СОДЕРЖАНИЕ

Том 61, номер 2, 2021 год

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

Неполиномиальная интерполяция функций с большими градиентами и ее применение
А. И. Задорин, Н. А. Задорин 179

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

Многометодные алгоритмы для решения сложных задач оптимального управления
А. И. Тятюшкин 189

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

Декомпозиция решения двумерного сингулярно возмущенного уравнения конвекции-диффузии с переменными коэффициентами в квадрате; оценки в гёльдеровых нормах
В. Б. Андреев, И. Г. Белухина 206

Об аппроксимации слабых решений уравнения Лапласа гармоническими многочленами
М. Е. Боговский 217

МАТЕМАТИЧЕСКАЯ ФИЗИКА

Численный анализ трехмерных задач магнитной маскировки на основе оптимизационного метода
Г. В. Алексеев, Ю. Э. Спивак 224

Аналитическое исследование хаотической динамики двумерной модели Лотки–Вольтерра с сезонностью
Ю. В. Бибик 239

Угловой пограничный слой в краевых задачах для сингулярно возмущенных параболических уравнений с кубическими нелинейностями
И. В. Денисов 256

Моделирование ламинарно-турбулентного перехода с применением диссипативных численных схем
И. В. Егоров, Н. К. Нгуен, Т. Ш. Нгуен, П. В. Чувахов 268

Численное моделирование нестационарных дозвуковых течений вязкого газа на основе составных компактных схем высокого порядка
А. Д. Савельев 281

Обратная задача для уравнений сложного теплообмена с френелевскими условиями сопряжения
А. Ю. Чеботарев 303

ИНФОРМАТИКА

Численные методы для задачи распределения ресурсов в компьютерной сети
Е. А. Воронцова, А. В. Гасников, П. Е. Двуреченский, А. С. Иванова, Д. А. Пасечнюк 312

О верхней границе сложности сортировки
И. С. Сергеев 345

Памяти Ивана Станиславовича Меньшикова (1952–2020) 363

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 179–188

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.652

НЕПОЛИНОМИАЛЬНАЯ ИНТЕРПОЛЯЦИЯ ФУНКЦИЙ С БОЛЬШИМИ ГРАДИЕНТАМИ И ЕЕ ПРИМЕНЕНИЕ[1)]

© 2021 г. А. И. Задорин[1,*], Н. А. Задорин[1,]**

[1] 630090 Новосибирск, пр-т Акад. Коптюга, 4, Ин-т матем. СО РАН, Россия

**e-mail: zadorin@ofim.oscsbras.ru*

***e-mail: nik-zadorin@yandex.ru*

Поступила в редакцию 02.06.2020 г.
Переработанный вариант 20.08.2020 г.
Принята к публикации 16.09.2020 г.

Исследуется вопрос интерполяции функции одной переменной с большими градиентами в области пограничного слоя. Проблема в том, что применение классических полиномиальных интерполяционных формул на равномерной сетке к функциям с большими градиентами может приводить к погрешностям порядка $O(1)$, несмотря на малость шага сетки. Исследована интерполяционная формула, построенная на основе подгонки к составляющей, задающей погранслойный рост функции. Получена оценка погрешности, зависящая от числа узлов интерполяции и равномерная по погранслойной составляющей и ее производным. Показано, как построенная интерполяционная формула может быть применена для построения формул численного дифференцирования и интегрирования, в двумерном случае. Получены соответствующие оценки погрешности. Библ. 21. Табл. 2.

Ключевые слова: пограничный слой, функция с большими градиентами, неполиномиальная интерполяционная формула, оценка погрешности.

DOI: 10.31857/S0044466921020150

1. ВВЕДЕНИЕ

Многочлен Лагранжа широко используется для интерполяции функций. Однако в случае функций с большими градиентами применение интерполяции Лагранжа может приводить к погрешностям порядка O(1) (см. [1]). Следовательно, актуален вопрос построения интерполяционных формул для функций с большими градиентами в пограничном слое. Интерполяционная формула должна строиться таким образом, чтобы ее погрешность была равномерной по резким изменениям функции в пограничном слое. Для построения таких формул можно выделить два подхода: применение интерполяции Лагранжа на сетке, сгущающейся в области пограничного слоя и построение специальных интерполяционных формул, основанных на подгонке к погранслойной составляющей функции.

Формула линейной интерполяции при наличии экспоненциального пограничного слоя на сетках Г.И. Шишкина (см. [2]) и Н.С. Бахвалова (см. [3]) исследовалась в [4]. В [5] доказано, что в случае экспоненциального пограничного слоя многочлен Лагранжа можно применять на сетке Шишкина. Для многочлена Лагранжа с произвольно заданным числом узлов интерполяции получены оценки погрешности, равномерные по малому параметру.

Подход, основанный на подгонке интерполяционной формулы к быстро растущей составляющей, менее исследован. В [6] рассмотрен вопрос интерполяции функции, представимой в виде

$$u(x) = p(x) + \gamma\Phi(x), \quad x \in [a, b], \tag{1.1}$$

где функция $u(x)$ является достаточно гладкой, погранслойная составляющая $\Phi(x)$ известна и имеет большие градиенты на интервале $[a, b]$, регулярная составляющая $p(x)$ ограничена вместе

[1)]Работа выполнена при финансовой поддержке РФФИ (работа по секциям 1, 2, 4 поддержана проектом № 20-01-00650, по секциям 3, 5, 6 – проектом № 19-31-60009).

с производными до некоторого порядка, постоянная γ не задана. В частности, декомпозиция (1.1) строилась в [7] для решения сингулярно возмущенной краевой задачи, при этом

$$\Phi(x) = e^{-mx/\varepsilon}, \quad x \in [0,1], \quad m > 0, \quad \varepsilon \in (0,1]. \tag{1.2}$$

Производные функции $\Phi(x)$ неограниченно растут с уменьшением параметра ε, из-за чего погрешность полиномиальных интерполяционных формул становится порядка $O(1)$.

В [6] построена интерполяционная формула на произвольном сеточном интервале $[x_{n-1}, x_n]$ с двумя узлами интерполяции x_{n-1} и x_n, точная на составляющей $\Phi(x)$. Доказано, что если $\Phi'(x) \neq 0$, то погрешность построенной формулы порядка $O(h)$ равномерна по составляющей $\Phi(x)$. Здесь h – шаг сетки.

В [8] для функции вида (1.1) построена интерполяционная формула с произвольно заданным числом узлов интерполяции, точная на составляющей $\Phi(x)$. Однако в [8] нет оценки погрешности, равномерной по погранслойной составляющей $\Phi(x)$.

В данной работе получим оценку погрешности интерполяционной формулы из [8] с k узлами интерполяции. Рассмотрим применение этой формулы для построения формул численного дифференцирования и интегрирования, а также в двумерном случае.

2. АНАЛИЗ ИНТЕРПОЛЯЦИОННОЙ ФОРМУЛЫ

Пусть Ω^h – равномерная сетка интервала $[a,b]$:

$$\Omega^h = \{x_n : x_n = a + (n-1)h,\ x_1 = a,\ x_k = b,\ n = 1,2,\ldots,k\}.$$

Предполагаем, что функция $u(x)$ вида (1.1) задана в узлах сетки, $u_n = u(x_n)$.

Пусть $L_n(u,x)$ – многочлен Лагранжа для функции $u(x)$ с узлами интерполяции $x_1,\ldots,x_n$. Покажем, что применение многочлена Лагранжа к функции вида (1.1) может приводить к значительным погрешностям. Для этого зададим $u(x) = e^{-x/\varepsilon}$ при $x \in [0,1]$. Пусть $\varepsilon = h$, тогда при интерполяции на интервале $[0,h]$ выполнится $L_2(u,h/2) - u(h/2) \approx 0.075$. Итак, точность интерполяции не повышается с уменьшением шага h, если $\varepsilon = h$.

В [8] для интерполяции функции вида (1.1) построена интерполяционная формула

$$L_{\Phi,k}(u,x) = L_{k-1}(u,x) + \frac{[x_1,\ldots,x_k]u}{[x_1,\ldots,x_k]\Phi}[\Phi(x) - L_{k-1}(\Phi,x)], \tag{2.1}$$

где $[x_1,\ldots,x_k]u$ – разделенная разность для функции $u(x)$ (см. [9]).

Пусть

$$\Phi^{(k-1)}(x) \neq 0, \quad x \in (a,b). \tag{2.2}$$

Тогда знаменатель в (2.1) не обращается в нуль и формула задана корректно.

Покажем, что формула (2.1) является интерполяционной. Преобразуем формулу (2.1). В соответствии с [9], справедливо соотношение

$$L_k(u,x) = L_{k-1}(u,x) + r_{k-1}(x)[x_1,x_2,\ldots,x_k]u, \tag{2.3}$$

где $r_{k-1}(x) = (x-x_1)(x-x_2)\cdots(x-x_{k-1})$. Учитывая (2.3), из (2.1) получаем

$$L_{\Phi,k}(u,x) = L_k(u,x) + \frac{[x_1,\ldots,x_k]u}{[x_1,\ldots,x_k]\Phi}[\Phi(x) - L_k(\Phi,x)]. \tag{2.4}$$

Очевидно, что формула (2.4) является интерполяционной с узлами интерполяции $x_1, \ldots, x_k$. Следовательно, и формула в виде (2.1) является интерполяционной.

Учитывая, что, согласно [9, с. 44],

$$\Phi(x) - L_{k-1}(\Phi,x) = r_{k-1}(x)[x_1,x_2,\ldots,x_{k-1},x]\Phi \tag{2.5}$$

и

$$u(x) - L_k(u,x) = \frac{u^{(k)}(s)}{k!}r_k(x), \quad \exists s \in (a,b), \tag{2.6}$$

получаем, что формула (2.1) является точной на многочленах степени $(k-2)$ и на функции $\gamma\Phi(x)$.

Лемма 1. *Пусть выполнено условие* (2.2),

$$M_k(\Phi, x) = \frac{\Phi(x) - L_{k-1}(\Phi, x)}{\Phi(x_k) - L_{k-1}(\Phi, x_k)}. \tag{2.7}$$

Тогда

$$\max_x \left|L_{\Phi,k}(u, x) - u(x)\right| \le \max_x \left|L_{k-1}(p, x) - p(x)\right| (1 + \max_x \left|M_k(\Phi, x)\right|). \tag{2.8}$$

Доказательство. Интерполяционная формула (2.1) точна на составляющей $\Phi(x)$, поэтому

$$L_{\Phi,k}(u, x) - u(x) = L_{k-1}(p, x) - p(x) + \frac{[x_1, \ldots, x_k]p}{[x_1, \ldots, x_k]\Phi}[\Phi(x) - L_{k-1}(\Phi, x)].$$

Учитывая (2.5), получаем

$$L_{\Phi,k}(u, x) - u(x) = [L_{k-1}(p, x) - p(x)] - [L_{k-1}(p, x_k) - p(x_k)]M_k(\Phi, x), \tag{2.9}$$

где $M_k(\Phi, x)$ соответствует (2.7). Теперь из (2.9) получаем (2.8). Лемма доказана.

Следствие 1. Учитывая (2.6), из (2.8) получаем

$$\max_x \left|L_{\Phi,k}(u, x) - u(x)\right| \le \max_x \left|p^{(k-1)}(x)\right| (1 + \max_x \left|M_k(\Phi, x)\right|) h^{k-1}, \quad x \in [a, b].$$

Лемма 2. *Пусть*

$$\Phi^{(k-1)}(x) \ne 0, \quad \Phi^{(k)}(x) \ne 0, \quad k \ge 2, \quad x \in (a, b). \tag{2.10}$$

Тогда

$$\max_x \left|L_{\Phi,k}(u, x) - u(x)\right| \le 2 \max_x \left|L_{k-1}(p, x) - p(x)\right|, \quad x \in [a, b]. \tag{2.11}$$

Доказательство. Рассмотрим случай, когда производные $\Phi^{(k-1)}(x)$, $\Phi^{(k)}(x)$ одного знака:

$$\Phi^{(k-1)}(x) > 0, \quad \Phi^{(k)}(x) > 0, \quad x \in (a, b), \tag{2.12}$$

или

$$\Phi^{(k-1)}(x) < 0, \quad \Phi^{(k)}(x) < 0, \quad x \in (a, b). \tag{2.13}$$

Остановимся на условиях (2.12), условия (2.13) рассматриваются аналогично. Учитывая (2.5) и (2.7), получаем

$$M_k(\Phi, x) = \frac{r_{k-1}(x)[x_1, x_2, \ldots, x_{k-1}, x]\Phi}{r_{k-1}(x_k)[x_1, x_2, \ldots, x_{k-1}, x_k]\Phi}. \tag{2.14}$$

В соответствии с [9] для некоторого $s \in (a, b)$

$$[x_1, x_2, \ldots, x_{k-1}, x]\Phi = \Phi^{(k-1)}(s)/(k-1)!. \tag{2.15}$$

Согласно (2.12), $\Phi^{(k-1)}(x) > 0$. С учетом (2.15) имеем $z(x) = [x_1, x_2, \ldots, x_{k-1}, x]\Phi > 0$. В соответствии с [9, с. 82] для производной разделенной разности справедливо соотношение

$$z'(x) = [x_1, x_2, \ldots, x_{k-1}, x, x]\Phi.$$

Согласно (2.12), $\Phi^{(k)}(x) > 0$. Учитывая (2.15), получаем $z'(x) \ge 0$, $x \in [a, b]$. Итак, функция $z(x)$ на интервале (a, b) является положительной и возрастающей. Учитывая неравенство $\left|r_{k-1}(x)\right| \le r_{k-1}(x_k)$, из (2.14) получаем

$$\left|M_k(\Phi, x)\right| \le 1, \quad x \in [a, b]. \tag{2.16}$$

Теперь из (2.8) следует (2.11).

Остановимся на случае, когда производные $\Phi^{(k-1)}(x)$ и $\Phi^{(k)}(x)$ разных знаков. Представление (1.1) для $u(x)$ может быть записано в виде

$$u(a + b - x) = p(a + b - x) + \gamma\Phi(a + b - x), \quad x \in [a, b]. \tag{2.17}$$

Зададим $v(x) = u(a + b - x)$, $\Psi(x) = \Phi(a + b - x)$. Тогда (2.17) принимает вид

$$v(x) = p(a + b - x) + \gamma\Psi(x), \quad x \in [a, b].$$

Пусть k четно. Тогда

$$\Psi^{(k-1)}(x) = -\Phi^{(k-1)}(a + b - x), \quad \Psi^{(k)}(x) = \Phi^{(k)}(a + b - x).$$

Следовательно, производные $\Psi^{(k-1)}(x)$ и $\Psi^{(k)}(x)$ одного знака. Итак, ограничения (2.12) или (2.13) справедливы для функции $\Psi(x)$. Мы доказали, что в этом случае $|M_k(\Psi, x)| \le 1$, поэтому в соответствии с (2.8) справедливо неравенство

$$|L_{\Psi,k}(v, x) - v(x)| \le 2\max_s |L_{k-1}(p, s) - p(s)|, \quad x, s \in [a, b].$$

Это неравенство можно записать в виде

$$|L_{\Psi,k}(v, a + b - x) - v(a + b - x)| \le 2\max_s |L_{k-1}(p, s) - p(s)|, \quad x, s \in [a, b]. \tag{2.18}$$

Далее учитываем, что $v(a + b - x) = u(x)$, $\Psi(a + b - x) = \Phi(x)$, и из (2.18) получаем требуемую оценку (2.11).

Случай нечетного k рассматривается аналогично. Лемма доказана.

В соответствии с леммой 2 при ограничениях (2.10) оценка погрешности построенной интерполяционной формулы (2.1) сведена к оценке погрешности интерполяции многочленом Лагранжа $L_{k-1}(p, x)$ на регулярной составляющей $p(x)$. Для оценки погрешности интерполяции многочленом Лагранжа $L_{k-1}(p, x)$ известны оценки через $\max |p^{(k-1)}(x)|$ и в интегральной форме.

С учетом известной оценки прогрешности интерполяции многочленом Лагранжа на равномерной сетке (см. [9]):

$$|L_k(p, x) - p(x)| \le \max_s |p^{(k)}(s)| h^k, \quad x \in [a, b], \tag{2.19}$$

из (2.11) получаем

$$\max_x |L_{\Phi,k}(u, x) - u(x)| \le 2\max_x |p^{(k-1)}(x)| h^{k-1}, \quad x \in [a, b]. \tag{2.20}$$

Для отдельных значений k можно выписать оценку погрешности интерполяции многочленом Лагранжа в интегральной форме. Например,

$$|L_2(p, x) - p(x)| \le h\int_a^b |p''(s)|\, ds, \quad x \in [a, b]. \tag{2.21}$$

Тогда из (2.11) получаем

$$\max_x |L_{\Phi,3}(u, x) - u(x)| \le 2h\int_a^b |p''(s)|\, ds, \quad x \in [a, b].$$

Замечание 1. Условия (2.10) выполнены для пограничных слоев следующих видов:

экспоненциального пограничного слоя, когда $\Phi(x)$ соответствует (1.2);

степенного пограничного слоя, $\Phi(x) = (x + \varepsilon)^\alpha$, $0 < \alpha < 1$, $x > 0$, $\varepsilon > 0$;

слоя с логарифмической особенностью, $\Phi(x) = \ln x$, $x \ge \varepsilon > 0$.

