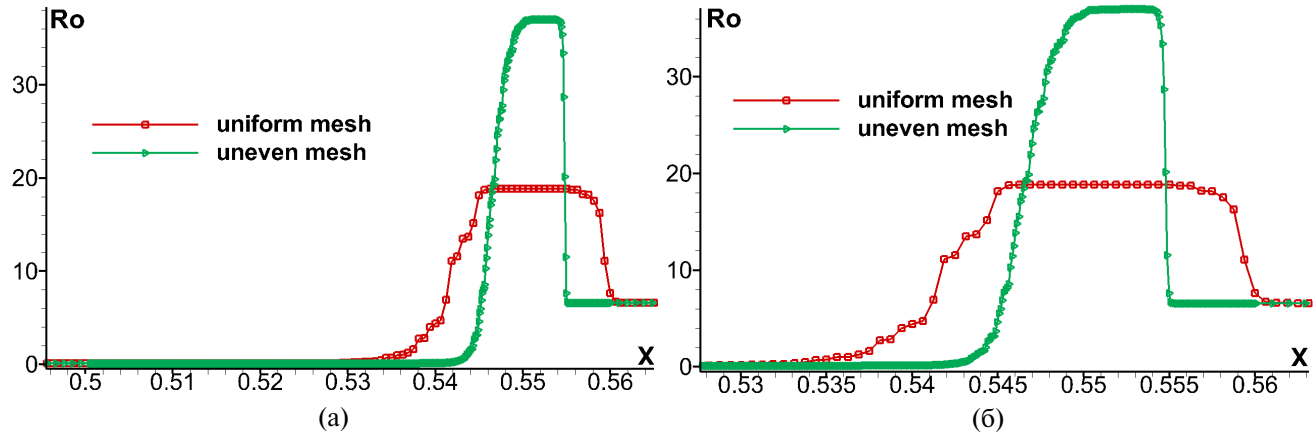
В качестве второго теста была выбрана задача о распаде сильного разрыва при весьма существенных перепадах плотности и давления, что приводит к возникновению сильной ударной волны (задача Сода, см. [11], [12]). В начальный момент времени перепад плотности составлял 8 безразмерных единиц, а давления 320 единиц. В начальный момент газ покоится во всей расчетной области. Разрыв плотности и давления в начальный момент находился в точке $x = 150$. В данном случае важно правильно рассчитать положение разрывов и подавить осцилляции численной природы в областях постоянства параметров потока. Использование неравномерной сетки приводит к уточнению результатов расчета. В данном случае их отклонение от точного решения не превышает 1–2 %. Результаты расчетов задачи Сода (распределение плотности) представлены на фиг. 3: на неравномерной сетке – красным цветом, а на равномерной сетке – зеленым цветом. Сгущение сетки (в 2 раза) осуществлялось в двух областях: 1) $140 < x < 160$, 2) $250 < x < 600$.

В качестве третьего теста была выбрана задача Римана о возникновении “пика плотности” (test peak), описание которой можно найти в [12]. Задача представляет интерес тем, что ее решение описывает ударную волну и контактный разрыв с весьма большим перепадом плотности (около 300). Как отмечено в [12] и других работах, при решении задачи в такой постановке немонотонность разностной схемы может вызвать неустойчивость счета. Также в этой публикации отмечено, что, поскольку профиль плотности характеризуется узким “пиком”, расчеты по многим численным методикам повышенной точности не позволяют достичь референсного значения амплитуды пика плотности $\|\rho_{\mathrm{ref}}(x)\|_c = 38.6$.

В двух вариантах решения задачи “test peak” были использованы равномерная и неравномерная сетки, содержавшие одинаковое количество ячеек – 1600. Согласно описанию теста из [12], расчеты выполнялись до момента времени $t = 0.0039$. Особенность тестового решения такова, что область пика в нем отделяет ударную волну от контактного разрыва. В этом случае для расчета течения с высоким разрешением структуры необходима точная регулировка уровня численной диссипации. Для повышения точности расчета по схеме с АИВ было



Фиг. 4. Задача 3: (а) – решение во всей расчетной области; (б) – увеличенное изображение решения в области контактного разрыва и ударной волны.



Фиг. 5. Задача 3: (а) – изображение фрагмента решения, точками отмечены положения узлов расчетной сетки; (б) – увеличенное изображение решения в области контактного разрыва и ударной волны.

существенно увеличено количество расчетных точек в области формирования пика распределения. В варианте расчета на неравномерной сетке уменьшение пространственного шага в области $0.5 \leqslant x \leqslant 0.56$ в 10 раз позволило повысить значение амплитуды плотности почти в 2 раза. Получены следующие числовые данные. Расчет на равномерной сетке дал значение плотности в максимуме $\rho_1 = 19.10$.Относительное отклонение плотности, вычисленное в сеточной норме C $\|\Delta\|_c = \|(\rho_{\text{ref}}(x) - \rho(x))/\rho_{\text{ref}}(x)\|_c$ в варианте равномерной сетки оказалось весьма значительным: $\|\Delta_1\|_c \approx 0.51$. В варианте расчета на неравномерной сетке со сгущением в области резкого увеличения градиента плотности получено значение пиковой плотности $\rho_2 = 37.13$, при этом относительное отклонение существенно уменьшилось: $\|\Delta_2\|_c \approx 0.04$. Таким образом, для данного теста переменный пространственный шаг является принципиально важным требованием, которое позволяет качественно улучшить результат расчета при фиксированном числе разностных ячеек.

Результаты расчетов задачи о распространении сильной ударной волны представлены на фиг. 4: в варианте равномерной сетки они выделены красным цветом, в варианте неравномерной – зеленым цветом. Шаг разностной сетки измельчен в 10 раз в области $0.5 \leqslant x \leqslant 0.56$.

На фиг. 5 показано распределение узлов разностной сетки в области больших градиентов. Узлы равномерной разностной сетки отображаются квадратиками, а неравномерной разностной сетки – треугольниками.

ЗАКЛЮЧЕНИЕ

Разработанная нами модификация метода АИВ продемонстрировала при выполнении вычислительных экспериментов достаточно хорошую точность при использовании неравномерных расчетных сеток. Для сеток, которые не удовлетворяют условию квазиравномерности, метод позволяет сохранить качественные и количе-

ственные характеристики структур в рассчитываемом течении. Число Маха при этом может меняться в весьма широких пределах. Вычислительные эксперименты, приведенные в настоящей работе, показали, что при выполнении условия Куранта адаптивная вязкость позволяет производить вычисления с сохранением свойств монотонности и устойчивости разностных решений при весьма большой разнице размеров соседних расчетных ячеек. Так, например, один из тестов рассчитан на сетке, в которой размеры соседних ячеек отличаются на порядок величины.

Предложенный в настоящей работе численный метод достаточно прост в реализации, при этом он позволяет создавать экономичные алгоритмы расчетов течений при наличии локальных областей больших градиентов газодинамических величин. Адаптивная искусственная вязкость может использоваться в более сложных компьютерных моделях течений сжимаемых сред. Так, например, для расчетов высокоскоростных течений плазмы в астрофизических условиях возникает необходимость рассчитывать отдельно динамику ионов и нейтральных частиц (см. [13]). В такой модели движения частично ионизованной среды адаптивная вязкость может непосредственно использоваться для расчета балансов массы, импульса и энергии нейтральной и ионизованной компонент среды.

СПИСОК ЛИТЕРАТУРЫ

1. *Попов И.В., Фрязинов И.В.* Метод адаптивной искусственной вязкости численного решения уравнений газовой динамики. М.: Красанд, 2015. 200 с.
2. *Калиткин Н.Н., Альшин А.Б, Альшина Е.А., Рогов Б.В.* Вычисления на квазиравномерных сетках. М.: Физматлит, 2005. 224 с.
3. *Калиткин Н.Н., Кузнецов И.О., Панченко С.Л.* Метод квазиравномерных сеток в бесконечной области // ДАН. 2000. Т. 374. № 5. С. 598–601.
4. *Дарьин Н.А., Мажукин В.И., Самарский А.А.* Конечно-разностный метод решения одномерных уравнений газовой динамики на адаптивных сетках // Докл. АН СССР. 1988. Т. 302. № 5. С. 1078–1081.
5. *Berger M.J., Oliger J.* Adaptive mesh refinement for hyperbolic partial differential equations // J. Comput. Phys. 1984. V. 53. № 3. P. 484–512. doi 10.1016/0021-9991(84)90073-1.
6. *Berger M.J., Colella P.* Local adaptive mesh refinement for shock hydrodynamics // J. Compu. Phys. 1989. V. 82. № 1. P. 64–84. doi:10.1016/0021-9991(89)90035-1.
7. *Василевский В.Ф., Вязников К.В., Тишкин В.Ф., Фаворский А.П.* Квазимонотонные разностные схемы повышенного порядка точности на адаптивных сетках нерегулярной структуры // Препринт ИПМ № 124. Москва, 1990. 31 с.
8. *Годунов С.К., Забродин А.В., Иванов М.Я., Крайко А.Н., Прокопов Г.П.* Численное решение многомерных задач газовой динамики. М.: Наука, 1976. 400 с.
9. *Годунов С.К., Рябенький В.С.* Разностные схемы. М.: Наука, 1977. 440 с.
10. *Самарский А.А.* Теория разностных схем. М.: Наука, 1971. 553 с.
11. *Shu C., Osher S.* Efficient implementation of essentially non-oscillatory shook-capturing schemes II // J. Comput. Phys. 1989. V. 83. P. 32–78.
12. *Liska R., Wendroff B.* Comparison of several difference schemes on 1D and 2D test problems for the Euler equations // SIAM J. Sci. Comput. 2003. V. 25. № 3. P. 31–84. DOI: 10.1137/S1064827502402120
13. *Шарова Ю.С., Глазырин С.И., Гасилов В.А.* Исследование влияния фоновой нейтральной компоненты на динамику оболочки в остатках сверхновых // Письма в Астрон. журн. 2021. Т. 47. № 11. С. 773–781. DOI: 10.31857/S032001082111005X

DEVELOPMENT OF THE METHOD OF ADAPTIVE ARTIFICIAL VISCOSITY FOR FLUID DYNAMICS COMPUTATIONS ON NONUNIFORM DIFFERENCE GRIDS

A. Yu. Krukovsky[a], I. V. Popov[a,*], V. A. Gasilov[a]

[a] *Keldysh Institute of Applied Mathematics, RAS, Moscow, Russia*
**e-mail: piv2964@mail.ru*

Received 5 July, 2024
Revised 27 July, 2024
Accepted 23 August, 2024

Abstract. The method of adaptive artificial viscosity is generalized to construct difference schemes for fluid dynamics that ensure high resolution of the structure of flows both on uniform and nonuniform grids. Difference schemes approximating the one-dimensional system of fluid dynamics equations are considered. Bounds on the magnitude of adaptive viscosity obtained in this paper take into account the nonuniformity of the distribution of gas-dynamic quantities in the computational domain and the nonuniformity of the difference grid. The constructed schemes with adaptive artificial viscosity are homogeneous and conservative. These schemes are evaluated on model problems the solutions to which describe various smooth gas-dynamic structures, as well as strong and weak discontinuities. The possibility of obtaining highly accurate solutions on grids with significant difference of geometric size of adjacent difference cells is demonstrated.

Keywords: fluid dynamics, adaptive artificial viscosity, homogeneous conservative difference scheme, computational experiment.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2401–2410

УДК 519.635

МОДЕЛИРОВАНИЕ НЕЛИНЕЙНЫХ ВОЛНОВЫХ ПРОЦЕССОВ В СВЧ-ГЕНЕРАТОРЕ С МАГНИТНОЙ ИЗОЛЯЦИЕЙ

© 2024 г. С.В. Поляков[1,*], Н.И. Тарасов[1,], Т.А. Кудряшова[1,***]**

[1] 125047 Москва, Миусская пл., 4, ИПМ им. М.В. Келдыша РАН, Россия
**e-mail: polyakov@imamod.ru*
***e-mail: nikita_tarasov@imamod.ru*
****e-mail: kudryasova@imamod.ru*

Поступила в редакцию 11.06.2024 г.
Переработанный вариант 28.07.2024 г.
Принята к публикации 23.08.2024 г.

Рассмотрена актуальная задача моделирования нелинейных волновых процессов в СВЧ-генераторе с магнитной изоляцией. Для ее численного анализа предложена новая компьютерная модель, включающая уравнения Максвелла и уравнения движения релятивистских заряженных частиц, совместное их интегрирование методом сеток и методом облачных частиц, а также параллельную программную реализацию. В численных экспериментах получены пространственно-временные характеристики релятивистских электронных пучков и плазмы, а также параметры выходного излучения генератора. Анализ полученных результатов подтвердил корректность разработанного численного подхода. Библ. 17. Фиг. 7.

Ключевые слова: математическое моделирование, волновые процессы, СВЧ-генераторы излучения, релятивистские пучки, низкотемпературная плазма.

DOI: 10.31857/S0044466924120139, **EDN**: KBMIPO

ВВЕДЕНИЕ

С середины прошлого века одним из интенсивно развивающихся научных и инженерных направлений является СВЧ-электроника [1–7]. Сфера приложений СВЧ-приборов охватывает практически весь ареал высокотехнологичных отраслей промышленности, а также энергетику, связь, медицину и другие отрасли хозяйства. Актуальным и востребованным направлением СВЧ-электроники является разработка широкополосных плазменных генераторов излучения [6–9]. Применение этих приборов часто связывается с обработкой поверхностей твердых материалов. Например, такие генераторы применяются для закаливания металлических поверхностей в станкостроении. Однако плазменные генераторы используются в системах радиосвязи, а также являются основой нанотехнологий отжига и напыления в современной микро- и наноэлектронике. Известно применение СВЧ-генераторов и в медицинских целях при облучении раковых клеток на коже и в приповерхностном слое тканей различных органов. Применительно к нанотехнологиям и медицине наиболее востребованными являются короткоимпульсные плазменные генераторы, которые оказывают лишь локальное неразрушающее действие на материал поверхности. В данной работе основное внимание уделяется расчету параметров физических процессов, происходящих в подобных СВЧ-приборах.

Целью настоящей работы является развитие и применение новой компьютерной модели для расчета функционирования релятивистского плазменного СВЧ-генератора с магнитной изоляцией [8, 9] и проведение цикла исследований, связанных с ее тестированием и апробацией. Заметим, что в основе таких приборов лежит генерация релятивистского электронного пучка (РЭП) и его взаимодействие с плазменным слоем. В результате этого взаимодействия энергия поступающего в прибор электромагнитного импульса сначала передается электронам пучка, затем преобразуется в энергию пучка и плазмы, и наконец преобразуется в энергию выходного излучения.

В работах [10, 11] нами уже представлялся прототип компьютерной модели, позволяющий рассчитывать процессы эмиссии электронов с поверхности металлов и композитных материалов и генерации релятивистских электронных пучков. В основу модели была положена полная система уравнений Максвелла [12] и метод облачных частиц [13]. Общая система уравнений Максвелла интегрировалась известным явным сеточным

методом FDTD [14] на декартовых разнесенных сетках. Уравнения для облачных частиц интегрировались по явной симметричной схеме Адамса [15]. В работе [10] центральным местом была численная реализация механизма электронной эмиссии и формирования с его помощью РЭП с адекватными экспериментам параметрами. В работе [11] была предложена и верифицирована компьютерная модель плазмы. В настоящей работе рассмотрено объединение моделей и решение практической задачи о генерации излучения в цилиндрически симметричной геометрической конструкции с параметрами близкими к реальным экспериментам. Результаты расчетов качественно сравнивались с результатами, полученными с помощью кода KARAT [16].

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

В данном разделе коротко опишем используемую нами математическую модель. Для определенности рассмотрим далее случай цилиндрической симметрии. Основу модели составляют уравнения Максвелла, которые в международной системе единиц (СИ) с учетом материальных уравнений имеют следующий вид:

$$\begin{aligned} &\operatorname{div}(\varepsilon_a \mathbf{E}) = \rho, \quad \operatorname{div}\mathbf{B} = 0; \\ &\operatorname{rot}\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}; \quad \operatorname{rot}\left(\frac{1}{\mu_a}\mathbf{B}\right) = \mathbf{j} + \frac{\partial}{\partial t}(\varepsilon_a \mathbf{E}). \end{aligned} \tag{1}$$

Здесь $\mathbf{E}$ – вектор напряженности электрического поля, $\mathbf{B}$ – вектор магнитной индукции, div и rot – дифференциальные операторы дивергенции и ротора в случае цилиндрической симметрии в координатах (r, φ, z), определенные соотношениями:

$$\begin{aligned} &\mathbf{A} \equiv A_r \cdot \mathbf{i}_r + A_\varphi \cdot \mathbf{i}_\varphi + A_z \cdot \mathbf{i}_z, \quad \operatorname{div}\mathbf{A} \equiv \frac{1}{r}\frac{\partial}{\partial r}(rA_r) + \frac{\partial A_z}{\partial z}, \\ &\operatorname{rot}\mathbf{A} \equiv \left(-\frac{\partial A_\varphi}{\partial z}\right)\cdot \mathbf{i}_r + \left(\frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r}\right)\cdot \mathbf{i}_\phi + \left(\frac{1}{r}\frac{\partial}{\partial r}\left(rA_\varphi\right)\right)\cdot \mathbf{i}_z, \end{aligned}$$

где $\mathbf{i}_r, \mathbf{i}_\varphi, \mathbf{i}_z$ – направляющие векторы, $\frac{\partial}{\partial t}$ – производная по времени t, $\rho = \rho_- + \rho_+$ – объемная плотность зарядов, ρ_- и ρ_+ – суммарные плотности отрицательно и положительно заряженных частиц, $\mathbf{j} = \mathbf{j}_- + \mathbf{j}_+$ – суммарная плотность токов частиц, $\mathbf{j}_- = \rho_- \mathbf{v}_-$ и $\mathbf{j}_+ = \rho_+ \mathbf{v}_+$ – плотности токов частиц, $\mathbf{v}_-$ и $\mathbf{v}_+$ – средние скорости частиц, $\varepsilon_a = \varepsilon\varepsilon_0$ и $\mu_a = \mu\mu_0$ – абсолютные диэлектрическая и магнитная проницаемости среды, μ_0 и ε_0 – магнитная и диэлектрическая проницаемости вакуума, $c = \sqrt{\varepsilon_0\mu_0}$ – скорость света.

Уравнения (1) рассматриваются в области цилиндра Ω, не занятой металлическими элементами конструкции – катодом Ω_C и анодом Ω_A, т.е. в подобласти $\Omega_D = \Omega / (\Omega_C \cup \Omega_A)$.

Граничные условия для векторных полей $\mathbf{E}$ и $\mathbf{B}$ рассматриваются в приближении идеальной проводимости на металлических поверхностях и задаются в виде:

$$[\mathbf{E} \times \mathbf{n}] = 0, \quad (\mathbf{B} \cdot \mathbf{n}) = 0, \quad \mathbf{r} \in \partial\Omega_A \cup \partial\Omega_C. \tag{2}$$

Граничные условия на свободных границах $\partial\Omega_F$ имеют вид:

$$(\mathbf{E} \cdot \mathbf{n}) = 0, \quad (\mathbf{B} \cdot \mathbf{n}) = 0, \quad \mathbf{r} \in \partial\Omega_F. \tag{3}$$

Граничные условия на оси симметрии ($r = 0$) задаются следующим образом:

$$[\mathbf{E} \times \mathbf{l}_z] = 0, \quad (\mathbf{B} \cdot \mathbf{l}_z) = 0, \quad \mathbf{l}_z = (0, 0, 1)^T. \tag{4}$$

В начальный момент динамические компоненты электромагнитного поля отсутствуют.

В дополнение к системе (1)–(4) рассматривается следующая релятивистская модель частиц:

$$\begin{aligned} &\frac{d\mathbf{r}_{\alpha,k}}{dt} = \mathbf{v}_{\alpha,k}, \quad \frac{d\mathbf{p}_{\alpha,k}}{dt} = q_{\alpha,k}\left(\mathbf{E} + [\mathbf{v}_{\alpha,k} \times \mathbf{B}]\right), \\ &\mathbf{p}_{\alpha,k} = m_{\alpha,k}\mathbf{v}_{\alpha,k}\gamma_{\alpha,k}, \quad \gamma_{\alpha,k} = 1/\sqrt{1 - (v_{\alpha,k}/c)^2}, \quad k = 1, \ldots, N_\alpha; \end{aligned} \tag{5}$$

$$\begin{aligned} &\rho_\alpha = \sum_{k=1}^{N_\alpha} q_{\alpha,k}\delta(\mathbf{r} - \mathbf{r}_{\alpha,k}), \quad \mathbf{j}_\alpha = \sum_{k=1}^{N_\alpha} q_{\alpha,k}\delta(\mathbf{r} - \mathbf{r}_{\alpha,k})\mathbf{v}_{\alpha,k}, \\ &\mathbf{v}_\alpha = \frac{1}{N_\alpha}\sum_{k=1}^{N_\alpha}\mathbf{v}_{\alpha,k}, \quad \mathbf{p}_\alpha = \frac{1}{N_\alpha}\sum_{k=1}^{N_\alpha}\mathbf{p}_{\alpha,k}. \end{aligned} \tag{6}$$

Здесь $\mathbf{r}_{\alpha,k}$, $\mathbf{v}_{\alpha,k}$, $\mathbf{p}_{\alpha,k}$, $q_{\alpha,k}$, $m_{\alpha,k}$, $\gamma_{\alpha,k}$ – радиус вектор, вектор скорости, вектор импульса, заряд, масса и релятивистский фактор частицы сорта α с номером k, $\delta\left(\mathbf{r}-\mathbf{r}_{\alpha,k}\right)$ – модифицированная в соответствии с [10, 11] дельта-функция Дирака, описывающая плотность заряда частицы сорта α, N_α – число частиц сорта α, $v_{\alpha,k}=|\mathbf{v}_{\alpha,k}|$ – модуль скорости частицы.

Граничные условия на эмиссионной поверхности для вновь появляющихся в результате эмиссии отрицательно заряженных частиц имеют вид:

$$q_{-,k}=\left[\left|(\varepsilon_a\mathbf{E}\cdot\mathbf{n})\right|\Delta S_C\right],\quad \mathbf{r}_{-,k}=\mathbf{r},\quad \mathbf{v}_{-,k}=\mathbf{v}^0_{-,k},\quad \mathbf{p}_{-,k}=\mathbf{p}^0_{-,k},\quad \mathbf{r}\in\partial\Omega_C. \tag{7}$$

Здесь k – номер новой частицы, $q_{-,k}$ – ее заряд, $\mathbf{r}$ – радиус-вектор элемента эмиссионной поверхности (эмиссионного центра), $\mathbf{n}$ – вектор нормали к поверхности, ΔS_C – площадь элемента, $\mathbf{v}^0_{-,k}$ и $\mathbf{p}^0_{-,k}$ – начальная скорость и импульс новой частицы, которые вычисляются по температуре эмиттера T_{emi}:

$$\mathbf{v}^0_{-,k}=\sqrt{3k_BT_{emi}/m_{-,k}},\quad \mathbf{p}^0_{-,k}=m_{-,k}\mathbf{v}^0_{-,k}\gamma^0_{-,k}.$$

Здесь k_B – константа Больцмана, $\gamma^0_{-,k}$ – начальный релятивистский фактор.

Граничные условия свободного выхода частиц из расчетной области или поглощения металлическими поверхностями реализуются с помощью геометрического анализа их положения и соответствующей процедуры удаления. На границах диэлектриков ставится условие либо непротекания, либо полного или частичного отражения.

В начальный момент времени в рассматриваемой области либо частиц нет, либо в отдельных подобластях имеются частицы квазинейтральной холодной плазмы с общей температурой T_0 и нулевым суммарным импульсом:

$$\begin{gathered}\mathbf{v}_{\alpha,k}=\mathbf{v}^0_{\alpha,k},\quad \mathbf{p}_{\alpha,k}=\mathbf{p}^0_{\alpha,k},\quad k=1,\ldots,N^0_\alpha,\quad \rho_\alpha=\rho^0_\alpha=\sum_{k=1}^{N^0_\alpha}q^0_{\alpha,k}\delta\left(\mathbf{r}-\mathbf{r}_{\alpha,k}\right),\\ \sum_\alpha\frac{1}{2N^0_\alpha}\sum_k m_{\alpha,k}\left(\mathbf{v}^0_{\alpha,k}\right)^2=\frac{3}{2}k_BT_0,\quad \sum_\alpha\sum_k\mathbf{p}^0_{\alpha,k}=0.\end{gathered} \tag{8}$$

Здесь N^0_α, $\mathbf{v}^0_{\alpha,k}$, $\mathbf{p}^0_{\alpha,k}$ $q^0_{\alpha,k}$ и $m_{\alpha,k}$ – начальные количества заряженных частиц, их скорости, импульсы, заряды и массы, ρ^0_α – плотности заряда, соответствующие количеству и распределению частиц по расчетной области. Распределения частиц плазмы по пространству в начальный момент времени могут задаваться различными способами: равномерно по всему или по части объема в соответствии с конкретной физической постановкой задачи.

2. ЧИСЛЕННАЯ РЕАЛИЗАЦИЯ

Численная реализация модели (1)–(8) для случая аксиальной симметрии в координатах (r,z) достаточно подробно обсуждалась в [10, 11]. Здесь сделаем несколько важных замечаний.

Во-первых, использовавшаяся в расчетах сетка является трехмерной. В плоскости (r,z) она представляет собой произведение неравномерной сетки со средним шагом h_r по радиусу r и равномерной сетки с шагом h_z по продольной координате z. При этом в ней присутствует зависимость $r\left(z\right)$, которая позволяет задавать достаточно произвольную реальную геометрию. Для реализации движения частиц используется дополнительно сетка по азимутальной координате φ. Она также имеет параметризацию $\varphi=\varphi\left(r\left(z\right)\right)$ и подчинена условию согласования шагов: $h_\varphi\sim h=\min\left(h_r,h_z\right)$. В результате такого подхода в плоскости (r,φ) возникает так называемая мортарная сетка с различным количеством ячеек на каждом радиусе $N_\varphi\left(r\right)\approx 2\pi r/h$.

Во-вторых, численная реализация нашего подхода была бы не столь эффективной без коррекции модели частиц. В ее основе лежит аппроксимация дельта-функции с помощью гауссова профиля, который при определенных условиях является точным решением уравнений (5). В проинтегрированном по углу варианте данная аппроксимация имеет следующий вид:

$$\begin{gathered}\delta_{\alpha,k}=A_{\alpha,k}\exp\left[-R^2\left(r,z;r_{\alpha,k},z_{\alpha,k}\right)/R_p^2\right],\\ R\left(r,z;r_{\alpha,k},z_{\alpha,k}\right)=\sqrt{\left(r-r_{\alpha,k}\right)^2+\left(z-z_{\alpha,k}\right)^2},\end{gathered} \tag{9}$$

здесь $(r_{\alpha,k}=r_{\alpha,k}\left(t\right),z_{\alpha,k}=z_{\alpha,k}\left(t\right))$ – текущие (r,z) координаты частицы, R_p – эффективный радиус частиц, составляющий несколько шагов сетки ($R_p\sim n\cdot h$, $h=\sqrt{h_r^2+h_z^2}$, $n=1,2,3,\ldots$), $A_{\alpha,k}=A_{\alpha,k}\left(t,\mathbf{r}_{\alpha,k}\left(t\right)\right)$ –

параметр нормировки. Важным моментом здесь является зависимость параметра нормировки от времени и места положения облачной частицы. При рождении частицы на эмиттере учитывается, например, что облако электронов появляется в свободном пространстве постепенно. При подлете к металлическому контакту оно поглощается им тоже по частям. При взаимодействии с диэлектриком облако электронов может растекаться по его поверхности. В разработанной численной схеме эти физические процессы учтены посредством деформации и усечения облака частицы, имеющего изначально сферическую форму. Данная методика позволила добиться гладкости функций, описывающих плотность заряда и плотность тока.

В-третьих, особое внимание уделено в численной схеме закону сохранения заряда. Для этого используются принцип детального равновесия и принцип суперпозиции полей. В основе их применения лежит понимание, что каждая отдельная движущаяся частица формирует свое элементарное электромагнитное поле, а также свои вклады в плотность заряда и в плотность тока. Поэтому закон сохранения заряда должен работать независимо для каждой частицы в любом месте ее расположения в любой момент времени при любой конфигурации геометрии облака. Это реализовано в численной схеме с помощью специального итерационного процесса, относящегося к каждой частице.