3. ПОСТРОЕНИЕ КВАДРАТУРНОЙ ФОРМУЛЫ ДЛЯ ФУНКЦИИ С БОЛЬШИМИ ГРАДИЕНТАМИ

Рассмотрим вопрос численного интегрирования функции вида (1.1). В [10], [11] показано, что применение составной квадратурной формулы Ньютона–Котеса при наличии экспоненциального пограничного слоя при достаточно малых значениях параметра ε приводит к погрешностям порядка $O(h)$, несмотря на увеличение числа узлов базовой квадратурной формулы. Например, составная формула Симпсона при $\varepsilon = 1$ имеет погрешность порядка $O(h^4)$, а при $\varepsilon \le h$ погрешность становится порядка $O(h)$.

Таким образом, в случае равномерной сетки неприемлемо применять формулы Ньютона–Котеса для численного интегрирования функций вида (1.1). В [10]–[12] обоснованы аналоги формул Ньютона–Котеса с числом узлов от двух до пяти, построенные на основе замены подынтегральной функции $u(x)$ интерполянтом (2.1) вместо многочлена Лагранжа. В этих работах доказано, что построенные составные квадратурные формулы обладают погрешностью порядка $O(h^{k-1})$ равномерно по составляющей $\Phi(x)$ и ее производным, где k – число узлов базовой квадратурной формулы. При оценке погрешности накладывается ограничение $\Phi^{(k-1)}(x) \neq 0$ на каждом интервале с k узлами, на котором строится базовая квадратурная формула. Это условие выполнено для всех функций из замечания 1. Доказано, что если выделить область пограничного слоя и вне этой области строить формулы Ньютона–Котеса, то точность составной квадратурной формулы повышается на порядок и при этом погрешность становится такой же, как в регулярном случае, когда интегрируемая функция имеет ограниченные производные.

В [13] на основе интерполянта (2.1) построен и обоснован аналог формулы Ньютона–Котеса в общем случае, когда квадратурная формула содержит k узлов. При обосновании оценки погрешности потребовались ограничения на знак остаточного члена квадратурной формулы в случае функции $\Phi(x)$. Выполнение требуемых ограничений можно проверить для отдельных значений k на основе задаваемых в ряде работ таблиц, в которых указан вид остаточного члена квадратурной формулы.

Полученные оценки погрешности интерполяции (2.11), (2.20) можно применить для оценивания погрешности квадратурной формулы, построенной в [13]. При этом накладываемые ограничения (2.10) имеют более простой для проверки вид, чем ограничения в [13].

Итак, применяем интерполяционную формулу в виде (2.4) для построения квадратурной формулы с k узлами:

$$\int_a^b u(x)dx \approx \int_a^b L_{\Phi,k}(u,x)dx.$$

Учитывая (2.4), полученную квадратурную формулу можно записать в виде

$$\int_a^b u(x)dx \approx S_{\Phi,k}(u) = S_k(u) + \frac{[x_1,\dots,x_k]u}{[x_1,\dots,x_k]\Phi}\left[\int_a^b \Phi(x)dx - S_k(\Phi)\right],$$

где $S_k(u)$ – замкнутая формула Ньютона–Котеса с k узлами,

$$S_k(u) = \int_a^b L_k(u,x)dx.$$

Предполагается, что интеграл от функции $\Phi(x)$ можно вычислить в явном виде. Это условие выполнено для функций из замечания 1.

Лемма 3. *Пусть выполнены условия* (2.10). *Тогда*

$$\left|\int_a^b u(x)dx - S_{\Phi,k}(u)\right| \le 2(b-a)\max_{x\in[a,b]}\left|L_{k-1}(p,x) - p(x)\right| \le 2(b-a)\max_{x\in[a,b]}\left|p^{(k-1)}(x)\right|h^{k-1}.$$

Доказательство леммы следует из оценок (2.11), (2.19). Итак, получена оценка погрешности квадратурной формулы, равномерная по составляющей $\Phi(x)$. Погрешность $\left|L_{k-1}(p,x) - p(x)\right|$ для отдельных значений k может быть оценена более точно в интегральной форме, например, как в (2.21).

4. ФОРМУЛЫ ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ

Покажем необходимость построения специальных формул численного дифференцирования в случае функций с большими градиентами. Рассмотрим классическую формулу

$$u'(x) \approx \frac{u_n - u_{n-1}}{h}, \quad x \in [x_{n-1}, x_n].$$

Тогда при $u(x) = e^{-x/\varepsilon}$, $x \in [0,1]$, и $\varepsilon = h$ выполнится

$$\varepsilon\left|\frac{u_1 - u_0}{h} - u'(0)\right| = e^{-1}.$$

Получаем, что точность формулы не повышается при $h \to 0$, если $\varepsilon = h$, и требуется разработка формул численного дифференцирования при наличии пограничного слоя. Умножением на параметр ε вычисляется относительная погрешность, так как производная $u'(x)$ порядка $O(1/\varepsilon)$.

Известно, что классические разностные формулы для производных, построенные дифференцированием многочлена Лагранжа, можно применять на сетках, сгущающихся в области пограничного слоя. В [14]–[17] на сетках Шишкина и Бахвалова получены оценки относительной погрешности, равномерные по параметру ε.

Построение специальных формул численного дифференцирования функций с большими градиентами на равномерной сетке менее исследовано.

Интерполяционную формулу (2.1) можно применить для построения формул численного дифференцирования. Дифференцируя интерполянт (2.1), получаем

$$u^{(j)}(x) \approx L^{(j)}_{\Phi,k}(u,x) = L^{(j)}_{k-1}(u,x) + \frac{[x_1,\dots,x_k]u}{[x_1,\dots,x_k]\Phi}[\Phi^{(j)}(x) - L^{(j)}_{k-1}(\Phi,x)], \quad x \in [x_1,x_k]. \tag{4.1}$$

В [18] в случае экспоненциального пограничного слоя были получены равномерные по ε оценки относительной погрешности при вычислении первой производной при $k = 2,3$ и второй производной при $k = 3,4$, где k – число узлов в формуле (4.1), и $\Phi(x)$ соответствует (1.2).

В случае погранслойной составляющей $\Phi(x)$ общего вида оценки относительной погрешности при вычислении первой производной при $k = 2,3$ и второй производной при $k = 3$ были получены в [19]. Для пояснения остановимся на случае вычисления первой производной по формуле

$$u'(x) \approx L'_{\Phi,2}(u,x) = \frac{u_n - u_{n-1}}{\Phi_n - \Phi_{n-1}}\Phi'(x), \quad x \in [x_{n-1},x_n],$$

соответствующей (4.1). В [19] доказана следующая лемма.

Лемма 4. *Предположим, что*

$$|\Phi'(x)| \le B_n, \quad x \in [x_{n-1},x_n]$$

и для некоторой постоянной G_n

$$\frac{\int_{x_{n-1}}^{x_n}|\Phi''(s)|\,ds}{B_n|\Phi_n - \Phi_{n-1}|} \le G_n.$$

Тогда

$$\left|\frac{L'_{\Phi,2}(u,x) - u'(x)}{B_n}\right| \le G_n\int_{x_{n-1}}^{x_n}|p'(s)|\,ds + \frac{1}{B_n}\int_{x_{n-1}}^{x_n}|p''(s)|\,ds, \quad x \in [x_{n-1},x_n]. \tag{4.2}$$

В случае, когда $\Phi(x)$ соответствует (1.2), выполнится $B_n = m/\varepsilon$, $G_n = 1$. Тогда из (4.2) получаем

$$\varepsilon\left|L'_{\Phi,2}(u,x) - u'(x)\right| \le \int_{x_{n-1}}^{x_n}\left[|p'(s)| + \frac{\varepsilon}{m}|p''(s)|\right]ds, \quad x \in [x_{n-1},x_n]. \tag{4.3}$$

С помощью леммы 4 получена оценка погрешности (4.3) порядка $O(h)$, равномерная по составляющей $\Phi(x)$.

Аналогично можно воспользоваться леммой 4 для оценивания погрешности в случае другой функции $\Phi(x)$, например, из замечания 1.

5. ДВУМЕРНАЯ ИНТЕРПОЛЯЦИОННАЯ ФОРМУЛА

Исследуем формулу для интерполяции функции двух переменных с большими градиентами. Эта формула является обобщением формулы (2.1) и для получения оценки погрешности будет использована лемма 2.

Итак, пусть для достаточно гладкой функции $u(x, y)$ справедливо представление

$$u(x, y) = p(x, y) + d_1(y)\Phi(x) + d_2(x)\Theta(y) + d_3\Phi(x)\Theta(y), \tag{5.1}$$

где $(x, y) \in \bar{\Omega}$, $\bar{\Omega} = [a, b]\times[c, d]$. Предполагаем, что в (5.1) регулярная составляющая $p(x, y)$ и функции $d_1(y)$, $d_2(x)$ не заданы в явном виде и имеют ограниченные производные до некоторого порядка, а функции $\Phi(x)$, $\Theta(y)$ известны и имеют большие градиенты в области пограничного слоя. Остановимся на примере такой функции $u(x, y)$.

Рассмотрим сингулярно возмущенную задачу для эллиптического уравнения

$$\begin{gathered}\varepsilon u_{xx} + \varepsilon u_{yy} + a(x)u_x + b(y)u_y - c(x, y)u = f(x, y), \quad (x, y) \in \Omega; \\ u(x, y) = g(x, y), \quad (x, y) \in \Gamma,\end{gathered} \tag{5.2}$$

где $\Gamma = \bar{\Omega}\backslash\Omega$. Предполагается, что функции a, b, c, f, g – достаточно гладкие и угловые пограничные слои отсутствуют:

$$a(x) \geq \alpha > 0, \quad b(y) \geq \beta > 0, \quad c(x, y) \geq 0, \quad \varepsilon > 0.$$

В соответствии с [20] решение задачи (5.2) представимо в виде (5.1) при

$$\Phi(x) = e^{-a(0)x/\varepsilon}, \quad \Theta(y) = e^{-b(0)y/\varepsilon}.$$

Зададим сетку $\Omega^h = \Omega_x^{h_1} \times \Omega_y^{h_2}$ в исходной области $\bar{\Omega}$:

$$\Omega_x^{h_1} = \{x_i : x_i = a + (i-1)h_1,\ i = 1, 2, \ldots, k_1\}, \quad h_1 = (b-a)/(k_1 - 1),$$

$$\Omega_y^{h_2} = \{y_j : y_j = c + (j-1)h_2,\ j = 1, 2, \ldots, k_2\}, \quad h_2 = (d-c)/(k_2 - 1), \quad k_1 \geq 2, \quad k_2 \geq 2.$$

Построим интерполяционную формулу для функций вида (5.1), точную на погранслойных составляющих.

Сначала при заданном значении y в соответствии с (2.1) зададим интерполяцию по x :

$$L_x(u, x, y) = L_{k_1-1}(u, x, y) + \frac{[x_1, x_2, \ldots, x_{k_1}]u}{[x_1, x_2, \ldots, x_{k_1}]\Phi}[\Phi(x) - L_{k_1-1}(\Phi, x)]. \tag{5.3}$$

В (5.3) $L_{k_1-1}(u, x, y)$ соответствует интерполяции по x функции $u(x, y)$ многочленом Лагранжа с узлами интерполяции $x_1, x_2, \ldots, x_{k_1-1}$ при заданном y.

По аналогии с (5.3) зададим интерполяционную формулу по y:

$$L_y(u, x, y) = L_{k_2-1}(u, x, y) + \frac{[y_1, y_2, \ldots, y_{k_2}]u}{[y_1, y_2, \ldots, y_{k_2}]\Theta}[\Theta(y) - L_{k_2-1}(\Theta, y)]. \tag{5.4}$$

Используя (5.4), после интерполяции по x осуществляем интерполяцию по y:

$$L_{\Phi,\Theta,k_1,k_2}(u, x, y) = L_{k_2-1}(L_x(u, x, y), x, y) + \frac{[y_1, y_2, \ldots, y_{k_2}]L_x(u, x, y)}{[y_1, y_2, \ldots, y_{k_2}]\Theta}[\Theta(y) - L_{k_2-1}(\Theta, y)]. \tag{5.5}$$

Итак, построена двумерная интерполяционная формула (5.3)–(5.5).

Формула (5.3)–(5.5) задана корректно, если знаменатель в (5.3) и (5.5) не обращается в нуль. В соответствии с соотношением (2.15) это условие выполняется, если

$$\Phi^{(k_1-1)}(x) \neq 0, \quad x \in (a, b), \quad \Theta^{(k_2-1)}(y) \neq 0, \quad y \in (c, d). \tag{5.6}$$

Несложно получить, что двумерная интерполяционная формула (5.3)–(5.5) является точной на функциях

$$x^i, \quad x^i\Theta(y), \quad i = 0, 1, \ldots, k_1 - 2, \quad y^j, \quad y^j\Phi(x), \quad j = 0, 1, \ldots, k_2 - 2, \quad \Phi(x)\Theta(y).$$

Формула (5.3)–(5.5) исследовалась в [21], где доказана следующая лемма.

Лемма 5. *Пусть выполнены условия* (5.6). *Тогда для некоторой постоянной C, не зависящей от функций $\Phi(x)$, $\Theta(y)$ и их производных, справедлива оценка погрешности*

$$\left|u(x,y) - L_{\Phi,\Theta,k_1,k_2}(u,x,y)\right| \le C(1+\max_x\left|M_{k_1}(\Phi,x)\right|)(1+\max_y\left|M_{k_2}(\Theta,y)\right|)[h_1^{k_1-1}+h_2^{k_2-1}], \tag{5.7}$$

где $(x,y)\in\bar{\Omega}$, $M_{k_1}(\Phi,x)$, $M_{k_2}(\Theta,y)$ *определяются согласно* (2.7).

Оценка погрешности (5.7) зависит от погранслойных составляющих $\Phi(x)$, $\Theta(y)$. Улучшим эту оценку.

Лемма 6. *Пусть*

$$\Phi^{(k_1-1)}(x)\ne 0,\quad \Phi^{(k_1)}(x)\ne 0,\quad k_1\ge 2,\quad x\in(a,b), \tag{5.8}$$

$$\Theta^{(k_2-1)}(y)\ne 0,\quad \Theta^{(k_2)}(y)\ne 0,\quad k_2\ge 2,\quad y\in(c,d). \tag{5.9}$$

Тогда

$$\left|u(x,y) - L_{\Phi,\Theta,k_1,k_2}(u,x,y)\right| \le 4C[h_1^{k_1-1}+h_2^{k_2-1}],\quad x\in[a,b],\quad y\in[c,d]. \tag{5.10}$$

Доказательство. По аналогии с леммой 2, где обоснована оценка (2.16) при выполнении условий (2.10), получаем, что при выполнении условий (5.8) и (5.9) справедливы оценки

$$\left|M_{k_1}(\Phi,x)\right|\le 1,\quad \left|M_{k_2}(\Theta,y)\right|\le 1.$$

Теперь из (5.7) следует (5.10), что доказывает лемму.

Замечание 2. Если исходную область $[a,b]\times[c,d]$ разбить на непересекающиеся прямоугольные ячейки с k_1 узлами по x и k_2 узлами по y, то при интерполяции функции $u(x,y)$ можно применить интерполяционную формулу (5.3)–(5.5) в каждой ячейке. Если ячейка не пересекается с областью больших градиентов функции $u(x,y)$, то в ней можно применять классические интерполяционные формулы, основанные на многочленах Лагранжа.

6. РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Рассмотрим функцию вида (1.1):

$$u(x)=\cos(\pi x)+e^{-x/\varepsilon},\quad x\in[0,1],\quad \varepsilon>0.$$

При этом $\Phi(x)=e^{-x/\varepsilon}$. Предполагаем, что число сеточных интервалов N четно и разобьем интервал $[0,1]$ на непересекающиеся интервалы вида $[x_{n-1},x_{n+1}]$, где $n=1,3,\ldots,N-1$. На каждом таком интервале зададим интерполяционную формулу (2.1) при $k=3$:

$$L_{\Phi,3}(u,x)=u_{n-1}+\frac{u_n-u_{n-1}}{h}(x-x_{n-1})+\frac{u_{n+1}-2u_n+u_{n-1}}{\Phi_{n+1}-2\Phi_n+\Phi_{n-1}}\times$$
$$\times\left[\Phi(x)-\Phi_{n-1}-\frac{\Phi_n-\Phi_{n-1}}{h}(x-x_{n-1})\right],\quad x\in[x_{n-1},x_{n+1}].$$

Зададим погрешность интерполяции многочленом Лагранжа

$$\Delta_{\varepsilon,N}=\max_{1\le n\le N}\left|u(\tilde{x}_n)-L_3(u,\tilde{x}_n)\right|,\quad \tilde{x}_n=(x_n+x_{n-1})/2.$$

В табл. 1 приведена погрешность интерполяции $\Delta_{\varepsilon,N}$ многочленом Лагранжа $L_3(u,x)$ в зависимости от ε и N. При малых значениях ε погрешность не убывает при уменьшении шага сетки. Это подтверждает неприемлемость применения для интерполяции многочлена Лагранжа на равномерной сетке при наличии пограничного слоя.