Общий численный алгоритм расчета представляет собой вариант схемы расщепления по физическим процессам [17]. Задав начальное состояние системы частиц и стартовые распределения электрического и магнитного полей, далее в цикле по времени производятся вычисления, состоящие из следующих четырех этапов.

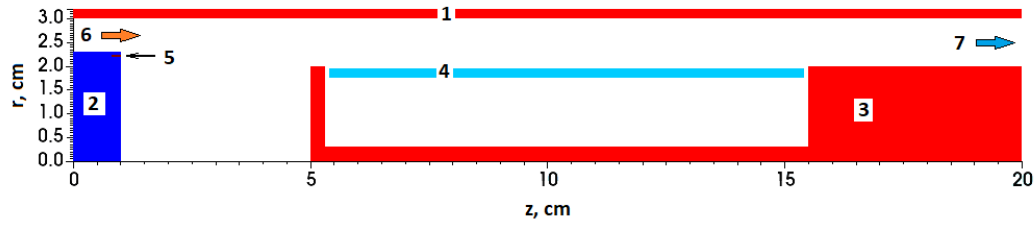
На первом этапе по уравнениям (1) при фиксированной плотности заряда ρ рассчитываются квазистатические части распределений компонент векторов $\mathbf{E}$ и $\mathbf{B}$. На втором этапе по уравнениям (5) рассчитываются новые положения частиц и их импульсы, а также геометрические структуры их облаков. На этом же этапе в систему добавляются эмитированные электроны при условии выполнения критерия эмиссии. На третьем этапе по уравнениям (6) рассчитываются новые плотность заряда и тока. На четвертом этапе по уравнениям (1) вычисляются динамические компоненты векторов $\mathbf{E}$ и $\mathbf{B}$.

На основе разработанной численной модели была создана параллельная программа. При ее реализации использовались языки программирования ANSI C и C++, а также стандарты параллельных вычислений MPI и OpenMP. Детали параллельного алгоритма представлены в [10, 11].

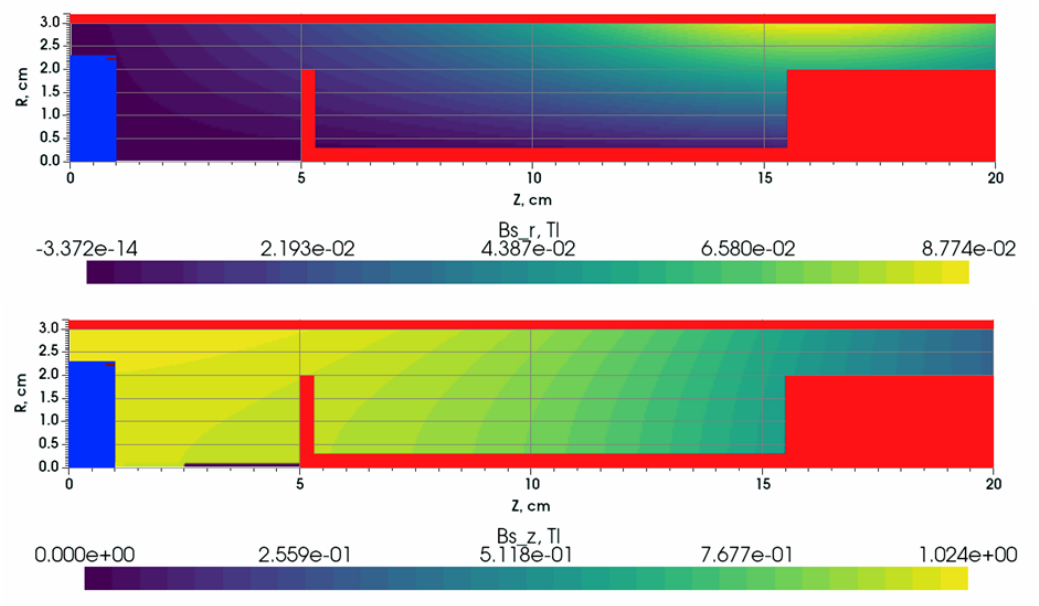
3. РЕЗУЛЬТАТЫ РАСЧЕТОВ

Рассмотрим далее результаты моделирования процессов генерации РЭП и его взаимодействия с плазменным слоем. На фиг. 1 показана расчетная геометрия прибора и отдельные элементы конструкции. Длина прибора составляет 20 см, радиус 3.2 см. Толщина анода 0.2 см, радиус катода 2.3 см. Толщина графитового эмиттера 0.05 см. Толщина плазменного слоя 0.2 см. Плазменный слой удерживается магнитным полем соленоида с параметрами: длина соленоида 32 см (центр соленоида по продольной координате находится в сечении $z = 0$ см, правый конец – в сечении $z = 16$ см), радиус соленоида 10 см, амплитуда поля соленоида $B_0 = 1$ Тл (фиг. 2). Эмиссионный процесс запускался с помощью ТЕМ-волны, набегающей слева, с напряженностью поля 511 кВ/см. Форма импульса показана на фиг. 3. Константа эмиссии равна 0.004, температура эмиттера 2731 K. Концентрация аргоновой плазмы 10^{12} см$^{-3}$.

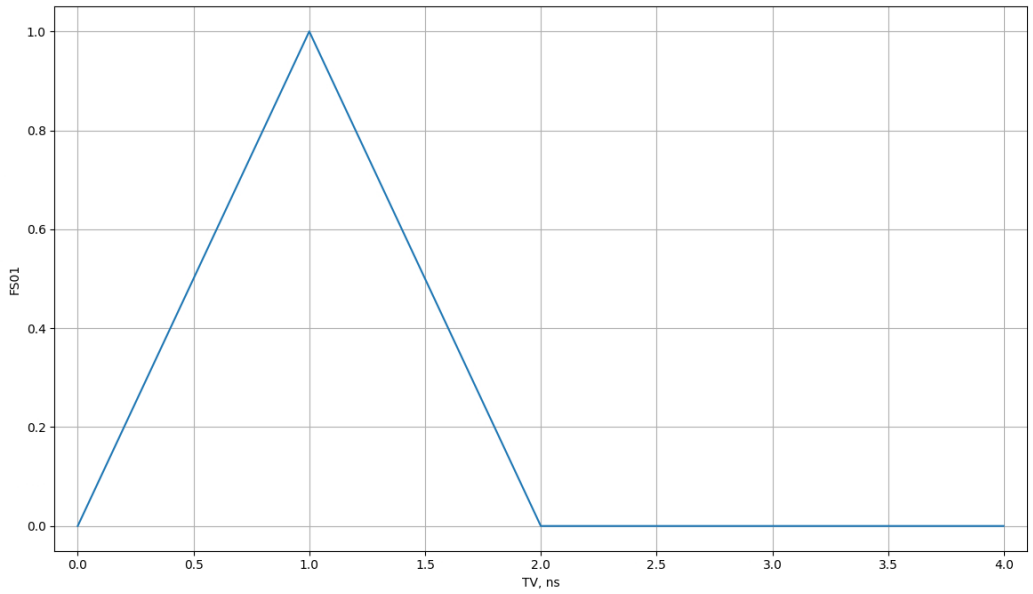
Параметры численной схемы были следующие. Все расчеты проводились в цилиндрической геометрии в координатах (r, z) на равномерной декартовой сетке с шагом $h = 0.05$ см. При реализации механизма эмиссии количество частиц, рождающихся одновременно на боковой части соответствующей ячейки сетки равнялось 4. Заметим, что в численных экспериментах это число варьировалось, однако оказалось, что его увеличение слабо влияет на точность расчета, но резко увеличивает вычислительные затраты. Поэтому управление точностью лучше производить через величину шага сетки h.



Фиг. 1. Расчетная геометрия генератора. Цифрами 1, 2, 3, 4, 5, 6, 7 обозначены – анод, катод, коллектор, слой плазмы, эмиттер, вход ТЕМ волны, выход результирующей электромагнитной волны.



Фиг. 2. Распределения компонент поля соленоида B_φ (а) и B_z (б).

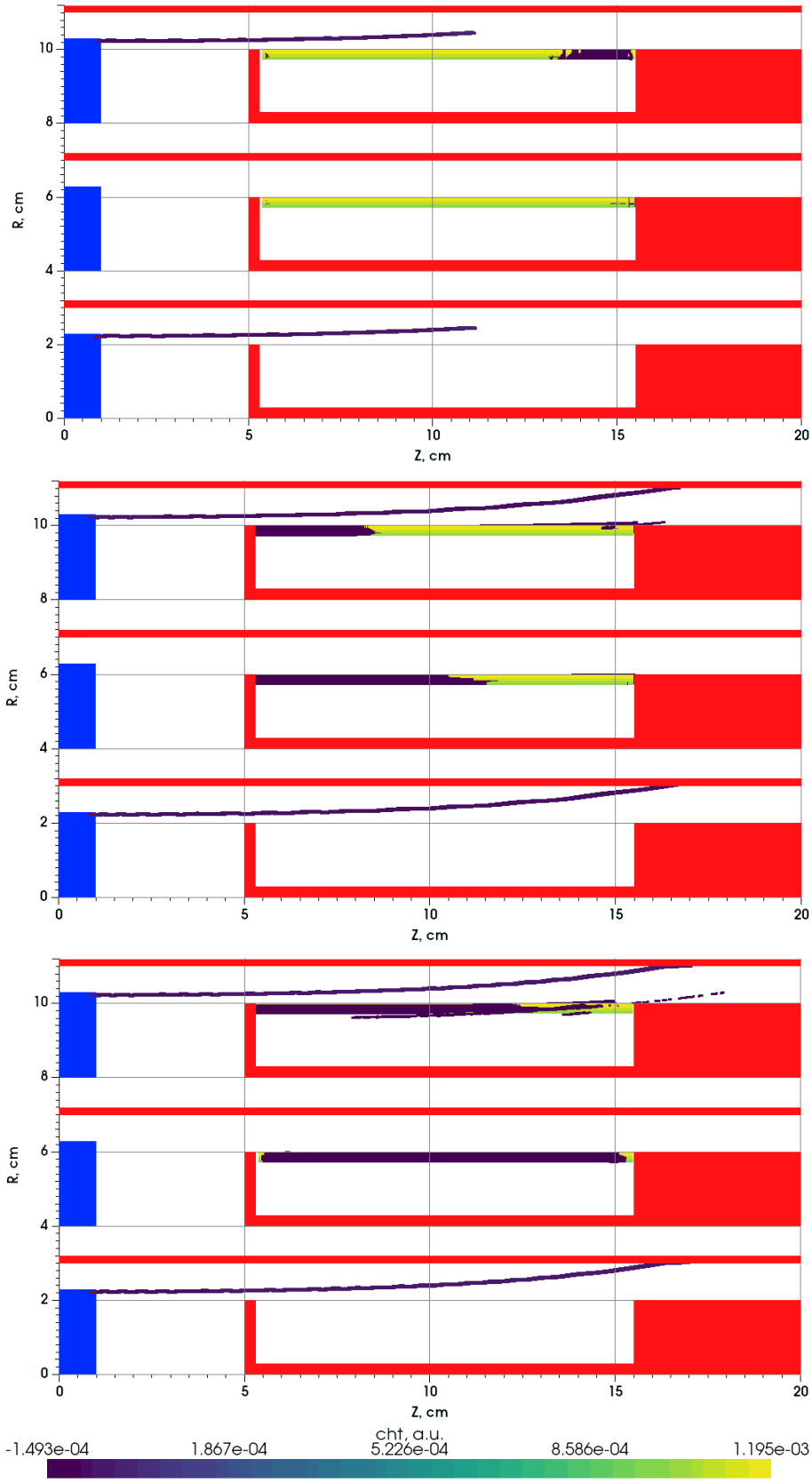


Фиг. 3. Форма импульса.

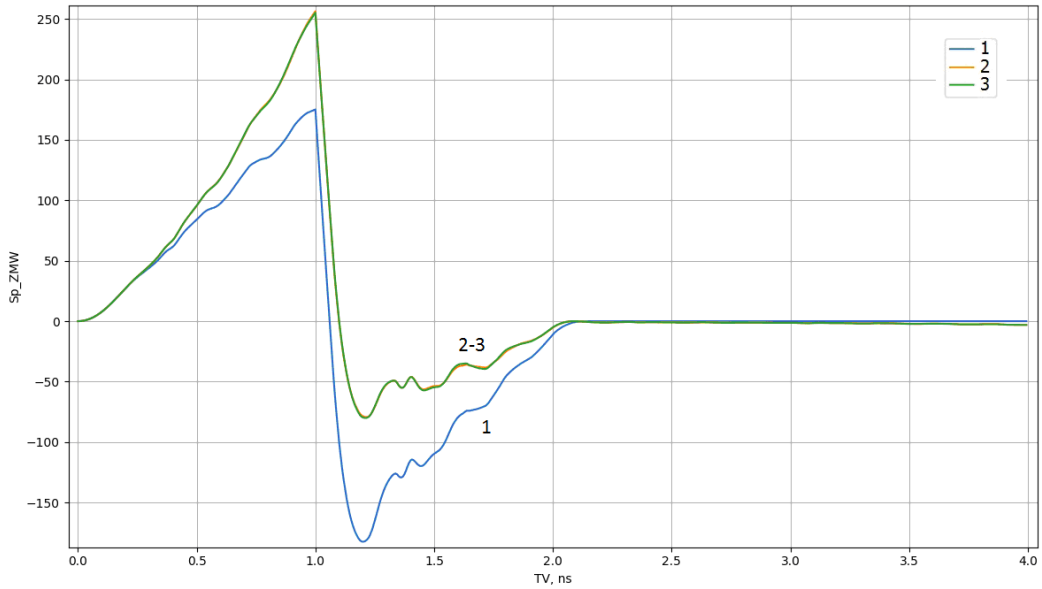
В соответствии с работой [11] параметры модели плазмы были следующие: число мультиионов в каждой ячейке сетки задавалось равным 25, каждый ион находится в окружени восьми мультиэлектронов. В [11] нами давались подробные пояснения к выбору этих параметров.

Перейдем теперь к описанию результатов расчетов. Было выполнено три варианта расчета. В варианте 1 отсутствовал плазменный слой, но с помощью механизма эмиссии генерировался РЭП. В варианте 2 РЭП был выключен, но присутствовала плазма. В варианте 3 присутствовали РЭП и плазма.

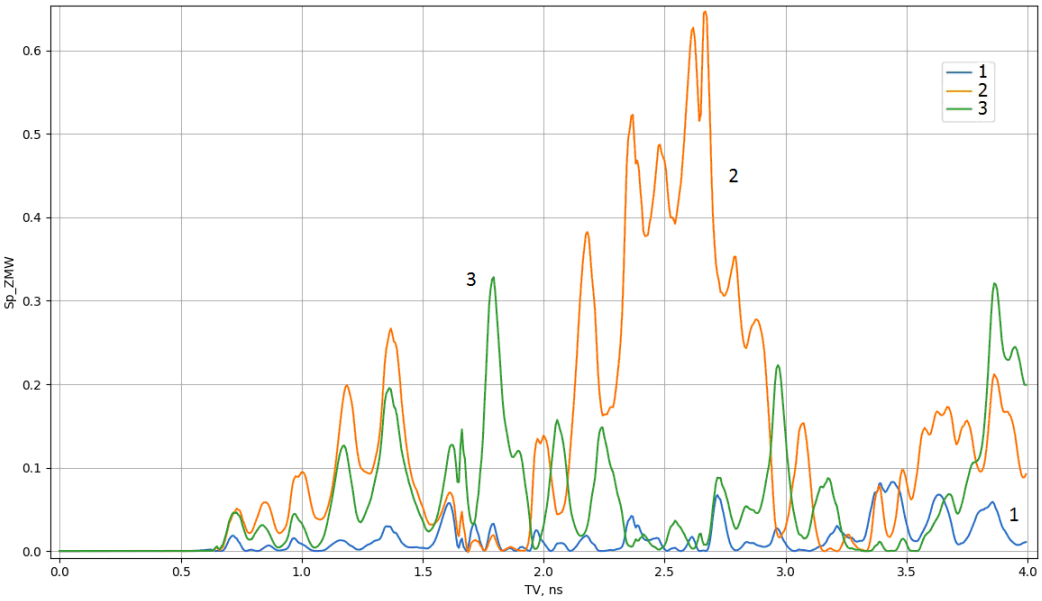
Первая серия результатов касается динамики облачных частиц. На фиг. 4 показаны распределения частиц в РЭП и в плазменном слое в характерные моменты времени, связанные с динамикой импульса ТЕМ волны. На каждом из рисунков а, б, в показаны распределения, соответствующие трем вариантам расчета. Приведенные данные в целом демонстрируют эволюцию РЭП и плазмы во времени и их реакцию на суммарную электромагнитную волну. Поведение РЭП является наиболее устойчивым и демонстрирует лишь отклонение от прямолинейной траектории, соответствующее действию магнитного поля. Поведение плазмы со временем в отсутствие



Фиг. 4. Распределения частиц в моменты времени 1, 2, 3 нс (фиг. а, б, в). Снизу на каждом рисунке показан расчет по варианту 1, в середине – по варианту 2, сверху – по варианту 3. Темно синим цветом отображены электроны РЭП и плазмы, желтым – ионы плазмы.



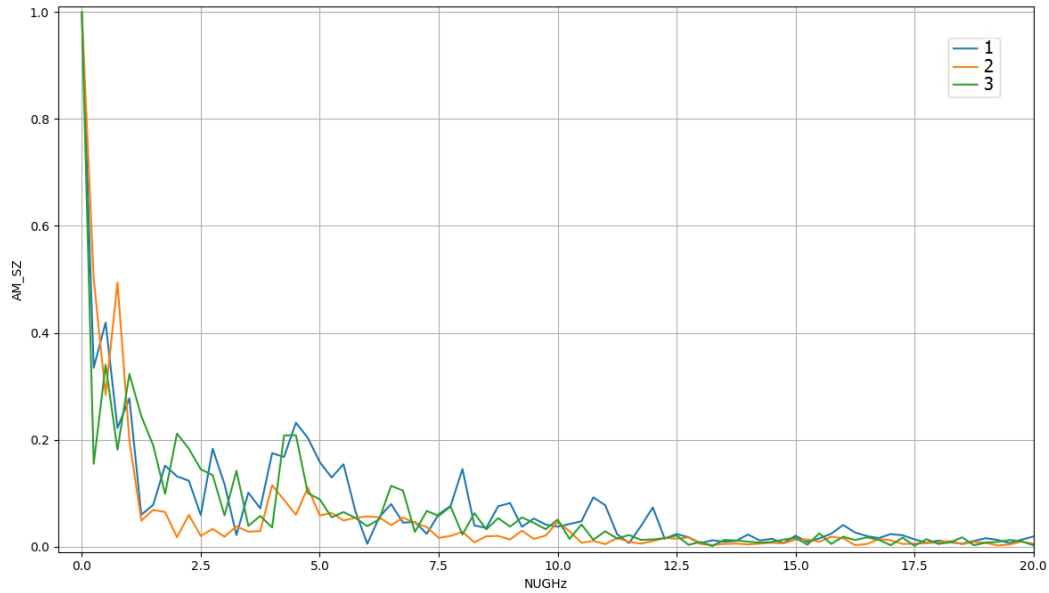
Фиг. 5. Интеграл от z-компоненты вектора Умова-Пойнтинга на входе в генератор ($z = 0.025$ см). Кривые 1–3 соответствуют номеру варианта расчета.



Фиг. 6. Интеграл от z-компоненты вектора Умова-Пойнтинга на выходе из генератора ($z = 19.025$ см). Кривые 1–3 соответствуют номеру варианта расчета.

РЭП проявляется в разделении электронной и ионной частей по всему слою и реализации плазменных колебаний. В присутствии РЭП эти процессы развиваются сильнее. К тому же в этом случае имеется большой пространственный сдвиг электронной части плазмы в сторону движения РЭП.

На фиг. 5 и 6 показана эволюция интеграла z-компоненты вектора Умова-Пойнтинга на входе и на выходе из генератора. Фиг. 5 иллюстрирует известный эффект опрокидывания (смены знака) z-компоненты вектора Умова-Пойнтинга, вызванный появлением обратной волны, отраженной от левой части коллектора. Фиг. 5 также показывает, что при наличии плазмы имеется определенное увеличение интеграла за счет обратной связи с плазменным слоем. На выходе из генератора (фиг. 6) это влияние прослеживается сильнее и безусловно зависит от концентрации плазмы. Также фиг. 6 иллюстрирует уменьшение амплитуды колебаний интеграла в варианте 3. По-видимому, это связано с частичной компенсацией колебание РЭП и плазмы в зоне коллектора.



Фиг. 7. Нормированный спектр интеграла от z-компоненты вектора Умова-Пойнтинга на выходе из генератора ($z = 19.025$ см). Кривые 1–3 соответствуют номеру варианта расчета.

На фиг. 7 показаны нормированные спектры интеграла z-компоненты вектора Умова-Пойнтинга на выходе из генератора. Из рисунка видно, что во всех трех вариантах расчета в диапазоне частот от 1 до 12.5 ГГц имеется множество гармоник с заметной амплитудой. При этом наличие РЭП усиливает некоторые из них, а наличие плазмы сглаживает спектр.

Оценивая полученные численные результаты в целом можно отметить следующее. Представленные результаты расчетов иллюстрируют несколько физических эффектов, которые наблюдаются в натурных экспериментах с использованием выбранного класса СВЧ-устройств. Анализ представленных численных данных показывает, что качественно разработанная методика воспроизводит основную физику процесса генерации электромагнитного излучения в заданном диапазоне частот. Количественные совпадения представленной численной методики проверялись при сравнении с расчетами по коду КАРАТ [16]. Они показали, что на данном классе задач методика позволяет получить результаты с погрешностью порядка 10%. Однако здесь следует подчеркнуть, что в коде КАРАТ не реализован самосогласованный расчет эмиссии, и РЭП задается на входе с помощью постоянного потока электронов. Фактически в нем игнорируется эффект экранирования РЭП вблизи эмиттера. Это может быть одной из причин вышеуказанного отличия.

ЗАКЛЮЧЕНИЕ

Рассмотрена задача расчета волновых процессов в СВЧ-генераторе с магнитной изоляцией. Для аксиально-симметричной геометрии представлена новая численная методика, сочетающая метод частиц и метод сеток. Методика использует представление облачных частиц и реализует расчеты электромагнитных полей на декартовых пространственных сетках. В методику внедрена ранее разработанная авторами модель плазмы. В численных экспериментах изучен процесс генерации выходного излучения для трех вариантов задачи. В численных расчетах получены пространственно-временные характеристики релятивистских электронных пучков и частиц плазмы. Предложенная численная методика и разработанный на ее основе программный код подтвердили свою общую корректность. Дальнейшее развитие результатов работы будет связано с рассмотрением СВЧ-систем реальной геометрии и уточнением моделей эмиссии и плазмы.

СПИСОК ЛИТЕРАТУРЫ

1. *Левитский С.М., Кошевая С.В.* Вакуумная и твердотельная электроника СВЧ. Киев: Вища школа, 1986. С. 272.

2. *Шахнович И.* Твердотельные СВЧ-приборы и технологии. Состояние и перспективы // ЭЛЕКТРОНИКА: Наука, Технология, Бизнес. 2005. № 5. С. 58-64.

3. *Викулов И.* СВЧ-электроника сегодня: направления и вызовы // ЭЛЕКТРОНИКА: наука, технология, бизнес. 2015. № 3. С. 64-72.

4. *Белоус А.И., Мерданов М.К., Шведов С.В.* СВЧ-электроника в системах радиолокации и связи. Техническая энциклопедия. В 2-х книгах. Книга 1. М.: ТЕХНОСФЕРА, 2016. С. 688.

5. *Белоус А.И., Мерданов М.К., Шведов С.В.* СВЧ-электроника в системах радиолокации и связи. Техническая энциклопедия. В 2-х книгах. Книга 2. М.: ТЕХНОСФЕРА, 2016. С. 728.

6. *Кузелев М.В., Лоза О.Т., Рухадзе А.А., Стрелков П.С., Шкварунец А.Г.* Плазменная релятивистская СВЧ-электроника // Физика плазмы. 2001. Т. 27. № 8. С. 710–733.

7. *Кузелев М.В., Рухадзе А.А., Стрелков П.С.* Плазменная релятивистская СВЧ-электроника. М.: Изд-во МГТУ им. Н. Баумана, 2018. С. 624.

8. *Литвин В.О., Лоза О.Т.* Плазменный сильноточный генератор мощных широкополосных СВЧ-импульсов с магнитной самоизоляцией // Труды ИОФАН. 2016. Т. 72. С. 134–139.

9. *Булейко А.Б., Бахтин В.П., Лоза О.Т., Раваев А.А., Быков А.Г., Коновальцева Л.В.* Плазменный мазер с магнитной самоизоляцией // Прикладная физика. 2023. № 1. С. 72–77.

10. *Поляков С.В., Тарасов Н.И., Кудряшова Т.А.* Моделирование эмиссионных процессов в сильных электромагнитных полях // Ж. вычисл. матем. и матем.физ. 2023. Т. 63. № 8. С. 1354–1366.

11. *Galstyan E.A., Kudryashova T.A, Polyakov S.V., Tarasov N.I.* Computer Simulation of Explosive Emission Processes in Strong Electromagnetic Fields // J. of Physics: Conference Series 2701 (2024) 012062.

12. *Ландау Л.Д., Лифшиц Е.М.* Курс теоретической физики. Том 8. Электродинамика сплошных сред. М.: Наука, 1982. С. 621.

13. *Birsdall C.K., Fuss D.* Clouds-in-clouds, clouds-in-cells physics for many-body plasma simulation // J. of Comput. Phys. 1969. V. 3. Issue 4. April 1969. P. 494–511.

14. *Taflove Allen, Hagness Susan C.* Computational Electrodynamics. The Finite-Difference Time-Domain Method. Third Edition. Artech House. 2005. P. 1038.

15. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. М.: БИНОМ, лаборатория знаний, 2011. С. 636.

16. *Tarakanov V. P.* User's Manual for Code KARAT. Springfield, VA: Berkeley Research. VA, 1992. С. 262.

17. *Марчук Г.И.* Методы расщепления. М.: Наука, 1980. С. 264.

MODELING OF NONLINEAR WAVE PROCESSES IN A MICROWAVE GENERATOR WITH MAGNETIC INSULATION

S. V. Polyakov*, N. I. Tarasov, T. A. Kudryashova*****

125047 Moscow, Miusskaya sq., 4, Keldysh IAM RAS, Russia,
**e-mail: polyakov@imamod.ru*
***e-mail: nikita_tarasov@imamod.ru*
****e-mail: kudryashova@imamod.ru*

Received: 11.06.2024
Revised: 28.07.2024
Accepted: 23.08.2024

Abstract. The actual problem of modeling nonlinear wave processes in a microwave generator with magnetic isolation is considered. For its numerical analysis, a new computer model is proposed, including Maxwell's equations and equations of motion of relativistic charged particles, their joint integration by the grid method and the cloud particle method, as well as a parallel software implementation. In numerical experiments, the space-time characteristics of relativistic electron beams and plasma, as well as the parameters of the output radiation of the generator, are obtained. The analysis of the obtained results confirmed the correctness of the developed numerical approach.

Keywords: mathematical modeling, wave processes, microwave radiation generators, relativistic beams, low-temperature plasma.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2411–2419

УДК 519.63

“СУПЕР-БЫСТРЫЙ” АЛГОРИТМ РЕШЕНИЯ ПРЯМОЙ ЗАДАЧИ РАССЕЯНИЯ ДЛЯ СИСТЕМЫ МАНАКОВА[1)]

© 2024 г. Л. Л. Фрумин[1,*], А. Е. Чернявский[1,], О. В. Белай[1,***]**

[1] 630090 Новосибирск, пр-т Акад. Коптюга, 1, ИАиЭ СО РАН, Россия
**e-mail: lfrumin@iae.nsk.su*
***e-mail: alexander.cher.99@gmail.com*
****e-mail: ovbelai@gmail.com*

Поступила в редакцию 10.05.2024 г.
Переработанный вариант 29.07.2024 г.
Принята к публикации 23.08.2024 г.