В табл. 2 аналогичным образом представлена погрешность интерполянта $L_{\Phi,3}(u,x)$ и вычисленный порядок точности $M_{\varepsilon,N}=\log_2(\Delta_{\varepsilon,N}/\Delta_{\varepsilon,2N})$. Из табл. 2 следует, что порядок точности интерполяционной формулы понижается с 3 до 2 при уменьшении ε, результаты вычислений согласуются с оценкой (2.20) при $k=3$.

Другие результаты вычислений по всем исследуемым вопросам содержатся в публикациях авторов, приведенных в списке литературы настоящей статьи, и согласуются с полученными в данной работе оценками погрешностей.

Таблица 1. Погрешность интерполяции многочленом Лагранжа $L_3(u, x)$

ε	N					
	3×2^3	3×2^4	3×2^5	3×2^6	3×2^7	3×2^8
1	1.36*e* – 4	1.72*e* – 5	2.15*e* – 6	2.68*e* – 7	3.36*e* – 8	4.19*e* – 9
10^{-1}	3.15*e* – 3	4.71*e* – 4	6.45*e* – 5	8.43*e* – 6	1.08*e* – 6	1.36*e* – 7
10^{-2}	2.62*e* – 1	1.14*e* – 1	3.00*e* – 2	5.68*e* – 3	8.82*e* – 4	1.23*e* – 4
10^{-3}	3.75*e* – 1	3.75*e* – 1	3.70*e* – 1	3.05*e* – 1	1.58*e* – 1	4.82*e* – 2
10^{-4}	3.75*e* – 1	3.75*e* – 1	3.75*e* – 1	3.75*e* – 1	3.75*e* – 1	3.74*e* – 1

Примечание. Здесь и в табл. 2 $e - m$ обозначает 10^{-m}.

Таблица 2. Погрешность и вычисленный порядок точности интерполянта $L_{\Phi,3}(u, x)$

ε	N					
	3×2^3	3×2^4	3×2^5	3×2^6	3×2^7	3×2^8
1	1.47*e* – 4	1.84*e* – 5	2.30*e* – 6	2.87*e* – 7	3.59*e* – 8	4.49*e* – 9
	3.0	3.0	3.0	3.0	3.0	
10^{-1}	4.87*e* – 4	6.00*e* – 5	7.40*e* – 6	9.19*e* – 7	1.15*e* – 7	1.43*e* – 8
	3.0	3.0	3.0	3.0	3.0	
10^{-2}	4.61*e* – 3	6.34*e* – 4	7.69*e* – 5	9.23*e* – 6	1.12*e* – 6	1.38*e* – 7
	2.9	3.0	3.0	3.0	3.0	
10^{-3}	6.38*e* – 3	1.60*e* – 3	3.96*e* – 4	8.26*e* – 5	1.23*e* – 5	1.52*e* – 6
	2.0	2.0	2.3	2.7	3.0	
10^{-4}	6.38*e* – 3	1.60*e* – 3	4.01*e* – 4	1.00*e* – 4	2.51*e* – 5	6.25*e* – 6
	2.0	2.0	2.0	2.0	2.0	
10^{-5}	6.38*e* – 3	1.60*e* – 3	4.01*e* – 4	1.00*e* – 4	2.51*e* – 5	6.27*e* – 6
	2.0	2.0	2.0	2.0	2.0	

7. ЗАКЛЮЧЕНИЕ

Исследована неполиномиальная интерполяционная формула для функции одной переменной с большими градиентами в области пограничного слоя. Формула построена так, чтобы она была точной на погранслойной составляющей, отвечающей за рост функции в пограничном слое. Доказано, что при достаточно легко проверяемых ограничениях, которые выполнены в случаях экспоненциального и степенного пограничных слоев, при наличии логарифмической особенности построенная интерполяционная формула обладает погрешностью, равномерной по погранслойной составляющей и ее производным. Показано, как исследуемая интерполяционная формула может быть применена для построения формул численного интегрирования и дифференцирования, а также в двумерном случае. Получены соответствующие оценки погрешности.

СПИСОК ЛИТЕРАТУРЫ

1. *Задорин А.И.* Метод интерполяции для задачи с пограничным слоем // Сиб. ж. вычисл. матем. 2007. Т. 10. № 3. С. 267–275.
2. *Шишкин Г.И.* Сеточные аппроксимации сингулярно возмущенных эллиптических и параболических уравнений. Екатеринбург: УрО РАН, 1992.
3. *Бахвалов Н.С.* К оптимизации методов решения краевых задач при наличии пограничного слоя // Ж. вычисл. матем. и матем. физ. 1969. Т. 9. № 4. С. 841–859.

4. *Linβ T.* Layer-Adapted Meshes for Reaction-Convection-Diffusion Problems. Berlin: Springer-Verlag, 2010.
5. *Задорин А.И.* Интерполяция Лагранжа и формулы Ньютона–Котеса для функций с погранслойной составляющей на кусочно-равномерных сетках // Сиб. ж. вычисл. матем. 2015. Т. 18. № 3. С. 289–303.
6. *Zadorin A.I.* Interpolation method for a function with a singular component // Lect. Notes in Comput Sci. 2009. V. 5434. P. 612–619.
7. *Kellogg R.B., Tsan A.* Analysis of some difference approximations for a singular perturbation problems without turning points // Math. Comput. 1978. V. 32. P. 1025–1039.
8. *Zadorin A.I., Zadorin N.A.* Interpolation formula for functions with a boundary layer component and its application to derivatives calculation // Sib. Electronic Math. Reports. 2012. V. 9. P. 445–455.
9. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. М.: Наука, 1987.
10. *Задорин А.И., Задорин Н.А.* Квадратурные формулы для функций с погранслойной составляющей // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. № 11. С. 1952–1962.
11. *Задорин А.И., Задорин Н.А.* Аналог формулы Ньютона–Котеса с четырьмя узлами для функции с погранслойной составляющей // Сиб. ж. вычисл. матем. 2013. Т. 16. № 4. С. 313–323.
12. *Zadorin A., Zadorin N.* Quadrature formula with five nodes for functions with a boundary layer component // Lect. Notes in Comput. Sci. 2013. V. 8236. P. 540–546.
13. *Задорин А.И., Задорин Н.А.* Аналог формул Ньютона–Котеса для численного интегрирования функций с погранслойной составляющей // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 3. С. 368–376.
14. *Shishkin G.I.* Approximations of solutions and derivatives for a singularly perturbed elliptic convection-diffusion equations // Mathematical Proceedings of the Royal Irish Academy. 2003. V. 103A. P. 169–201.
15. *Kopteva N.* Error expansion for an upwind scheme applied to a two-dimensional convection-diffusion problem // SIAM Journal on Numerical Analysis. 2003. V. 41. P. 1851–1869.
16. *Gracia J.L., O'Riordan E.* Numerical approximation of solution derivatives of singularly perturbed parabolic problems of convection-diffusion type // Mathematics of Computation. 2016. V. 85. P. 581–599.
17. *Задорин А.И.* Анализ формул численного дифференцирования на сетке Шишкина при наличии пограничного слоя // Сиб. ж. вычисл. матем. 2018. Т. 21. № 3. С. 243–254.
18. *Zadorin A., Tikhovskaya S.* Formulas of numerical differentiation on a uniform mesh for functions with the exponential boundary layer // Int. J. of Num. Analysis and Modeling. 2019. V. 16. № 4. P. 590–608.
19. *Il'in V.P., Zadorin A.I.* Adaptive formulas of numerical differentiation of functions with large gradients // J. of Physics: Conference Series. 2019. V. 1260. 042003.
20. *Roos H.G., Stynes M., Tobiska L.* Numerical Methods for Singularly Perturbed Differential Equations, Convection-Diffusion and Flow Problems. Berlin: Springer, 2008.
21. *Задорин А.И.* Интерполяция функции двух переменных с большими градиентами в пограничных слоях // Ученые записки Казанского университета. Физ.-мат. науки. 2015. Т. 157. Кн. 2. С. 55–67.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 189–205

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 519.68

МНОГОМЕТОДНЫЕ АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ СЛОЖНЫХ ЗАДАЧ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ

© 2021 г. А. И. Тятюшкин

664033 Иркутск, ул. Лермонтова, 134, Институт динамики систем и теории управления им. В.М. Матросова СО РАН, Россия

e-mail: tjat@icc.ru

Поступила в редакцию 19.12.2019 г.
Переработанный вариант 20.08.2020 г.
Принята к публикации 16.09.2020 г.

Рассматриваются задачи оптимального управления с терминальными условиями без ограничений на управление, задачи со свободным правым концом траектории с ограничениями на управление и задачи оптимизации с параметрами при ограничениях на параметры и управление. Для каждого из этих классов задач конструируются многометодные алгоритмы, состоящие из наиболее эффективных для данных классов численных методов оптимального управления. Работа предложенных алгоритмов подтверждается численным решением сложных прикладных задач. Библ. 14. Фиг. 4.

Ключевые слова: многометодные алгоритмы оптимизации, оптимальное управление, задачи с параметрами, градиентные методы, принцип максимума, численные методы, итерации.

DOI: 10.31857/S0044466921020137

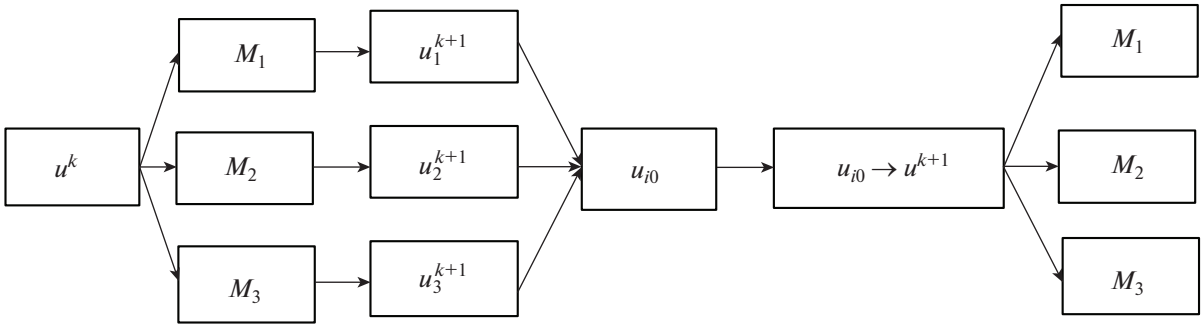
ВВЕДЕНИЕ

Многометодная технология решения задач оптимального управления заключается в параллельном использовании сразу нескольких итерационных методов оптимизации для поиска решения одной и той же задачи. Основной проблемой применения многометодной технологии при численном решении задач оптимального управления является выбор метода для эффективного продолжения процесса оптимизации с того момента, когда ухудшилась сходимость текущего метода. Современные операционные системы позволяют обеспечить решение задачи путем организации параллельных вычислительных потоков для одновременного проведения расчетов несколькими методами. В каждом таком потоке можно реализовывать итерационный процесс одного из методов оптимизации, и решение одной задачи вести несколькими методами одновременно. На многопроцессорных компьютерах для реализации каждого метода удобнее использовать отдельный процессор. После нахождения очередного приближения все методы оцениваются, например, по полученному приращению функционала и из них выбирается наиболее эффективный метод для продолжения оптимизации, а полученное этим методом приближение передается остальным методам в качестве начального для выполнения следующей итерации.

Продолжая итерационный процесс до получения приближения, на котором с заданной точностью будет выполнен критерий оптимальности, найдем приближенное решение задачи. При этом решение будет найдено многометодным алгоритмом, состоящим из последовательности шагов разных методов, подключаемых к процессу оптимизации с целью ускорения его сходимости. Например, в случае параллельного использования трех методов (см. фиг. 1) лучшее приближение будет определяться по максимуму приращения функционала, полученного на данной итерации каждым из трех методов:

$$u_{i_0} = \operatorname*{argmax}_{i\in\{1,2,3\}} (I(u_i^k) - I(u_i^{k-1})).$$

Затем это приближение передается всем трем методам для выполнения следующей итерации: $u_i^{k+1} = u_{i_0}$, $i = 1, 2, 3$.



Фиг. 1. Схема выполнения $(k+1)$-й итерации многометодным алгоритмом для группы из трех методов M_1, M_2, M_3.

Таким образом, многометодная технология решения прикладных задач оптимального управления, реализованная в виде параллельных итерационных процессов оптимизации с выбором лучшего приближения, находит решение задачи с автоматическим применением разных методов оптимизации и тем самым существенно повышает эффективность поиска и надежность получения численного решения с заданной точностью в прикладных задачах оптимального управления.

1. ЗАДАЧА БЕЗ ОГРАНИЧЕНИЙ НА УПРАВЛЕНИЕ

Рассмотрим сначала задачу оптимального управления (см. [1]) с условиями типа равенств и без ограничений на управление

$$\dot{x} = f(x,u,t), \quad t \in T = [t_0,t_1], \quad x(t) \in R^n, \quad u(t) \in R^r, \quad x(t_0) = x^0, \tag{1.1}$$

$$I_0(u) \to \min, \tag{1.2}$$

$$I_j(u) = 0, \quad j = \overline{1,m}, \tag{1.3}$$

где

$$I_j(u) = \varphi^j(x(t_1)) + \int_{t_0}^{t_1} F^j(x,u,t)dt, \quad m \le n.$$

Градиенты функционалов (1.2), (1.3)

$$\nabla I_j(u) = -H_u^j(\psi,x,u,t), \tag{1.4}$$

где (функция Понтрягина из [1]) $H^j(\psi,x,u,t) = \psi_j'(t)f(x,u,t) - F^j(x,u,t)$; $\psi_j(t)$ – решение сопряженной системы

$$\dot{\psi}_j = -f_x'(x,u,t)\psi_j + F_x^j(x,u,t), \quad \psi(t_1) = -\psi_x^j(x(t_1)). \tag{1.5}$$

Рассмотрим численный метод решения задачи (1.1)–(1.3), основанный на применении первой и второй вариаций. На первой фазе этого метода, когда итерационный процесс реализуется с сильным нарушением терминальных условий, минимизируется штрафной функционал

$$J(u) = \varphi(x(t_1)) + \int_{t_0}^{t_1} F^0(x,u,t)dt, \tag{1.6}$$

где

$$\varphi(x(t_1)) = \varphi^0(x(t_1)) + \sum_{j=1}^{m} K_j[\varphi^j(x(t_1)) + x_{n+j}(t_1)]^2,$$

$K_j \ge 0$, $x_{n+j}(t_1)$, $j = \overline{1,m}$ – решения дополнительных к системе (1.1) уравнений

$$\dot{x}_{n+j} = F^j(x,u,t), \quad x_{n+j}(t_0) = 0, \quad j = \overline{1,m}.$$

После прекращения сходимости первой фазы выполняется вторая фаза метода, на итерациях которой минимизируется исходный функционал (1.2), а вариация $\delta u(t)$ строится уже с учетом линеаризованных краевых условий.

Пусть теперь $H(\psi, x, u, t) = \psi' f(x, u, t) - F^0(x, u, t)$,

$$\dot{\psi} = -H_x(\psi, x, u, t), \quad \psi(t_1) = -\varphi_x(x(t_1)) \tag{1.7}$$

и для заданного управления $u^k(t)$ найдены решения x^k и ψ^k систем (1.1) и (1.7). Тогда проблему построения подходящей вариации $\delta u^k(t)$ сформулируем в виде следующей линейно-квадратичной задачи:

$$I(\delta u) = \frac{1}{2}\delta x'(t_1)\varphi_{xx}(x^k(t_1))\delta x(t_1) - \int_{t_0}^{t_1} H_u'\delta u dt - \frac{1}{2}\int_{t_0}^{t_1}[\delta u' H_{uu}\delta u + 2\delta u' H_{ux}\delta x + \delta x' H_{xx}\delta x]dt \to \min, \tag{1.8}$$

$$\delta\dot{x} = f_x(x^k, u^k, t)\delta x + f_u(x^k, u^k, t)\delta u, \quad \delta x(t_1) = 0. \tag{1.9}$$

Здесь H_u, H_{uu}, H_{ux}, H_{xx} – $r\times 1$-, $r\times r$-, $r\times n$-, $n\times n$-матрицы частных производных функции H, вычисленные на управлении $u^k(t)$ и траекториях $x^k(t)$, $\psi^k(t)$.

Для этой задачи построим гамильтониан

$$\mathcal{H}(\delta\psi, \delta x, \delta u, t) = \delta\psi' f_x\delta x + \delta\psi' f_u\delta u + H_u'\delta u + \frac{1}{2}(\delta u' H_{uu}\delta u + 2\delta u' H_{ux}\delta x + \delta x' H_{xx}\delta x)$$

и сопряженную систему

$$\delta\dot{\psi} = -f_x'\delta\psi - H_{ux}'\delta u - H_{xx}\delta x, \quad \delta\psi(t_1) = -\varphi_{xx}\delta x(t_1). \tag{1.10}$$

Из условия $\mathcal{H}_{\delta u} = 0$ найдем решение вариационной задачи (1.8), (1.9):

$$\delta u = -H_{uu}^{-1}(f_u'\delta\psi + H_{ux}\delta x + H_u). \tag{1.11}$$

Подставив последнюю формулу в уравнения (1.9), (1.10), получим следующую линейную двухточечную краевую задачу:

$$\delta\dot{x} = C\delta x - f_u H_{uu}^{-1} f_u'\delta\psi - f_u H_{uu}^{-1} H_u, \tag{1.12}$$

$$\delta\dot{\psi} = -(H_{xx} - H_{ux}' H_{uu}^{-1} H_{ux})\delta x - C'\delta\psi + H_{ux} H_{uu}^{-1} H_u, \tag{1.13}$$

где

$$\begin{gathered} C = f_x - f_u H_{uu}^{-1} H_{ux}, \\ \delta x(t_0) = 0, \quad \delta\psi(t_1) = -\varphi_{xx}\delta x(t_1). \end{gathered} \tag{1.14}$$

Стандартный способ решения задачи состоит в применении формулы Коши, устанавливающей связь между краевыми условиями с помощью переходной матрицы $\Phi(t_0, t_1)$ размерности $2n\times 2n$:

$$\begin{pmatrix} \delta x(t_1) \\ \delta\psi(t_1) \end{pmatrix} = \Phi(t_0, t_1)\begin{pmatrix} \delta x(t_0) \\ \delta\psi(t_0) \end{pmatrix} + \begin{pmatrix} \rho(t_1) \\ \eta(t_1) \end{pmatrix},$$

где $(\rho(t), \eta(t))$ – решение задачи Коши для системы (1.12), (1.13) при $\delta x(t_1) = \delta x_0 = 0$, $\delta\psi(t_0) = \delta\psi_0 = 0$.

Матрицу $\Phi(t_0, t_1)$ разобьем на четыре равных блока и, учитывая равенство $\delta x(t_0) = 0$, последнее уравнение перепишем в следующем виде:

$$\begin{aligned} \delta x(t_1) &= \Phi_{12}\delta\psi_0 + \rho(t_1), \\ \delta\psi(t_1) &= \Phi_{22}\delta\psi_0 + \eta(t_1). \end{aligned} \tag{1.15}$$

Подставляя краевое условие (1.14) в систему (1.15), получаем уравнение для определения начальных условий $\delta\psi_0$:

$$(\Phi_{22} + \varphi_{xx}\Phi_{12})\delta\psi_0 = -\varphi_{xx}\rho(t_1) - \eta(t_1). \tag{1.16}$$

Блоки Φ_{12} и Φ_{22} можно вычислить, проинтегрировав матричное уравнение

$$\dot{\Phi} = A(t)\Phi, \quad \Phi(t_0, t_0) = E,$$

где $A(t)$ есть $2n\times 2n$-матрица коэффициентов системы (1.12), (1.13).

Рассмотрим другой способ, требующий в два раза меньше процессорного времени, который состоит в следующем.

Полагая $\delta x(t_0) = 0$, $\delta\psi(t_0) = e^i$, $i = \overline{1,n}$ (e^i – векторы стандартного ортонормированного базиса), n раз интегрируем $2n$-мерную систему (1.12), (1.13). Каждый из полученных в результате интегрирования векторов возьмем в качестве $(n+i)$-го столбца матрицы Φ; в результате получим ее блоки Φ_{12} и Φ_{22}.

Решив систему линейных алгебраических уравнений (1.16), определим вектор $\delta\psi_0$. Затем, интегрируя систему (1.12), (1.13) в прямом времени, на решении $(\delta x(t), \delta\psi(t))$ по формуле (1.11) найдем искомую вариацию $\delta u(t)$. Формула (1.11) применима только при достаточно хорошей обусловленности матрицы H_{uu}. Чтобы сохранить вычислительную устойчивость метода для общего случая, вместо H_{uu} на практике применяется матрица $H_{uu} + W$, где W – положительно-определенная матрица. Найденная вариация используется для построения нового приближения $u^k + \alpha_k \delta u^k$ при $\alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}} J(u^k + \alpha \delta u^k)$. Итерационный процесс первой фазы метода прекращается при выполнении неравенства $J(u^k) - J(u^{k+1}) \leq \varepsilon$, $\varepsilon \geq 0$. Поскольку при этом минимизировался штрафной функционал (1.6), требуемая точность выполнения краевых условий может оказаться не достигнутой, тогда выполняется вторая фаза метода, которая учитывает линеаризованные краевые условия (1.3) при решении задачи (1.12), (1.13).

Предположим, что с помощью введения дополнительных уравнений в систему (1.1), краевые условия (1.3) сведены к следующему виду:

$$\varphi^i(x(t_1)) = 0, \quad i = \overline{1,m}, \quad m \leq n, \tag{1.17}$$

или $\varphi(x(t_1)) = 0$, где $\varphi = (\varphi^1, \varphi^2, \ldots, \varphi^m)'$, а функционал

$$I_0(u) = \varphi^0(x(t_1)).$$

Линеаризуем краевые условия (1.17) в окрестности точки $x^k(t_1)$:

$$\varphi(x^k(t_1)) + \varphi_x(x^k(t_1))\delta x(t_1) = 0. \tag{1.18}$$

Часть компонент $\delta x_i^f, i = \overline{1,m}$, вектора $\delta x(t_1)$ связана соотношениями (1.18), а оставшиеся $n - m$ свободных компонент должны удовлетворять терминальным условиям

$$\delta\psi^c(t_1) = -\varphi^0_{xx}(x^k(t_1))\delta x^0(t_1). \tag{1.19}$$

Разобьем введенные выше матрицы Φ_{12} и Φ_{22} на блоки в соответствии с составляющими векторов $\delta x(t_1) = (\delta x^f, \delta x^c)$, $\delta\psi(t_1) = (\delta\psi^f, \delta\psi^c)$ и перепишем уравнения (1.15) в следующем виде:

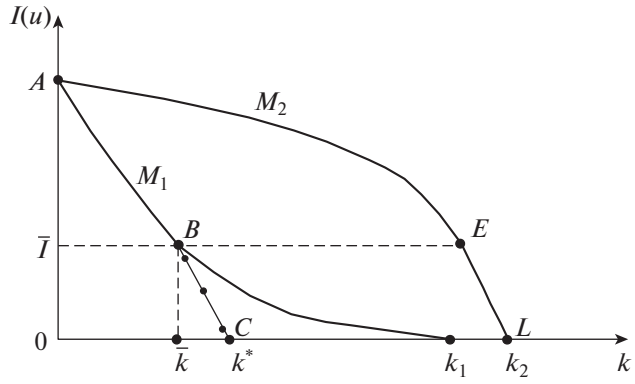
$$\begin{gathered}
\Phi_{12}^{11}\delta\psi_0^f + \Phi_{12}^{12}\delta\psi_0^c + \rho^f = \delta x^f, \\
\Phi_{12}^{21}\delta\psi_0^f + \Phi_{12}^{22}\delta\psi_0^c + \rho^c = \delta x_1^c, \\
\Phi_{12}^{11}\delta\psi_0^f + \Phi_{12}^{12}\delta\psi_0^c + \eta^f = \delta\psi_1^f, \\
\Phi_{12}^{21}\delta\psi_0^f + \Phi_{12}^{22}\delta\psi_0^c + \eta^c = \delta\psi_1^c.
\end{gathered}$$

Подставив в эти уравнения краевые условия (1.18), (1.19), получим систему линейных алгебраических уравнений относительно вектора начальных условий $\delta\psi_0 = (\delta\psi_0^f, \delta\psi_0^c)$:

$$\begin{gathered}
(\Phi_{22}^{21} + \varphi^0_{xx}\Phi_{12}^{21})\delta\psi_0^f + (\Phi_{22}^{22} + \varphi^0_{xx}\Phi_{12}^{22})\delta\psi_0^c = -\varphi^0_{xx}\rho^0 - \eta^c, \\
(\varphi_{x^f}\Phi_{12}^{11} + \varphi_{x^c}\Phi_{12}^{21})\delta\psi_0^f + (\varphi_{x^f}\Phi_{12}^{12} + \varphi_{x^c}\Phi_{12}^{22})\delta\psi_0^c = -\varphi_x\rho - \varphi,
\end{gathered} \tag{1.20}$$

где

$$\varphi^0_{xx} = \varphi^0_{xx}(x^k(t_1)), \quad \varphi_x = (\varphi_{x^f}, \varphi_{x^c}) = \varphi_x(x^k(t_1)), \quad \varphi = \varphi(x^k(t_1)).$$



Фиг. 2. Убывание функционала на итерациях методов M_1, M_2 и l многометодного алгоритма.

Следовательно, вариация $\delta u(t)$, построенная на решении $(\delta x(t), \delta\psi(t))$ системы (1.12), (1.13) с начальными условиями $(0, \delta\psi_0)$, будет выполнять в линейном приближении краевые условия (1.17) и одновременно минимизировать квадратичное приближение функционала (1.2).

Алгоритм 1. Приведем вычислительную схему, включающую обе фазы описанного метода. На итерациях первой фазы выполняются следующие операции.

1. При заданном управлении $u^k(t)$ интегрируется уравнение (1.1), в узлах интегрирования запоминается траектория $x^k(t)$.

2. В обратном времени интегрируется уравнение (1.7), в узлах интегрирования запоминается решение $\psi^k(t)$.

3. С коэффициентами, вычисленными на управлении $u^k(t)$ и решениях $x^k(t)$, $\psi^k(t)$, $n+1$ раз интегрируется система (1.12), (1.13) при начальных условиях

$$\delta x(t_0) = 0, \quad \delta\psi(t_0) = e^i, \quad i = \overline{1,n}; \quad \delta x(t_0) = 0, \quad \delta\psi(t_0) = 0;$$

$(n+j)$-е столбцы матрицы Φ принимаются равными $(\delta x'(t_1), \delta\psi'(t_1))'$, $j = \overline{1,n}$.

4. С помощью блоков Φ_{12} и Φ_{22} матрицы Φ формируется система (1.16) и находится ее решение $\delta\psi^0$.

5. Интегрируется система (1.12), (1.13) при $\delta x(t_0) = 0$, $\delta\psi(t_0) = \delta\psi^0$, в каждом узле интегрирования по формуле (1.11) вычисляется и запоминается вариация $\delta u^k(t)$.

6. С помощью процедуры одномерного поиска за ν задач Коши (1.1) находится параметр $\alpha_k = \operatorname{argmin}_{\alpha\geq 0} J(u^k + \alpha\delta u^k)$.

7. Строится новое приближение $u^{k+1}(t) = u^k(t) + \alpha_k\delta u^k(t)$, $t \in T$, и если $J(u^k) - J(u^{k+1}) > \varepsilon$, то повторяется итерация при $k = k+1$ с п. 1. В противном случае выполняется вторая фаза алгоритма.

Основное отличие второй фазы алгоритма состоит в том, что вместо системы (1.16) для нахождения вектора $\delta\psi^0$ применяется система (1.20). На итерациях второй фазы выполняются те же операции 1–7, кроме 4, где вместо (1.16) формируется система (1.20) и операции 2, где вместо (1.7) интегрируется уравнение (1.5) при $j = 0$.

Вторая фаза описанного алгоритма 1 по сути является вариантом метода квазилинеаризации (см. [2]), обладающего квадратичной скоростью сходимости, но требующего достаточно хорошее начальное приближение, которое в описанном алгоритме обеспечивается первой фазой. Таким образом, наиболее эффективным будет алгоритм, последовательность приближений которого будет построена двумя методами оптимизации, если удастся уловить момент переключения с одного метода на другой.

Графически убывание функционала $I(u)$ на итерациях многометодного алгоритма будет изображаться ломаной линией, составленной из графиков отдельных методов. На фиг. 2 показана работа многометодного алгоритма в случае использования двух методов M_1 и M_2. Приведенные здесь графики показывают убывание функционала на итерациях этих методов. Затем из них со-

ставлен график убывания функционала на итерациях многометодного алгоритма – кривая ABC, участок BC которой получен параллельным переносом кривой EL. Согласно фиг. 2, нулевое значение функционала методом M_1 достигается за k_1 итераций, а методом M_2 – за k_2 итераций. Многометодный алгоритм до $\overline{k}$-й итерации работает по методу M_1 (кривая AB), а далее – по методу M_2 (кривая BC), так как, начиная с $\overline{k}$, скорость убывания функционала по методу M_2 будет выше. В результате нулевое значение функционала многометодным алгоритмом достигается за k^* итераций, т.е. за существенно меньшее число итераций, чем каждым из методов M_1 и M_2.

2. ЗАДАЧИ С ОГРАНИЧЕНИЯМИ НА УПРАВЛЕНИЕ

Для решения наиболее важного с прикладной точки зрения класса задач с ограничениями на управление

$$u(t) \in U, \quad t \in T, \tag{2.1}$$

где U – компактное множество из R^r, построим многометодные алгоритмы, основанные на принципе максимума (см. [3], [4]) и методах градиентного типа (см. [5]–[8]).

Будем предполагать, что с помощью введения штрафного функционала уже имеем задачу со свободным правым концом и что при некотором допустимом управлении $u^k(t) \in U$, $t \in T$, найдено решение уравнения (1.1) $x^k(t)$. Решая уравнение (1.5) при $u = u^k(t)$, $x = x^k(t)$, $j = 0$, найдем $\psi^k = \psi_0(t)$ и вычислим

$$\overline{u}^k(t) = \operatorname*{argmax}_{u \in U} H(\psi^k, x^k, u, t), \quad t \in T. \tag{2.2}$$

Построим скалярную функцию

$$w_k(u(t), t) = H(\psi^k, x^k, u, t) - H(\psi^k, x^k, u^k, t), \tag{2.3}$$

которая при $u = \overline{u}(t)$, очевидно, удовлетворяет неравенству

$$w_k(\overline{u}^k(t), t) \geq 0, \quad t \in T. \tag{2.4}$$

Пусть $\tau_k \in T$ – точка максимума функции $w_k(\overline{u}^k(t), t)$:

$$w_k(\overline{u}^k(\tau_k), \tau_k) = \max_{t \in T} w_k(\overline{u}^k(t), t). \tag{2.5}$$

Тогда необходимое условие первого порядка (принцип максимума (см. [1], [2])) формулируется так: если $u^k(t)$ – оптимальное управление в задаче (1.2), (1.1), (2.1), то

$$w_k(\overline{u}^k(\tau_k), \tau_k) = 0. \tag{2.6}$$

Предположим, что для заданного $u^k(t) \in U$ и найденных $x^k(t)$, $\psi^k(t)$, $\overline{u}^k(t)$ условие (2.6) не выполняется:

$$w_k(\overline{u}^k(\tau_k), \tau_k) > 0.$$

Тогда можно найти новое управление, на котором значение функционала (1.2) будет меньше, чем $I_0(u^k)$.

Построим отрезок $T_\varepsilon^k \subseteq T$ по следующему правилу:

$$T_\varepsilon^k = [\tau_k - \varepsilon(\tau_k - t_0^k), \tau_k + \varepsilon(t_1^k - \tau_k)], \quad \varepsilon \in [0,1], \tag{2.7}$$

где t_0^k и t_1^k – ближайшие слева и справа точки разрыва функции $w_k(\overline{u}^k(t), t)$. Этот отрезок с мерой $\operatorname{mes} T_\varepsilon^k = \alpha_k(\varepsilon) = \alpha_k \varepsilon$, $0 \leq a_k \leq t_1 - t_0$, $\varepsilon \in [0,1]$, обладает следующими свойствами:

1) $\alpha_k(\varepsilon) \to 0$, когда $\varepsilon \to 0$;

2) при $\varepsilon \to 0$ он стягивается в точку τ_k;

3) при всех $\varepsilon \in [0,1]$ функция $w_k(\overline{u}^k(t), t)$ непрерывна на T_ε.

Далее находим параметр

$$\varepsilon_k = \operatorname*{argmin}_{\varepsilon\in[0,1]} I_0(u_\varepsilon^k), \tag{2.8}$$

где

$$u_\varepsilon^k = \begin{cases} \overline{u}^k(t), & t \in T_\varepsilon, \\ u^k(t), & t \in T\setminus T_\varepsilon, \end{cases} \tag{2.9}$$

и определяем новое приближение

$$u^{k+1}(t) = u_{\varepsilon_k}^k(t), \quad t \in T, \quad k = 0,1,\ldots. \tag{2.10}$$

Алгоритм 2. Построим вычислительную схему описанного метода, сходимость которого доказана в [4].

1. Задаем граничное управление $u^0(t), t \in T$; полагаем $k = 1$.

2. В прямом времени интегрируем систему (1.1) при $u = u^k(t)$, запоминая в узлах интегрирования $x^k(t)$.

3. В обратном времени интегрируем сопряженную систему (1.5). При этом в каждом узле интегрирования находим и запоминаем управление (2.2), вычисляем значение функции $w_k(\overline{u}^k(t),t)$, определяем точку максимума τ_k.