Рассматривается построение ускоренного алгоритма решения прямой задачи рассеяния для непрерывного спектра системы Манакова, ассоциированной с векторным нелинейным уравнением Шрёдингера модели Манакова. Численная постановка задачи приводит к проблеме быстрого расчета произведений полиномов, зависимых от спектрального параметра задачи. Для локализованных решений представлен так называемый “супер-быстрый” алгоритм решения прямой задачи рассеяния второго порядка точности, основанный на теореме о свертке и быстром преобразование Фурье, требующий для дискретной сетки размером N асимптотически всего $O\left(N\mathrm{Log}^2 N\right)$ арифметических операций. Для ускорения расчета спектров коэффициентов отражения предложен и апробирован матричный вариант быстрого преобразования Фурье, когда коэффициенты ряда дискретного преобразования Фурье представляют собой некоммутирующие матрицы. Численное моделирование на примере точного решения системы Манакова (гиперболического секанса) подтвердило высокую скорость расчетов и второй порядок точности аппроксимации алгоритма. Библ. 24. Фиг. 1.

Ключевые слова: уравнение Шрёдингера, система Манакова, прямая задача рассеяния, трансфер-матрица, свертка, преобразование Фурье.

DOI: 10.31857/S0044466924120143, **EDN**: KBMEIL

1. ВВЕДЕНИЕ

Нелинейное уравнение Шрёдингера (НУШ) представляет собой фундаментальную математическую модель, описывающую эффекты нелинейно-дисперсионного взаимодействия волн, порождающие уединенные волны – солитоны (см. [1], [2]). Векторное обобщение НУШ, учитывающее поляризацию волн, известно как модель Манакова (см. [3]). Эта модель одновременно учитывает эффекты нелинейности, дисперсии и поляризации волн, что представляет исключительный интерес для теоретической физики и нелинейной оптики. В последние годы НУШ и модель Манакова нашли важное применение для описания нелинейно-дисперсионных и поляризационных эффектов в процессах передачи информации по волоконно-оптическим линиям связи, составляющим технологическую основу глобальной сети Интернет. Наблюдавшийся в начале века стремительный взлет производительности (пропускной способности) волоконно-оптических линий, обусловленный прогрессом современных оптических технологий, в последние годы практически прекратился из-за усиления роли нелинейно-дисперсионных эффектов, заметно искажающих информационный сигнал (см. [4], [5]). Именно эти эффекты и описывает в скалярном случае НУШ, а в более близком к реальности векторном случае, т.е. с учетом поляризации – модель Манакова. Нелинейное векторное уравнение Шрёдингера модели Манакова имеет следующий нормализованный вид (см. [3]):

$$\mathrm{i}\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial^2 \mathbf{q}}{\partial x^2} - 2\sigma\left|\mathbf{q}^2\right|\mathbf{q} = 0, \tag{1.1}$$

где $\mathbf{q}(t,x) = (q,p)$ – вектор решения модели Манакова, содержащий две поляризационные нормированные компоненты поля $q(t,x)$ и $p(t,x)$, x – пространственная координата, t – временной параметр, i – мнимая единица. Параметр σ принимает значения $+1$ или -1 соответственно для дефокусирующего и фокусирующего слу-

[1)] Работа выполнена при финансовой поддержке РНФ, грант 24-22-00183.

чаев. Заметим, что в задачах, описывающих эволюцию модулированной огибающей оптического сигнала в оптическом волокне, координаты (t, x) заменяются парой (z, τ), где z – расстояние вдоль оптического волокна, а τ – временной параметр в сопутствующей оптическому импульсу системе отсчета (см. [4]).

Как и НУШ, модель Манакова принадлежит к числу так называемых интегрируемых уравнений, в исследование которых значительный вклад вносит метод обратной задачи рассеяния (МОЗР) (см. [1], [2]). Этот метод сводит решение нелинейного уравнения к решению прямых и обратных спектральных задач рассеяния для некоторой ассоциированной с этим уравнением линейной системы, коэффициенты которой определяются решениями исходного нелинейного уравнения. Система двух уравнений Захарова–Шабата служит такой ассоциированной линейной системой для скалярного НУШ. В случае векторного НУШ модели Манакова задачи рассеяния ставятся для системы трех линейных уравнений, называемых системой Манакова (см. [3]):

$$\frac{\partial u(x)}{\partial x} + \mathrm{i}\lambda u(x) = q(x)\, v(x) + p(x)\, w(x),$$
$$\frac{\partial v(x)}{\partial x} - \mathrm{i}\lambda v(x) = \sigma q^*(x)\, u(x), \quad \frac{\partial w(x)}{\partial x} - \mathrm{i}\lambda w(x) = \sigma p^*(x)\, u(x), \tag{1.2}$$

где λ – собственное значение, (u, v, w) – трехмерный собственный вектор волновых амплитуд, причем u имеет смысл амплитуды падающей волны, v, w – амплитуды отраженных волн, i – мнимая единица, а символ * обозначает комплексное сопряжение. Система Манакова (1.2) рассматривается при фиксированном параметре времени t, который здесь и далее для краткости не указывается в аргументах функций. Система Манакова в фокусирующем случае ($\sigma = -1$) содержит дискретный и непрерывный спектры, а в дефокусирующем ($\sigma = +1$) – только непрерывный спектр. Дискретный спектр системы с комплексным спектральным параметром λ соответствует солитонным решениям НУШ. Непрерывный спектр системы Манакова имеет действительные собственные значения λ и описывает диспергирующие волны.

С вычислительной точки зрения метод обратной задачи рассеяния, позволивший аналитически исследовать целый ряд нелинейных эволюционных волновых уравнений (см. [1]), оказался мощным инструментом для эффективного численного решения задачи Коши для этих нелинейных уравнений, причем, без каких-либо итераций. Его численная реализация требует развития эффективных алгоритмов решения как обратных, так и прямых задач рассеяния, которым и посвящена настоящая работа, применительно к системе Манакова.

Задачи рассеяния (ЗР) оперируют данными рассеяния, в качестве которых в общем случае служат компоненты матрицы рассеяния (см. [1], [2], [6], [7]). В случае прямой ЗР для непрерывного спектра системы Манакова данные рассеяния содержат амплитуду $a(\lambda)$ падающей волны, амплитуды $b_1(\lambda), b_2(\lambda)$ отраженных волн для каждой из двух поляризаций, а также и амплитуду прошедшей волны $d(\lambda)$. Две поляризационные компоненты вектора решения $q(x), p(x)$ следует рассматривать при этом как компоненты рассеивающего потенциала. Прямая ЗР представляет собой определение данных рассеяния, в частности, коэффициентов отражения и импульсных откликов, по заданному вектору решения (рассеивающему потенциалу) $\mathbf{q}$. Поиск вектора решения (рассеивающего потенциала) $\mathbf{q}(x)$ по данным рассеяния, при фиксированном эволюционном параметре t, составляет обратную ЗР для системы Манакова. В настоящей работе мы будем придерживаться более узкой, однако, более распространенной в приложениях постановки, где в качестве искомых для прямой ЗР принимаются коэффициенты отражения

$$r_1(\lambda) = b_1(\lambda)/a(\lambda), \quad r_2(\lambda) = b_2(\lambda)/a(\lambda). \tag{1.3}$$

Преобразование Фурье от коэффициентов отражения определяет импульсные отклики $R_{1,2}(\xi)$ пары компонент рассеивающего потенциала:

$$R_{1,2}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} r_{1,2}(\lambda)\, e^{-\mathrm{i}\lambda\xi} d\lambda. \tag{1.4}$$

В недавних работах авторов [8]–[11] для решения обратных задач рассеяния системы Манакова были предложены “быстрые” алгоритмы, основанные на методах окаймления и требующие всего $O(N^2)$ арифметических операций, где N – размер расчетной сетки. Они представляют собой блочно-матричные алгоритмы второго порядка точности аппроксимации (порядком точности мы называем порядок степенной зависимости погрешности вычислений (остаточного члена) от шага разностной сетки). В одном из алгоритмов применен блочный вариант окаймления Левинсона, восстанавливающий тёплицеву структуру блочных матриц, путем переноса некоторых слагаемых дискретных систем уравнений в правую часть. Другой алгоритм был основан на тёплицевом разложении (см. [12]) блочной матрицы, близкой к блочно-тёплицевой, и алгоритме окаймления Тыртышникова (см. [13]). Сравнительно недавно в работе [14] были представлены так называемые “супер-быстрые” алгоритмы решения прямой и обратной ЗР для системы Манакова первого порядка точности аппроксимации.

Эти алгоритмы для расчетной сетки размером N требуют асимптотически всего порядка $O\left(N\mathrm{Log}^2 N\right)$ арифметических операций и потому названы “супер-быстрыми”. Аналогичные алгоритмы ранее были предложены для системы Захарова–Шабата в работе [15]. Однако “супер-быстрые” алгоритмы первого порядка точности при заданной точности расчетов (норме ошибки) ε заметно менее эффективны, так как асимптотически уступают в быстродействии “быстрым” алгоритмам второго порядка точности, требующим порядка $O\left(N^2\right)$ арифметических операций. Действительно, “супер-быстрый” алгоритм первого порядка точности требует порядка $O(\varepsilon^{-1}\left(\mathrm{Log}\,\varepsilon^{-1}\right)^2)$ арифметических операций, тогда как “быстрый” алгоритм второго порядка точности при той же точности расчетов ε потребует всего $O\left(\varepsilon^{-1}\right)$ операций, поскольку при заданной точности ему требуется дискретная сетка меньшего размера N. К настоящему времени порядок точности аппроксимации “супер-быстрых” алгоритмов (см. [15]) для системы Захарова–Шабата достиг второго. Последующие разделы посвящены построению аналогичного эффективного (точного и быстрого) алгоритма решения прямой задачи для системы Манакова.

2. ПОСТАНОВКА ЗАДАЧИ РАССЕЯНИЯ

В этой работе рассматривается “супер-быстрый” алгоритм решения прямой ЗР для непрерывного спектра системы Манакова, для случая локализованных (убывающих на бесконечности) решений, второго порядка точности аппроксимации. Этот алгоритм основан на методе матриц переноса, называемом также методом трансфер-матриц (МТМ) или методом Т-матриц. МТМ служит основным, базовым методом решения задач рассеяния в оптических, радиофизических и акустических приложениях (см. [16], [17]). В рамках МТМ непрерывная среда разбивается на N дискретных слоев, каждый из которых характеризуется трансфер-матрицей T_m, где m – номер слоя. Эта матрица связывает вектор волновых амплитуд (u_m, v_m, w_m) в m-м слое с аналогичным вектором в соседнем слое.

Для решения трех линейных уравнений системы Манакова требуется три граничных условия. Прямая ЗР для локализованных потенциалов модели Манакова при постановке граничных условий опирается на асимптотики волновых решений системы Манакова на бесконечности. Рассматривая, для определенности, правую ЗР, когда исходная волна падает справа на рассеивающий потенциал, запишем асимптотики решения системы Манакова в следующем виде (см. [1], [3]):

$$(u,v,w)_{x\to-\infty} = \left(e^{-i\lambda x},0,0\right), \quad (u,v,w)_{x\to+\infty} = \left(a\left(\lambda\right)e^{-i\lambda x}, b_1\left(\lambda\right)e^{i\lambda x}, b_2\left(\lambda\right)e^{i\lambda x}\right).$$

Подобные асимптотики определяют линейно независимые фундаментальные решения системы Манакова, называемые функциями Йоста.

В численных подходах, где используется конечный носитель, к примеру, интервал $[x_0, x_N]$, асимптотики принимают вид граничных условий:

$$(u_0,v_0,w_0) = \left(e^{-i\lambda x_0},0,0\right), \quad (u_N,v_N,w_N) = \left(a\left(\lambda\right)e^{-i\lambda x_N}, b_1\left(\lambda\right)e^{i\lambda x_N}, b_2\left(\lambda\right)e^{i\lambda x_N}\right), \tag{2.1}$$

где $u_0 = u\left(x_0\right), v_0 = v\left(x_0\right), w_0 = w\left(x_0\right), u_N = u\left(x_N\right), v_N = v\left(x_N\right), w_N = w\left(x_N\right)$.

Решение прямой ЗР сводится к расчету неизвестных спектральных коэффициентов $a\left(\lambda\right), b_1\left(\lambda\right), b_2\left(\lambda\right)$ путем последовательного умножения трансфер-матриц на вектор-столбец амплитуды прошедшей волны:

$$\begin{pmatrix} u_N \\ v_{1,N} \\ v_{2,N} \end{pmatrix} = T_{N-1}T_{N-2}\ldots T_1T_0 \begin{pmatrix} u_0 \\ v_{1,0} \\ v_{2,0} \end{pmatrix}.$$

Используя граничные условия (2.1), для определения данных рассеяния получим

$$\begin{pmatrix} a\left(\lambda\right)e^{-i\lambda x_N} \\ b_1\left(\lambda\right)e^{i\lambda x_N} \\ b_2\left(\lambda\right)e^{i\lambda x_N} \end{pmatrix} = T_{N-1}T_{N-2}\ldots T_1T_0 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} e^{-i\lambda x_0}. \tag{2.2}$$

Спектральные данные $a\left(\lambda\right), b_1\left(\lambda\right), b_2\left(\lambda\right)$, полученные в результате перемножения матриц в (2.2), далее используются для расчета коэффициентов отражения $r_{1,2}\left(\lambda\right)$ по формулам (1.3), а затем и для импульсных откликов задачи рассеяния $R_{1,2}\left(\xi\right)$, с помощью численного расчета преобразования Фурье (1.4).

3. ДИСКРЕТИЗАЦИЯ ЗАДАЧИ РАССЕЯНИЯ

Определим по координате x равномерную расчетную сетку

$$x_m, \quad m = 0,1,\ldots,N : x_m \in [-L/2 = x_0, x_1, \ldots, x_N = L/2],$$

с шагом L/N, где $L = x_N - x_0$ – размер интервала. Половину этого шага $L/(2N)$ будем обозначать далее, как h. В узлах сетки находятся дискретные отсчеты поляризационных компонент решения (рассеивающего потенциала) q_m, p_m, $m = 0, 1, \ldots, N$. Для локализованных потенциалов предполагается, что основная часть потенциала сосредоточена в пределах выбранного интервала, а на его границах значения компонент потенциала пренебрежимо малы. Для численного расчета дискретных значений импульсных откликов $R_{1,2}(\xi)$ требуется сетка $\xi_m = 2x_m$, $m = 0, 1, \ldots, N$, на вдвое большем интервале $[-L, L]$ с шагом $\tilde{h} = \frac{2L}{N} = 4h$. Импульсные отклики определяют непрерывную часть спектра данных рассеяния (ядер) обратной ЗР в интегральной постановке, опирающейся на уравнения Гельфанда–Левитана–Марченко (ГЛМ). Уравнения ГЛМ представляют собой интегральный вариант системы Манакова (см. [18]). В дефокусирующем случае импульсные отклики полностью определяют интегральные ядра этих уравнений. Фурье-преобразование (1.4) при расчете импульсных откликов требует применения (быстрого) дискретного преобразования Фурье (ДПФ) и соответствующей дискретной сетки в спектральной области $\lambda_k = k\Delta\lambda$, $k = 0, 1, \ldots, N$. Шаг сетки по спектральной переменной $\Delta\lambda$выбирается из условия, налагаемого ДПФ: $\tilde{h}\Delta\lambda = 2\pi/N$, что дает $\Delta\lambda = \pi/L$.

Точность численного решения прямой ЗР определяется точностью дискретной аппроксимации трансфер-матриц T_m. В работе [19] был представлен алгоритм решения прямой ЗР для системы Захарова–Шабата со вторым глобальным порядком точности аппроксимации. Используя аналогичный подход для системы Манакова (1.2), перепишем ее в следующем виде:

$$\begin{aligned} \left(e^{i\lambda x}u\right)' &= (qv + pw) * e^{i\lambda x}, \\ \left(e^{-i\lambda x}v\right)' &= \sigma q^* u * e^{-i\lambda x}, \\ \left(e^{-i\lambda x}w\right)' &= \sigma p^* u * e^{-i\lambda x}, \end{aligned} \tag{3.1}$$

где штрих обозначает дифференцирование по x. Интегрируя систему уравнений (3.1) на отрезке $[x_m, x_{m+1}]$, получим

$$\begin{aligned} e^{i\lambda x_{m+1}}u(x_{m+1}) - e^{i\lambda x_m}u(x_m) &= \int\limits_{x_m}^{x_{m+1}} e^{i\lambda x}[qv + pw]dx, \\ e^{-i\lambda x_{m+1}}v(x_{m+1}) - e^{-i\lambda x_m}v(x_m) &= \int\limits_{x_m}^{x_{m+1}} \sigma e^{-i\lambda x}q^* u dx, \\ e^{-i\lambda x_{m+1}}w(x_{m+1}) - e^{-i\lambda x_m}w(x_m) &= \int\limits_{x_m}^{x_{m+1}} \sigma e^{-i\lambda x}p^* u dx. \end{aligned} \tag{3.2}$$

Заменим интегралы в (3.2) по формуле трапеций с третьим локальным порядком точности аппроксимации $O\left(h^3\right)$:

$$\begin{aligned} u_{m+1} - e^{-2i\lambda h}u_m &= h\big(q_{m+1}v_{m+1} + p_{m+1}w_{m+1} + e^{-2i\lambda h}(q_m v_m + p_m w_m)\big), \\ v_{m+1} - e^{2i\lambda h}v_m &= \sigma h(q^*_{m+1}u_{m+1} + e^{2i\lambda h}q^*_m u_m), \\ w_{m+1} - e^{2i\lambda h}w_m &= \sigma h(p^*_{m+1}u_{m+1} + e^{2i\lambda h}p^*_m u_m). \end{aligned} \tag{3.3}$$

Обозначим $z = e^{2i\lambda h}$, и запишем систему (3.3) в матричном виде:

$$\begin{bmatrix} 1 & -hq_{m+1} & -hp_{m+1} \\ -\sigma h q^*_{m+1} & 1 & 0 \\ -\sigma h p^*_{m+1} & 0 & 1 \end{bmatrix} \begin{pmatrix} u_{m+1} \\ v_{m+1} \\ w_{m+1} \end{pmatrix} = \begin{bmatrix} z^{-1} & hz^{-1}q_m & hz^{-1}p_m \\ \sigma h z q^*_m & z & 0 \\ \sigma h z p^*_m & 0 & z \end{bmatrix} \begin{pmatrix} u_m \\ v_m \\ w_m \end{pmatrix}. \tag{3.4}$$

Обратной к матрице, стоящей первой слева в (3.4), является матрица

$$\frac{1}{1 - \sigma h^2\left(|q_{m+1}|^2 + |p_{m+1}|^2\right)} \begin{bmatrix} 1 & hq_{m+1} & hp_{m+1} \\ \sigma h q^*_{m+1} & 1 - \sigma h^2|p_{m+1}|^2 & \sigma h^2 q^*_{m+1}p_{m+1} \\ \sigma h p^*_{m+1} & \sigma h^2 q_{m+1}p^*_{m+1} & 1 - \sigma h^2|q_{m+1}|^2 \end{bmatrix}.$$

Умножив слева систему (3.4) на эту матрицу, опуская члены с h^3 и приближенно принимая $h^2 p_{m+1} = h^2 p_m + O\left(h^3\right)$ и $h^2 q_{m+1} = h^2 q_m + O\left(h^3\right)$, запишем итоговую трансфер-матрицу T_m:

$$T_m = \begin{bmatrix} 1 + \sigma h^2\zeta(|q_m|^2 + |p_m|^2) & h(q_m + \zeta q_{m+1}) & h(p_m + \zeta p_{m+1}) \\ \sigma h(q^*_{m+1} + \zeta q^*_m) & \zeta - \sigma h^2(\zeta|p_m|^2 - |q_m|^2) & \sigma h^2 q^*_m p_m(1 + \zeta) \\ \sigma h(p^*_{m+1} + \zeta p^*_m) & \sigma h^2 p^*_m q_m(1 + \zeta) & \zeta - \sigma h^2(\zeta|q_m|^2 - |p_m|^2) \end{bmatrix}, \tag{3.5}$$

где обозначено $\zeta = z^2 = e^{4i\lambda h} = e^{i\lambda\tilde{h}}$. У этой матрицы опущен общий множитель $\frac{z^{-1}}{1-\sigma h^2\left(|q_{m+1}|^2+|p_{m+1}|^2\right)}$. При расчете коэффициентов отражения в виде отношения коэффициентов $r_{1,2}(\lambda) = b_{1,2}(\lambda)/a(\lambda)$ общий множитель $b_{1,2}(\lambda)$ и $a(\lambda)$ сокращается, однако его необходимо учитывать, если требуется расчет коэффициента прохождения $d(\lambda)$.

4. “СУПЕР-БЫСТРЫЙ” АЛГОРИТМ РЕШЕНИЯ ПРЯМОЙ ЗАДАЧИ

Все элементы $t_{m;ij}$ полученной матрицы в (3.5) T_m представляют собой биномы вида $t_{m;ij} = \alpha_{m;ij} + \zeta\beta_{m;ij}$, т.е. полиномы первого порядка от спектрального параметра $\zeta = e^{i\lambda\tilde{h}}$. Эти биномы можно представить в виде матричных биномов:

$$T_m = A_m + \zeta B_m.$$

Здесь A_m и B_m – матрицы размером 3×3 с элементами соответственно $\alpha_{m;ij}, \beta_{m;ij}$.

Расчет произведения трансфер-матриц (2.2) сводится к вычислению произведений полиномов. Спектральные данные рассеяния $a(\lambda), b_1(\lambda), b_2(\lambda)$ в итоге принимают вид полиномов N-й степени от параметра $\zeta = e^{i\lambda\tilde{h}}$. В результате мы получаем полиномиальную аппроксимацию спектральных данных. Расчет этих полиномов на дискретной спектральной сетке определяет дискретизованные спектры данных рассеяния $a(\lambda_k), b_1(\lambda_k), b_2(\lambda_k), k = 0, 1, \ldots, N$. Дискретизованные спектры отражения представляют собой соответствующие отношения полиномов $r_{1,2}(\lambda_k) = b_{1,2}(\lambda_k)/a(\lambda_k)$ для аргумента λ_k.

Последовательное перемножение всех трансфер-матриц T_m размерностью 3×3, элементы которых являются биномами, приводит к алгоритму, основным вычислительным этапом которого является умножение полиномов (элементов матриц) на бином. При каждом таком перемножении степень результирующих полиномов будет увеличиваться на единицу. На k-м шаге процесса перемножения матриц для каждого из девяти элементов результирующей матрицы требуется умножить $k+1$ коэффициент полинома на два коэффициента бинома. Суммарное число операций умножения при этом составит для каждого элемента $\sum_{k=1}^{N-1} 2(k+1) = N^2 + N - 2$. Число сложений имеет такой же порядок величины $O\left(N^2\right)$.

В работе [15], применительно к системе Захарова–Шабата, был предложен более быстрый способ перемножения $N = 2^M$ матриц, основанный на алгоритме дублирования и быстром дискретном преобразовании Фурье. В процессе перемножения трансфер-матриц они разбиваются на соседние пары. Сначала перемножаются эти пары, и получается $N/2$ матриц с элементами в виде полиномов 2-го порядка. Результаты такого парного перемножения снова разбиваются на пары, затем эти новые пары перемножаются. Процедура перемножения полиномов повторяется $M - 1 = \mathrm{Log}_2(N) - 1$ раз. Степень результирующего полинома при каждом перемножении удваивается. На k-м шаге перемножается 2^{M-k} пар полиномов. Каждый из них имеет степень 2^{k-1} и после приведения общих членов содержит $2^{k-1}+1$ коэффициентов. Перемножение пары таких полиномов потребует $\left(2^{k-1}+1\right)^2$ умножений. Полное число умножений пар полиномов составит $\sum_{k=1}^{M-1}\left(2^{k-1}+1\right)^2 2^{M-k} = M2^M + 2^{2(M-1)} - 2^{M-1} - 2$. Таким образом, по порядку величины этот способ расчета приводит к оценке числа операций $O\left(2^{2(M-1)}\right) = O\left(N^2/4\right)$.

Для системы Захарова–Шабата в работе [15] для еще большего ускорения расчетов произведений полиномов было предложено использовать теорему о свертке. Пусть на k-м шаге попарного произведения матриц полином степени $n = 2^{k-1}$ умножается на другой полином той же степени. Произведение полиномов $a_0 + a_1\zeta + a_2\zeta^2 + \ldots + a_n\zeta^n$ и $b_0 + b_1\zeta + b_2\zeta^2 + \ldots + b_n\zeta^n$ дает полином степени $2n = 2^k$ с коэффициентами $c_j, j = 0, 1, \ldots, 2n$. Если дополнить нулевыми коэффициентами оба исходных полинома до степени $2n$, то коэффициенты c_j результирующего полинома можно представить в виде дискретной свертки:

$$c_j = \sum_{s=0}^{j} a_s b_{j-s}, \quad j = 0, 1, \ldots, 2n. \tag{4.1}$$

Прямое вычисление свертки потребует $\sum_{j=0}^{j=2n}(j+1) = 2n^2 + 3n + 1$ умножений и почти столько же сложений. Число арифметических операций можно значительно уменьшить, если применить теорему о свертке. Согласно этой теореме свертку (4.1) можно вычислить как обратное дискретное преобразование Фурье от почленного произведения векторов дискретных преобразований Фурье от коэффициентов исходных массивов:

$$c = IFFT\big(FFT(a)\cdot FFT(b)\big). \tag{4.2}$$

Здесь точка “$\cdot$” обозначает почленное произведение векторов, дискретных фурье-образов полиномов a и b, а прямое и обратное дискретные преобразования Фурье обозначены как FFT и $IFFT$ соответственно, что указывает на использование быстрого дискретного преобразования Фурье (БПФ – fast Fourier transform) (см. [20]).

Как известно, быстрые дискретные преобразования Фурье для массива размером $N = 2^M$ требуют всего порядка $O(NM) = O(N\mathrm{Log}_2 N)$ умножений. Оценку полного числа умножений пары полиномов, с учетом почленного умножения фурье-образов в (4.2), дает сумма $\sum_{k=1}^{M-1}\left(2^k(k+2)\right)2^{M-k} = NM^2/2 + 3/2NM - 2N$. В результате для итогового числа операций умножения для такого "супер-быстрого" алгоритма получаем асимптотическую оценку

$$O\left(N(\mathrm{Log}_2^2 N + 3\,\mathrm{Log}_2 N\right). \tag{4.3}$$

В оценке (4.3) помимо квадратичного оставлено линейное по $\mathrm{Log}_2 N$ слагаемое, которое для диапазона умеренных значений $M = \mathrm{Log}_2 N \sim 7-13$ дает заметный вклад в итоговую оценку числа операций умножения. Мы не приводим здесь оценку числа операций сложения, полагая, что оценка числа умножений дает вполне удовлетворительную асимптотическую оценку полного числа арифметических операций.

В процессе расчета произведения пары матричных полиномов на каждом шаге перемножения требуется с помощью БПФ вычислять девять полиномов, по числу элементов матрицы размером 3×3. В работе [21], где рассматривалось быстрое перемножение полиномов с матричными коэффициентами с помощью теоремы о свертке и ДПФ, отмечается, что число вычислений ДПФ можно уменьшить, учитывая, что каждый полином как элемент матрицы размером, например, 3×3 в процессе вычисления произведений матриц трижды подвергается преобразованию Фурье. Использование памяти для запоминания результата БПФ позволяет выполнить только одно преобразование, и тем самым заметно сократить число требуемых арифметических операций. Такую же экономию числа операций может давать матричный вариант быстрого ДПФ для матричных полиномов, коэффициенты которых представляют собой некоммутирующие в общем случае матрицы. Матричный вариант БПФ кратко рассмотрен в разд. 6.

5. ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ ПРЯМОЙ ЗР

Для тестирования алгоритма использовалось следующее точное решение системы Манакова – гиперболический секанс (см. [22], [23]):

$$q(x, t=0) = A\,\mathrm{sch}(x - x_c)\sin(\varphi),$$
$$p(x, t=0) = A\,\mathrm{sch}(x - x_c)\cos(\varphi),$$

где φ – угол, определяющий поляризацию, A — амплитуда секанса, x_c — координата его центра. Спектральные коэффициенты отражения для такого решения для непрерывного спектра (вещественные λ) имеют следующий вид:

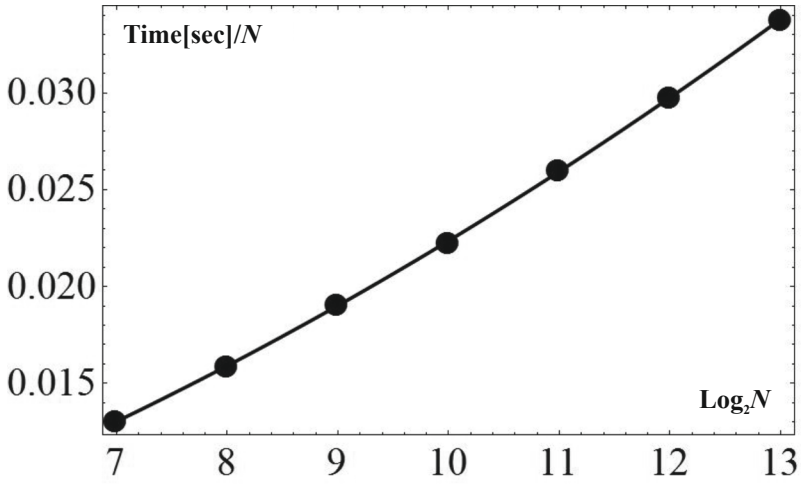
$$r_1(\lambda) = -\frac{\sin(\pi A)}{\mathrm{ch}(\pi\lambda)}\frac{\Gamma(-i\lambda + A + \frac{1}{2})\Gamma(-i\lambda - A + \frac{1}{2})}{\Gamma(-i\lambda + \frac{1}{2})^2}e^{-2i\lambda x_c}\sin(\varphi), \tag{5.1}$$

$$r_2(\lambda) = -\frac{\sin(\pi A)}{\mathrm{ch}(\pi\lambda)}\frac{\Gamma(-i\lambda + A + \frac{1}{2})\Gamma(-i\lambda - A + \frac{1}{2})}{\Gamma(-i\lambda + \frac{1}{2})^2}e^{-2i\lambda x_c}\cos(\varphi). \tag{5.2}$$

Здесь Γ – гамма-функция Эйлера. Гиперболический секанс – это, по существу, решение системы Захарова–Шабата, "повернутое" на поляризационный угол φ. Помимо непрерывного спектра гиперболический секанс порождает также и дискретный спектр (см. [22], [23]). Прямая ЗР для секанс-потенциала рассматривалась на интервале $x \in [-30, 30]$ при следующем наборе параметров: $A = 0.25, \varphi = 2\pi/3$. Численное моделирование прямой задачи рассеяния с помощью представленного супер-быстрого алгоритма и сравнение результатов расчетов с точными формулами (5.1), (5.2) показало, что при увеличении размера N задачи (размера дискретной сетки) вдвое, среднеквадратичная норма ошибки расчета спектральных коэффициентов отражения падает в 4 раза, что подтверждает второй порядок точности расчетов. В качестве примера приведем значения евклидовых норм ε абсолютной ошибки расчета вектора импульсного отклика (5.1), (5.2) при $N = 2^{11}$ и $N = 2^{12}$: $\varepsilon = 0.00014385$ и $\varepsilon = 0.00003597$ соответственно. Отношение этих ошибок равно 3.99917.

В ходе численного моделирования рассматривалась зависимость времени решения прямой задачи рассеяния Time от размера задачи N. Эта зависимость иллюстрируется на фиг. 1.

График на фиг. 1 показывает, что зависимость времени расчета Time (в секундах), деленного на размер задачи N, хорошо аппроксимируется квадратичной зависимостью $a + bM + cM^2$. Выше, на примере операций умножения, уже отмечалось, что при умеренных значениях M линейное по M слагаемое дает заметный вклад, сравнимый с вкладом квадратичного. Асимптотически, с ростом M, основной вклад станет давать квадратичное по M слагаемое, и итоговая оценка числа операций будет соответствовать оценке $O\left(N\mathrm{Log}_2^2 N\right)$.



Фиг. 1. Зависимость времени расчета Time (в секундах) от размера задачи N для прямой задачи рассеяния для системы Манакова “супер-быстрым” алгоритмом. Кружки – дискретные значения Time/N. Сплошная кривая – аппроксимация параболой от аргумента $M = \mathrm{Log}_2 N$.

6. МАТРИЧНЫЙ АЛГОРИТМ БПФ

Для локализованных потенциалов в предыдущих разделах был представлен быстрый алгоритм решения прямой задачи рассеяния, второго порядка точности, основанный на теореме о свертке и быстром преобразование Фурье. В поисках еще большего ускорения расчетов был реализован матричный вариант быстрого преобразования Фурье, когда коэффициенты ряда дискретного преобразования Фурье представляют собой некоммутирующие матрицы, возникающие при парном перемножении матриц, начиная с биномов $A_m + \varsigma B_m$.

Детальный анализ алгоритма БПФ по основанию 2 для варианта, предложенного Кули и Тьюки (см. [20]), показывает, что этот алгоритм не требует коммутативности коэффициентов дискретного ряда Фурье. Следовательно, его можно распространить на конечные ряды Фурье, коэффициенты которых представляют собой некоммутирующие объекты, например, матрицы. Для некоммутирующих коэффициентов рядов Фурье теорема о свертке выполняется, если не изменяется исходный порядок умножения. Такой матричный алгоритм БПФ был реализован и протестирован, на примере теоремы о свертке, для конечных рядов с коэффициентами в виде случайных матриц. Матричный алгоритм заметно упрощает схему алгоритма, и ожидалось, что он позволит уменьшить число арифметических операций и вызовов тригонометрических функций. Однако больших преимуществ в скорости расчета матричный алгоритм БПФ не дает, если применение “обычного” скалярного алгоритма БПФ достаточно оптимизировано.

Ускорение расчетов прямой ЗР в перспективе можно достичь с помощью так называемых теоретико-числовых преобразований Ферма, Мерсенна и др (см. [24]) в суррогатных полях, например, в полях Галуа, где тригонометрические функции заменяются на целочисленные выражения. Теоретико-числовые “спектры” хотя и не имеют того физического смысла, какой имеют фурье-спектры, однако для них тоже выполняется теорема о свертке. Использование целочисленной арифметики при расчете сверток полиномов позволит в будущем еще более ускорить расчеты прямой ЗР.

7. ЗАКЛЮЧЕНИЕ

Предложен алгоритм “супер-быстрого” решения прямых задач рассеяния для непрерывного спектра системы Манакова, ассоциированной с векторным нелинейным уравнением Шрёдингера модели Манакова, требующий для дискретной сетки размером N асимптотически порядка $O\left(N\mathrm{Log}_2^2 N\right)$ арифметических операций. Алгоритм имеет второй порядок точности для гладких локализованных (убывающих на бесконечности) решений прямой задачи рассеяния для непрерывного спектра системы Манакова. Этот алгоритм находит полиномиальную аппроксимацию зависимости спектральных данных (коэффициентов рассеяния $a\left(\lambda\right), b_{1,2}\left(\lambda\right)$) от спектрального параметра $\zeta = e^{i\lambda\tilde{h}}$, которые определяют спектральные коэффициенты отражения $r_{1,2}\left(\lambda\right)$. Для ускорения расчетов прямой задачи рассеяния для системы Манакова используется теорема о свертке и быстрое преобразование Фурье. Предложен и апробирован матричный вариант быстрого преобразования Фурье, когда коэффициенты ряда дискретного преобразования Фурье представляют собой некоммутирующие матри-

цы. При применении матричного алгоритма заметно упрощается структура алгоритма и оптимизируется число вычислений тригонометрических функций и арифметических операций. Впрочем, скалярный вариант алгоритма, при должной оптимизации, не уступает в быстродействии матричному алгоритму. Полученная полиномиальная аппроксимация спектральной зависимости коэффициентов рассеяния может быть использована в дальнейшем для поиска дискретного спектра системы Манакова.

СПИСОК ЛИТЕРАТУРЫ

1. *Захаров В. Е., Манаков С. В. , Новиков С. П., Питаевский Л. П.* Теория солитонов. Метод обратной задачи. М.: Наука, 1980. 319 с.
2. *Захаров В. Е., Шабат А. Б.* Точная теория двумерной самофокусировки и одномерной автомодуляции волн в нелинейных средах // Ж. эксперим. и теор. физ. 1971. Т. 61. С. 118.
3. *Манаков С. В.* К теории двумерной стационарной самофокусировки электромагнитных волн // Ж. эксперим. и теор. физ. 1973. Т. 65. № 2. С. 505.
4. *Агравал Г.* Нелинейная волоконная оптика. М.: Мир, 1995. 848 с.
5. *Richardson D. J.* Filling the Light Pipe // Science. 2010. V. 330 (6002). P. 327.
6. *Boffetta G. and Osborne A. R.* Computation of the direct scattering transform for the nonlinear Schroedinger equation // J. Comput. Phys. 1992. V. 102. P. 252.
7. *Burtsev S., Camassa R., Timofeyev I.* Numerical algorithms for the direct spectral transform with applications to nonlinear Schroedinger type systems // J. Comp. Phys. 1998. V. 147. № 1. P. 166.
8. *Белай О. В., Фрумин Л. Л., Чернявский А. Е.* Алгоритмы решения обратной задачи рассеяния для модели Манакова. // Ж. вычисл. матем. и матем. физ. 2024. Т. 64. № 2.
9. *Chernyavsky A. E., Frumin L. L.* Inverse scattering transform algorithm for the Manakov system // Comput. Optic. 2023. V. 47. № 6. P. 856.
10. *Белай О. В.* Быстрый численный метод второго порядка точности решения обратной задачи рассеяния // Квант. электроника. 2022. Т. 52. № 11. С. 1039.
11. *Frumin L.L.* Algorithms for solving scattering problems for the Manakov model of nonlinear Schrodinger equations // J. Inv. and Ill-posed Probl. 2021. V. 29. № 2. P. 369.
12. *Воеводин В.В., Тыртышников Е.Е.* Вычислительные процессы с тёплицевыми матрицами. М.: Наука, 1987. 320 с.
13. *Тыртышников Е.Е.* Тёплицевы матрицы, некоторые их аналоги и приложения. М.: Изд. АН СССР, 1989. 310 с.
14. *Долматов А.Р., Коняев Д.А.* Обобщение сверхбыстрых алгоритмов LayerPeeling для системы уравнений Манакова // Вестн. Московского ун-та. 2022. № 1. С. 23.
15. *Wahls S., Poor H.V.* Introducing the fast nonlinear Fourier transform. // Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP). 2013. P. 5780.
16. *Борн М., Вольф Л.Ф.* Основы оптики. М.: Наука, 1977. 720 с.
17. *Бреховских Л.М.* Волны в слоистых средах. М.: Наука, 1973. 343 с.
18. *Maimistov A.I., Basharov A.M.* Nonlinear optical waves. Dordrecht: Springer Science and Business Media, 2013. 670 p.
19. *Горбенко Н.И., Ильин В.П., Фрумин Л.Л.* Расчет рассеяния на Брэговской решетке рекурсией трансферматриц на неравномерной сетке // Автометрия. 2019. Т. 55. № 1. С. 40.
20. *Cooley, J.W., Tukey, J.W.* An algorithm for the machine calculation of complex Fourier series // Math. Comput. 1965. V. 19. P. 297.

21. *Бусленко А.С., Икрамов Х.Д.* Об умножении числовых и матричных степенных рядов // Ж. вычисл. матем. и матем. физ. 2005. Т 45. № 1. С. 3.

22. *Satsuma J., Yajima N. B.* Initial value problems of one-dimensional self-modulation of nonlinear waves in dispersive media // Progress Theor. Phys. Suppl. 1974. T. 55. C. 284.

23. *Мулляджанов Р.И., Гелаш А.А.* Разложение Магнуса для прямой задачи рассеяния: схемы высокого порядка // Изв. высш. уч. заведений. Радиофизика. 2020. Т. 63. № 9–10. С. 874.

24. *Маккеллан Дж.Х., Рейдер Ч.М.* Применение теории чисел в цифровой обработке сигналов. М.: Радио и связь, 1983. 264 с.

A “SUPER-FAST” ALGORITHM FOR SOLVING THE DIRECT SCATTERING PROBLEM FOR THE MANAKOV SYSTEM

L. L. Frumin*, A. E. Chernyavsky, O. V. Belay*****

630090 Novosibirsk, Akademika Koptyuga Ave., 1, Institute of Automation and Electrometry, SB RAS, Russia
**e-mail: lfrumin@iae.nsk.su*
***e-mail: alexander.cher.99@gmail.com*
****e-mail: ovbelai@gmail.com*

Received: 10.05.2024
Revised: 29.07.2024
Accepted: 23.08.2024

Abstract. The construction of an accelerated algorithm for solving the direct scattering problem for the continuous spectrum of the Manakov system associated with the vector nonlinear Schrodinger equation of the Manakov model is considered. The numerical formulation of the problem leads to the problem of quickly calculating the products of polynomials dependent on the spectral parameter of the problem. For localized solutions, the so-called “super-fast” algorithm for solving the direct scattering problem of the second order of accuracy is presented, based on the convolution theorem and the fast Fourier transform, which requires asymptotically only $O\left(N\ \mathrm{Log}^2 N\right)$ arithmetic operations for a discrete grid of size N. To speed up the calculation of the reflection coefficient spectra, a matrix variant of the fast Fourier transform is proposed and tested, when the coefficients of a series of discrete Fourier transforms are non-commuting matrices. Numerical simulation using the example of the exact solution of the Manakov system (hyperbolic secant) confirmed the high calculation speed and the second order of accuracy of the algorithm approximation.

Keywords: Schrodinger equation, Manakov system, direct scattering problem, transfer matrix, convolution, Fourier transform.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2024, том 64, № 12, с. 2420–2448

ИНФОРМАТИКА

УДК 519.86

ГИПОТЕЗА РАМСЕЯ О СОЦИАЛЬНОЙ СТРАТИФИКАЦИИ КАК ПРИНЦИП ОТБОРА ПО ФИШЕРУ[1)]

© 2024 г. Г.С. Парастаев[1,2,*], А.А. Шананин[1,2,3,4,5,]**

[1] 119991 Москва, Ленинские горы, 1, МГУ, Россия
[2] 119333 Москва, ул. Вавилова, 44, ФИЦ ИУ РАН, Россия
[3] 141701 М.о., Долгопрудный, Институтский пер., 9, МФТИ, Россия
[4] 119991 Москва, Ленинские горы, 1, Московский центр фундаментальной и прикладной математики, Россия
[5] 117198 Москва, ул. Миклухо-Маклая, 6, РУДН, Россия
**e-mail: parastaew1996@yandex.ru*
***e-mail: alexshan@yandex.ru*

Поступила в редакцию 23.08.2024 г.
Переработанный вариант 23.08.2024 г.
Принята к публикации 23.08.2024 г.

Гипотеза Рамсея о социальной стратификации утверждает, что в популяции домашних хозяйств богатство концентрируется у наиболее бережливых агентов, которые дисконтируют потребительские расходы с наименьшим коэффициентом дисконтирования. Гипотезу Рамсея можно рассматривать как утверждение о справедливости принципа естественного отбора по Фишеру в популяции домашних хозяйств. В статье на основе гипотезы Дж. Дьюзенберри коэффициенты дисконтирования формируются в зависимости от распределения капитала между агентами. Поведение домашних хозяйств описывается моделями рационального репрезентативного потребителя рамсеевского типа. Для соответствующих задач оптимального управления построены решения в форме синтеза, которые использованы при моделировании динамики популяции домашних хозяйств. Доказаны теоремы для популяции домашних хозяйств, обосновывающие справедливость гипотезы Рамсея. Исследовано влияние потребительского кредита на социальную стратификацию домашних хозяйств. Библ. 28. Фиг. 2.

Ключевые слова: оптимальное управление в форме синтеза, коэффициент дисконтирования, гипотеза относительного дохода, гипотеза Рамсея, функция Ляпунова.

DOI: 10.31857/S0044466924120156, **EDN**: KBERNL

1. ВВЕДЕНИЕ

Проблемы неравенства в распределении богатства, доходов и расходов в последние десятилетия приобретают такую же актуальность, как и сто лет назад. Тома Пикетти в монографии [1] утверждает, что такие явления, как социальные лифты, сокращение неравенства в доходах и расходах являются особенностью XX века и в современных условиях неравенство будет определяться возрастающим неравенством в распределении богатства, характерным для XVIII и XIX веков. В книге Филиппа Агийона и Джеффри Уильямсона [2], так же как и в монографии Т. Пикетти [1] обсуждается влияние возрастания неравенства на замедление темпов экономического роста мировой экономики. Ведущие экономисты, такие как Энтони Б. Аткинсон [3], ищут стратегии решения этих проблем.

Истоки современной теории экономического роста принято отсчитывать с конца 20-х годов XX века, когда на свет появилась работа английского математика Ф. П. Рамсея [4]. В ней Рамсей исследовал несколько вопросов, касающихся сбережений. Рамсеем было выдвинуто предположение о том, что в равновесном состоянии совокупность экономических агентов разделяется на два класса – тех, кто увеличивает свои сбережения и в конечном итоге оказывается собственником всего капитала, и тех, кто сокращает свои сбережения и впоследствии живет на уровне заработной платы. При этом всем богатством завладевает тот экономический агент, что имеет наименьший коэффициент дисконтирования среди всех остальных. Вопрос справедливости данного утверждения для более широкого класса моделей стал со временем приобретать все больший интерес среди

[1)] Работа выполнена при финансовой поддержке РНФ (грант 24-11-00329).

специалистов в области экономики и в смежных с ней областях и впоследствии закрепился как гипотеза Рамсея.

Глубина рассмотренной Рамсеем модели позволила развить ее в самых разных вариациях [5–7]. Одной из таких вариаций стала статья Р. Беккера [8], в которой изучается модель рамсеевского типа в дискретном времени с наложением запрета на заимствование капитала. Запрет был наложен с целью исключения ситуации, когда агент может брать определенную сумму в долг под свою будущую заработную плату. Вывод, полученный автором, также подтверждает справедливость гипотезы Рамсея.

Не менее весомый вклад внес Т. Бьюли, совместивший в своей работе рассмотрением модели рамсеевского типа теорию общего равновесия с классом теорем о магистрали [9]. Построенная Бьюли модель в дискретном времени допускает заимствование капитала, в отличие от модели, предложенной Беккером.

В совместной работе Г. Зоргера и Т. Митры [10] была рассмотрена модель Рамсея–Беккера в непрерывном времени, где коэффициенты дисконтирования заданы константными значениями. Помимо справедливости гипотезы Рамсея, авторам удалось установить единственность стационарного равновесного состояния, его глобальную асимптотическую устойчивость, а также следующее свойство: самое терпеливое домашнее хозяйство, т.е. то хозяйство, которое обладает наименьшим значением коэффициента дисконтирования, в пределе завладевает всем богатством, располагаемым в экономике.

В статье К. Ю. Борисова и М. А. Пахнина [7], посвященной обзору моделей экономического роста с неоднородным дисконтированием, приведены ссылки на эмпирические исследования, говорящие о наличии зависимости значения коэффициента дисконтирования от неравенства в распределении доходов, что приводит к выводу о необходимости рассмотрения альтернативных способов формирования временного предпочтения выбора. Одним из таких способов является эндогенное формирование межвременных предпочтений, которое было впервые рассмотрено Купмансом [11] и Удзавой [12]. В обзорной статье [7], в частности, приводятся ссылки на две работы одного из авторов [13, 14], где рассматриваются модели с эндогенностью временных предпочтений в дискретном времени.

Стоит также отметить, что в работе [7] обсуждается вопрос существования равновесных траекторий для моделей рамсеевского типа в случае, когда коэффициенты дисконтирования экономических агентов являются социально обусловленными, то есть, зависят от отношения располагаемого конкретным агентом богатства к средней величине богатства в экономике. Такой подход восходит к гипотезе относительного дохода американского экономиста Дж. Дьюзенберри [15]. В свое время Дж. М. Кейнс [16] заметил, что по мере увеличения своего текущего дохода индивиды сберегают все большую долю дохода, а тратят на потребление все меньшую долю дохода. Иными словами, функция потребления является вогнутой по доходу, а функция сбережений — выпуклой по доходу. Дж. Дьюзенберри [15] уточнил мысль Кейнса и выдвинул гипотезу относительного дохода, согласно которой предельная и средняя склонность индивида к сбережению растет с ростом его относительного дохода. К сожалению, теория потребления, основанная на идеях Дьюзенберри, была практически забыта, хотя, как отмечает известный специалист по поведенческой экономике Р. Фрэнк [17], она превосходит вытеснившую ее теорию перманентного дохода как в теоретическом, так и в эмпирическом разрезе.

Модели, формализующие гипотезу относительного дохода, были построены в работах [18] и [19]. В этих моделях естественным образом возникают двухклассовые равновесия. Однако эти модели обладают тем недостатком, что в них функции потребления и сбережения задаются непосредственным образом без микроэкономических оснований. Попытки преодолеть этот недостаток были предприняты в [13] и [14], где рассматривается модель, в которой каждый отдельный потребитель максимизирует межвременную полезность с коэффициентом дисконтирования, зависящим от его относительного уровня благосостояния, а также в [20], где предполагается, что функция полезности отдельного индивида подвержена отрицательной потребительской экстерналии (или, по-простому, зависти). В этих моделях тоже естественным образом возникают двухклассовые равновесия.

Данная работа предлагает несколько иной подход, когда в модели экономического роста стратегия потребления каждого агента определена оптимальным управлением в форме синтеза в задаче о поведении рационального репрезентативного потребителя. Благодаря такому подходу становится возможным использовать технику теории устойчивости. Кроме того, гипотеза Рамсея поддается интерпретации на языке популяционной динамики, что позволяет выявить аналогию гипотезы с принципом естественного отбора, изложенным в монографии Рональда Фишера (см. [21, Глава 2]).

Дальнейшее изложение устроено следующим образом. В разд. 2 описывается модель репрезентативного экономического агента на бесконечном горизонте. Разд. 3 посвящен построению синтеза оптимального управления в соответствующих экстремальных задачах. Модель социальной динамики и формулировки основных результатов приведены в разд. 4, в то время как их доказательства изложены в разд. 5. В разд. 6 установлена связь функции Ляпунова для модели популяционной динамики из разд. 4 с индексом неравенства Джини. В разд. 7 приводится заключение с указанием дальнейших направлений работы.

2. МОДЕЛИРОВАНИЕ ПОВЕДЕНИЯ ЭКОНОМИЧЕСКИХ АГЕНТОВ

В основу модели экономического поведения, описываемой задачами оптимального управления, положим концепцию Ф. Рамсея (см. [4]) и будем предполагать денежный рынок совершенным, т.е. равенство процентных ставок по кредитам и депозитам. Будем считать, что экономические решения, принимаемые типичным агентом, являются ограниченно рациональными, т.е. агент определяет потребительские расходы, сбережения и заимствования по потребительскому кредиту так, чтобы максимизировать дисконтированное потребление с учетом бюджетных ограничений. При этом предполагается, что он не может спрогнозировать изменения экономической конъюнктуры и принимает решения, считая, что его заработная плата w и процентная ставка на денежном рынке r будут оставаться на неизменном уровне. Будем также считать, что текущее финансовое состояние агента в момент времени t описывается его капиталом $k(t)$. Капитал агента изменяется в результате поступления заработной платы w, доходов $rk(t)$ и потребительских доходов $c(t) \geq 0$. Таким образом, ожидаемая агентом динамика капитала описывается уравнением

$$\dot{k}(t) = rk(t) + w - c(t)$$

с начальным условием

$$k(0) = k_0.$$

Тогда задание программы потребительских расходов $c(t)$ определяет предполагаемую динамику капитала. Предполагается, что агент стремится максимизировать дисконтированную функцию полезности с постоянным отвращением к риску

$$J(c(\cdot)) = \int_0^{+\infty} e^{-\rho s} (c(s))^{\beta}\, ds \to \max_{c(\cdot)},$$

где $\rho > 0$ – коэффициент дисконтирования, $1-\beta$ – коэффициент отвращения к риску, $\beta \in (0,1)$. Потребительские расходы $c(t)$ являются кусочно-непрерывной функцией.

Будем рассматривать две постановки задачи. В первой постановке допускается потребительский кредит, т.е. величина капитала может принимать отрицательные значения, и в этом случае агент, обслуживая кредит, выплачивает процентные платежи. Величина задолженности по потребительскому кредиту должна быть обеспечена будущими поступлениями заработной платы, т.е.

$$k(t) \geq -\frac{w}{r},\ t \in [0, +\infty).$$

Во второй постановке потребительский кредит не допускается, и капитал агента должен быть неотрицательным, т.е.

$$k(t) \geq 0,\ t \in [0, +\infty).$$

Первая постановка, названная в [7] моделью Рамсея–Бьюли, приводит к задаче оптимального управления вида

$$\int_0^{+\infty} e^{-\rho t} (c(t))^{\beta}\, dt \to \max_{c(\cdot)}, \tag{1}$$

$$\dot{k}(t) = rk(t) + w - c(t),\ k(0) = k_0, \tag{2}$$

$$c(t) \geq 0, \tag{3}$$

$$k(t) \geq -\frac{w}{r}. \tag{4а}$$

Вторая постановка, названная в [7] моделью Рамсея–Беккера, приводит к задаче оптимального управления

$$\int_0^{+\infty} e^{-\rho t} (c(t))^{\beta}\, dt \to \max_{c(\cdot)}, \tag{1}$$

$$\dot{k}(t) = rk(t) + w - c(t),\ k(0) = k_0, \tag{2}$$

$$c(t) \geq 0, \tag{3}$$

$$k(t) \geq 0. \tag{4б}$$

Поскольку далее описание поведения агента будет использоваться в модели «популяции» агентов, в которой коэффициенты дисконтирования формируются эндогенно в зависимости от распределения капитала между агентами, требуется получить из решения этих задач оптимального управления выбор размеров потребления агента в форме синтеза оптимального управления $c(k,\rho)$, т.е. как функцию текущего значения капитала (фазовой переменной задачи оптимального управления) и параметра ρ.

3. СИНТЕЗ В ЗАДАЧАХ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ

Решать задачи (1)–(3), (4а) и (1)–(3), (4б) мы будем, используя подход, который был использован в работах [22, 23] для доказательства утверждений, связанных с принципом максимума Понтрягина на бесконечном полуинтервале. Он заключается в рассмотрении последовательности вспомогательных задач оптимального управления на конечных отрезках $[0,T_n]$, $n=1,2,\ldots$, где $\{T_n\}_{n=1}^{\infty}$ – произвольная возрастающая последовательность положительных чисел, такая что $T_n \xrightarrow[n\to\infty]{} +\infty$. Следовательно, для любого $T>0$ найдется такое $n_1\in\mathbb{N}$, что $T_{n_1}\geq T$.

Будем рассматривать вспомогательную задачу оптимального управления с фиксированным временным горизонтом $[0,T]$ при функционале полезности вида

$$\int_0^T e^{-\rho t}\,(c(t))^{\beta}\,dt \to \max_{c(\cdot)} \tag{1_T}$$

с ограничениями (2), (3) на отрезке $[0,T]$, а также ограничением (4а) или (4б) соответственно.

Будем называть пару (c,k) *допустимой* для задачи оптимального управления, если управление $c(t)$ является кусочно-непрерывной функцией, удовлетворяющей ограничению (3), а соответствующая ему траектория $k(t)$ удовлетворяет задаче Коши (2) и ограничению (4а) (или (4б) соответственно).

Определение 1. Пусть для любого $T>0$ $c_{T_n}(t)$ сходится к $c_*(t)$ по норме в $L^1[0,T]$, а $k_{T_n}(t)$ равномерно сходится к функции $k_*(t)$ на отрезке $[0,T]$, где (c_{T_n},k_{T_n}) – оптимальная допустимая пара для задачи оптимального управления (1_{T_n}), (2), (3), (4а) (или (4б) соответственно) на отрезке $[0,T_n]$, $n=1,2,\ldots$ Пусть также пара (c_*,k_*) является допустимой для задачи (1)–(3), (4а) (или (4б) соответственно). Тогда пара (c_*,k_*) называется *обобщенным* решением задачи оптимального управления на бесконечном полуинтервале (1)–(3), (4а) (или (4б) соответственно).

Введенное выше понятие обобщенного решения позволяет разрешить вопрос с определением понятия оптимальности в случае, когда функционал (1) является расходящимся несобственным интегралом. При этом если значение функционала (1) конечно, то обобщенное решение задачи на бесконечном полуинтервале времени совпадает с классическим решением (см. [22]). С другими понятиями оптимальности при расходящемся интеграле в задаче оптимального управления на бесконечном полуинтервале можно ознакомиться, к примеру, в [24, 25].