4. Если $w_k(\overline{u}^k(\tau_k),\tau_k) \le \varepsilon$, то процесс прекращаем. В противном случае выполним операцию 5.

5. С помощью процедуры одномерного поиска решаем задачу (2.8), (2.9), (2.7). Поскольку в точках $t \in T$, удовлетворяющих неравенству $t < \tau_k - \varepsilon(\tau_k - t_0^k)$, управление (2.9) будет равным $u^k(t)$, интегрирование системы (1.1) следует начинать с ближайшего к $\tau_k - \varepsilon(\tau_k - t_0^k)$ левого узла t_k, в котором $x(t_k) = x^k(t_k)$ было найдено в п. 2. Учитывая структуру управления (2.9), в качестве начальной точки можно брать также “наибольший” узел $t_k \in \{t_k \in T_\varepsilon^k : u^k(t_k) = \overline{u}^k(t_k)\}$. Так как при поиске ε_k система (1.1) интегрируется несколько раз, то такой выбор начальной точки может существенно сократить время решения задачи.

6. Если при найденном ε_k, точности δ вычисления $I_0(u^k)$, шаге интегрирования h выполняются неравенства $I_0(u_{\varepsilon_k}^k) \ge I_0(u^k) - \delta$ и $t_1^s - t_0^s > h$, то сокращаем отрезок T_ε^k, полагая $\varepsilon = 2^{-s}$ (s – число сокращений T_ε^k), и вновь выполняем операцию 5.

Если $I_0(u_{\varepsilon_k}^k) < I_0(u^k) - \delta$, то, приняв $u^{k+1} = u_{\varepsilon k}^k$, $k = k+1$, повторяем цикл с п. 2. В противном случае итерационный процесс прекращается.

Поскольку итерации алгоритма идут в классе кусочно-постоянных управлений, на полученном управлении условие оптимальности (2.8) может не выполняться с заданной точностью (хотя величина $\int_T w_k(u^k(t),t)dt$ будет достаточно мала). В отличие от градиентных методов данный алгоритм применим и к таким задачам оптимального управления, в которых вектор-функция $f(x,u,t)$ не дифференцируема по u, а множество U не является выпуклым или даже связным.

Как уже отмечалось в [4], [8], [9], алгоритмы принципа максимума нередко приводят к “прилипанию” управления к границе, что является причиной ухудшения сходимости. Этот эффект обусловлен тем, что в некоторых системах (например, линейных по управлению) решение задачи (1.2) достигается на границе и, следовательно, приближения по управлению вида (2.9), (2.10) также имеют граничные значения. При таком приближении сходимость алгоритма на последних итерациях обеспечивается очень малым шагом дискретизации и влечет большие затраты процессорного времени. Восстановить или улучшить сходимость итерационного процесса в этой ситуации часто удается более простым способом – построением выпуклой комбинации управлений $u^k(t)$ и $\overline{u}^k(t)$:

$$u^{k+1}(t) = u^k(t) + \alpha_k[\overline{u}^k(t) - u^k(t)], \quad \alpha_k \in [0,1], \tag{2.11}$$

на отрезке T_ε^k.

Причиной прекращения сходимости может служить и то обстоятельство, что при сокращении отрезка T_ε^k малое число узлов интегрирования, оставшихся в нем, не обеспечивает построение такой вариации управления, на которой приращение функционала достигает значения, большего (по модулю) величины погрешностей интегрирования.

Рассмотрим численный метод, в котором в качестве нового приближения берется управление, полученное из принципа максимума на множестве

$$T_\varepsilon = \{t \in T : w_k(\bar{u}^k(t), t) \geq \varepsilon w_k(\bar{u}^k(t_k), \tau_k)\}, \quad \varepsilon \in [0,1]. \tag{2.12}$$

Множество T включает все точки $t \in T$ нарушения принципа максимума и состоит из нескольких непересекающихся отрезков, если функция $w_k(\bar{u}^k(t), t)$ многоэкстремальна. Варьируя ε, можно найти такое значение ε_k, при котором управление (2.9) доставит наименьшее значение функционалу. Применение множества (2.12) вместо отрезка (2.7) во многих задачах улучшает сходимость алгоритма 2. В случае, когда максимум $w_k(\bar{u}^k(t), t)$ достигается на множестве T_{ε_k} с положительной мерой, возможно прекращение сходимости алгоритма. Тогда итерации алгоритма 2 продолжаются с выбором T_ε^k по формуле (2.7).

Алгоритм 3. Таким образом, итоговая вычислительная схема для построения нового приближения будет иметь следующий вид.

1. Интегрируется уравнение (1.1) при $u = u^k(t)$, в узлах интегрирования запоминается траектория $x^k(t)$.

2. В обратном времени интегрируется сопряженная система (1.5), в каждом узле интегрирования вычисляются и запоминаются управление $\bar{u}^k(t)$ и скалярная функция

$$w_k(t) = w_k(\bar{u}^k(t), t).$$

3. Находится точка $\tau_k = \operatorname{argmax}_{t \in T} w_k(t)$. Если $w_k(\tau_k) \leq \varepsilon$, то итерационный процесс прекращается.

4. Решается задача одномерного поиска $I(u_\varepsilon^k) \to \min$. При этом для каждого $\varepsilon \in [0,1]$, используемого методом одномерного поиска, интегрирование уравнения (1.1) начинается с узла t_k, в котором впервые выполняется неравенство $w_k(t) > \varepsilon w_k(\tau_k)$, а управление выбирается по формуле

$$u_\varepsilon^k(t) = \begin{cases} \bar{u}^k(t), & \text{если} \quad w_k(t) \geq \varepsilon w_k(\tau_k), \\ u^k(t), & \text{если} \quad w_k(t) < \varepsilon w_k(\tau_k), \quad t \in T. \end{cases}$$

5. Если $I_0(u_{\varepsilon_k}^k) \geq I_0(u^k) - \delta$ и $\operatorname{mes} T_{\varepsilon_k} > h$, выполняется операция 5 алгоритма 2, где T_ε^k строится по формуле (2.7). При этом в качестве первого приближения берется T_{ε_k}.

6. Если итерация алгоритма 2 тоже не улучшает управление $u^k(t)$, то на полученном в п. 4 отрезке $T_{\varepsilon_k}^k$ выполняется итерация метода условного градиента (2.11), где для определения α_k снова применяется одномерный поиск.

7. Если в п. 4–6 будет найдено значение ε_k или α_k, при которых новое приближение для управления обеспечивает меньшее значение функционалу, то выполняется новая итерация (при $k := k + 1$) с п. 1. В противном случае итерационный процесс прекращается.

Заметим, что данный алгоритм, как и описанный выше, ориентирован на решение задач с ограничениями на управление, но со свободным правым концом.

3. ОБЩАЯ ЗАДАЧА ОПТИМАЛЬНОГО УПРАВЛЕНИЯ С ПАРАМЕТРАМИ

Рассмотрим теперь более общую задачу оптимального управления – с фазовыми ограничениями, и когда правая часть системы зависит не только от управлений, но и от параметров. Начальные условия системы также могут зависеть от параметров, и их выбором обычно обеспечивается, например, оптимальный “старт” процесса.

Для решения этой сложной задачи применим сначала редукцию к конечномерной задаче, а затем построим многометодный алгоритм поиска оптимального управления.

Пусть задан управляемый процесс, зависящий от параметров,

$$\dot{x} = f(x,u,w,t), \quad x(t) \in E^n, \quad u(t) \in E^r, \quad t \in T = [t_0,t_1],$$
$$x(t_0) = \Theta(v), \quad w \in R^p, \quad v \in R^n, \tag{3.1}$$

с терминальными условиями

$$I_i(u) = h_i(x(t_1)) = 0, \quad i = \overline{1,m}, \tag{3.2}$$

и фазовыми ограничениями

$$J_i(u,v) = g_i(x(t),t) = 0, \quad t \in T, \quad i = \overline{1,s}. \tag{3.3}$$

Управление и параметры стеснены следующими ограничениями:

$$c_i(u,t) = 0, \quad t \in T, \quad i = \overline{1,l}, \tag{3.4}$$

$$u^{\text{н}}(t) \le u(t) \le u^{\text{в}}(t), \quad t \in T, \tag{3.5}$$

$$v^{\text{н}} \le v \le v^{\text{в}}, \quad w^{\text{н}} \le w \le w^{\text{в}}, \tag{3.6}$$

где $c_i(u,t)$, $i = \overline{1,l}$ – непрерывно дифференцируемые по u и кусочно-непрерывные по t функции; $\Theta(v)$ – непрерывно дифференцируемая вектор-функция. Относительно функций, определяющих условия (3.1)–(3.3), справедливы предположения, оговоренные ранее, к которым добавляется также их непрерывная дифференцируемость по параметрам.

Требуется среди управлений и параметров, удовлетворяющих ограничениям (3.4)–(3.6), найти такие, которые обеспечивают выполнение условий (3.3) для управляемого процесса (3.1) и приводят его в точку фазового пространства, где с заданной точностью будут выполнены условия (3.2), а функционал

$$I_0(u) = \varphi(x(t_1)) \tag{3.7}$$

достигнет наименьшего значения.

3.1. Редукция к конечномерной задаче

Для построения конечномерной задачи на заданном интервале T вводится сетка дискретизации с узлами $t_0, t^1, \ldots, t^N$ такими, что

$$t_0 = t^0 < t^1 < \ldots < t^N = t_1. \tag{3.8}$$

Эта сетка может быть и неравномерной.

Управляющие функции $u^i(t)$, $i = \overline{1,r}$, ищутся только в узлах (3.8), а для получения промежуточных значений $u^i(t)$, $i = \overline{1,r}$, используется либо кусочно-постоянная аппроксимация

$$u^i(t) = u^i(t^j) = u^i_j, \quad t \in [t^j, t^{j+1}],$$

либо кусочно-линейная

$$u^i(t) = [(t^{j+1} - t)u_j + (t - t^j)u_{j+1}](t^{j+1} - t^j), \quad t \in [t^j, t^{j+1}]. \tag{3.9}$$

Тогда конечномерная задача, аппроксимирующая задачу (3.1)–(3.7), будет иметь следующий вид:

$$\dot{x} = f(x,u,w,t), \quad t \in T = [t_0,t_1], \quad x(t_0) = \Theta(v),$$
$$h_i(x(t^N)) = 0, \quad i = \overline{1,m},$$
$$g_i(x(t^j),t^j) = 0, \quad i = \overline{1,s}, \quad j = \overline{0,N},$$
$$c_i(u_j,t^j) = 0, \quad i = \overline{1,l}, \quad j = \overline{0,N}, \tag{3.10}$$
$$v^{\text{н}} \le v \le v^{\text{в}}, \quad w^{\text{н}} \le w \le w^{\text{в}},$$

$$\varphi(x(t^N)) \to \min, \quad u_j^{\text{н}} \le u_j \le u_j^{\text{в}}, \quad j = \overline{0, N},$$

где

$$u_j^{\text{н}} = u^{\text{н}}(t^j), \quad u_j^{\text{в}} = u^{\text{в}}(t^j), \quad j = \overline{0, N}.$$

Заметим, что в аппроксимирующей задаче (3.10) управляемый процесс (3.1) остается непрерывным, а в процессе счета он с требуемой точностью (достаточно высокой) моделируется численным методом интегрирования.

3.2. Численное решение конечномерной задачи

Градиенты функционалов $I_j(u)$, $j = \overline{0, m}$, с помощью функций $H^j(\psi_j, x, u, t) = \psi_j'(t) f(x, u, t)$ и сопряженной системы

$$\dot{\psi}_j = -f_x(x, u, t)'\psi_j(t), \quad \psi_j(t_1) = -\varphi_x^j(x(t_1))$$

традиционно определяются по формулам

$$\nabla I_j(u) = -H_u^j(\psi_j, x, u, t), \quad j = \overline{0, m}.$$

Для каждого $t \in T$ можно аналогично вычислить градиенты $J_j(u, t)$, $j = \overline{1, s}$:

$$\nabla I_j(u, t) = -\bar{H}_u^j(\Phi_j, x, u, t, \tau), \quad t_0 \le \tau \le t \le t_1,$$

где $\bar{H}^j(\Phi_j, x, u, t, \tau) = \Phi_j'(t, \tau) f(x, u, \tau), \Phi_j(t, \tau)$, $j = \overline{1, s}$ – решения сопряженной системы

$$\frac{\partial \Phi_j(t, \tau)}{\partial \tau} = -\frac{\partial f(x, u, \tau)}{\partial x}\Phi_j(t, \tau), \quad \tau \in [t_0, t],$$

с краевыми условиями

$$\Phi_j(t, t) = -\frac{\partial g^j(x(t))}{\partial x}, \quad j = \overline{1, s}.$$

Линеаризуем ограничения в аппроксимирующей задаче. Матрица-якобиан линеаризованных составляется из градиентов $\nabla I_i, i = \overline{1, m}$, и $\nabla J_j(t)$, $j = \overline{1, s}$, $t \in T$. Так как правые части и начальные условия системы (3.1) зависят еще и от параметров, то необходимо иметь также градиенты функционалов $I_i, i = \overline{1, m}$, и $J_j(t)$, $j = \overline{1, s}$, $t \in T$, по этим параметрам (см. [3], [8]):

$$\begin{aligned} \nabla_v I_i(u^k, w^k, v^k) &= -\psi_i(t_0)'\Theta_v(v^k), \quad i = \overline{1, m}, \\ \nabla_w I_i(u^k, w^k, v^k) &= -\int_{t_0}^{t_1} \psi_i(t)' f_w(x^k, u^k, w^k, t)dt, \end{aligned} \tag{3.11}$$

$$\nabla_w J_i(u^k, w^k, v^k, t^j) = -\int_{t_0}^{t^j} \Phi_i(t)' f_w(x^k, u^k, w^k, t)dt, \tag{3.12}$$

$$\nabla_v J_i(u^k, w^k, v^k, t^j) = -\Phi_i(t_0)'\Theta(v^k), \quad i = \overline{1, s}, \quad j = \overline{1, N}. \tag{3.13}$$

Пусть теперь на k-й итерации внешнего метода на сетке (3.8) найдено $u^k(t^j)$ и соответствующее ему $x^k(t^j)$, $j = \overline{1, N}$. Для расчета градиентов по управлению $\nabla_u I_i, i = \overline{1, m}$, система (3.10) m раз интегрируется от t_1 до t_0 с разными начальными условиями. Попутно вычисляются градиенты (3.11) с использованием квадратурных формул для расчета интегралов. Далее ищутся градиенты функционалов $J_i(t^j)$, $j = \overline{1, N}$, $i = \overline{1, s}$. Для этого нужно s раз решить задачу Коши для каждого узла сетки (3.8), т.е. проинтегрировать систему $s \cdot N$ раз в среднем на половине отрезка T.

На полученных решениях вычисляются компоненты градиентов $\nabla_u I_i, i = \overline{1,m}$, и $\nabla_u J_i(t^j)$, $i = \overline{1,s}$, $j = \overline{1,N}$; с учетом аппроксимации управления их значения равны

$$\int_{t^j}^{t^{j+1}} \psi^k(t)' f_u(x^k, u_j^k, w^k, t)dt$$

в случае кусочно-постоянной аппроксимации и

$$\frac{1}{t^{j+1}-t^j}\left[\int_{t^{j-1}}^{t^j} \psi^k(t)' f_u(x^k, \bar{u}^k(t), w^k, t)(t-t^{j-1})dt + \int_{t^j}^{t^{j+1}} \psi^k(t)' f_u(x^k, \bar{u}^k(t), w^k, t)(t^{j+1}-t)dt\right] \tag{3.14}$$

в случае кусочно-линейной аппроксимации (3.9). При этом $\bar{u}^k(t)$ вычисляется по формуле (3.9) при $u_j = u_j^k$, $u_{j+1} = u_{j+1}^k$.

Из полученных значений градиентов по управлению$\nabla I_i, i = \overline{1,m}$, и $\nabla J_j(t)$, $j = \overline{1,s}$, $t \in T$, и вычисленных по формулам (3.11)–(3.13) градиентов по параметрам составляется матрица коэффициентов линеаризованных ограничений. Она дополняется также блоком элементов $\partial c_i/\partial u_j$, соответствующим ограничениям на управление, и приобретает вид матрицы специальной блочной структуры, которую обозначим через A.

3.3. Алгоритм метода приведенного градиента из [6]

Вводя векторные обозначения для равенств (3.2)–(3.4), построим модифицированную функцию Лагранжа (см. [7]) для задачи (3.1)–(3.7):

$$\begin{aligned} L = {} & \varphi(x(t_1)) - \lambda^{k'}[h(x(t_1)) - \bar{h}^L] + \frac{\rho}{2}[h(x(t_1)) - \bar{h}^L]'[h(x(t_1)) - \bar{h}^L] - \int_{t_0}^{t_1} \mu^{k'}(t)[g(x(t),t) - \bar{g}^L]dt + \\ & + \frac{\rho}{2}\int_{t_0}^{t_1}[g(x(t),t) - \bar{g}^L]'[g(x(t),t) - \bar{g}^L]dt - \int_{t_0}^{t_1}\gamma^k(t)[c(u,t) - \bar{c}^L]dt + \frac{\rho}{2}\int_{t_0}^{t_1}[c(u,t) - \bar{c}^L]'[c(u,t) - \bar{c}^L]dt, \end{aligned} \tag{3.15}$$

где

$$\bar{h}^L = h(x^k(t_1)) + h_x(x^k(t_1))\delta x(t_1), \quad \bar{g}^L = g(x^k(t),t) + g_x(x^k(t),t)\delta x(t),$$

$$\bar{c}^L = c(u^k(t),t) + c_u(u^k(t),t)\delta u(t), \quad \delta u = u - u^k, \quad \delta x = x - x^k.$$

Далее линеаризуем ограничения (3.2), (3.3) на k-м приближении:

$$I^k + \sum_{j=0}^{N} \nabla_u I^k(t^j)'(u_j - u_j^k) + \nabla_w I^{k'}(w - w^k) + \nabla_v I^{k'}(v - v^k) = 0, \tag{3.16}$$

$$J_j^k + \sum_{i=0}^{j}[\nabla_u J^k(t^j)'(u_i - u_i^k) + \nabla_w J^k(t^j)'(w - w^k) + \nabla_v J^k(t^j)'(v - v^k)] = 0, \quad j = \overline{0,N}. \tag{3.17}$$

Здесь $I = (I_1, I_2, \ldots, I_m)$, $J = (J_1, J_2, \ldots, J_s)$. Следовательно, имеем m ограничений (3.16) и $(N+1)s$ ограничений (3.17), которые представляют собой явную форму (через u, w, v) линеаризованных (h^L, g_j^L) ограничений (3.2), (3.3), причем вместо равенств (3.3), заданных для каждого момента $t \in T$, имеем N равенств, определенных в узлах сетки (3.8).

Линеаризуем также условия (3.4):

$$c(u^k, t^j) + \nabla_u c(u^k, t^j)'(u_j - u_j^k) = 0, \quad j = \overline{0,N}, \tag{3.18}$$

где $c = (c_1, c_2, \ldots, c_l)$. Прямые ограничения на управление и параметры оставим без изменений:

$$u_j^{\text{н}} \le u_j \le u_j^{\text{в}}, \quad j = \overline{1,N}, \tag{3.19}$$

$$v_j^{\text{н}} \le v_j \le v_j^{\text{в}}, \quad j = \overline{1,n}, \quad w_i^{\text{н}} \le w_i \le w_i^{\text{в}}, \quad i = \overline{1,p}. \tag{3.20}$$

Конечномерная аппроксимация функционала (3.15), в котором переменные $x(t)$ определены через систему (3.1) по заданному $u(t)$, $t \in T$, имеет следующий вид:

$$\begin{aligned} L = \varphi(x^N) - \lambda^{k'}[h(x(t^N)) - \overline{h}^L] + \frac{\rho}{2}[h(x(t^N)) - \overline{h}^L]'[h(x(t^N)) - \overline{h}^L] - \sum_{j=0}^{N} \mu_j^{k'}[g(x(t^j), t^j) - \overline{g}^L] + \\ + \frac{\rho}{2}\sum_{j=0}^{N}[g(x(t^j), t^j) - \overline{g}^L]'[g(x(t^j), t^j) - \overline{g}^L] - \sum_{j=0}^{N}\gamma_j^{k'}[c(u_j, t^j) - \overline{c}^L] + \\ + \frac{\rho}{2}\sum_{j=0}^{N}[c(u_j, t^j) - \overline{c}^L]'[c(u_j, t^j) - \overline{c}^L], \end{aligned} \tag{3.21}$$

где

$$\begin{aligned} \overline{h}^L &= h(x^k(t^N)) + h_x(x^k(t^N))(x(t^N) - x^k(t^N)), \\ \overline{g}^L &= g(x^k(t^j), t^j) + g_x(x^k(t^j), t^j)(x(t^j) - x^k(t^j)), \\ \overline{c}^L &= c(u^k(t^j), t^j) + c_u(u^k(t^j), t^j)(u_j - u_j^k), \quad j = \overline{0,N}. \end{aligned}$$

Для минимизации функционала (3.21), который теперь по сути является функцией многих переменных, при линейных ограничениях (3.16)–(3.20) применяется метод приведенного градиента (см. [6]). Заметим, что функционал (3.21) предполагает использование исходной системы (3.1) для расчета траектории $\{x(t^1), x(t^2), \ldots, x(t^N)\}$ по заданным параметрам v, w и управлению $u(t^0), u(t^1), \ldots, u(t^N)$, т.е. полная модель вспомогательной задачи описывается соотношениями (3.1), (3.16)–(3.21).

Обозначив через $A[m + (l+s)(N+1)] \times [r(N+1) + p + n]$ матрицу коэффициентов линейных равенств (3.16)–(3.18), через b – вектор их свободных членов размерности $m + (l+s)(N+1)$ и через z – вектор искомых переменных $(u_j, j = \overline{0,N};\ v;\ w)$ размерности $r(N+1) + p + n$ соответственно, поставленную задачу запишем в виде

$$\begin{aligned} L(z) &\to \min, \\ Az &= b, \\ z^{\text{н}} &\le z \le z^{\text{в}}. \end{aligned} \tag{3.22}$$

Для решения задачи (3.22) применяется метод приведенного градиента (см. [6]), который отличается от известного в линейном программировании симплекс-метода тем, что в силу нелинейности целевой функции его последовательные приближения необязательно будут находиться в вершинах многогранника линейных ограничений, а могут быть и его внутренними точками.

3.4. Алгоритм метода спроектированного лагранжиана (см. [6]–[9])

Рассмотрим теперь полный алгоритм решения исходной задачи (3.1)–(3.6).

1. С заданным управлением u_j^k, $j = \overline{0,N}$, интегрируется система (3.1), и в узлах сетки (3.8) запоминаются точки фазовой траектории x_j^k, $j = \overline{0,N}$. Здесь k – номер итерации (первый раз $k = 0$).

На полученном решении линеаризуются ограничения задачи (3.10) и строится вспомогательная задача (3.16)–(3.21).

2. Методом приведенного градиента решается вспомогательная задача минимизации модифицированной функции Лагранжа (3.21) при линейных ограничениях (3.16)–(3.20).

В результате будут найдены новые приближения для управления u_j^{k+1}, $j = \overline{0,N}$, параметров w^{k+1} и v^{k+1}, а также для двойственных переменных λ^{k+1} и μ_j^{k+1}, $j = \overline{0,N}$.

3. Проверяется критерий окончания итерационного процесса как по прямым, так и по двойственным переменным:

$$\left|I_i(u^{k+1}, w^{k+1}, v^{k+1})\right| / (1 + \alpha^{k+1}) \le \varepsilon, \quad i = \overline{1, m},$$
$$\left|J_i(u^{k+1}, w^{k+1}, v^{k+1})\right| / (1 + \alpha^{k+1}) \le \varepsilon, \quad i = \overline{1, s},$$

где

$$\alpha^{k+1} = \max\{\|u_j^{k+1}\|, j = \overline{0, N}; |w_i|, i = \overline{1, p}; |v_l|, l = \overline{1, n}\};$$
$$\left|\lambda_j^k - \lambda_j^{k+1}\right| / (1 + \Theta^{k+1}) \le \varepsilon, \quad j = \overline{1, m};$$
$$\left|\mu_{ij}^k - \mu_{ij}^{k+1}\right| / (1 + \Theta^{k+1}) \le \varepsilon, \quad i = \overline{1, s}, \quad j = \overline{0, N};$$
$$\Theta^{k+1} = \max\{|\lambda_j^{k+1}|, j = \overline{1, m}; |\mu_{ij}^{k+1}|, i = \overline{1, s}, j = \overline{0, N}\}.$$

При нарушении хотя бы одного из этих условий выполняется новая $(k+1)$-я итерация с п. 1. Если же эти неравенства выполняются для заданного $\varepsilon > 0$, то итерационный процесс прекращается, а найденные u_j^{k+1}, $j = \overline{0, N}$, w^{k+1} и v^{k+1} выдаются в качестве приближенного решения задачи оптимального управления.

4. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ. ПРИКЛАДНЫЕ ЗАДАЧИ С РЕШЕНИЯМИ

Приведем примеры прикладных задач, относящихся к рассмотренным в разд. 1–3 классам задач оптимизации, решения которых найдены с помощью многометодных алгоритмов, изложенных в этих разделах.

4.1. Задача оптимального управления сферическим мобильным роботом с трехмерными управляющими воздействиями

Задача сформулирована М.М. Свининым и приведена в [10]. Рассматривается мобильный сферический робот, перемещающийся по плоскости. Конструкция робота представляет собой оболочку с размещенными в ней тремя роторами-двигателями. Динамика робота, редуцированная к контактным координатам, описывается следующей системой обыкновенных дифференциальных уравнений:

$$\dot{x} = G(x)J^{-1}(x)J_r\sum_{k=1}^{n} n_k(x)u_k,$$

где векторы состояний и управлений определены как

$$x \triangleq [u_a, v_a, u_o, v_o, \psi]^{\mathrm{T}}, \quad u \triangleq [\dot{\varphi}_1, \dot{\varphi}_2, \dot{\varphi}_3]^{\mathrm{T}},$$

а φ_i, $i = \overline{1,3}$, обозначают углы поворота двигателей.

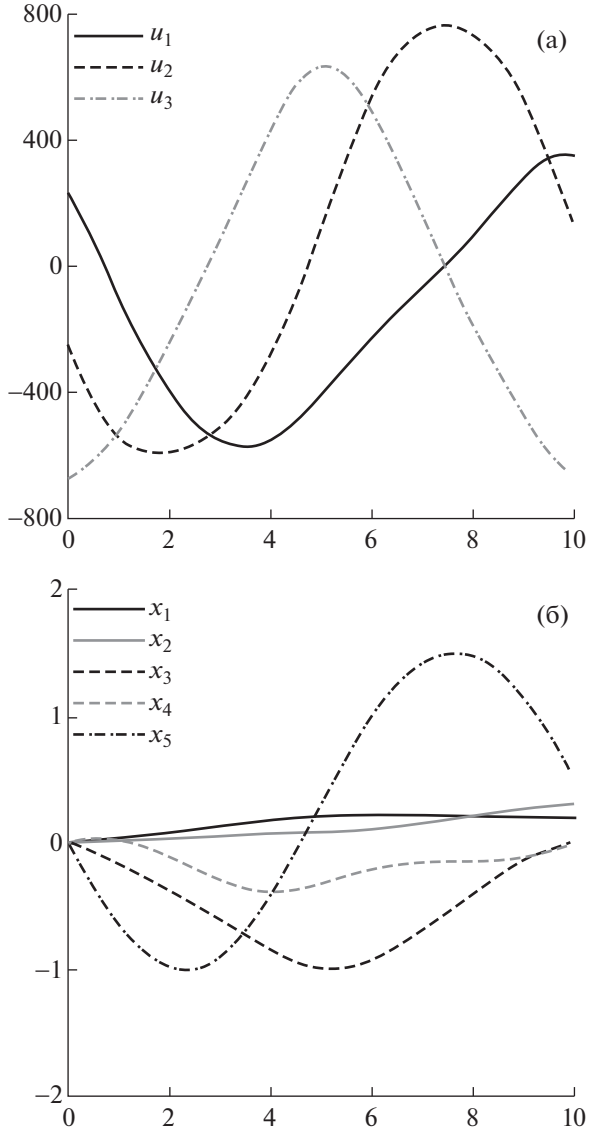
Положение точки контакта на плоскости задается координатами u_a, v_a, а ее координаты на сфере задаются углами u_o, v_o.

Матричные и векторные величины определяются следующим образом:

$$G = \begin{bmatrix} 0 & -R & 0 \\ R & 0 & 0 \\ \sin\psi/\cos v_o & \cos\psi/\cos v_o & 0 \\ \cos\psi & -\sin\psi & 0 \\ \sin\psi\,\mathrm{tg}\,v_o & \cos\psi\,\mathrm{tg}\,v_o & 1 \end{bmatrix},$$

а векторы n_1, n_2, n_3 – столбцы матрицы

$$R = \begin{bmatrix} \cos u_o\cos\psi + \sin u_o\sin v_o\sin\psi & \cos v_o\sin\psi & -\sin u_o\cos\psi + \cos u_o\sin v_o\sin\psi \\ -\cos u_o\sin\psi + \sin u_o\sin v_o\cos\psi & \cos v_o\cos\psi & \sin u_o\sin\psi + \cos u_o\sin v_o\cos\psi \\ \sin u_o\cos v_o & -\sin v_o & \cos u_o\cos v_o \end{bmatrix}.$$



Фиг. 3. Найденные управления и траектории для мобильного робота.

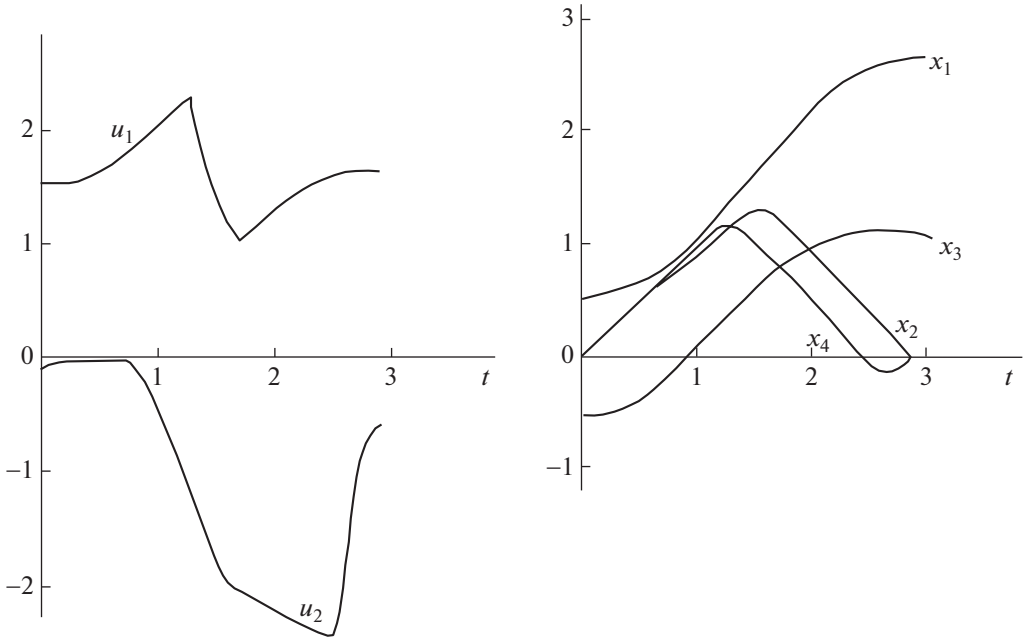
Матрица инерции системы определяется как

$$J = \begin{bmatrix} (2/3m_o + M)R^2 & 0 & 0 \\ 0 & (2/3m_o + M)R^2 & 0 \\ 0 & 0 & 2/3m_oR^2 \end{bmatrix} + (2J_p + J_r)E,$$

где M – общая масса робота, а m_o и m_r обозначают, соответственно, массу сферической оболочки и массу отдельного ротора.

Задача оптимального управления сферическим роботом состоит в переводе из точки $x(0)$ в точку $x(T)$ при условии минимизации энергии управления $J = \int_0^T u^T u dt$.

Пусть, например, $x(0) = [0, 0, 0, 0, 0, 0]$, а $x(10) = \left[0, 2, 0, 3, 0, 0, \frac{\pi}{6}\right]$, тогда полученные решения представлены на фиг. 3 и являются физически реализуемыми. Точность выполнения ограничений порядка 10^{-5}, значение функционала $J(10) = 5769001$.



Фиг. 4. Графики управлений и фазовых координат для задачи 4.2.

4.2. Задача об оптимальном управлении манипулятором робота

Динамика движения манипулятора промышленного робота описывается системой дифференциальных уравнений

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = \frac{[M_1(x,u) - F_1(x)]a_{22} - [M_2(x,u) - F_2(x)]a_{12}(x)}{a_{11}a_{22} - a_{12}(x)a_{21}(x)},$$

$$\dot{x}_3 = x_4,$$

$$\dot{x}_4 = \frac{[M_2(x,u) - F_2(x)]a_{11} - [M_1(x,u) - F_1(x)]a_{21}(x)}{a_{11}a_{22} - a_{12}(x)a_{21}(x)},$$

где

$$M_1(x,u) = -c_1(x_1 - u_1), \quad M_2(x,u) = -c_2(x_3 - x_1 - u_2),$$

$$F_1(x) = -m_2 l_1 R_2 \sin(x_3 - x_1)x_2^2, \quad F_2(x) = m_2 l_2 R_2 \sin(x_3 - x_1)x_4^2,$$

$$a_{11} = m_1\rho_1^2 + m_2 l_1^2, \quad a_{12} = a_{21} = m_2 R_1 l_1 \cos(x_3 - x_1), \quad a_{22} = m_2\rho_2^2.$$

Для рассматриваемой модели робота $m_1 = 7.62$, $m_2 = 8.73$, $R_1 = 0.239$, $R_2 = 0.251$, $\rho_1 = 0.968$, $\rho_2 = 0.973$, $l_1 = 0.5$, $l_2 = 0.67$, $c_1 = c_2 = 10$. На траекторию движения накладываются ограничения $|M_i(x,u)| \le 10$, $i = 1,2$, $\pi/6 \le x_1(t) \le 5/6\pi$, $\pi/3 \le x_1(t) - x_3(t) \le 5/6\pi$, $t \in [0,T]$.