Для вспомогательной задачи первого типа предлагается рассмотреть альтернативное ограничение. Оно заключается в том, что мы допускаем возможность домашнему хозяйству брать потребительский кредит, но обязываем его к конечному моменту времени погасить все накопленные долги. Иными словами, капитал домашнего хозяйства в конечный момент времени должен быть неотрицательным:

$$k(T)\geq 0. \tag{4^{a}_T}$$

Отметим, что на решении задачи (1_T), (2), (3), (4^{a}_T) условие (4а) выполняется автоматически. Мы далее покажем, что в пределе при $T\to+\infty$ решение этой задачи даст оптимальное управление на бесконечном горизонте времени для задачи (1)–(3) с ограничением (4а).

Далее под записью p_+ будем понимать $p_+=\max\{p,0\}$.

Теорема 3.1. *Синтез в задаче оптимального управления* (1)–(3) *с ограничением* (4а) *имеет вид*

$$c_1(k,\rho)=\frac{(\rho-\beta r)_+}{1-\beta}\left[k+\frac{w}{r}\right]. \tag{4}$$

Теорема 3.2. *Синтез в задаче оптимального управления* (1)–(3) *с ограничением* (4б) *имеет вид*

$$c_2(k,\rho)=\begin{cases} c_1(k,\rho), & \rho\leq r,\\ \hat{c}(k,\rho), & \rho>r,\end{cases} \tag{5}$$

где функция $\hat{c} = \hat{c}\,(k, \rho) \geq w$ *является решением уравнения*

$$\frac{\rho - \beta r}{1 - \beta}\left(k + \frac{w}{r}\right) = \hat{c} + \frac{\rho - r}{(1-\beta)\,r} w \left(\frac{w}{\hat{c}}\right)^{\frac{(1-\beta)r}{\rho - r}}. \tag{6}$$

Перед доказательством теорем напомним одно важное утверждение, которое позволяет проверить полученное решение на оптимальность. Рассмотрим функцию цены

$$V\,(t,k) = \sup\left\{\int_t^T e^{-\rho(s-t)}\,(c\,(s))^{\beta}\,ds\right\},$$

соответствующую задаче Коши (2) с условием $k\,(t) = k$, ограничению (3) и ограничению $(4^{\text{а}}_T)$ или (4б) соответственно. Задаче (1_T), (2), (3), $(4^{\text{а}}_T)$ (или (4б) соответственно) соответствует уравнение Гамильтона–Якоби–Беллмана

$$\frac{\partial V}{\partial t} - \rho V + \sup\left\{\frac{\partial V}{\partial k}\,[rk + w - c] + c^{\beta}\,\middle|\, c \geq 0\right\} = 0 \tag{7}$$

с граничным условием $V\,(T, k\,[T]) = 0$.

Сформулируем и докажем вариант теоремы верификации (см. [26, с.12]).

Лемма 3.1. *Пусть функция* $\mathcal{V}\,(t,k)$ *непрерывно дифференцируема по* (t,k) *в области* $[0,T] \times (a + \mathbb{R}_+)\,, a \leq 0$, *и удовлетворяет уравнению Гамильтона–Якоби–Беллмана* (7) *с граничным условием* $\mathcal{V}\,(T,k) = 0$. *Тогда* $\mathcal{V}\,(t,k) \geq$ $\geq V\,(t,k)$. *Пусть также управление* $c^0\,(s)$ *и отвечающая ему траектория* $k^0\,[s] = k^0\,(s,t,k)\,, s \geq t, k^0\,[t] = k$, *удовлетворяют соотношению*

$$\frac{\partial \mathcal{V}\left(s, k^0\,(s)\right)}{\partial k}\left[rk^0\,(s) + w - c^0\,(s)\right] + \left(c^0\,(s)\right)^{\beta} =$$
$$= \sup_{c \geq 0}\left\{\frac{\partial \mathcal{V}\left(s, k^0\,(s)\right)}{\partial k}\left[rk^0\,(s) + w - c\right] + c^{\beta}\right\}.$$

Тогда управление $c^0\,(\cdot)$ *и соответствующая ему траектория* $k^0\,[\cdot]$ *оптимальны, и* $\mathcal{V}\,(t,k) = V\,(t,k)$.

Доказательство. Возьмем произвольное управление $\tilde{c}\,(\cdot) \geq 0$ и отвечающую ему траекторию $\tilde{k}\,[s] = \tilde{k}\,(s,t,k)$ с начальным значением $\tilde{k}\,[t] = k$. Из уравнения Гамильтона–Якоби–Беллмана для функции $\mathcal{V}$ следует, что для управления $\tilde{c}\,(\cdot)$ выполняется неравенство

$$\frac{\partial \mathcal{V}\left(t, \tilde{k}\,[t]\right)}{\partial t} - \rho \mathcal{V}\left(t, \tilde{k}\,[t]\right) + \frac{\partial \mathcal{V}\left(t, \tilde{k}\,[t]\right)}{\partial k}\left[r\tilde{k}\,[t] + w - \tilde{c}\,(t)\right] + (\tilde{c}\,(t))^{\beta} \leq 0.$$

Представив разность

$$e^{-\rho T}\mathcal{V}\left(T, \tilde{k}\,[T]\right) - e^{-\rho t}\mathcal{V}\,(t,k) = \int_t^T \frac{d\left(e^{-\rho s}\mathcal{V}\left(s, \tilde{k}\,(s)\right)\right)}{ds}\,ds,$$

имеем

$$e^{-\rho T}\mathcal{V}\left(T, \tilde{k}\,[T]\right) - e^{-\rho t}\mathcal{V}\,(t,k) = \int_t^T e^{-\rho s}\left(-\rho \mathcal{V}\left(s, \tilde{k}\,(s)\right) + \frac{\partial \mathcal{V}\left(s, \tilde{k}\,(s)\right)}{\partial s} + \right.$$
$$\left. + \frac{\partial \mathcal{V}\left(s, \tilde{k}\,(s)\right)}{\partial k}\left[r\tilde{k}\,(s) + w - \tilde{c}\,(s)\right]\right)\,ds \leq -\int_t^T e^{-\rho s}\,(\tilde{c}\,(s))^{\beta}\,ds.$$

Следовательно, так как $\mathcal{V}\left(T, \tilde{k}\,[T]\right) = 0$, имеем

$$\mathcal{V}\,(t,k) \geq e^{\rho t}\int_t^T e^{-\rho s}\,(\tilde{c}\,(s))^{\beta}\,ds,$$

откуда следует, что $\mathcal{V}(t,k) \geq V(t,k)$. С другой стороны, для управления $c^0(\cdot)$ и отвечающей ему траектории $k^0(\cdot)$ имеем

$$e^{-\rho T}\mathcal{V}\left(T,k^0[T]\right) - e^{-\rho t}\mathcal{V}(t,k) = \int\limits_t^T e^{-\rho s}\left(-\rho\mathcal{V}\left(s,k^0(s)\right) + \frac{\partial\mathcal{V}\left(s,k^0(s)\right)}{\partial s} + \right.$$

$$\left. + \frac{\partial\mathcal{V}\left(s,k^0(s)\right)}{\partial k}\left[rk^0(s) + w - c^0(s)\right]\right) ds = -\int\limits_t^T e^{-\rho s}\left(c^0(s)\right)^{\beta} ds \Rightarrow$$

$$\Rightarrow \mathcal{V}(t,k) = e^{\rho t}\int\limits_t^T e^{-\rho s}\left(c^0(s)\right)^{\beta} ds \leq V(t,k),$$

поэтому $\mathcal{V}(t,k) = V(t,k)$, и управление $c^0(\cdot)$ является оптимальным. Лемма доказана.

Лемма 3.2. *Задача оптимального управления* (1_T), (2), (3) *с ограничением* (4_T^{a}) *имеет решение вида*

$$c_{T,1}(t,k) = \frac{\rho-\beta r}{1-\beta}\frac{k + \frac{w}{r}\left(1 - e^{-r(T-t)}\right)}{1 - e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}}. \tag{8}$$

Доказательство. Функция Гамильтона–Понтрягина задачи (1_T), (2), (3), (4_T^{a}) имеет вид

$$\mathcal{H}_1(k,p,c) = c^{\beta} + p(rk + w - c).$$

Первое условие – достижение максимума функции Гамильтона–Понтрягина по c. Поскольку функция $\mathcal{H}_1$ дифференцируема по c, для нее справедливо необходимое условие экстремума:

$$\frac{\partial\mathcal{H}_1(k,c,p)}{\partial c}\bigg|_{\substack{k=k_1(s),\,c=c_1(s),\\ p=p_1(s)}} = \beta\left(c_1(s)\right)^{\beta-1} - p_1(s) = 0 \Rightarrow c_1(s) = \left(\frac{\beta}{p_1(s)}\right)^{\frac{1}{1-\beta}}.$$

Полученное выражение имеет смысл при $p_1(s) > 0$. При $p_1(s) \leq 0$ мы имеем $\frac{\partial\mathcal{H}_1}{\partial c} > 0$, поэтому максимум функции Гамильтона–Понтрягина достигается при $c_1 = +\infty$. Поскольку управление должно быть конечным, условие $p_1(s) > 0$ является необходимым. Следовательно,

$$c_{T,1}[s] = \left(\frac{\beta}{p_1(s)}\right)^{\frac{1}{1-\beta}}.$$

Сопряженное уравнение для функции $p_1(s)$ имеет вид

$$\dot{p}_1(s) = \rho p_1(s) - \frac{\partial\mathcal{H}_1(k,p,c)}{\partial k}\bigg|_{\substack{k=k_1(s),p=p_1(s),\\ c=c_1(s)}} = (\rho - r)p_1(s),$$

решением которого является функция $p_1(s) = p_1(t)\,e^{(\rho-r)(s-t)}$.

Также имеем условие трансверсальности в момент времени T:

$$p_1(T)\,k_1(T) = 0.$$

Из формулы для $p_1(s)$ очевидно, что сопряженная переменная p_1 не меняет знака на отрезке $[t,T]$. В силу конечности значения капитала в момент времени T имеем $p_1(t) > 0$, поэтому $p_1(T) > 0$. Следовательно, по условию трансверсальности должно выполняться равенство $k_1(T) = 0$, при этом мы имеем следующую систему ОДУ для пары (k_1, p_1):

$$\dot{k_1}(s) = rk_1(s) + w - \left(\frac{\beta}{p_1(s)}\right)^{\frac{1}{1-\beta}},\ k_1(t) = k,$$
$$\dot{p_1}(s) = (\rho - r)p_1(s),\ p_1(t) = p_{1t}.$$

Подставляя найденное решение второго уравнения в первое, находим формулу для капитала:

$$k_1(s) = ke^{r(s-t)} + \frac{w}{r}\left(e^{r(s-t)} - 1\right) - \left(\frac{\beta}{p_{1t}}\right)^{\frac{1}{1-\beta}}\frac{1-\beta}{\rho-\beta r}e^{r(s-t)}\left(1 - e^{-\frac{\rho-\beta r}{1-\beta}(s-t)}\right).$$

Следовательно, используя условие трансверсальности $k_1(T)=0$, мы можем выразить

$$\left(\frac{\beta}{p_{1t}}\right)^{\frac{1}{1-\beta}}=\frac{\rho-\beta r}{1-\beta}\frac{k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)}{1-e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}}.$$

Тогда управление в программной форме для первой задачи имеет вид

$$c_{T,1}[s]=\frac{\beta r-\rho}{1-\beta}\frac{k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)}{e^{\frac{\beta r-\rho}{1-\beta}(T-t)}-1}e^{\frac{r-\rho}{1-\beta}(s-t)},$$

а функционал принимает значение, равное

$$J(c_{T,1}[\cdot])=\int_t^T e^{-\rho s}\left(c_{T,1}[s]\right)^{\beta}ds=e^{-\rho t}\left(\frac{\beta r-\rho}{1-\beta}\frac{k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)}{e^{\frac{\beta r-\rho}{1-\beta}(T-t)}-1}\right)^{\beta}\times$$

$$\times\int_t^T e^{\frac{\beta r-\rho}{1-\beta}(s-t)}ds=e^{-\rho t}\left(k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)\right)^{\beta}\left(\frac{1-\beta}{\beta r-\rho}\left(e^{\frac{\beta r-\rho}{1-\beta}(T-t)}-1\right)\right)^{1-\beta}.$$

Воспользуемся теоремой верификации (леммой 3.1). Значение функционала имеет вид

$$J(c_{T,1}[\cdot])=e^{-\rho t}\left(k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)\right)^{\beta}\left[\frac{1-\beta}{\rho-\beta r}\left(1-e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}\right)\right]^{1-\beta}.$$

Тогда $\mathcal{V}(t,k)=e^{\rho t}J(c_{T,1}[\cdot])$. Функция $\mathcal{V}(t,k)$ по построению является непрерывно дифференцируемой по (t,k). Найдем частные производные функции $\mathcal{V}$, обозначив $\mathcal{V}_1(t,k)=k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)$, $\mathcal{V}_2(t)=$ $=\frac{1-\beta}{\rho-\beta r}\left(1-e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}\right)$:

$$\mathcal{V}(t,k)=\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{1-\beta}\Rightarrow$$

$$\Rightarrow\frac{\partial\mathcal{V}}{\partial t}=-we^{-r(T-t)}\beta\mathcal{V}_1(t,k)^{\beta-1}\mathcal{V}_2(t)^{1-\beta}-e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}(1-\beta)\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{-\beta},$$

$$\frac{\partial\mathcal{V}}{\partial k}=\beta\mathcal{V}_1(t,k)^{\beta-1}\mathcal{V}_2(t)^{1-\beta}.$$

Найдем максимизатор:

$$c=\left(\frac{\beta}{\frac{\partial\mathcal{V}}{\partial k}}\right)^{\frac{1}{1-\beta}}=\frac{\mathcal{V}_1(t,k)}{\mathcal{V}_2(t)}.$$

Тогда

$$\frac{\partial\mathcal{V}}{\partial t}-\rho\mathcal{V}+\sup_{c\geq0}\left\{\frac{\partial\mathcal{V}}{\partial k}[rk+w-c]+c^{\beta}\right\}=$$

$$=-we^{-r(T-t)}\beta\mathcal{V}_1(t,k)^{\beta-1}\mathcal{V}_2(t)^{1-\beta}-e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}(1-\beta)\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{-\beta}-$$

$$-\rho\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{1-\beta}+\beta\mathcal{V}_1(t,k)^{\beta-1}\mathcal{V}_2(t)^{1-\beta}[rk+w]-$$

$$-\beta\mathcal{V}_1(t,k)^{\beta-1}\mathcal{V}_2(t)^{1-\beta}\frac{\mathcal{V}_1(t,k)}{\mathcal{V}_2(t)}+\left(\frac{\mathcal{V}_1(t,k)}{\mathcal{V}_2(t)}\right)^{\beta}=$$

$$=\underbrace{(1-\beta)\left(1-e^{-\frac{\rho-\beta r}{1-\beta}(T-t)}\right)}_{(\rho-\beta r)\mathcal{V}_2(t)}\left(\frac{\mathcal{V}_1(t,k)}{\mathcal{V}_2(t)}\right)^{\beta}+$$

$$+\beta r\underbrace{\left(k+\frac{w}{r}\left(1-e^{-r(T-t)}\right)\right)}_{\mathcal{V}_1(t,k)}\mathcal{V}_1(t,k)^{\beta-1}\mathcal{V}_2(t)^{1-\beta}-\rho\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{1-\beta}=$$

$$=(\rho-\beta r)\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{1-\beta}+\beta r\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{1-\beta}-\rho\mathcal{V}_1(t,k)^{\beta}\mathcal{V}_2(t)^{1-\beta}=0.$$

Таким образом, по теореме верификации (см. лемму 3.1) построенные с помощью принципа максимума Понтрягина траектория и управление (8) являются оптимальным решением задачи (1_T), (2), (3), (4_T^{a}). Лемма доказана.

Доказательство теоремы 3.1. Искомое управление находится в результате перехода к пределу в управлении, полученном в лемме 3.2, при $T \to +\infty$:

$$c_1[s] = \lim_{T\to+\infty} c_{T,1}[s] = \begin{cases} \dfrac{\rho-\beta r}{1-\beta}\left(k+\dfrac{w}{r}\right)e^{\frac{r-\rho}{1-\beta}(s-t)}, & \rho > \beta r, \\ 0, & \rho \le \beta r, \end{cases}$$

$$\lim_{T\to+\infty} J(c_{T,1}[\cdot]) = \begin{cases} \left(k+\dfrac{w}{r}\right)^{\beta}\left(\dfrac{1-\beta}{\rho-\beta r}\right)^{1-\beta}, & \rho > \beta r, \\ +\infty, & \rho \le \beta r. \end{cases}$$

Заметим, что при $T \to +\infty$ выражение капитала для первой задачи принимает вид

$$k_1(s) = \left(k+\frac{w}{r}\right)e^{\frac{r-\rho}{1-\beta}(s-t)} - \frac{w}{r}.$$

Данное обстоятельство позволяет нам выразить $c_1[s]$ в форме синтеза $c_1(k,\rho)$ по формуле (4).

Для доказательства теоремы 3.2 докажем четыре вспомогательные леммы.

Лемма 3.3. *Пусть выполняется неравенство* $\rho \le r$. *Тогда решения задач* (1_T), (2), (3), (4_T^{a}) *и* (1_T), (2), (3), (4б) *совпадают.*

Доказательство. Так как фазовое ограничение (4б) включает в себя ограничение (4_T^{a}), то если управление $c_{T,1}[s]$ является допустимым и для задачи с ограничением (4б), то доставляемое им значение функционала оказывается оптимальным. Возьмем оптимальное управление $c_{T,2}[s] = c_{T,1}(s,t,k)$ и убедимся, что доставляемая этим управлением траектория $k_2(s)$ не нарушает фазовых ограничений. Так как $k_2(t) = k, k_2(T) = 0$, то рассмотрим $s \in (t,T)$. Введя обозначение $F(x) = \frac{\rho-\beta r}{1-\beta}\frac{k+\frac{w}{r}\left(1-e^{-rx}\right)}{1-e^{-\frac{\rho-\beta r}{1-\beta}x}}$, запишем выражение для траектории капитала $k_2(s)$ в следующем виде:

$$k_2(s) = ke^{r(s-t)} + \frac{w}{r}\left(e^{r(s-t)}-1\right) - F(T-t)\,e^{r(s-t)}\frac{1-e^{-\frac{\rho-\beta r}{1-\beta}(s-t)}}{\frac{\rho-\beta r}{1-\beta}} =$$

$$= e^{r(s-t)}\frac{1-e^{-\frac{\rho-\beta r}{1-\beta}(s-t)}}{\frac{\rho-\beta r}{1-\beta}}\left(F(s-t) - F(T-t)\right).$$

Вычислим производную функции $F(x)$:

$$F'(x) = \frac{\rho-\beta r}{1-\beta}\frac{we^{-rx}\left(1-e^{-\frac{\rho-\beta r}{1-\beta}x}\right) - \frac{\rho-\beta r}{1-\beta}e^{-\frac{\rho-\beta r}{1-\beta}x}\left(k+\frac{w}{r}\left(1-e^{-rx}\right)\right)}{\left(1-e^{-\frac{\rho-\beta r}{1-\beta}x}\right)^2} =$$

$$= e^{-\frac{\rho+(1-2\beta)r}{1-\beta}x}\left(\frac{\frac{\rho-\beta r}{1-\beta}}{1-e^{-\frac{\rho-\beta r}{1-\beta}x}}\right)^2\left(w\left[\frac{e^{\frac{\rho-\beta r}{1-\beta}x}-1}{\frac{\rho-\beta r}{1-\beta}} - \frac{e^{rx}-1}{r}\right] - ke^{rx}\right).$$

Исследуем теперь поведение функции $f(a,x) = \frac{e^{ax}-1}{a}$ с параметром $x > 0$:

$$\frac{\partial f(a,x)}{\partial a} = \frac{(ax-1)e^{ax}+1}{a^2} = \frac{e}{a^2}\left((ax-1)e^{ax-1} + \frac{1}{e}\right).$$

Так как функция $w(s) = se^s$ достигает минимума в точке $s_0 = -1$ со значением, равным $-\frac{1}{e}$, то числитель дроби положителен при $a \ne 0$. Разложим экспоненту e^{ax} в ряд Тейлора с точностью до второго порядка малости по a:

$$\frac{\partial f(a,x)}{\partial a} = \frac{(ax-1)\left(1+ax+\frac{a^2x^2}{2}+o\left(a^2\right)\right)+1}{a^2} = \left(\frac{x^2}{2}+\frac{ax^3}{2}+o(1)\right) \xrightarrow[a\to 0]{} \frac{x^2}{2}.$$

Следовательно, для любого $x > 0$ $\frac{\partial f(a,x)}{\partial a} > 0, \forall a \in \mathbb{R}$. Но тогда $f\left(\frac{\rho-\beta r}{1-\beta}, x\right) \le f(r,x)$, и поэтому $F'(x) < 0, \forall x > 0$, и по теореме Лагранжа о среднем значении отсюда следует, что найдется такая точка $\xi = \xi(s-t, T-t) \in (s-t, T-t)$, что $F(s-t) - F(T-t) = F'(\xi)(s-T) > 0, \forall s \in (t,T)$. Следовательно, $k_2(s) > 0$. При $k = 0$ и $\rho = r$ имеет место равенство $F(x) = w\ \forall x \ge 0$, поэтому $k_2(s) = 0$. Лемма доказана.

Лемма 3.4. *Пусть* $\rho > r$. *Тогда уравнение* (6) *при каждом фиксированном значении* $k > 0$ *имеет два корня* $c_1 \in (0, w)$, $c_2 > w$, *а при* $k = 0$ *обладает единственным корнем* $c = w$.

Доказательство. Исследуем правую часть уравнения (6) как функцию от c:

$$f(c) = c + \frac{\rho - r}{(1-\beta)\, r} w \left(\frac{w}{c}\right)^{\frac{(1-\beta)r}{\rho - r}}.$$

Найдем ее производную: $f'(c) = 1 - \left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho - r}}$. Нетрудно заметить, что $f'(w) = 0$, причем $f'(c) < 0$, $0 < c < w$ и $f'(c) > 0$, $c > w$. Следовательно, в точке $c = w$ функция достигает минимального значения, равного $f_{\min} = \frac{\rho-\beta r}{(1-\beta)r} w$. Левая часть уравнения (6), которую мы обозначим через $h(k)$, достигает значения $f_{\min}$ тогда и только тогда, когда $k = 0$. Так как $f(c)$ и $h(k)$ являются непрерывными функциями своих аргументов, причем $\lim\limits_{c \to 0+} f(c) = \lim\limits_{c \to +\infty} f(c) = +\infty = \lim\limits_{k \to +\infty} h(k)$, то в силу теоремы о прохождении непрерывной функции через промежуточное значение для любого $f \in (f_{\min}, +\infty)$ найдутся $c_1 \in (0, w)$ и $c_2 \in (w, +\infty)$ такие, что $f(c_1) = f(c_2) = f$. В свою очередь, для функции h найдется такое $\tilde{k}$, что $h\left(\tilde{k}\right) = f$. Таким образом, имеем для каждого $k \in (0, +\infty)$ два корня: $c_1 \in (0, w)$ и $c_2 > w$. Лемма доказана.

Лемма 3.5. *Пусть* $\rho > r$. *Тогда задача оптимального управления* (1_T), (2), (3) *с ограничением* (4б) *при* $k_0 = 0$ *имеет решение* $c(t) = w$.

Доказательство. Так как x^β – вогнутая функция, в силу неравенства Иенсена имеем

$$\int_0^T \frac{\rho e^{-\rho t}}{1 - e^{-\rho T}} (c(t))^\beta \, dt \le \left(\int_0^T \frac{\rho e^{-\rho t}}{1 - e^{-\rho T}} c(t)\, dt \right)^\beta = \left(\frac{\rho}{1 - e^{-\rho T}} \right)^\beta \left(\int_0^T e^{-\rho t} c(t)\, dt \right)^\beta.$$

Отсюда следует, что

$$\int_0^T e^{-\rho t} (c(t))^\beta \, dt \le \left(\frac{1 - e^{-\rho T}}{\rho} \right)^{1-\beta} \left(\int_0^T e^{-\rho t} c(t)\, dt \right)^\beta,$$

причем равенство достигается тогда и только тогда, когда $c(t) \equiv \text{const}, t \in [0, T]$.

Рассмотрим теперь вспомогательную задачу оптимального управления с функционалом

$$\int_0^T e^{-\rho t} c(t)\, dt \to \max_{c(\cdot)},$$

задачей Коши (2) при начальном условии $k_0 = 0$ и ограничениями (3), (4б). Тогда траектория капитала имеет вид

$$k(t) = \int_0^t e^{r(t-\tau)} (w - c(\tau))\, d\tau \ge 0.$$

Из формулы для траектории капитала и ограничения (4б) следует, что справедливо неравенство

$$\int_0^t e^{-r\tau} c(\tau)\, d\tau \le \frac{w}{r} \left(1 - e^{-rt}\right).$$

Тогда

$$\int_0^T e^{-\rho t} c(t)\, dt = \int_0^T e^{-(\rho - r)t} d \left(\int_0^t e^{-rs} c(s)\, ds \right) = e^{-(\rho-r)t} \int_0^t e^{-rs} c(s)\, ds \Big|_0^T +$$
$$+ (\rho - r) \int_0^T e^{-(\rho-r)t} \left(\int_0^t e^{-rs} c(s)\, ds \right) dt = e^{-(\rho-r)T} \int_0^T e^{-rs} c(s)\, ds +$$
$$+ (\rho - r) \int_0^T e^{-(\rho-r)t} \left(\int_0^t e^{-rs} c(s)\, ds \right) dt.$$

Положив $x(t) = \int\limits_0^t e^{-rs} c(s)\, ds$, получаем задачу оптимизации

$$(\rho - r) \int\limits_0^T e^{-(\rho-r)t} x(t)\, dt + e^{-(\rho-r)T} x(T) \to \max_{x(\cdot)},$$

$$0 \le x(t) \le \frac{w}{r}\left(1 - e^{-rt}\right),$$

решение которой достигается равенством $x(t) = \frac{w}{r}\left(1 - e^{-rt}\right), t \in [0, T]$. Тогда, взяв производную по времени функции $x(t)$, получаем, что для любого $t \in [0, T]$

$$\frac{dx}{dt} = we^{-rt} = e^{-rt} c(t).$$

Следовательно, $c(t) = w, t \in [0, T]$. Но на функции, равной константе, неравенство Иенсена обращается в равенство, поэтому функционал (1_T) достигает на функции $c(t) = w$ максимального значения. Следовательно, управление $c(t) = w$ оптимально. Лемма доказана.

Лемма 3.6. *Задача оптимального управления* (1_T), (2), (3) *с ограничением* (4б) *имеет решение вида*

$$c_{T,2}(t, k) = \begin{cases} c_{T,1}(t, k), & \rho \le r, \\ \hat{c}(k, \rho), & \rho > r, \end{cases} \tag{9}$$

где функция $\hat{c} = \hat{c}(k, \rho) \ge w$ *является решением уравнения* (6).

Доказательство. Функция Гамильтона–Понтрягина для задачи (1_T), (2), (3), (4б) имеет вид

$$\mathcal{H}_2(k, c, p, \mu) = c^\beta + p[rk + w - c] + \mu k,\ \mu \ge 0,\ \mu k = 0.$$

Первое условие – достижение максимума функций Гамильтона–Понтрягина по c. Поскольку функция $\mathcal{H}_2$ дифференцируема по c, для нее справедливо необходимое условие экстремума:

$$\left.\frac{\partial \mathcal{H}_2(k, c, p, \mu)}{\partial c}\right|_{\substack{k=k_2(s),\, c=c_2(s),\\ p=p_2(s),\, \mu=\mu(s)}} = \beta (c_2(s))^{\beta-1} - p_2(s) = 0 \Rightarrow c_2(s) = \left(\frac{\beta}{p_2(s)}\right)^{\frac{1}{1-\beta}}.$$

Полученное выражение имеет смысл при $p_2(s) > 0$. При $p_2(s) \le 0$ мы имеем $\frac{\partial \mathcal{H}_2}{\partial c} > 0$, поэтому максимум функции Гамильтона–Понтрягина достигается при $c_2 = +\infty$. Поскольку управление должно быть конечным, условие $p_2(s) > 0$ является необходимым. Следовательно,

$$c_{T,2}[s] = \left(\frac{\beta}{p_2(s)}\right)^{\frac{1}{1-\beta}}.$$

Сопряженное уравнение для функции $p_2(s)$ имеет вид

$$\dot{p}_2(s) = \rho p_2(s) - \left.\frac{\partial \mathcal{H}_2(k, c, p, \mu)}{\partial k}\right|_{\substack{k=k_2(s),\, c=c_2(s),\\ p=p_2(s),\, \mu=\mu(s)}} = (\rho - r) p_2(s) - \mu(s),$$

решением которого являются функция $p_2(s) = p_2(t) e^{(\rho-r)(s-t)} - e^{(\rho-r)(s-t)} \int\limits_t^s \mu(\tau) e^{-(\rho-r)(\tau-t)}\, d\tau$.