Необходимо найти управление, переводящее систему из точки $x'(0) = (\pi/6, 0, -\pi/6, 0)$ в точку $x'(T) = (5/6\pi, 0, \pi/3, 0)$ за минимальное время T.

Начиная с приближения $T = 4$, $u_1(t) = 0$, $u_2(t) = 0$, $t \in [0,T]$, было получено решение, на котором невязки по ограничениям не превысили 10^{-3}, а значение функционала составило 2.88. Вид оптимального управления и соответствующих ему фазовых координат приведен на фиг. 4.

4.3. Задача оптимизации электроэнергетической системы (ЭЭС)

Математическая модель ЭЭС разработана в Институте систем энергетики им. Л.А. Мелентьева СО РАН и представляет собой совокупность подсистем, описывающих генерирование и по-

требление электроэнергии, которые объединены в единую систему уравнениями электрической сети, и много лет успешно применяется для расчета различных режимов работы проектируемых ЭЭС. Рассмотрим небольшую модель ЭЭС, состоящую из m синхронных генераторов и m паровых турбин. Каждая синхронная машина описывается дифференциальными уравнениями Парка–Горева (без учета переходных процессов в обмотке статора), которые после приведения их к стандартной форме принимают следующий вид:

$$\dot{x}_i = x_{m+i}, \quad i = \overline{1,m},$$

$$\dot{x}_{m+i} = \omega_0/(T_{ji}P_{n_i})[x_{5m+i} - U_i^1/X_{d_i}^1(x_{2m+i}\sin x_i + x_{3m+i}\cos x_i) +$$
$$+ U_i^2/X_{d_i}^1(x_{2m+i}\cos x_i - x_{3m+i}\sin x_i) - D_i x_{m+i}], \quad i = \overline{1,m},$$

$$\dot{x}_{2m+i} = 1/T_{d_0 i}[x_{4m+i} - x_{2m+i} + (X_{d_i} - X_{d_i}^1)/X_{d_i}^1(U_i^2\sin x_i - x_{2m+i} + U_i^1\cos x_i)], \quad i = \overline{1,m},$$

$$\dot{x}_{4m+i} = 1/TR_i[E_{R_0 i} - x_{4m+i} + K_{Ui}(U_{0i} - \sqrt{(U_i^1)^2 + (U_i^2)^2} + K_{I_i}/X_{d_i}^1(-I_{0i}X_{d_i}^1 + ((x_{2m+i})^2 +$$
$$+ (x_{3m+i})^2 + (U_i^1)^2 + (U_i^2)^2 - 2\cos x_i(U_i^2 x_{3m+i} + U_i^1 x_{2m+i}) + 2\sin x_i(U_i^1 x_{3m+i} - U_i^2 x_{2m+i}))^{1/2}) +$$
$$+ K_{f_i}/2\pi x_{m+i} + K_{f_i}^1/2\pi x_{2m+i}], \quad i = \overline{1,m}.$$

Уравнения динамики паровых турбин следующие:

$$\dot{x}_{5m+i} = 1/T_{si}[-P_{ni}/\omega_0\sigma_{1i}x_{m+i} + x_{6m+i} - x_{5m+i}], \quad i = \overline{1,m},$$

$$\dot{x}_{6m+i} = 1/T_{p2i}[-P_{ni}/\omega_0\sigma_{2i}x_{m+i} + u_i], \quad i = \overline{1,m}.$$

Переменными состояния x_{jm+i}, $j = \overline{0,6}$, являются угол ротора генератора, скольжение, составляющие переходной ЭДС машины в продольной и поперечной осях и напряжение обмотки возбуждения соответственно для каждого $i = \overline{1,m}$. Управление U_i изменяет установку регулятора скорости, чтобы обеспечить устойчивый динамический переход в заданное послеаварийное состояние при аварийных сбросах нагрузки. В правые части уравнений входят также технические параметры генераторов и турбин, смысл которых здесь приводить не будем. Число генераторов и турбин было задано равным пяти. Таким образом, при $m = 5$ число дифференциальных уравнений равно 35, т.е. $n = 35$.

Модель электрической сети составляют алгебраические уравнения на узловые напряжения, в которые входят и переменные состояния. Эти уравнения обычно задаются в комплексных переменных, а при численном решении выполняется переход к действительным переменным. Так, например, для эксперимента было задано $N = 14$ уравнений в комплексных переменных, а после перехода была получена система из 28 алгебраических уравнений

$$C_i^1 U_i^1 - C_i^2 U_i^2 + \sum_{k=1,k\neq i}^{N}(-U_k^1 Y_{ik}^1 + U_k^2 Y_{ik}^2) = 1/X_{d_i}^1(x_{2m+i}\sin x_i + x_{3m+i}\cos x_i), \quad i = \overline{1,N},$$

$$C_i^1 U_i^2 + C_i^2 U_i^1 + \sum_{k=1,k\neq i}^{N}(-U_k^1 Y_{ik}^2 - U_k^2 Y_{ik}^1) = 1/X_{d_i}^1(x_{2m+i}\cos x_i + x_{3m+i}\sin x_4), \quad i = \overline{1,N},$$

где $C_i^1 = P_{ni}/(U_i)^2 + A_i^1$, $C_i^2 = A_i^2 - Q_{ni}/(U_i)^2$, $(U_i)^2 = (U_i^1)^2 + (U_i^2)^2$.

Кроме того, заданы ограничения на фазовые координаты и управление

$$|x_i(t) - x_j(t)| \le \delta_{\max}, \quad i,j = \overline{1,m},$$
$$x_{\min_i} \le x_{4m+i}(t) \le x_{\max_i}, \quad i = \overline{1,2m},$$
$$U_{\min_i} \le U_i(t) \le U_{\max_i}, \quad i = \overline{1,m}.$$

Целевым функционалом является функция конечного состояния системы (при $t = 10$ с), измеряющая отклонение некоторых фазовых координат от заданных величин (например, мощностей).

При численном решении поставленной задачи методом спроектированного лагранжиана было сделано 11 внешних итераций, каждая из которых содержала около 20 внутренних итераций метода приведенного градиента. Заданные равенства были выполнены с точностью 10^{-6}, при этом ни одна из переменных не вышла на заданные ограничения. Полученное оптимальное

управление обеспечивает вывод ЭЭС в требуемый режим работы за 10 с после резкого сброса нагрузки.

Выводы. Многометодная вычислительная технология, реализованная в виде параллельных итерационных процессов оптимизации с выбором лучшего приближения, находит решение задачи с автоматическим применением разных методов оптимизации и тем самым существенно повышает эффективность поиска и надежность получения численного решения в прикладных задачах оптимального управления. Получение численного решения при наименьших затратах вычислительных ресурсов актуально при проектировании робототехнических и электроэнергетических систем с автоматическим управлением.

ЗАКЛЮЧЕНИЕ

Практика показывает, что последовательность приближений многометодного алгоритма в сложных задачах управления, как правило, состоит из приближений нескольких (3–5) численных методов, которые выбирались по заданному критерию автоматически в процессе оптимизации. Проведенные вычислительные эксперименты подтверждают эффективность предложенной технологии при решении прикладных задач оптимального управления. Установлено, что применение многометодной технологии нередко является единственным способом получения численного решения в сложной задаче оптимального управления, так как сходимость каждого из методов в отдельности прекращалась до получения оптимального решения. Современные информационные технологии и многопроцессорная вычислительная техника допускают достаточно эффективную реализацию многометодных алгоритмов. Программное обеспечение, разработанное на основе данного подхода и реализующее многометодную технологию расчета оптимального управления и оптимальных параметров (см. [9]–[12]), успешно применяется для решения сложных прикладных задач оптимального управления из различных областей науки и техники (см. [10]–[14]). Применение эффективной технологии расчета управления особенно актуально в управляемых системах реального времени, например, в системах управления летательными аппаратами, обладающими высокой маневренностью. Например, при проектировании СУ-57 (мирового лидера по маневренности) для решения серии задач оптимального маневрирования использовалось данное программное обеспечение (см. [11]). Известно также, что наличие программы выбора оптимального начального приближения в бортовом компьютере беспилотного космического аппарата Буран позволило ему выбрать наименее зависимую от бокового ветра стартовую точку для начала посадки на аэродром и успешно завершить свой беспрецедентный полет, реализовав программу оптимальной посадки (глиссаду) с высокой точностью.

СПИСОК ЛИТЕРАТУРЫ

1. *Понтрягин Л.С., Болтянский В.Н., Гамкрелидзе Р.В., Мищенко Е.Ф.* Математическая теория оптимальных процессов. М.: Наука, 1969.
2. *Беллман Р., Калаба Р.* Квазилинеаризация и нелинейные краевые задачи. М.: Мир, 1968.
3. *Габасов Р., Кириллова Ф.М.* Качественная теория оптимальных процессов. М.: Наука, 1971.
4. *Васильев О.В., Тятюшкин А.И.* Об одном методе решения задач оптимального управления, основанном на принципе максимума // Ж. вычисл. матем. и матем. физ. 1981. Т. 21. № 6. С. 1376–1384.
5. *Габасов Р., Кириллова Ф.М., Тятюшкин А.И.* Конструктивные методы оптимизации. Ч. 1: Линейные задачи. Минск: Университетское, 1984.
6. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация. М.: Мир, 1985.
7. *Евтушенко Ю.Г.* Методы решения экстремальных задач и их применение в системах оптимизации. М.: Наука, 1982.
8. *Федоренко Р.П.* Приближенное решение задач оптимального управления. М.: Наука, 1978.
9. *Тятюшкин А.И.* ППП КОНУС для оптимизации непрерывных управляемых систем // Пакеты прикладных программ: Опыт использования. М.: Наука, 1989. С. 63–83.
10. *Горнов А.Ю., Тятюшкин А.И., Финкельштейн Е.А.* Численные методы для решения терминальных задач оптимального управления // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 2. С. 224–237.
11. *Тятюшкин А.И., Федунов Б.Е.* Возможности защиты от атакующей ракеты задней полусферы самолета вертикальным маневром // Изв. РАН, ТиСУ. 2006. № 1. С. 111–125.
12. *Тятюшкин А.И.* Многометодная технология оптимизации управляемых систем. Новосибирск: Наука, 2006.
13. *Тятюшкин А.И.* Численные методы решения задач оптимального управления с параметрами // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 10. С. 1615–1630.
14. *Тятюшкин А.И.* Многометодная оптимизация управления в сложных прикладных задачах // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 2. С. 235–246.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 206–216

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.958

ДЕКОМПОЗИЦИЯ РЕШЕНИЯ ДВУМЕРНОГО СИНГУЛЯРНО ВОЗМУЩЕННОГО УРАВНЕНИЯ КОНВЕКЦИИ-ДИФФУЗИИ С ПЕРЕМЕННЫМИ КОЭФФИЦИЕНТАМИ В КВАДРАТЕ; ОЦЕНКИ В ГЁЛЬДЕРОВЫХ НОРМАХ

© 2021 г. В. Б. Андреев[1,*], И. Г. Белухина[1,]**

[1] *119991 Москва, Ленинские горы, 1, МГУ им. М.В. Ломоносова, Россия*

**e-mail: andreev@cs.msu.su*

***e-mail: belukh@cs.msu.su*

Поступила в редакцию 12.05.2020 г.
Переработанный вариант 23.07.2020 г.
Принята к публикации 16.09.2020 г.

В единичном квадрате плоскости Oxy рассматривается первая краевая задача для линейного стационарного сингулярно возмущенного уравнения конвекции-диффузии с переменными коэффициентами. Предполагается, что при заданном коэффициенте конвекции задача имеет один регулярный и два характеристических пограничных слоя, каждый из которых расположен в окрестности одной из сторон квадрата. В работе построена декомпозиция решения задачи, для регулярной составляющей которой получены априорные оценки в гёльдеровых нормах. Библ. 9.

Ключевые слова: сингулярно возмущенное уравнение, конвекция-диффузия, переменные коэффициенты, двумерная задача, априорные оценки, пространства Гёльдера.

DOI: 10.31857/S0044466921020046

1. ВВЕДЕНИЕ

При анализе численных методов решения дифференциальных уравнений обычно необходима информация о величинах производных приближаемого решения. В сингулярно возмущенном случае максимумы модулей производных порядка k оценивается величиной $O(\varepsilon^{-\sigma k})$, где $\varepsilon > 0$ – малый параметр. Несмотря на то что эта оценка, как правило, является точной, она мало эффективна. Связано это с тем, что указанные значения производные принимают только в малой части области, называемой пограничным слоем. Вне же пограничного слоя производные решения, как правило, ограничены (до порядка, определяемого гладкостью входных данных). Поэтому до проведения оценок решение полезно представить в виде суммы регулярной и сингулярной составляющих. Такое представление называется *декомпозицией решения*. Разумеется, декомпозиция не определяется однозначно, и тот или иной выбор декомпозиции связан со способом ее дальнейшего анализа и, в первую очередь, со способом анализа ее регулярной составляющей. Известные декомпозиции [1] (см. также литературу, цитированную в [2]), обладая большой общностью и широтой охвата, имеют существенный недостаток – они предъявляют довольно жесткие требования к гладкости исследуемого решения, которые далеко не всегда могут быть удовлетворены. Чтобы иметь возможность снизить требования к гладкости решения, нужны новые декомпозиции и новые методы их анализа. Естественно предположить, что для получения более точных результатов придется сузить класс исследуемых задач. Так, например, в [3] построена декомпозиция при существенно ослабленных по сравнению с [1] предположениях о гладкости решения для двумерного уравнения с постоянными коэффициентами. Неулучшаемые оценки для уравнения с постоянными коэффициентами получены в [2], [4]. Одномерный вариант этих оценок для уравнения с переменными коэффициентами содержится в [5].

В данной работе рассмотрена задача Дирихле в ограниченной области (единичном квадрате) для уравнения с переменными коэффициентами и конвекцией, направленной ортогонально одной из сторон квадрата. Для регулярной составляющей решения этой задачи с использованием

результатов [2], [4] получены априорные оценки в нормах гёльдеровых пространств через соответствующие нормы правой части уравнения и граничной функции.

2. ПОСТАНОВКА ЗАДАЧИ И ОСНОВНОЙ РЕЗУЛЬТАТ

Рассмотрим следующую задачу: в области $\Omega := (0,1)^2$ с границей $\partial\Omega = \bigcup_{\ell=1}^{4} \Gamma_\ell$ ищется решение задачи

$$\begin{gathered} Lu := -\varepsilon\Delta u + r(x,y)\frac{\partial u}{\partial x} + q(x,y)u = f(x,y), \quad (x,y) \in \Omega, \\ u\big|_{\partial\Omega} = g(x,y), \quad (x,y) \in \partial\Omega, \end{gathered} \tag{2.1}$$

где Δ – оператор Лапласа, $\varepsilon \in (0,1]$ – малый параметр, а коэффициенты удовлетворяют условиям

$$r(x,y) \geq 2r_0 = \text{const} > 0, \quad q(x,y) > 0.$$

Будем, кроме того, предполагать, что

$$r(x,y), q(x,y), f(x,y), g(x,y) \in C^{k,\lambda}, \quad k = 0,1,\ldots, \quad 0 < \lambda < 1. \tag{2.2}$$

При сделанных предположениях поставленная задача имеет три пограничных слоя: регулярный пограничный слой в окрестности правой границы квадрата и два характеристических слоя в окрестностях верхней и нижней сторон.

Представим решение задачи в виде

$$u(x,y) = U(x,y) + V(x,y), \tag{2.3}$$

где U – регулярная, а V – сингулярная составляющие, причем для регулярной составляющей выполняются условия

$$LU = f, \quad (x,y) \in \Omega, \quad U(0,y) = g(y) := g(0,y), \tag{2.4}$$

а

$$\begin{gathered} LV = 0, \quad (x,y) \in \Omega, \\ V\big|_{\partial\Omega} = g(x,y) - U(x,y), \quad (x,y) \in \partial\Omega. \end{gathered}$$

Основной результат работы содержит

Теорема 1. *Пусть $r(x,y)$, $q(x,y)$ и $f(x,y) \in C^{k,\lambda}(\bar{\Omega})$, $g(y) \in C^{k+2,\lambda}([0,1])$, $k = 0,1,\ldots$, $0 < \lambda < 1$. Тогда существует такая функция $U(x,y)$, удовлетворяющая* (2.4), (*регулярная составляющая решения задачи* (2.1)), *для которой справедлива оценка*

$$\varepsilon|U|_{C^{k+2,\lambda}} + \left|\frac{\partial U}{\partial x}\right|_{C^{k,\lambda}} + \sqrt{\varepsilon}\left|\frac{\partial U}{\partial y}\right|_{C^{k,\lambda}} + \|U\|_{C^{k,\lambda}} \leq c\left(\|f\|_{C^{k,\lambda}} + \varepsilon|g|_{C^{k+2,\lambda}} + \|g\|_{C^{k+1,\lambda}}\right). \tag{2.5}$$

Здесь c – положительная постоянная, не зависящая от U, ε и рядом стоящего сомножителя, $|\cdot|_{C^{k,\lambda}}$ – коэффициент Гёльдера (полунорма), а $\|\cdot\|_{C^{k,\lambda}}$ – норма в пространстве Гёльдера $C^{k,\lambda}$.