Условие трансверсальности в момент времени T имеет вид

$$p_2(T) k_2(T) = 0.$$

Теперь разберем решение второй задачи. Так как по лемме 3.3 решения задач совпадают при $\rho \le r$, нам осталось разобрать случай, когда $\rho > r$.

Для начала покажем, что ранее рассмотренная траектория капитала становится недопустимой. Для этого рассмотрим производную по времени функции k_2 в момент времени T, используя формулу (8) и условие трансверсальности $k_2(T) = 0$:

$$\dot{k}_2(T) = rk_2(T) + w - c_{T,2}[T] = w - F(T - t) e^{-\frac{\rho-r}{1-\beta}(T-t)} \vee 0.$$

Посредством арифметических преобразований нетрудно привести сравнение с нулем к эквивалентному сравнению вида

$$w\left(\frac{e^{\frac{\rho-r}{1-\beta}(T-t)}-1}{\frac{\rho-r}{1-\beta}}-\frac{1-e^{-r(T-t)}}{r}\right)\vee\frac{\rho-\beta r}{\rho-r}k.$$

Так как при $a>0$, $b>0$ у функции $g(x,a,b)=\frac{e^{ax}-1}{a}-\frac{1-e^{-bx}}{b}$ частная производная

$$\frac{\partial g}{\partial x}(x,a,b)=e^{ax}-e^{-bx}>0\,\forall\,x>0,$$

то левая часть возрастает по T, причем $g\left(T-t,\frac{\rho-r}{1-\beta},r\right)\xrightarrow[T\to+\infty]{}+\infty$, $g\left(0,\frac{\rho-r}{1-\beta},r\right)=0$, поэтому по теореме о прохождении непрерывной функции через любое промежуточное значение найдется такое $\hat{T}$, что полученное сравнение обращается в равенство при $T=\hat{T}$. Следовательно, выбрав $T>\hat{T}$, получаем, что $\dot{k}_2(T)>0$, что означает, что найдется $\Delta>0$ такое, что $k_2(t)<0, t\in(T-\Delta,T)$. Это свидетельствует о нарушении фазового ограничения задачи и недопустимости траектории $k_2(s)$. Следовательно, существует момент времени $T_1\in[t,T)$ такой, что $k_2(T_1)=0$. В силу леммы 3.5 выполняется равенство $c_2(s)=w, s\in[T_1,T]$, откуда следует, что $k_2(s)=0\,\forall s\in[T_1,T]$. Следовательно, в силу условия дополняющей нежесткости $p_2(s)=\beta w^{\beta-1}$, и $\mu(s)=$ $=(\rho-r)\beta w^{\beta-1}>0$. Тогда из соотношений $k_2(T_1)=0, c_2(T_1)=w$ можно вывести систему из двух уравнений с неизвестными моментом времени T_1 и значением сопряженной переменной $p(t)=p_{2t}$:

$$\left(\frac{\beta}{p_{2t}}\right)^{\frac{1}{1-\beta}}e^{-\frac{\rho-r}{1-\beta}(T_1-t)}=w,$$
$$k+\frac{w}{r}\left(1-e^{-r(T_1-t)}\right)-\left(\frac{\beta}{p_{2t}}\right)^{\frac{1}{1-\beta}}\frac{1-\beta}{\rho-\beta r}\left(1-e^{-\frac{\rho-\beta r}{1-\beta}(T_1-t)}\right)=0.$$

Так как $c_2(t)=\left(\frac{\beta}{p_{2t}}\right)^{\frac{1}{1-\beta}}$, нам удобнее выразить момент времени T_1 из первого уравнения как $T_1=t+\frac{1-\beta}{\rho-r}\ln\frac{c(t)}{w}$. Подставив T_1 во второе уравнение, получаем следующее уравнение относительно $c_2(t)$:

$$k+\frac{w}{r}\left(1-\left(\frac{w}{c_2(t)}\right)^{\frac{(1-\beta)r}{\rho-r}}\right)-c_2(t)\frac{1-\beta}{\rho-\beta r}\left(1-\left(\frac{w}{c_2(t)}\right)^{\frac{\rho-\beta r}{\rho-r}}\right)=0,$$

которое посредством арифметических преобразований приводится к уравнению (6) относительно $\hat{c}=c_2(t)$. По лемме 3.4 уравнение (6) при положительном значении k имеет два корня: $c^1\in(0,w)$ и $c^2>w$.

Теперь мы покажем, что момент времени T_1 является единственным, причем оптимальным управлением является функция

$$c_{T,2}[s]=\begin{cases}\left(\frac{\beta}{p_{2t}}\right)^{\frac{1}{1-\beta}}e^{-\frac{\rho-r}{1-\beta}(s-t)},\ s\in[t,T_1],\\ w,\ s\in(T_1,T],\end{cases}$$

со значением $c=\left(\frac{\beta}{p_{2t}}\right)^{\frac{1}{1-\beta}}>w$, являющимся корнем уравнения (6). Для этого нам следует проверить выполнение условий леммы 3.1. Вычислим функционал:

$$J(c_{T,2}[\cdot])=e^{-\rho t}\left(c^\beta\int\limits_t^{T_1}e^{-\frac{\rho-\beta r}{1-\beta}(s-t)}\,ds+\int\limits_{T_1}^{T}e^{\rho(t-s)}w^\beta\,ds\right)=$$
$$=e^{-\rho t}\left(\frac{1-\beta}{\rho-\beta r}\left(1-e^{-\frac{\rho-\beta r}{1-\beta}(T_1-t)}\right)c^\beta+\frac{w^\beta}{\rho}\left(e^{-\rho(T_1-t)}-e^{-\rho(T-t)}\right)\right).$$

Здесь T_1 – такой момент времени, что $k_2(T_1)=0$, $c_2(T_1)=w$, следовательно, $ce^{-\frac{r-\rho}{1-\beta}(T_1-t)}=w$, и $T_1-t=$ $=\frac{1-\beta}{\rho-r}\ln\frac{c}{w}$. Тогда

$$J(c_{T,2}[\cdot])=e^{-\rho t}\left(\frac{1-\beta}{\rho-\beta r}\left(1-\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right)c^{\beta}+\frac{w^{\beta}}{\rho}\left(\left(\frac{w}{c}\right)^{\frac{(1-\beta)\rho}{\rho-r}}-e^{-\rho(T-t)}\right)\right)=$$
$$=e^{-\rho t}\left(\frac{1-\beta}{\rho-\beta r}\left(1-\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right)c^{\beta}+\frac{c^{\beta}}{\rho}\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right)=$$
$$=e^{-\rho t}\left(c^{\beta}\left[\frac{1-\beta}{\rho-\beta r}+\frac{\beta}{\rho}\frac{\rho-r}{\rho-\beta r}\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right]-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right).$$

С другой стороны, используя уравнение (6), удобно получить эквивалентное представление:

$$J(c_{T,2}[\cdot])=e^{-\rho t}\left(c^{\beta-1}\left[\frac{1-\beta}{\rho-\beta r}c+\frac{\beta}{\rho}\frac{\rho-r}{\rho-\beta r}c\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right]-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right)=$$
$$=e^{-\rho t}\left(c^{\beta-1}\left[\frac{1-\beta}{\rho-\beta r}c+\frac{\beta}{\rho}\frac{\rho-r}{\rho-\beta r}w\left(\frac{w}{c}\right)^{\frac{(1-\beta)r}{\rho-r}}\right]-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right)=$$
$$=e^{-\rho t}\left(c^{\beta-1}\left[\frac{1-\beta}{\rho-\beta r}c+\frac{\beta}{\rho}\left(rk+w-\frac{(1-\beta)r}{\rho-\beta r}c\right)\right]-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right)=$$
$$=e^{-\rho t}\left(c^{\beta-1}\left[\frac{1-\beta}{\rho}c+\frac{\beta}{\rho}(rk+w)\right]-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right).$$

Тогда положим $\mathcal{V}(t,k)=e^{\rho t}J(c_{T,2}[\cdot])$. Так как по построению функция $\mathcal{V}(t,k)$ является непрерывно дифференцируемой, мы можем найти ее частные производные. Для начала продифференцируем уравнение (6) по k:

$$\frac{\rho-\beta r}{1-\beta}=\frac{\partial c}{\partial k}\left[1-\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right].$$

Тогда получим

$$\frac{\partial\mathcal{V}}{\partial k}=\beta c^{\beta-1}\frac{\partial c}{\partial k}\left[\frac{1-\beta}{\rho-\beta r}+\frac{\beta}{\rho}\frac{\rho-r}{\rho-\beta r}\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right]-c^{\beta-1}\frac{\beta}{\rho}\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\frac{\partial c}{\partial k}=$$
$$=\beta c^{\beta-1}\frac{\partial c}{\partial k}\left[\frac{1-\beta}{\rho-\beta r}+\frac{1}{\rho}\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\left(\beta\frac{\rho-r}{\rho-\beta r}-1\right)\right]=$$
$$=\beta c^{\beta-1}\frac{\partial c}{\partial k}\frac{1-\beta}{\rho-\beta r}\left[1-\left(\frac{w}{c}\right)^{\frac{\rho-\beta r}{\rho-r}}\right]=\beta c^{\beta-1}.$$

Отметим, что

$$\frac{\partial\mathcal{V}}{\partial t}=-w^{\beta}e^{-\rho(T-t)}.$$

Вычисляя супремум в уравнении Гамильтона–Якоби–Беллмана, имеем, что $c=\left(\frac{1}{\beta}\frac{\partial\mathcal{V}(t,k)}{\partial k}\right)^{\frac{1}{\beta-1}}$. Тогда

$$\frac{\partial\mathcal{V}}{\partial t}-\rho\mathcal{V}=-w^{\beta}e^{-\rho(T-t)}-\rho\left(c^{\beta-1}\left[\frac{1-\beta}{\rho}c+\frac{\beta}{\rho}(rk+w)\right]-\frac{w^{\beta}}{\rho}e^{-\rho(T-t)}\right)=$$
$$=-c^{\beta-1}\left[(1-\beta)c+\beta(rk+w)\right],$$

$$\sup_{c\geq0}\left\{\frac{\partial\mathcal{V}}{\partial k}\left[rk+w-c\right]+c^{\beta}\right\}=\beta c^{\beta-1}\left[rk+w-c\right]+c^{\beta}=$$
$$=c^{\beta-1}\left[\beta(rk+w)+(1-\beta)c\right].$$

Складывая полученные выражения, получаем, что

$$\frac{\partial\mathcal{V}}{\partial t}-\rho\mathcal{V}+\sup_{c\geq0}\left\{\frac{\partial\mathcal{V}}{\partial k}\left[rk+w-c\right]+c^{\beta}\right\}=-c^{\beta-1}\left[(1-\beta)c+\beta(rk+w)\right]+$$
$$+c^{\beta-1}\left[\beta(rk+w)+(1-\beta)c\right]=0,$$

то есть функция $\mathcal{V}(t,k)$ является решением уравнения Гамильтона–Якоби–Беллмана, а значит, в силу леммы 3.1 функция $\mathcal{V}(t,k)$ при $\rho > r$ является искомой функцией цены, а соответствующее ей управление $\hat{c}(k,\rho)$ является оптимальным. Лемма доказана.

Доказательство теоремы 3.2. Искомое оптимальное управление получается в результате перехода к пределу для управления, полученного в лемме 3.6, при $T \to +\infty$.

Замечание 1. Использование леммы 3.1 оказывается корректным при $\rho > \beta r$. При $\rho \le \beta r$ значение интегрального функционала полезности расходится при $T \to +\infty$, а предельное управление равно нулю. В силу расходимости несобственного интеграла пользоваться леммой 3.1 в качестве достаточного условия оптимальности при $T = +\infty$, вообще говоря, неверно, несмотря на оптимальность на любом конечном отрезке времени $[t,T]$. Поэтому полученное при $\rho \le \beta r$ решение понимается в обобщенном смысле. Экономическая же интерпретация данного результата такова, что при увеличении горизонта времени экономический агент откладывает потребление своего капитала на конец временного промежутка.

4. МОДЕЛЬ СОЦИАЛЬНОЙ ДИНАМИКИ. ФОРМУЛИРОВКИ ОСНОВНЫХ РЕЗУЛЬТАТОВ

Используя полученные решения задач оптимального управления в форме синтеза, опишем модель социальной динамики популяции из H домашних хозяйств. В стартовый момент времени $t_0 = 0$ у каждого домашнего хозяйства с индексом $h \in \{1,\dots,H\}$ имеется в распоряжении сбережений в размере $k_0^h \ge 0$, причем суммарные сбережения всех домашних хозяйств положительны: $K_0 = \sum\limits_{h=1}^{H} k_0^h > 0$.

Положим $K = \sum\limits_{h=1}^{H} k^h$. Используя синтез оптимального управления, построенный в теоремах 3.1 и 3.2, будем описывать динамику капитала с помощью задачи Коши

$$\frac{dk^h}{dt} = rk^h + w - c\left(k^h, \rho^h\left(k^h, K\right)\right),\ k^h(0) = k_0^h,$$

где функция $\rho^h\left(k^h,K\right)$ будет определена ниже. Основываясь на гипотезе относительного дохода Дж. Дьюзенберри [15], положим

$$\rho^h\left(k^h,K\right) = r\varphi\left(\frac{k^h}{K}\right), \tag{10}$$

где функция $\varphi(\cdot)$ удовлетворяет следующим предположениям.

Предположение 1. Функция $\varphi : \mathbb{R} \to \mathbb{R}_+$ непрерывна по Липшицу, строго монотонно убывает на отрезке $[0,1]$, $\varphi(x) = \varphi(1), x \ge 1$, $\varphi(x) = \varphi(0), x \le 0$, $\beta \le \varphi(1) < \varphi\left(\frac{1}{H}\right) \le 1 < \varphi(0)$.

Предположение 2. Функция $\varphi(x)$ вогнута на отрезке $[0,1]$.

Определение 2. Моделью Рамсея–Бьюли социальной динамики называется задача Коши для системы обыкновенных дифференциальных уравнений первого порядка вида

$$\frac{dk^h}{dt} = \frac{1 - \varphi\left(\frac{k^h}{K}\right)}{1-\beta}\left[rk^h + w\right],\ k^h(0) = k_0^h,\ h = 1,\dots,H. \tag{11}$$

Сформулируем теперь основной результат для модели Рамсея–Бьюли.

Теорема 4.1 (о гипотезе Рамсея–Бьюли). *Пусть для модели Рамсея–Бьюли выполнены предположения 1,2. Тогда для любого вектора начальных значений* $\mathbf{k_0}$ *капиталов домашних хозяйств в модели Рамсея–Бьюли, принадлежащего множеству*

$$\mathcal{K}_0^l = \left\{\mathbf{k} \in \mathbb{R}_+^H : k^1 = \dots = k^l > k^{l+1} \ge \dots \ge k^H\right\},\ 1 \le l \le H-1,$$

выполняется

$$k^h[t] \xrightarrow[t\to+\infty]{} \begin{cases} +\infty, & h = 1,\dots,l, \\ -\dfrac{w}{r}, & h = l+1,\dots,H, \end{cases}$$

$$\frac{k^h[t]}{K[t]} \xrightarrow[t\to+\infty]{} \begin{cases} \dfrac{1}{l}, & h = 1,\dots,l, \\ 0, & h = l+1,\dots,H, \end{cases}$$

где вектор-функция $\mathbf{k}[t] = \left(k^1[t],\dots,k^H[t]\right) = \mathbf{k}(t,0;\mathbf{k_0})$ *является решением задачи Коши* (11).

Определение 3. Моделью Рамсея–Беккера социальной динамики называется задача Коши для системы обыкновенных дифференциальных уравнений первого порядка вида

$$\frac{dk^h}{dt} = \frac{1-\varphi\left(\frac{k^h}{K}\right)}{1-\beta}\left[rk^h + w\right] + \frac{\left(\varphi\left(\frac{k^h}{K}\right)-1\right)_+}{1-\beta} w \left(\frac{w}{c_2\left(k^h, r\varphi\left(\frac{k^h}{K}\right)\right)}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^h}{K}\right)-1}}, \tag{12}$$

$$k^h(0) = k_0^h,\ h = 1,\ldots,H,$$

где функция $c_2(k,\rho)$ является решением уравнения (6) при $k = k^h(t)$, $\rho = r\varphi\left(\frac{k^h(t)}{K(t)}\right)$.

Для модели Рамсея–Беккера справедлива аналогичная

Теорема 4.2 (о гипотезе Рамсея–Беккера). *Пусть для модели Рамсея–Беккера выполнено предположение 1. Тогда для любого вектора начальных значений* $\mathbf{k_0}$ *капиталов домашних хозяйств в модели Рамсея–Беккера, принадлежащего множеству*

$$\mathcal{K}_0^l = \left\{\mathbf{k} \in \mathbb{R}_+^H : k^1 = \ldots = k^l > k^{l+1} \geq \ldots \geq k^H\right\},\ 1 \leq l \leq H-1,$$

выполняется

$$k^h[t] \xrightarrow[t\to+\infty]{} \begin{cases} +\infty, & h = 1,\ldots,l, \\ 0, & h = l+1,\ldots,H, \end{cases}$$

$$\frac{k^h[t]}{K[t]} \xrightarrow[t\to+\infty]{} \begin{cases} \dfrac{1}{l}, & h = 1,\ldots,l, \\ 0, & h = l+1,\ldots,H, \end{cases}$$

где вектор-функция $\mathbf{k}[t] = \left(k^1[t],\ldots,k^H[t]\right) = \mathbf{k}(t,0;\mathbf{k_0})$ *является решением задачи Коши* (12).

Замечание 2. Заметим, что можно выразить решение задачи Коши (11) в следующей форме:

$$rk^h(t) + w = \left[rk_0^h + w\right]\exp\left\{\frac{r}{1-\beta}\int\limits_0^t\left(1-\varphi\left(\frac{k^h(\tau)}{K(\tau)}\right)\right)d\tau\right\}. \tag{13}$$

Из данного представления следует, что $k^h(t) > -\frac{w}{r}$, $t \in [0,+\infty)$, $h = \overline{1,H}$.

Замечание 3. Из теоремы о единственности решения задачи Коши следует, что все неравенства, задаваемые начальными значениями капиталов для домашних хозяйств, сохраняются, т.е. если $k_0^i > k_0^j$, то $k^i(t) > k^j(t)$, $t \in [0,+\infty)$, $1 \leq i < j \leq H$. Тогда для любого $t \in [0,+\infty)$ выполняется неравенство $\frac{k^1(t)}{K(t)} > \frac{1}{H}$. Справедливость этого факта легко доказать от противного. Пусть $\frac{k^1(t)}{K(t)} \leq \frac{1}{H}$, тогда $\frac{k^h(t)}{K(t)} < \frac{1}{H}$, $h = l+1,\ldots,H$. Просуммировав все слагаемые $\frac{k^h(t)}{K(t)}$ по h от 1 до H, получаем, что $1 = \sum\limits_{h=1}^{H}\frac{k^h(t)}{K(t)} < \sum\limits_{h=1}^{H}\frac{1}{H} = 1$. Следовательно, $\varphi\left(\frac{k^1(t)}{K(t)}\right) < \varphi\left(\frac{1}{H}\right) \leq 1$.

5. ДОКАЗАТЕЛЬСТВА ГИПОТЕЗ РАМСЕЯ–БЬЮЛИ И РАМСЕЯ–БЕККЕРА

Нам понадобится доказать несколько вспомогательных утверждений.

Лемма 5.1. *В модели Рамсея–Бьюли на решении задачи Коши* (11) *с положительными начальными условиями справедливо неравенство* $K(t) = \sum\limits_{h=1}^{H} k^h(t) > 0$ *для любого* $t \in [0,+\infty)$.

Доказательство. Из (11) имеем, что

$$\dot{K} = \sum_{h=1}^{H}\frac{1-\varphi\left(\frac{k^h}{K}\right)}{1-\beta}\left[rk^h + w\right] = \frac{rK + wH}{1-\beta}\left(1 - \sum_{h=1}^{H}\varphi\left(\frac{k^h}{K}\right)\frac{rk^h + w}{rK + wH}\right).$$

В силу замечания 2 коэффициенты при $\varphi\left(\frac{k^h}{K}\right)$ неотрицательны. Тогда рассмотрим следующее выражение из внутренней скобки:

$$\sum_{h=1}^{H}\varphi\left(\frac{k^h}{K}\right)\frac{rk^h+w}{rK+wH}=\sum_{h:\,\frac{k^h}{K}<0}\varphi(0)\,\frac{rk^h+w}{rK+wH}+\sum_{h:0\le\frac{k^h}{K}\le1}\varphi\left(\frac{k^h}{K}\right)\frac{rk^h+w}{rK+wH}+$$

$$+\sum_{h:\,\frac{k^h}{K}>1}\varphi(1)\,\frac{rk^h+w}{rK+wH}\le\varphi\left(\sum_{h:0\le\frac{k^h}{K}\le1}\frac{k^h}{K}\frac{rk^h+w}{rK+wH}+\sum_{h:\,\frac{k^h}{K}>1}\frac{rk^h+w}{rK+wH}\right).$$

Покажем, что

$$\sum_{h:0\le\frac{k^h}{K}\le1}\frac{k^h}{K}\frac{rk^h+w}{rK+wH}+\sum_{h:\,\frac{k^h}{K}>1}\frac{rk^h+w}{rK+wH}\ge\frac{1}{H}.$$

Обозначив $N_1=\left|\left\{h:\frac{k^h}{K}>1\right\}\right|$, $N_2=\left|\left\{h:0\le\frac{k^h}{K}\le1\right\}\right|$, рассмотрим два случая:

- $N_1>0$. В этом случае достаточно оценить первое слагаемое снизу нулем, тогда

$$\sum_{h:\,\frac{k^h}{K}>1}\frac{rk^h+w}{rK+wH}=\sum_{h:\,\frac{k^h}{K}>1}\frac{\frac{k^h}{K}+\frac{w}{r}\frac{1}{K}}{1+\frac{w}{r}\frac{H}{K}}>N_1\frac{1+\frac{w}{r}\frac{1}{K}}{1+\frac{w}{r}\frac{H}{K}}>\frac{N_1}{H}\ge\frac{1}{H}.$$

- $N_1=0$. Тогда очевидно, что $\sum\limits_{h:0\le\frac{k^h}{K}\le1}\frac{k^h}{K}\ge1$, вследствие чего имеем

$$\sum_{h:0\le\frac{k^h}{K}\le1}\frac{k^h}{K}\frac{rk^h+w}{rK+wH}=\frac{1}{1+\frac{w}{r}\frac{H}{K}}\sum_{h:0\le\frac{k^h}{K}\le1}\left[\left(\frac{k^h}{K}\right)^2+\frac{w}{r}\frac{1}{K}\frac{k^h}{K}\right]\ge$$

$$\ge\left\{H\sum_{h:0\le\frac{k^h}{K}\le1}\left(\frac{k^h}{K}\right)^2\ge\sum_{h:0\le\frac{k^h}{K}\le1}\frac{k^h}{K}\ge1\right\}\ge\frac{1}{1+\frac{w}{r}\frac{H}{K}}\left[\frac{1}{H}+\frac{w}{r}\frac{1}{K}\right]=\frac{1}{H}.$$

Следовательно,

$$\sum_{h=1}^{H}\varphi\left(\frac{k^h}{K}\right)\frac{rk^h+w}{rK+wH}\le\varphi\left(\frac{1}{H}\right)\le1,$$

и мы получаем неотрицательность производной суммарного капитала K, откуда при положительности начального значения K_0 следует справедливость требуемого неравенства.

Замечание 4. В случае задачи Коши (12) для любого $t\in[0,+\infty)$, $h\in\{1,\ldots,H\}$ выполняются неравенства $k^h(t)\ge0$, $K(t)>0$. Отсюда также следует, что для любых $t\in[0,+\infty)$, $h=\overline{1,H}$ выполняется двойное неравенство $0\le\frac{k^h(t)}{K(t)}\le1$. Данное обстоятельство и положительность капитала, как будет выяснено позже, позволят нам провести доказательство теоремы 4.2 при более слабых условиях, нежели в случае с теоремой 4.1.

Лемма 5.2. *Для каждого $j=\overline{l+1,H}$ существует константа $M_j>0$ такая, что для любого $t\in[0,+\infty)$ выполняется неравенство*

$$\varphi\left(\frac{k^j(t)}{K(t)}\right)-\varphi\left(\frac{k^1(t)}{K(t)}\right)>M_j.$$

Доказательство. Пусть $1\le i\le l<j\le H$, тогда из (13) получаем, что

$$\frac{rk^j(t)+w}{rk^i(t)+w}=\frac{rk_0^j+w}{rk_0^i+w}\exp\left\{-\frac{r}{1-\beta}\int_0^t\left(\varphi\left(\frac{k^j(\tau)}{K(\tau)}\right)-\varphi\left(\frac{k^i(\tau)}{K(\tau)}\right)\right)d\tau\right\}.$$

В силу замечания 3 подынтегральное выражение неотрицательно, поэтому имеет место неравенство

$$\frac{rk^j(t)+w}{rk^i(t)+w} \le \frac{rk_0^j+w}{rk_0^i+w} = \frac{k_0^j+\frac{w}{r}}{k_0^i+\frac{w}{r}}.$$

Домножив обе части неравенства на множитель $\frac{rk^i(t)+w}{rK(t)}$, получаем неравенство

$$\frac{k^j(t)}{K(t)} + \frac{w}{r}\frac{1}{K(t)} \le \frac{k_0^j+\frac{w}{r}}{k_0^i+\frac{w}{r}}\left(\frac{k^i(t)}{K(t)} + \frac{w}{r}\frac{1}{K(t)}\right).$$

В свою очередь, домножив обе части полученного неравенства на -1 и прибавив слагаемое $\frac{k^i(t)}{K(t)} + \frac{w}{r}\frac{1}{K(t)}$, получаем неравенство

$$\frac{k^i(t)}{K(t)} - \frac{k^j(t)}{K(t)} \ge \frac{k_0^i-k_0^j}{k_0^i+\frac{w}{r}}\left(\frac{k^i(t)}{K(t)} + \frac{w}{r}\frac{1}{K(t)}\right).$$

Наконец, положив $i=1$ и пользуясь тем, что по замечанию 3 имеем $\frac{k^1(t)}{K(t)} > \frac{1}{H}$, а по лемме 5.1 $\frac{w}{r}\frac{1}{K(t)} > 0$, выводим оценку

$$\frac{k^1(t)}{K(t)} - \frac{k^j(t)}{K(t)} > \frac{k_0^1-k_0^j}{k_0^1+\frac{w}{r}}\frac{1}{H}, j=\overline{l+1,H}.$$

Тогда для каждого $j=\overline{l+1,H}$ имеем

$$\varphi\left(\frac{k^j(t)}{K(t)}\right) - \varphi\left(\frac{k^1(t)}{K(t)}\right) > \varphi\left(\frac{k^1(t)}{K(t)} - \frac{k_0^1-k_0^j}{k_0^1+\frac{w}{r}}\frac{1}{H}\right) - \varphi\left(\frac{k^1(t)}{K(t)}\right) \ge$$

$$\ge \min_{x\in\left[\frac{1}{H},1\right]}\left[\varphi\left(x - \frac{k_0^1-k_0^j}{k_0^1+\frac{w}{r}}\frac{1}{H}\right) - \varphi(x)\right] = M_j > 0.$$

Полученная константа M_j является положительной, так как в противном случае найдется точка $\tilde{x} \in \left[\frac{1}{H},1\right]$, такая что $M_j = \varphi\left(\tilde{x} - \frac{k_0^1-k_0^j}{k_0^1+\frac{w}{r}}\frac{1}{H}\right) - \varphi(\tilde{x}) \le 0$, откуда в силу строгого убывания функции $\varphi(\cdot)$ следует неравенство $k_0^1 \le k_0^j$, что противоречит условию $k_0^1 > k_0^j$. Полученное противоречие доказывает лемму.

Рассмотрим теперь задачу Коши (12). При $1 \le i \le l < j \le H$ в силу замечания 3 имеем $\frac{k^i(t)}{K(t)} > \frac{1}{H}$, поэтому $\varphi\left(\frac{k^i}{K}\right) < 1$, и для $k^i(t)$ справедлива формула (13). Рассмотрим теперь уравнение при $h=j$. Так как $c\left(k^h, r\varphi\left(\frac{k^h}{K}\right)\right)$ устроено как решение уравнения (6), причем $c\left(k^h, r\varphi\left(\frac{k^h}{K}\right)\right) \ge w$, то имеем следующую оценку:

$$\dot{k}^j(t) \le \frac{1-\varphi\left(\frac{k^j(t)}{K(t)}\right)}{1-\beta} rk^j(t) + \frac{\left(1-\varphi\left(\frac{k^j(t)}{K(t)}\right)\right)_+}{1-\beta} w \le$$

$$\le \left\{1-\varphi\left(\frac{k^j(t)}{K(t)}\right) \le \left(1-\varphi\left(\frac{k^j(t)}{K(t)}\right)\right)_+\right\} \le \frac{\left(1-\varphi\left(\frac{k^j(t)}{K(t)}\right)\right)_+}{1-\beta}\left[rk^j(t)+w\right].$$

Следовательно, по лемме Гронуолла–Беллмана справедливо неравенство

$$rk^j(t)+w \le \left[rk_0^j+w\right]\exp\left\{\frac{r}{1-\beta}\int_0^t\left(1-\varphi\left(\frac{k^j(\tau)}{K(\tau)}\right)\right)_+ d\tau\right\},$$

откуда, используя (13) при $h=i$, получаем неравенство

$$\frac{rk^j(t)+w}{rk^i(t)+w} \le$$

$$\le \frac{rk_0^j+w}{rk_0^i+w}\exp\left\{-\frac{r}{1-\beta}\int_0^t \min\left\{\left(\varphi\left(\frac{k^j(\tau)}{K(\tau)}\right) - \varphi\left(\frac{k^i(\tau)}{K(\tau)}\right), 1-\varphi\left(\frac{k^i(\tau)}{K(\tau)}\right)\right)\right\} d\tau\right\}.$$

В силу замечания 3 подынтегральное выражение неотрицательно, поэтому дальнейшие рассуждения в точности повторяют те, что были приведены при рассмотрении задачи Коши (11).

Напомним формулировку основной теоремы о дифференциальных неравенствах.

Теорема (Чаплыгина о дифференциальных неравенствах, см. [27]). *Если при $t \in [t_0, t_1]$ существует решение задачи Коши $\dot{x} = f(t,x)$, $x(t_0) = x_0$, являющееся однозначной функцией t, и если функция $x_u(t) \in C^1[t_0,t_1]$ такова, что $\dot{x}_u > f(t,x_u)$, $x_u(t_0) \geq x_0$, то имеет место неравенство $x(t) < x_u(t)$ $\forall\, t \in (t_0,t_1]$.*

Введем обозначение $K_{-l}(t) = \sum\limits_{h=l+1}^{H} k^h(t)$.

Лемма 5.3. *Для любого $t \in (0,+\infty)$ справедлива оценка*

$$\frac{K_{-l}(t)}{K(t)} \leq R(t), \tag{14}$$

где

$$R(t) = 1 - \left(\frac{1}{1 - \frac{K_{-l,0}}{K_0}} \exp\left\{ -\int\limits_0^t (a - b(\tau))\, d\tau \right\} + a \int\limits_0^t \exp\left\{ \int\limits_t^\tau (a - b(s))\, ds \right\} d\tau \right)^{-1}, \tag{15}$$

$$K_{-l,0} = \sum_{h=l+1}^{H} k_0^h, \quad K_0 = l k_0^1 + K_{-l,0}, \quad a = \frac{M_{l+1} r}{1-\beta},$$

$$b(t) = \frac{w}{(1-\beta) K(t)} \left((H-l) \left(1 - \varphi\left(\frac{k^1(t)}{K(t)} \right) - M_{l+1} \right)_+ + l \left(1 - \varphi\left(\frac{k^1(t)}{K(t)} \right) \right) \right).$$

Доказательство. Заметим, что так как по лемме 5.2 для каждого $h = \overline{l+1, H}$ существует константа $M_h > 0$ такая, что $\varphi\left(\frac{k^h}{K}\right) - \varphi\left(\frac{k^1}{K}\right) > M_h$, а также $\frac{k^{l+1}}{K} \geq \frac{k^h}{K}$, то для любого $h = \overline{l+1, H}$ справедливо также и неравенство $\varphi\left(\frac{k^h}{K}\right) - \varphi\left(\frac{k^1}{K}\right) > M_{l+1}$. Добавляя и вычитая единицу в левой части полученного неравенства, получаем $1 - \varphi\left(\frac{k^h}{K}\right) < 1 - \varphi\left(\frac{k^1}{K}\right) - M_{l+1}$.

Вычислим производную

$$\frac{d}{dt}\left(\frac{K_{-l}}{K}\right) = \frac{l}{K^2} \sum_{h=l+1}^{H} \left(\frac{dk^h}{dt} k^1 - k^h \frac{dk^1}{dt} \right).$$

В случае задачи Коши (11) имеем

$$\begin{aligned}
\frac{dk^h}{dt} k^1 - k^h \frac{dk^1}{dt} &= \frac{1 - \varphi\left(\frac{k^h}{K}\right)}{1-\beta} \left(r k^h + w\right) k^1 - \frac{1 - \varphi\left(\frac{k^1}{K}\right)}{1-\beta} \left(r k^1 + w\right) k^h < \\
&< \frac{1 - \varphi\left(\frac{k^1}{K}\right) - M_{l+1}}{1-\beta} \left(r k^h + w\right) k^1 - \frac{1 - \varphi\left(\frac{k^1}{K}\right)}{1-\beta} \left(r k^1 + w\right) k^h = \\
&= -\frac{M_{l+1} r}{1-\beta} k^1 k^h + w \left(\frac{1 - \varphi\left(\frac{k^1}{K}\right) - M_{l+1}}{1-\beta} k^1 - \frac{1 - \varphi\left(\frac{k^1}{K}\right)}{1-\beta} k^h \right),
\end{aligned}$$

откуда следует, что

$$\begin{aligned}
\frac{d}{dt}\left(\frac{K_{-l}}{K}\right) &< \frac{l}{K^2} \sum_{h=l+1}^{H} \left[-\frac{M_{l+1} r}{1-\beta} k^1 k^h + w \left(\frac{1 - \varphi\left(\frac{k^1}{K}\right) - M_{l+1}}{1-\beta} k^1 - \frac{1 - \varphi\left(\frac{k^1}{K}\right)}{1-\beta} k^h \right) \right] = \\
&= -\frac{M_{l+1} r}{1-\beta} \frac{l k^1}{K} \frac{K_{-l}}{K} + \frac{w}{K} \left((H-l) \frac{1 - \varphi\left(\frac{k^1}{K}\right) - M_{l+1}}{1-\beta} \frac{l k^1}{K} - l \frac{1 - \varphi\left(\frac{k^1}{K}\right)}{1-\beta} \frac{K_{-l}}{K} \right).
\end{aligned}$$

В случае задачи Коши (12) имеем

$$\frac{dk^h}{dt}k^1 - k^h\frac{dk^1}{dt} = \frac{1-\varphi\left(\frac{k^h}{K}\right)}{1-\beta}\left(rk^h+w\right)k^1 + \frac{\left(\varphi\left(\frac{k^h}{K}\right)-1\right)_+}{1-\beta}w\left(\frac{w}{c^h}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^h}{K}\right)-1}}k^1 -$$
$$-\frac{1-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}\left(rk^1+w\right)k^h \le \frac{1-\varphi\left(\frac{k^h}{K}\right)}{1-\beta}\left(rk^h+w\right)k^1 + \frac{\left(\varphi\left(\frac{k^h}{K}\right)-1\right)_+}{1-\beta}wk^1 -$$
$$-\frac{1-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}\left(rk^1+w\right)k^h = -\frac{\varphi\left(\frac{k^h}{K}\right)-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}rk^1k^h + \frac{\left(1-\varphi\left(\frac{k^h}{K}\right)\right)_+}{1-\beta}wk^1 -$$
$$-\frac{1-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}wk^h < -\frac{M_{l+1}r}{1-\beta}k^1k^h + w\left(\frac{\left(1-\varphi\left(\frac{k^1}{K}\right)-M_{l+1}\right)_+}{1-\beta}k^1 - \frac{1-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}k^h\right),$$

откуда получаем, что

$$\frac{d}{dt}\left(\frac{K_{-l}}{K}\right) < \frac{l}{K^2}\sum_{h=l+1}^{H}\left[-\frac{M_{l+1}r}{1-\beta}k^1k^h + w\left(\frac{\left(1-\varphi\left(\frac{k^1}{K}\right)-M_{l+1}\right)_+}{1-\beta}k^1 - \frac{1-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}k^h\right)\right] =$$
$$= -\frac{M_{l+1}r}{1-\beta}\frac{lk^1}{K}\frac{K_{-l}}{K} + \frac{w}{K}\left((H-l)\frac{\left(1-\varphi\left(\frac{k^1}{K}\right)-M_{l+1}\right)_+}{1-\beta}\frac{lk^1}{K} - l\frac{1-\varphi\left(\frac{k^1}{K}\right)}{1-\beta}\frac{K_{-l}}{K}\right).$$

Обратим внимание, что полученная оценка на производную в случае задачи Коши (12) отличается от оценки при (11) лишь вторым слагаемым. Применяя неравенство $1-\varphi\left(\frac{k^1}{K}\right)-M_{l+1} \le \left(1-\varphi\left(\frac{k^1}{K}\right)-M_{l+1}\right)_+$, считаем, что при рассмотрении обеих задач Коши выполняется одна и та же оценка.

Далее, заметив, что $\frac{lk^1}{K} = 1 - \frac{K_{-l}}{K}$, сгруппировав слагаемые и введя для удобства следующие обозначения:

$$x(t) = \frac{K_{-l}(t)}{K(t)}, \quad c(t) = \frac{w}{(1-\beta)K(t)}(H-l)\left(1-\varphi\left(\frac{k^1(t)}{K(t)}\right)-M_{l+1}\right)_+,$$

получаем дифференциальное неравенство

$$\dot{x}(t) \le a(x(t))^2 - (a+b(t))x(t) + c(t).$$

Правую часть неравенства можно преобразовать следующим образом:

$$a(x(t))^2 - (a+b(t))x(t) + c(t) = a(x(t)-1)^2 + (a-b(t))(x(t)-1) - b(t) + c(t).$$

Обратим внимание, что функция

$$b(t) - c(t) = l\frac{w}{K(t)}\frac{1-\varphi\left(\frac{k^1(t)}{K(t)}\right)}{1-\beta}$$

неотрицательна, в силу чего мы можем перейти к более грубому дифференциальному неравенству относительно функции $q(t) = x(t) - 1$, а именно $\dot{q}(t) < a(q(t))^2 + (a-b(t))q(t)$. По теореме Чаплыгина решением является функция, которая удовлетворяет дифференциальному уравнению Бернулли $\dot{y} = a(y(t))^2 + (a-b(t))y(t)$, сводимому заменой $z(t) = \frac{1}{y(t)}$ к линейному дифференциальному уравнению $\dot{z}(t) = -(a-b(t))z(t) - a$. Решение последнего, в свою очередь, определяется функцией

$$z(t) = z_0\exp\left\{-\int_0^t(a-b(\tau))\,d\tau\right\} - a\int_0^t\exp\left\{\int_t^\tau(a-b(s))\,ds\right\}d\tau.$$

При $t=0$ значение константы z_0 определяется выражением $z_0 = \frac{1}{y(0)} = -\frac{1}{1-\frac{K_{-l,0}}{K_0}}$, тогда решение дифференциального уравнения Бернулли имеет вид

$$y(t) = -\frac{1}{\frac{1}{1-\frac{K_{-l,0}}{K_0}}\exp\left\{-\int_0^t(a-b(\tau))\,d\tau\right\} + a\int_0^t\exp\left\{\int_t^\tau(a-b(s))\,ds\right\}d\tau} = R(t) - 1.$$

Следовательно, по теореме Чаплыгина для функции $q(t) = \frac{K_{-l}(t)}{K(t)} - 1$ справедлива оценка (14), что и требовалось.

Лемма 5.4. *Справедливы соотношения*

$$k^h(t) \xrightarrow[t\to+\infty]{} +\infty,\ h = \overline{1,l},\ K(t) \xrightarrow[t\to+\infty]{} +\infty,\ R(t) \xrightarrow[t\to+\infty]{} 0.$$

Доказательство. В силу замечания 2 и леммы 5.1 можно вывести неравенство

$$(rk^1 + w)^{\cdot} \geq r\frac{1-\varphi\left(\frac{1}{H}\right)}{1-\beta}\left(rk^1 + w\right) > 0,$$

откуда по лемме Гронуолла–Беллмана следует, что при $\varphi\left(\frac{1}{H}\right) < 1$ получим

$$rk^1(t) + w \geq \left(rk_0^1 + w\right)\exp\left\{r\frac{1-\varphi\left(\frac{1}{H}\right)}{1-\beta}t\right\} \xrightarrow[t\to+\infty]{} +\infty.$$

Следовательно, $k^1(t) \xrightarrow[t\to+\infty]{} +\infty$, а поскольку в случае задачи Коши (11) справедливо неравенство $K(t) \geq lk^1(t) - \frac{w}{r}(H-l)$ (задачи Коши (12) – $K(t) \geq lk^1(t)$), то и $K(t) \xrightarrow[t\to+\infty]{} +\infty$. Случай $\varphi\left(\frac{1}{H}\right) = 1$ повторяет предыдущие рассуждения с точностью до замены значения $\varphi\left(\frac{1}{H}\right)$ на $\varphi\left(\frac{1}{H}+\gamma\right)$, где $\gamma > 0$ – такая константа, что $\frac{k^1}{K} \geq \frac{1}{H} + \gamma, t \in [0, +\infty)$. Константа γ положительна, так как в противном случае вариант $\frac{k^1}{K} \xrightarrow[t\to+\infty]{} \frac{1}{H}$ возможен тогда и только тогда, когда $\frac{k^j}{K} \xrightarrow[t\to+\infty]{} \frac{1}{H}, j = \overline{2,H}$, т.е. $\frac{k^1}{K} - \frac{k^j}{K} \xrightarrow[t\to+\infty]{} 0$, что неверно в силу доказанной в лемме 5.2 оценки $\frac{k^1}{K} - \frac{k^j}{K} > \frac{k_0^1 - k_0^j}{k_0^1 + \frac{w}{r}}\frac{1}{H}$.

Заметим, что функция $1-\varphi\left(\frac{k^1(t)}{K(t)}\right)$ является ограниченной как сверху, так и снизу. Кроме того, $\lim\limits_{t\to+\infty} K(t) = +\infty$, поэтому $b(t) \xrightarrow[t\to+\infty]{} 0$. Следовательно, $a - b(t) \xrightarrow[t\to+\infty]{} a$, и отсюда следует, что найдется момент времени $\tau^* \in (0, +\infty)$ такой, что для любого $t > \tau^*$ выполняется неравенство $a - b(t) > \frac{a}{2}$. Тогда при $t > \tau^*$ имеем

$$\int_0^t (a - b(\tau))\, d\tau = \int_0^{\tau^*} (a - b(\tau))\, d\tau + \int_{\tau^*}^t (a - b(\tau))\, d\tau \geq \int_0^{\tau^*} (a - b(\tau))\, d\tau + \frac{a}{2}(t - \tau^*) \xrightarrow[t\to+\infty]{} +\infty,$$

следовательно,

$$\exp\left\{-\int_0^t (a - b(\tau))\, d\tau\right\} \xrightarrow[t\to+\infty]{} 0.$$

Исследуем теперь поведение функции $\int\limits_0^t \exp\left\{\int\limits_t^\tau (a - b(s))\, ds\right\} d\tau$ при $t \to +\infty$. Полагая $t > \tau^*$, имеем

$$\int_0^t \exp\left\{\int_0^\tau (a - b(s))\, ds\right\} d\tau = \overbrace{\int_0^{\tau^*} \exp\left\{\int_0^\tau (a - b(s))\, ds\right\} d\tau}^{A_1} +$$

$$+ \underbrace{\exp\left\{\int_0^{\tau^*} (a - b(s))\, ds\right\}}_{A_2} \int_{\tau^*}^t \exp\left\{\int_{\tau^*}^\tau (a - b(s))\, ds\right\} d\tau > A_1 + A_2 \int_{\tau^*}^t \exp\left\{\frac{a}{2}(\tau - \tau^*)\right\} d\tau =$$

$$= A_1 + \frac{2A_2}{a}\left(\exp\left\{\frac{a}{2}(t - \tau^*)\right\} - 1\right) \xrightarrow[t\to+\infty]{} +\infty.$$

Выражение $\int\limits_0^t \exp\left\{\int\limits_t^\tau (a - b(s))\, ds\right\} d\tau$ можно представить в виде дроби, числитель и знаменатель которой являются положительными и дифференцируемыми по t функциями, стремящимися к плюс бесконечности. Тогда,

применяя правило Лопиталя, получаем, что

$$\lim_{t\to+\infty}\frac{\int\limits_0^t \exp\left\{\int\limits_0^\tau (a-b(s))\,ds\right\}d\tau}{\exp\left\{\int\limits_0^t (a-b(s))\,ds\right\}}=\lim_{t\to+\infty}\frac{\exp\left\{\int\limits_0^t (a-b(\tau))\,d\tau\right\}}{(a-b(t))\exp\left\{\int\limits_0^t (a-b(\tau))\,d\tau\right\}}=$$
$$=\lim_{t\to+\infty}\frac{1}{a-b(t)}=\frac{1}{a}.$$

Но тогда

$$R(t)\xrightarrow[t\to+\infty]{}1-\left(\frac{1}{1-\frac{K_{-l,0}}{K_0}}0+a\frac{1}{a}\right)^{-1}=1-1=0,$$

что и требовалось доказать.

Теперь мы готовы доказать гипотезы Рамсея–Бьюли и Рамсея–Беккера.

Доказательство теоремы 4.1. Так как по лемме 5.4 мы имеем $k^h(t)\xrightarrow[t\to+\infty]{}+\infty$, $h=\overline{1,l}$, нам нужно показать,

$$k^h(t)\xrightarrow[t\to+\infty]{}-\frac{w}{r},\quad h=\overline{l+1,H}.$$

В силу леммы 5.1 суммарный капитал домашних хозяйств $K(t)$ положителен для любого $t\in[0,+\infty)$, поэтому по лемме 5.3 и замечанию 2 справедливо двойное неравенство

$$-\frac{w}{r}\frac{H-l}{K(t)}\le\frac{K_{-l}(t)}{K(t)}\le R(t).$$

Левая и правая части этого неравенства по лемме 5.4 сходятся к нулю, откуда следует, что $\frac{K_{-l}(t)}{K(t)}\xrightarrow[t\to+\infty]{}0$. Теперь докажем, что $\frac{k^h(t)}{K(t)}\xrightarrow[t\to+\infty]{}0, h=\overline{l+1,H}$. Используя двойное неравенство

$$-\frac{w}{r}\frac{1}{K(t)}\le\frac{k^H(t)}{K(t)}\le\frac{1}{H-l}\frac{K_{-l}(t)}{K(t)},$$

получаем сходимость к нулю для $\frac{k^H}{K}$ и, следовательно, для суммы $\sum\limits_{j=l+1}^{H-1}\frac{k^j}{K}$. Далее повторяем аналогичные рассуждения, применяя двойные неравенства вида

$$-\frac{w}{r}\frac{1}{K(t)}\le\frac{k^h(t)}{K(t)}\le\frac{1}{h-l}\sum_{j=l+1}^{h}\frac{k^j(t)}{K(t)},\quad h=H-1,\dots,l+2.$$

Следовательно, $\frac{k^h(t)}{K(t)}\xrightarrow[t\to+\infty]{}\frac{1}{l}$, $h=\overline{1,l}$.

Далее, так как $\frac{k^h(t)}{K(t)}\xrightarrow[t\to+\infty]{}0$, $h=\overline{l+1,H}$, то в силу непрерывности функции $\varphi(\cdot)$ имеем $\varphi\left(\frac{k^h(t)}{K(t)}\right)\xrightarrow[t\to+\infty]{}\xrightarrow[t\to+\infty]{}\varphi(0)>1$. Следовательно, найдется момент времени $\tau^{**}\in[0,+\infty)$ такой, что при $t>\tau^{**}$ выполняется неравенство $1-\varphi\left(\frac{k^h(t)}{K(t)}\right)<-\frac{\varphi(0)-1}{2}$.

Тогда, пользуясь представлением (13), имеем

$$0<rk^h(t)+w\le\left[rk_0^h+w\right]\exp\left\{\frac{r}{1-\beta}\int\limits_0^{\tau^{**}}\left(1-\varphi\left(\frac{k^h(\tau)}{K(\tau)}\right)\right)d\tau\right\}\times$$
$$\times\exp\left\{-\frac{r}{1-\beta}\frac{\varphi(0)-1}{2}(t-\tau^{**})\right\}\xrightarrow[t\to+\infty]{}0,$$

откуда следует, что $k^h(t)\xrightarrow[t\to+\infty]{}-\frac{w}{r}$.

Теорема 4.1 доказана.

Доказательство теоремы 4.2. Так как по лемме 5.4 мы имеем $k^h(t) \xrightarrow[t\to+\infty]{} +\infty$, $h = \overline{1,l}$, нам нужно показать, $k^h(t) \xrightarrow[t\to+\infty]{} 0$, $h = \overline{l+1,H}$. Дальнейшие рассуждения аналогичны рассуждениям доказательства теоремы 4.1 с той лишь разницей, что в силу леммы 5.3 и замечания 4 справедливо двойное неравенство $0 \le \frac{K_{-l}(t)}{K(t)} \le R(t)$, а сходимость $\frac{k^h(t)}{K(t)} \xrightarrow[t\to+\infty]{} 0$, $h = \overline{l+1,H}$, следует из двойного неравенства $0 \le \frac{k^h(t)}{K(t)} \le \frac{K_{-l}(t)}{K(t)}$. Следовательно, $\frac{k^h(t)}{K(t)} \xrightarrow[t\to+\infty]{} \frac{1}{l}$, $h = \overline{1,l}$.

Так как $\frac{k^h(t)}{K(t)} \xrightarrow[t\to+\infty]{} 0$, $h = \overline{l+1,H}$, то по непрерывности функции $\varphi(\cdot)$ имеем $\varphi\left(\frac{k^h(t)}{K(t)}\right) \xrightarrow[t\to+\infty]{} \varphi(0) > 1$. Следовательно, найдется момент времени $\tau^{**} \in [0,+\infty)$ такой, что при $t > \tau^{**}$ выполняется неравенство $1 - \varphi\left(\frac{k^h(t)}{K(t)}\right) < -\frac{\varphi(0)-1}{2}$. Но тогда при $t > \tau^{**}$ справедливо дифференциальное неравенство

$$\frac{dk^h}{dt} < -\frac{\varphi(0)-1}{2}\frac{r}{1-\beta}k^h,$$

из которого по лемме Гронуолла–Беллмана следует, что

$$0 \le k^h(t) \le k^h(\tau^{**})\exp\left\{-\frac{r}{1-\beta}\frac{\varphi(0)-1}{2}(t-\tau^{**})\right\} \xrightarrow[t\to+\infty]{} 0,$$

откуда получаем, что $k^h(t) \xrightarrow[t\to+\infty]{} 0$.

Теорема 4.2 доказана.

6. ИНДЕКС НЕРАВЕНСТВА ДЖИНИ И КРИВАЯ ЛОРЕНЦА

Проблемы расслоения населения по доходам обсуждаются в экономической литературе с начала XX века. В 1905 г. американский статистик М. Лоренц разработал подход к измерению неравенства в распределении доходов между домашними хозяйствами, который позволял анализировать изменение неравенства в динамике и сравнивать уровень неравенства в разных экономических сообществах. Для этого измерение неравенства не должно зависеть от абсолютных величин доходов, поскольку в разных экономических сообществах доходы могут измеряться в различных денежных единицах и в разные моменты времени покупательная способность денег также различна. Кроме того, неравенство должно определяться распределением доходов между домашними хозяйствами и не изменяться при перестановке доходов между ними. Следуя М. Лоренцу, рассмотрим сообщество из H домашних хозяйств. Обозначим через y_i доход i-го домашнего хозяйства. Вектор $y = (y_1, \ldots, y_H)$ задает распределение совокупного дохода $\sum\limits_{i=1}^{H} y_i$ между домашними хозяйствами. Поскольку измерение неравенства не зависит от перестановки доходов, т.е. компонент вектора y, построим перестановку σ, такую, что $y_{\sigma(1)} \le \ldots \le y_{\sigma(H)}$. Рассмотрим на плоскости точки с координатами $\left(\frac{m}{H}, u_m\right)$, $m = 1, \ldots, H$, где

$$u_m = \sum_{i=1}^{m} y_{\sigma(i)} \Big/ \sum_{j=1}^{H} y_j\,, \quad m = 1, \ldots, H.$$

Все эти точки лежат внутри единичного квадрата $\{(x,y)\,|x \in [0,1]\,,\ y \in [0,1]\,\}$. Соединяя вершины квадрата $(0,0)\,,(1,1)$ и соседние построенные точки отрезками, получим кривую Лоренца. На множестве кривых Лоренца вводится частичный порядок: если первая кривая Лоренца лежит ниже второй, то говорят, что распределение доходов, по которому построена первая кривая, мажорирует (по Лоренцу) распределение доходов, по которому построена вторая кривая. Равномерное распределение доходов, при котором

$$y_i = \frac{1}{H}\sum_{j=1}^{H} y_j,\ i = 1, \ldots, H,$$

мажорируется любым другим распределением доходов. Ему соответствует кривая Лоренца, совпадающая с диагональю квадрата, соединяющей вершины $(0,0)\,,\ (1,1)$. Распределение

$$y_{\sigma(1)} = \ldots = y_{\sigma(H-1)} = 0, \quad y_{\sigma(H)} = \sum_{j=1}^{H} y_j > 0,$$

при котором весь совокупный доход принадлежит самому богатому домашнему хозяйству, мажорирует любое другое распределение доходов. Кривые Лоренца используются для измерения неравенства в распределении доходов статистическими службами многих стран. В первой трети XX века была разработана математическая теория, связавшая мажоризацию по Лоренцу с перераспределением доходов (передачи Пигу-Дальтона), стохастическим доминированием и «отвращением к риску» (см. подробнее [28]). Для численного измерения неравенства используются индексы неравенства, которые согласованы с мажоризацией по Лоренцу, так что у мажорирующего распределения доходов значение индекса неравенства должно быть не меньшим. Самым распространенным индексом неравенства является индекс Джини, который равняется удвоенной площади между кривой Лоренца и диагональю единичного квадрата, соединяющей его вершины $(0,0)$, $(1,1)$. Для распределения доходов $y=(y_1,\ldots,y_H)$ индекс Джини равен

$$G(y_1,...,y_H)=\frac{\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{H}|y_i-y_j|}{2H^2\overline{y}},\quad \overline{y}=\frac{1}{H}\sum_{i=1}^{H}y_i.$$

Если перестановка σ такая, что $y_{\sigma(1)}\le y_{\sigma(2)}\le\ldots y_{\sigma(H)}$, то выражение для индекса имеет вид

$$G(y_1,...,y_H)=\frac{1}{H}\sum_{i=1}^{H}(2i-H-1)\,\frac{y_{\sigma(i)}}{\hat{y}},\quad \hat{y}=\sum_{j=1}^{H}y_j.$$

Индекс Джини равномерного распределения доходов равен нулю. Для распределения доходов $y^*=(y_1^*,\ldots,y_H^*)$, при котором весь доход достается самому богатому домашнему хозяйству, индекс Джини принимает максимальное значение $G^*=G\left(y_1^*,\ldots,y_H^*\right)=\frac{H-1}{H}$.

Используя обозначения из предыдущих разделов, положим

$$y_i=\frac{rk^{H+1-i}+w}{rK+Hw}\ge 0,\ i=1,\ldots,H.$$

Заметим, что $\hat{y}=\sum\limits_{j=1}^{H}y_j=1$. Согласно замечанию 3 область

$$\Gamma=\left\{\left(k^1,\ldots,k^H\right)\left|k^1=k^2=\ldots=k^l>k^{l+1}\ge\ldots\ge k^H\right.\right\}$$

является «ловушкой» как для задачи Коши (11), так и (12), т.е. если $\left(k_0^1,\ldots,k_0^H\right)\in\Gamma$, то $\left(k^1(t),\ldots,k^H(t)\right)\in\Gamma$ при $t\in[0,+\infty)$. Положим

$$\Lambda=\left\{\left(y^1,\ldots,y^H\right)\left|y_i=\frac{rk^{H+1-i}+w}{rK+Hw}\ge 0,\ i=1,\ldots,H,\ \left(k^1,\ldots,k^H\right)\in\Gamma\right.\right\}.$$

Динамической траектории $\left(k^1(t),\ldots,k^H(t)\right)\in\Gamma$ будем ставить в соответствие траекторию $\left(y_1(t),\ldots,y_H(t)\right)\in\Lambda$, по которой строится динамика кривых Лоренца. Это соответствие определяет в области Λ автономные системы обыкновенных дифференциальных уравнений, порождаемые соответственно системами (11) или (12). Обеим системам (11) и (12) соответствует единственная стационарная траектория $\hat{y}^*=(\hat{y}_1^*,\ldots,\hat{y}_H^*)$, где $\hat{y}_H^*=\ldots=\hat{y}_{H-l+1}^*=\frac{1}{l}$, $\hat{y}_i^*=0$, $i=1,\ldots,H-l$, в области Λ. Индекс Джини принимает максимальное в области Λ значение на распределении $\hat{y}^*$, равное $\hat{G}^*=\frac{H-l}{H}$.

Рассмотрим в области Λ функцию

$$V\left(y_1,\ldots,y_H\right)=\hat{G}^*-G\left(y_1,\ldots,y_H\right)=\hat{G}^*-\frac{1}{H}\sum_{i=1}^{H}(2i-H-1)\,y_i.$$

Функция $V\left(y_1,\ldots,y_H\right)>0$ при $\left(y_1,\ldots,y_H\right)\in\Lambda\backslash\{\hat{y}^*\}$, $V\left(\hat{y}^*\right)=0$, и непрерывно дифференцируема в области Λ.

Теорема 6.1. *Пусть выполнены предположения* 1 *и* 2, $\left(k^1(t),\ldots,k^H(t)\right)\in\Gamma$ *является решением задачи Коши* (11) *(модель Рамсея–Бьюли) или* (12) *(модель Рамсея–Беккера). Тогда*

$$\frac{d}{dt}V\left(\frac{rk^H(t)+w}{rK(t)+Hw},\ldots,\frac{rk^{H+1-i}(t)+w}{rK(t)+Hw},\ldots,\frac{rk^1(t)+w}{rK(t)+Hw}\right)<0.$$

Замечание 5. Положение равновесия $\hat{y}^*$ является глобально асимптотически устойчивым по Ляпунову в области Λ, а построенная по индексу Джини функция

$$V(y_1,\ldots,y_H)=\hat{G}^*-G(y_1,\ldots,y_H)$$

является функцией Ляпунова в области Λ.

Доказательство теоремы 6.1. Пусть

$$\left(\frac{rk_0^H+w}{rK_0+Hw},\ldots,\frac{rk_0^{H+1-i}+w}{rK_0+Hw},\ldots,\frac{rk_0^1+w}{rK_0+Hw}\right)\in\Gamma.$$

Тогда имеем

$$\left(\frac{rk^H(t)+w}{rK(t)+Hw},\ldots,\frac{rk^{H+1-i}(t)+w}{rK(t)+Hw},\ldots,\frac{rk^1(t)+w}{rK(t)+Hw}\right)\in\Gamma.$$

Вычислим производную

$$\frac{d}{dt}V\left(\frac{rk^H(t)+w}{rK(t)+Hw},\ldots,\frac{rk^{H+1-i}(t)+w}{rK(t)+Hw},\ldots,\frac{rk^1(t)+w}{rK(t)+Hw}\right)=$$
$$=-\frac{1}{H}\sum_{i=1}^{H}(H+1-2i)\frac{r}{(rK(t)+Hw)^2}\left(\frac{dk^i(t)}{dt}(rK(t)+Hw)-\frac{dK(t)}{dt}\left(rk^i(t)+w\right)\right).$$

В случае модели Рамсея–Бьюли, т.е. задачи Коши (11), имеем, что

$$\frac{d}{dt}V\left(\frac{rk^H+w}{rK+Hw},\ldots,\frac{rk^{H+1-i}+w}{rK+Hw},\ldots,\frac{rk^1+w}{rK+Hw}\right)=$$
$$=-\frac{r}{H(1-\beta)}\sum_{i=1}^{H}(H+1-2i)\frac{rk^i+w}{rK+Hw}\left(\left(1-\varphi\left(\frac{k^i}{K}\right)\right)-\sum_{j=1}^{H}\left(1-\varphi\left(\frac{k^j}{K}\right)\right)\frac{rk^j+w}{rK+Hw}\right).$$

В случае модели Рамсея–Беккера, т.е. задачи Коши (12), имеем, что

$$\frac{d}{dt}V\left(\frac{rk^H+w}{rK+Hw},\ldots,\frac{rk^{H+1-i}+w}{rK+Hw},\ldots,\frac{rk^1+w}{rK+Hw}\right)=$$
$$=-\frac{r}{H(1-\beta)(rK+Hw)}\sum_{i=1}^{H}(H+1-2i)\times$$
$$\times\left\{\left(1-\varphi\left(\frac{k^i}{K}\right)\right)(rk^i+w)+\left(\varphi\left(\frac{k^i}{K}\right)-1\right)_+w\left(\frac{w}{c_2\left(k^i,r\varphi\left(\frac{k^i}{K}\right)\right)}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^i}{K}\right)-1}}-\right.$$
$$\left.-\sum_{j=1}^{H}\left(\left(1-\varphi\left(\frac{k^j}{K}\right)\right)(rk^j+w)+\left(\varphi\left(\frac{k^j}{K}\right)-1\right)_+w\left(\frac{w}{c_2\left(k^j,r\varphi\left(\frac{k^j}{K}\right)\right)}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^j}{K}\right)-1}}\right)\frac{rk^i+w}{rK+Hw}\right\}.$$

Пусть $\alpha_i(t)=\frac{rk^i(t)+w}{rK(t)+Hw}$, $i=1,\ldots,H$. Поскольку

$$\left(\frac{rk^H(t)+w}{rK(t)+Hw},\ldots,\frac{rk^{H+1-i}(t)+w}{rK(t)+Hw},\ldots,\frac{rk^1(t)+w}{rK(t)+Hw}\right)\in\Gamma,$$

имеем $\alpha_1(t)=\alpha_2(t)=\ldots=\alpha_l(t)>\alpha_{l+1}(t)\geq\ldots\geq\alpha_H(t)\geq 0$, кроме того, $\sum\limits_{i=1}^{H}\alpha_i(t)=1$.

Положим в случае модели Рамсея–Бьюли

$$f_j(t)=1-\varphi\left(\frac{k^j(t)}{K(t)}\right),\quad j=1,\ldots,H,$$

а в случае модели Рамсея–Беккера имеем

$$f_i(t) = 1 - \varphi\left(\frac{k^i(t)}{K(t)}\right) +$$

$$+\left(\varphi\left(\frac{k^i(t)}{K(t)}\right) - 1\right)_+ w\left(\frac{w}{c_2\left(k^i(t), r\varphi\left(\frac{k^i(t)}{K(t)}\right)\right)}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^i(t)}{K(t)}\right)-1}} \frac{1}{rk^i(t)+w}, \quad i = 1, \ldots, H.$$

В обоих случаях

$$\frac{d}{dt}V\left(\frac{rk^H(t)+w}{rK(t)+Hw}, \ldots, \frac{rk^{H+1-i}(t)+w}{rK(t)+Hw}, \ldots, \frac{rk^1(t)+w}{rK(t)+Hw}\right) =$$

$$= -\frac{r}{H(1-\beta)}\sum_{i=1}^{H}(H+1-2i)\,\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t)\right).$$

Заметим, что

$$\sum_{i=1}^{H}\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t)\right) = 0.$$

Следовательно,

$$\frac{d}{dt}V\left(\frac{rk^H(t)+w}{rK(t)+Hw}, \ldots, \frac{rk^{H+1-i}(t)+w}{rK(t)+Hw}, \ldots, \frac{rk^1(t)+w}{rK(t)+Hw}\right) =$$

$$= \frac{2r}{H(1-\beta)}\sum_{i=1}^{H} i\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t)\right).$$

В силу предположения 1 в случае модели Рамсея–Бьюли имеем, что

$$f_1(t) = f_2(t) = \ldots = f_l(t) > f_{l+1}(t) \geq \ldots \geq f_H(t),$$

$$f_1(t) > \sum_{i=1}^{H}\alpha_i(t)\,f_i(t) > f_H(t),$$

откуда получаем, что $f_1(t) - \sum\limits_{j=1}^{H}\alpha_j(t)\,f_j(t) > 0$, $f_H(t) - \sum\limits_{j=1}^{H}\alpha_j(t)\,f_j(t) < 0$. Кроме того,

$$f_1(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t) = f_2(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t) = \ldots = f_l(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t) >$$

$$> f_{l+1}(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t) \geq \ldots \geq f_H(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t).$$

Таким образом, в случае модели Рамсея–Бьюли существует j^* такой, что

$$f_i(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t) \geq 0, \quad i \leq j^*; \quad f_i(t) - \sum_{j=1}^{H}\alpha_j(t)\,f_j(t) < 0, \quad i > j^*.$$

В лемме 5.1 с использованием предположения 2 доказано, что

$$\sum_{i=1}^{H}\left(1 - \varphi\left(\frac{k^i(t)}{K(t)}\right)\right)\frac{rk^i(t)+w}{rK(t)+Hw} \geq 1 - \varphi\left(\frac{1}{H}\right) \geq 0.$$

В модели Рамсея–Беккера с учетом предположения 1 имеем, что

$$\sum_{j=1}^{H}\alpha_j(t)\,f_j(t) = \sum_{j=1}^{H}\left(\left(1 - \varphi\left(\frac{k^j(t)}{K(t)}\right)\right)\frac{rk^j(t)+w}{rK(t)+Hw} +\right.$$

$$\left. +\left(\varphi\left(\frac{k^j(t)}{K(t)}\right) - 1\right)_+ w\left(\frac{w}{c_2\left(k^j(t), r\varphi\left(\frac{k^j(t)}{K(t)}\right)\right)}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^j(t)}{K(t)}\right)-1}} \frac{1}{rK(t)+Hw}\right) \geq 0.$$

В случае модели Рамсея–Беккера заметим, что если $\varphi\left(\frac{k^i(t)}{K(t)}\right) > 1$, то

$$f_i(t) = \frac{1}{rk^i(t)+w}\left(\left(1-\varphi\left(\frac{k^i(t)}{K(t)}\right)\right)\left(rk^i(t)+w\right)+\right.$$

$$\left.+\left(\varphi\left(\frac{k^i(t)}{K(t)}\right)-1\right)_+ w\left(\frac{w}{c_2\left(k^i(t), r\varphi\left(\frac{k^i(t)}{K(t)}\right)\right)}\right)^{\frac{1-\beta}{\varphi\left(\frac{k^i(t)}{K(t)}\right)-1}}\right) \leq$$

$$\leq \frac{1}{rk^i(t)+w}\left(1-\varphi\left(\frac{k^i(t)}{K(t)}\right)\right)rk^i(t) < 0.$$

Следовательно, если $\varphi\left(\frac{k^i(t)}{K(t)}\right) > 1$, то $f_i(t) - \sum\limits_{j=1}^{H}\alpha_j(t) f_j(t) < 0$. С другой стороны, из предположения 1 следует, что если $\varphi\left(\frac{k^i(t)}{K(t)}\right) \leq 1$, $k^s(t) \geq k^i(t)$, то $f_s(t) - \sum\limits_{j=1}^{H}\alpha_j(t) f_j(t) \geq f_i(t) - \sum\limits_{j=1}^{H}\alpha_j(t) f_j(t)$. Кроме того, поскольку $\frac{k^1(t)}{K(t)} > \frac{1}{H}$, имеем, что $\varphi\left(\frac{k^1(t)}{K(t)}\right) < \varphi\left(\frac{1}{H}\right) \leq 1$, $f_1(t) - \sum\limits_{j=1}^{H}\alpha_j(t) f_j(t) > 0$. Таким образом, в случае модели Рамсея–Беккера также существует j^* такой, что

$$f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t) \geq 0, \quad i \leq j^*; \quad f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t) < 0, \quad i > j^*.$$

Тогда получаем следующую оценку:

$$\frac{d}{dt}V\left(\frac{rk^H(t)+w}{rK(t)+Hw}, \ldots, \frac{rk^{H+1-i}(t)+w}{rK(t)+Hw}, \ldots, \frac{rk^1(t)+w}{rK(t)+Hw}\right) =$$

$$= \frac{2r}{H(1-\beta)}\sum_{i=1}^{H} i\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t)\right) =$$

$$= \frac{2r}{H(1-\beta)}\left\{\sum_{i=1}^{j^*} i\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t)\right) + \right.$$

$$\left. + \sum_{i=j^*+1}^{H} i\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t)\right)\right\} \leq$$

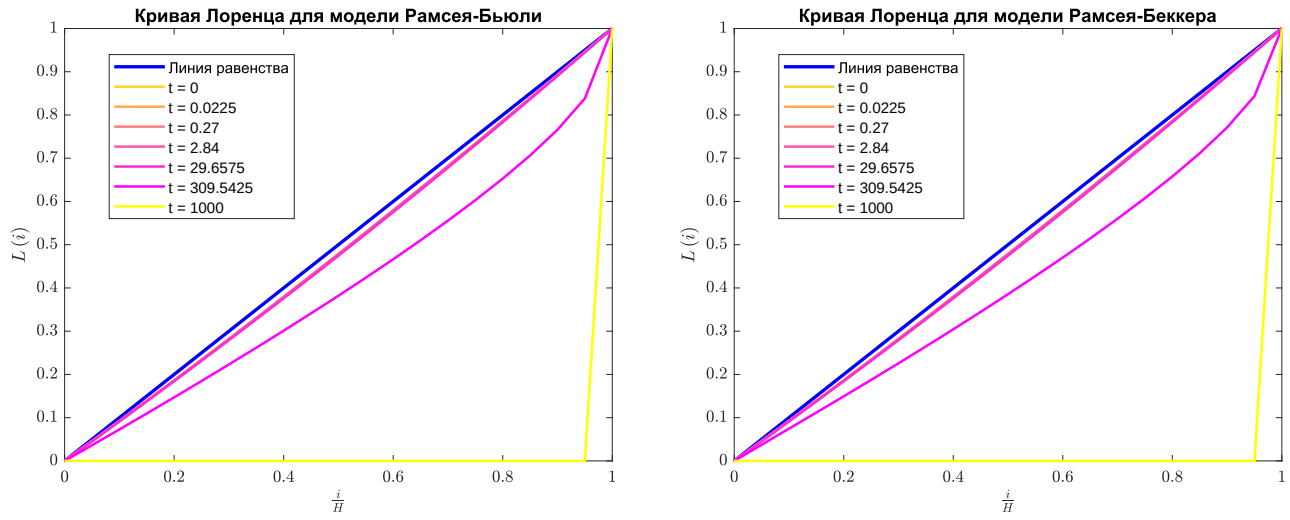
$$\leq \frac{2r}{H(1-\beta)}\left\{j^*\sum_{i=1}^{j^*}\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t)\right) + \right.$$

$$\left. + (j^*+1)\sum_{i=j^*+1}^{H}\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t)\right)\right\} =$$

$$= \frac{2r}{H(1-\beta)}\sum_{i=j^*+1}^{H}\alpha_i(t)\left(f_i(t) - \sum_{j=1}^{H}\alpha_j(t) f_j(t)\right) < 0.$$

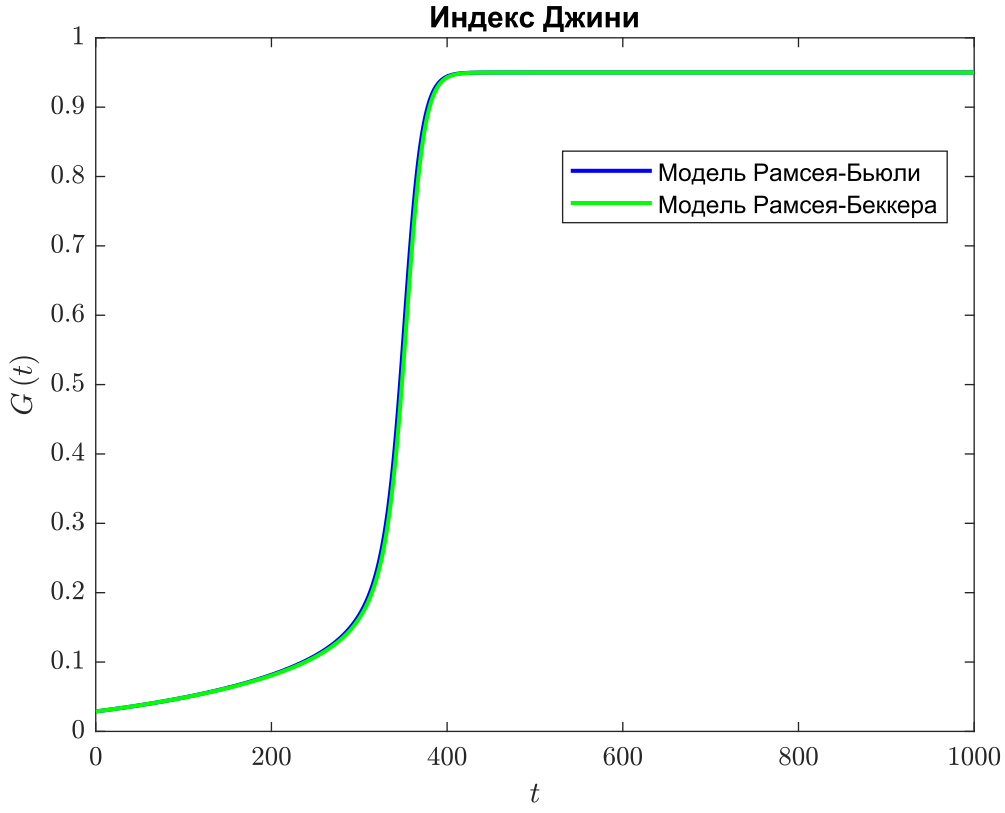
Теорема доказана.

На фиг. 1 приведена эволюция с течением времени кривой Лоренца для двух моделей – модели Рамсея–Бьюли (см. фиг. 1а) и модели Рамсея–Беккера (см. фиг. 1б). На фиг. 2 показана динамика индекса неравенства Джини. Сходимость распределения доходов к двухклассовому сообществу в модели Рамсея–Бьюли происходит быстрее.



(а) Модель Рамсея–Бьюли (б) Модель Рамсея–Беккера

Фиг. 1. Кривая Лоренца для популяции из $H = 20$ домашних хозяйств. Подсчитана при следующих функциях и значениях параметров: $T = 1000$, $\beta = 0.1$, $\varphi(x) = \varphi_{\max} - (\varphi_{\max} - \varphi_{\min})\, x^2$, $\varphi_{\max} = 1.001, \varphi_{\min} = \beta$, $r = 0.1$, $w = 5 \cdot 10^3$, $K_0 = 10^5$.



Фиг. 2. Индекс неравенства Джини в зависимости от времени t для популяции из $H = 20$ домашних хозяйств в моделях Рамсея–Бьюли и Рамсея–Беккера. Подсчитан при следующих функциях и значениях параметров: $T = 1000$, $\beta = 0.1$, $\varphi(x) = \varphi_{\max} - (\varphi_{\max} - \varphi_{\min})\, x^2$, $\varphi_{\max} = 1.001, \varphi_{\min} = \beta$, $r = 0.1$, $w = 5 \cdot 10^3$, $K_0 = 10^5$.

7. ЗАКЛЮЧЕНИЕ

Гипотеза Рамсея связывает формирование двухклассовой социальной структуры с различием в коэффициентах дисконтирования у экономических агентов. В большинстве работ, обосновывающих эту гипотезу, коэффициенты дисконтирования являются неизменными и заданными априори. В данной статье доказана справедливость гипотезы Рамсея с эндогенно формирующимися коэффициентами дисконтирования в соответствии с гипотезой относительного дохода Дьюзенберри. В статье исследуется влияние потребительского кредита на формирование социальной структуры общества. В модели Рамсея–Бьюли допускается потребительский кредит, обеспеченный будущими заработками. В модели Рамсея–Беккера потребительский кредит не допускается. В обоих моделях обоснована справедливость гипотезы Рамсея. Однако формирующаяся двухклассовая социальная структура оказывается различной. В модели Рамсея–Бьюли – это класс собственников и класс должников, отрабатывающих задолженность по ранее взятому потребительскому кредиту. В модели Рамсея–Беккера – это класс собственников и класс трудящихся, обеспеченных заработной платой.

В математических моделях двухклассовая структура понимается как распределение доходов, устанавливающееся в пределе на неограниченном временном горизонте. Для измерения экономического неравенства в обществе на конечных временных горизонтах статистические службы многих стран сравнивают кривые Лоренца, описывающие распределение доходов между экономическими агентами, вычисляют индексы неравенства, среди которых наиболее популярным является индекс Джини. В статье доказано, что индекс Джини связан с функциями Ляпунова в моделях популяционной динамики Рамсея–Бьюли и Рамсея–Беккера.

В завершение отметим интересные с нашей точки зрения направления дальнейших исследований гипотезы Рамсея и социальной динамики. Показанное в статье влияние потребительского кредита на социальную структуру актуализирует анализ влияния на социальную динамику ограничений ликвидности капитала и несовершенства рынка капитала, при котором различаются цена покупки и цена продажи капитала.

В статье исследовалась динамика популяции экономических агентов в условиях стационарных экономических условий (постоянство доходности капитала и заработной платы). Представляет интерес изучение социальной динамики и гипотезы Рамсея при изменении этих условий, описываемых моделями экономического роста или случайными процессами Леви.

СПИСОК ЛИТЕРАТУРЫ

1. *Piketty T.* Capital in the Twenty-First Century. Cambridge: The Belknap Press of Harvard University Press, 2014.
2. *Aghion P., Williamson J. G.* Growth, Inequality and Globalization: Theory, History and Policy. Cambridge: Cambridge University Press, 1999.
3. *Atkinson A. B.* Inequality: What Can Be Done? Cambridge: Harvard University Press, 2015.
4. *Ramsey F. P.* A Mathematical Theory of Saving // Econ. J. 1928. V. 38. № 152. P. 543–559.
5. *Acemoglu D.* Introduction to Modern Economic Growth. Princeton: Princeton University Press, 2009.
6. *Becker R. A.* Equilibrium Dynamics with Many Agents. In: Dana R.-A., Le Van C., Mitra T., Nishimura K. Handbook on Optimal Growth 1: Discrete Time. Berlin: Springer, 2006. P. 385–442.
7. *Борисов К. Ю., Пахнин М. А.* Модели экономического роста с неоднородным дисконтированием // Ж. вычисл. матем. и матем. физ. 2023. Т. 63. № 3. С. 355–379.
8. *Becker R. A.* On the Long-run Steady State in a Simple Dynamic Model of Equilibrium with Heterogeneous Households // Q. J. Econ. 1980. V. 95. № 2. P. 375–382.
9. *Bewley T. F.* An integration of equilibrium theory and turnpike theory // J. Math. Econ. 1982. V. 10. P. 233–267.
10. *Mitra T., Sorger G.* On Ramsey's conjecture // J. Econ. Theory. 2013. V. 148. № 5. P. 1953–1976.
11. *Koopmans T. C.* Stationary Ordinal Utility and Impatience // Econometrica. 1960. V. 28. № 2. P. 287–309.
12. *Uzawa H.* Time Preference, the Consumption Function, and Optimal Asset Holdings. In: Wolfe J. N. (ed.) Value, Capital and Growth: Papers in Honour of Sir John Hicks. Chicago: Aldine Publishing Company, 1968. P. 485–505.
13. *Borissov K.* Growth and Distribution in a Model with Endogeneous Time Preferences and Borrowing Constraints // Math. Soc. Sci. 2013. V. 66. № 2. P. 117–128.

14. *Borissov K., Lambrecht S.* Growth and Distribution in an AK-model with Endogeneous Impatience // Econ. Theory. 2009. V. 39. № 1. P. 93–112.

15. *Duesenberry J. S.* Income, Saving and the Theory of Consumer Behavior. Cambridge: Harvard University Press, 1949.

16. *Keynes J. M.* The General Theory of Employment, Interest and Money. London: Macmillan, 1936.

17. *Frank R. H.* Falling Behind: How Rising Inequality Harms the Middle Class. Berkeley: University of California Press, 2007.

18. *Schlicht E.* A Neoclassical Theory of Wealth Distribution // Jahrb. Natl. Stat. 1975. V. 189. P. 78–96.

19. *Bourguignon F.* Pareto Superiority of Unegalitarian Equilibria in Stiglitz' Model of Wealth Distribution with Convex Saving Function // Econometrica. 1981. V. 49. P. 1469–1475.

20. *Borissov K.* The Rich and the Poor in a Simple Model of Growth and Distribution // Macroecon. Dyn. 2016. V. 20. № 7. P. 1934–1952.

21. *Fisher R. A.* The Genetical Theory of Natural Selection. Oxford: Clarendon Press, 1930.

22. *Асеев С. М., Кряжимский А. В.* Принцип максимума Понтрягина и задачи оптимального экономического роста // Тр. МИАН. 2007. Т. 257. С. 3–271.

23. *Асеев С. М., Бесов К. О., Кряжимский А. В.* Задачи оптимального управления на бесконечном интервале времени в экономике // Успехи матем. наук. 2012. Т. 67. Вып. 2 (404). С. 3–64.

24. *Carlson D. A., Haurie A. B., Leizarowitz A.* Infinite Horizon Optimal Control: Deterministic and Stochastic Systems. Berlin: Springer-Verlag, 1991.

25. *Seierstad A., Sydsæter K.* Optimal Control Theory with Economic Applications. Amsterdam: North-Holland, 1987.

26. *Fleming W. H., Soner H. M.* Controlled Markov Processes and Viscosity Solutions. New York: Springer, 2006.

27. *Тихонов А. Н., Васильева А. Б., Свешников А. Г.* Дифференциальные уравнения. М.: Физматлит, 2005. 256 с.

28. *Marshall A. W., Olkin I., Arnold B. C.* Inequalities: Theory of Majorization and Its Applications. Second Edition. New York: Springer, 2011.

RAMSEY'S CONJECTURE OF SOCIAL STRATIFICATION AS FISHER'S SELECTION PRINCIPLE

G. S. Parastaev[a,b,*], **A. A. Shananin**[a,b,c,d,e,**]

[a] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, 119991 Russia*
[b] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, 119333 Russia*
[c] *Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow oblast, 141701 Russia*
[d] *Moscow Center for Fundamental and Applied Mathematics, Moscow, 119991 Russia*
[e] *Peoples' Friendship University of Russia (RUDN University), Moscow, 117198 Russia*
**e-mail: parastaew1996@yandex.ru*
***e-mail: alexshan@yandex.ru*

Received: 23.08.2024
Revised: 23.08.2024
Accepted: 23.08.2024

Abstract. Ramsey's conjecture of social stratification states that wealth in a population of households is concentrated among the most frugal agents, who discount consumer spending with the lowest discount factor. Ramsey's conjecture can be viewed as stating that Fisher's principle of natural selection holds in a population of households. In this paper, based on Duesenberry's hypothesis, discount factors are formed depending on the capital distribution among the agents. The behavior of households is described by Ramsey-type models of a rational representative consumer. For the corresponding optimal control problems, we construct solutions in the form of synthesis, which are used to model the dynamics of a household population. Theorems for a household population are proved that justify the validity of Ramsey's conjecture. The influence of consumer loans on the social stratification of households is studied.

Keywords: optimal control synthesis, discount factor, relative income hypothesis, Ramsey's conjecture, Lyapunov function.