Последующее изложение данной работы посвящено доказательству этой теоремы.

3. УРАВНЕНИЕ С ПОСТОЯННЫМИ КОЭФФИЦИЕНТАМИ В ПОЛУПЛОСКОСТИ. НЕУЛУЧШАЕМЫЕ ОЦЕНКИ РЕШЕНИЯ

Рассмотрим следующую задачу: в правой полуплоскости $\mathbb{R}^2_+$ плоскости Oxy найти ограниченное решение задачи

$$-\varepsilon\Delta u + 2\alpha\frac{\partial u}{\partial x} + qu = f(x,y), \quad (x,y) \in \mathbb{R}^2_+, \tag{3.1}$$

$$u(0,y) = g(y), \quad -\infty < y < \infty, \tag{3.2}$$

где α и q – коэффициенты, которые предполагаются постоянными и положительными.

В [2] для решения этой задачи при $g(y) \equiv 0$ получена оценка

$$\varepsilon |u|_{C^{2,\lambda}} + \left|\frac{\partial u}{\partial x}\right|_{C^{\lambda}} + \sqrt{\varepsilon}\left|\frac{\partial u}{\partial y}\right|_{C_y^{\lambda}} + \varepsilon^{\frac{1-\lambda}{2}}\left|\frac{\partial u}{\partial y}\right|_{C_x^{\lambda}} + \|u\|_{C^{\lambda}} \le c\|f\|_{C^{\lambda}}, \quad \lambda \in (0,1),$$

а в [4] при $f(x,y) \equiv 0$ получена оценка

$$\|u\|_{C^{k,\lambda}} \le c\|g\|_{C^{k,\lambda}}, \quad k = 0,1,2, \quad \lambda \in (0,1).$$

В силу линейности задачи (3.1), (3.2) отсюда следует, что для решения $u(x,y)$ задачи (3.1), (3.2) справедлива априорная оценка

$$\varepsilon |u|_{C^{2,\lambda}} + \left|\frac{\partial u}{\partial x}\right|_{C^{\lambda}} + \sqrt{\varepsilon}\left|\frac{\partial u}{\partial y}\right|_{C_y^{\lambda}} + \varepsilon^{\frac{1-\lambda}{2}}\left|\frac{\partial u}{\partial y}\right|_{C_x^{\lambda}} + \|u\|_{C^{\lambda}} \le c\left(\|f\|_{C^{\lambda}} + \varepsilon |g|_{C^{2,\lambda}} + \|g\|_{C^{1,\lambda}}\right), \quad \lambda \in (0,1). \tag{3.3}$$

Из (3.3) следуют оценки для любых $k = 0,1,\ldots$

$$\varepsilon |u|_{C^{k+2,\lambda}} + \left|\frac{\partial u}{\partial x}\right|_{C^{k,\lambda}} + \sqrt{\varepsilon}\left|\frac{\partial u}{\partial y}\right|_{C_y^{k,\lambda}} + \varepsilon^{\frac{1-\lambda}{2}}\left|\frac{\partial u}{\partial y}\right|_{C_x^{k,\lambda}} + \|u\|_{C^{k,\lambda}} \le c\left(\|f\|_{C^{k\lambda}} + \varepsilon |g|_{C^{k+2,\lambda}} + \|g\|_{C^{k+1,\lambda}}\right), \quad k = 0,1,\ldots. \tag{3.4}$$

4. ПРЕДСТАВЛЕНИЕ РЕШЕНИЯ ИСХОДНОЙ ЗАДАЧИ

Построим теперь регулярную составляющую решения задачи (2.1) и установим оценку (2.5). Для этого сначала сделаем замену переменной $u(x,y) = e^{\beta x} v(x,y)$. При $\beta > 0$ для новой функции $v(x,y)$ получим задачу

$$\begin{gathered} -\varepsilon \Delta v + \hat{r}(x,y)\frac{\partial v}{\partial x} + \hat{q}(x,y)v = \hat{f}(x,y), \quad (x,y) \in \Omega, \\ v\big|_{\partial\Omega} = \hat{g}(x,y), \quad (x,y) \in \partial\Omega, \end{gathered} \tag{4.1}$$

где

$$\begin{gathered} \hat{r}(x,y) = r(x,y) - 2\beta\varepsilon, \\ \hat{q}(x,y) = q(x,y) + \beta r(x,y) - \beta^2\varepsilon, \\ \hat{f}(x,y) = e^{-\beta x} f(x,y). \end{gathered} \tag{4.2}$$

Пусть

$$\varepsilon_1 = \frac{r_0}{2\beta}. \tag{4.3}$$

Тогда при $0 < \varepsilon \le \varepsilon_1$ для коэффициентов $\hat{r}(x,y)$, $\hat{q}(x,y)$ справедливы оценки

$$\hat{r}(x,y) \ge r_0 > 0, \quad \hat{q}(x,y) \ge \frac{3}{2}\beta r_0 > 0. \tag{4.4}$$

Продолжим $\hat{r}(x,y)$, $\hat{q}(x,y)$ и $\hat{f}(x,y)$ гладко с Ω на $\Omega^* = (0,3/2)\times(-1/2,3/2)$, ($\hat{g}(0,y) = g(y)$ с $(0,1)$ на $(-1/2,3/2)$), а затем на всю полуплоскость $\mathbb{R}^2_+$ (всю ось Oy), с сохранением класса и нормы. Продолженные функции будем обозначать теми же буквами, но со звездочкой, т.е. для всех $\hat{\varphi}(x,y)$, где под $\hat{\varphi}(x,y)$ понимаются функции ($\hat{r}(x,y)$, $\hat{q}(x,y)$, $\hat{f}(x,y)$)

$$\begin{gathered} \varphi^*(x,y) = \hat{\varphi}(x,y) \quad (x,y) \in \Omega, \\ \varphi^*(x,y) \in C^{k,\lambda}(\mathbb{R}^2_+), \quad \|\varphi^*\|_{C^{k,\lambda}(\mathbb{R}^2_+)} \le c\|\hat{\varphi}\|_{C^{k,\lambda}(\Omega)}, \end{gathered}$$

и аналогично для одномерной функции $g^*(y) = g(y)$ при $y \in [0,1]$. Будем также предполагать, что

$$r^*(x,y) \ge r_0/2 > 0, \quad q^*(x,y) \ge \beta r_0/2 > 0, \quad (x,y) \in \mathbb{R}^2_+,$$

и, более того, будем предполагать, что

$$r^*(x,y) = 2\alpha > 0, \quad q^*(x,y) = Q > 0, \quad f^*(x,y) = 0,$$
$$\{(x \geq 3/2) \cup (-\infty < y \leq -1/2) \cup (3/2 \leq y < \infty)\},$$
$$g^*(y) = 0 \quad (-\infty < y \leq -1/2) \cup (3/2 \leq y < \infty).$$

Применительно к несвязным областям существование таких продолжений функций с указанными свойствами в двумерном и одномерном случаях следует, например, из [7, Дополнение, с. 587–597].

Теперь в правой полуплоскости $x \geq 0$ рассмотрим задачу

$$L^*U^*(x,y) = -\varepsilon\Delta U^* + r^*(x,y)\frac{\partial U^*}{\partial x} + q^*(x,y)U^* = f^*(x,y), \quad (x,y) \in \mathbb{R}^2_+,$$
$$U^*\big|_{x=0} = g^*(y), \quad -\infty < y < \infty, \quad \lim_{\sqrt{x^2+y^2}\to\infty} U^*(x,y) = 0. \tag{4.5}$$

Из связи функций $u(x,y)$, $v(x,y)$ и $U^*(x,y)$, описанной ранее в этом разделе, очевидно, что сужение функции $U^*e^{\beta x}$ на Ω можно рассматривать как регулярную составляющую U декомпозиции (2.3).

Для доказательства теоремы 1 аналогично [6, гл. III, § 2] будем пользоваться методом разбиения единицы и техникой явного применения такого метода в одномерном случае в [5]. Для начала построим разбиение единицы для полуплоскости. Напомним, что разбиение единицы на полуоси Ox в [5] задается при помощи бесконечно дифференцируемой функции

$$\omega(\xi) = \begin{cases} 1, & |\xi| \leq 1/4, \\ \dfrac{1}{2}\left[1 - \operatorname{th}\dfrac{|\xi| - 1/2}{(|\xi| - 1/4)(3/4 - |\xi|)}\right], & \dfrac{1}{4} \leq |\xi| \leq \dfrac{3}{4}, \\ 0, & |\xi| \geq \dfrac{3}{4}, \end{cases}$$

с носителем $[-3/4, 3/4]$ следующим образом:

$$\sum_{m=0}^{2N-1} \xi_m(x) + \xi^+(x) = 1 \quad \text{при} \quad x \in [0,\infty), \tag{4.6}$$

где

$$\xi_m(x) := \omega(xN - m), \quad \operatorname{supp}\xi_m(x) = \left[\frac{m - 3/4}{N}, \frac{m + 3/4}{N}\right] =: \Delta_m, \quad \operatorname{mes}\Delta_m = \frac{3}{2N} =: \delta,$$

а

$$\xi^+(x) = \begin{cases} \xi_{2N}(x), & 0 \leq x \leq 2, \\ 1, & 2 \leq x < \infty. \end{cases}$$

Аналогично построим разбиение единицы на оси Oy.

Очевидно, что искомое разбиение есть

$$\sum_{n=-N+1}^{2N-1} \eta_n(y) + \eta^+(y) + \eta^-(y) = 1 \quad \text{при} \quad y \in (-\infty,\infty), \tag{4.7}$$

если

$$\eta_n(y) := \omega(yN - n), \quad \operatorname{supp}\eta_n(y) = \left[\frac{n - 3/4}{N}, \frac{n + 3/4}{N}\right] =: \Delta_n, \quad \operatorname{mes}\Delta_n = \frac{3}{2N} = \delta,$$

а

$$\eta^-(y) = \begin{cases} \eta_{-N}(y), & -1 \leq y < \infty, \\ 1, & -\infty < y \leq -1, \end{cases} \qquad \eta^+(y) = \begin{cases} \eta_{2N}(y), & -\infty < y \leq 2, \\ 1, & 2 \leq y < \infty. \end{cases}$$

Теперь разбиение единицы $\zeta_{m,n}$ на всей полуплоскости $\mathbb{R}^2_+$ зададим как тензорное произведение построенных одномерных разбиений (4.6), (4.7):

$$\sum_{m=0}^{2N}\sum_{n=-N}^{2N}\zeta_{m,n}=1,\quad \zeta_{m,n}(x,y)=\xi_m(x)\eta_n(y).$$

В соответствии с этим разбиением представим функцию $U^*(x,y)$ в виде

$$U^*(x,y)=\sum_{m=0}^{2N}\sum_{n=-N}^{2N}u_{m,n}(x,y),\quad \text{где}\quad u_{m,n}(x,y)=U^*(x,y)\zeta_{m,n}(x,y). \tag{4.8}$$

5. ДОКАЗАТЕЛЬСТВО ОСНОВНОЙ ТЕОРЕМЫ

Умножим уравнение (4.5) для $U^*(x,y)$ на соответствующие функции $\zeta_{m,n}(x,y)$ и, поступая аналогично одномерному случаю [5], получим для функций $u_{m,n}(x,y)$ задачи с постоянными коэффициентами в полуплоскости типа (3.1), (3.2). Затем воспользуемся оценкой (3.3) для решений каждой из этих задач, и, наконец, получим для $U^*(x,y)$, как для суммы решений $u_{m,n}(x,y)$, сначала оценку (3.3), а затем оценку (2.5) при $k=0$, и, наконец, для любых $k=0,1,\ldots$.

Опишем этот процесс более подробно. Умножим уравнение (4.5) на $\zeta_{m,n}(x,y)$. Будем иметь

$$-\varepsilon\zeta_{m,n}\Delta U^*+r^*\zeta_{m,n}\frac{\partial U^*}{\partial x}+q^*\zeta_{m,n}U^*=f^*\zeta_{m,n}. \tag{5.1}$$

Очевидно, что

$$\frac{\partial}{\partial x}(\zeta_{m,n}U^*)=\zeta_{m,n}\frac{\partial U^*}{\partial x}+\frac{\partial\zeta_{m,n}}{\partial x}U^*,$$
$$\frac{\partial^2}{\partial x^2}(\zeta_{m,n}U^*)=\zeta_{m,n}\frac{\partial^2U^*}{\partial x^2}+2\frac{\partial\zeta_{m,n}}{\partial x}\frac{\partial U^*}{\partial x}+\frac{\partial^2\zeta_{m,n}}{\partial x^2}U^*,$$
$$\frac{\partial^2}{\partial y^2}(\zeta_{m,n}U^*)=\zeta_{m,n}\frac{\partial^2U^*}{\partial y^2}+2\frac{\partial\zeta_{m,n}}{\partial y}\frac{\partial U^*}{\partial y}+\frac{\partial^2\zeta_{m,n}}{\partial y^2}U^*.$$

С учетом того, что $\zeta_{m,n}U^*=u_{m,n}$ (см. (4.8)), отсюда следует, что

$$\zeta_{m,n}\frac{\partial U^*}{\partial x}=\frac{\partial u_{m,n}}{\partial x}-\frac{\partial\zeta_{m,n}}{\partial x}U^*,$$
$$\zeta_{m,n}\frac{\partial^2U^*}{\partial x^2}=\frac{\partial^2u_{m,n}}{\partial x^2}-2\frac{\partial\zeta_{m,n}}{\partial x}\frac{\partial U^*}{\partial x}-\frac{\partial^2\zeta_{m,n}}{\partial x^2}U^*,$$
$$\zeta_{m,n}\frac{\partial^2U^*}{\partial y^2}=\frac{\partial^2u_{m,n}}{\partial y^2}-2\frac{\partial\zeta_{m,n}}{\partial y}\frac{\partial U^*}{\partial y}-\frac{\partial^2\zeta_{m,n}}{\partial y^2}U^*.$$

Теперь (5.1) можно записать в виде

$$-\varepsilon\Delta u_{m,n}+r^*\frac{\partial u_{m,n}}{\partial x}+q^*u_{m,n}=f_{m,n}-\varepsilon\left[2\frac{\partial\zeta_{m,n}}{\partial x}\frac{\partial U^*}{\partial x}+2\frac{\partial\zeta_{m,n}}{\partial y}\frac{\partial U^*}{\partial y}+\Delta\zeta_{m,n}U^*\right]+r^*\frac{\partial\zeta_{m,n}}{\partial x}U^*.$$

Далее преобразуем полученное уравнение для $u_{m,n}$ аналогично тому, как это было сделано в [5] для одномерного случая. Будем иметь

$$-\varepsilon\Delta u_{m,n}(x,y)+2\alpha_{m,n}\frac{\partial u_{m,n}}{\partial x}(x,y)+q_{m,n}u_{m,n}(x,y)=f_{m,n}(x,y)+[2\alpha_{m,n}-r^*(x,y)]\frac{\partial u_{m,n}}{\partial x}(x,y)+$$
$$+[q_{m,n}-q^*(x,y)]u_{m,n}(x,y)-\varepsilon\left[2\frac{\partial\zeta_{m,n}}{\partial x}\frac{\partial U^*}{\partial x}+2\frac{\partial\zeta_{m,n}}{\partial y}\frac{\partial U^*}{\partial y}+\Delta\zeta_{m,n}U^*\right]+r^*\frac{\partial\zeta_{m,n}}{\partial x}U^*,$$
$$m=0,\ldots,2N,\quad n=-N,\ldots,2N,$$

где $\alpha_{m,n}$ и $q_{m,n}$ – некоторые положительные постоянные, которые будут выбраны в дальнейшем. Для указанных значений m, n носители $\Delta_{m,n} = \Delta_m(x) \times \Delta_n(y)$ функции $\zeta_{m,n}(x,y)$ принадлежат $\mathbb{R}^2_+$, поэтому функции $u_{m,n}(x,y)$ можно рассматривать как решения задачи (3.1), (3.2) с соответствующими коэффициентами и правыми частями и граничными условиями

$$g_n(y) = g^*(y)\zeta_{0,n}(0,y), \quad m = 0 \quad \text{и} \quad g_n = 0, \quad m \neq 0.$$

Применим к каждому такому решению $u_{m,n}$ оценку (3.3). Получим

$$\varepsilon |u_{m,n}|_{C^{2,\lambda}} + \left|\frac{\partial u_{m,n}}{\partial x}\right|_{C^\lambda} + \sqrt{\varepsilon}\left|\frac{\partial u_{m,n}}{\partial y}\right|_{C_y^\lambda} + \varepsilon^{\frac{1-\lambda}{2}}\left|\frac{\partial u_{m,n}}{\partial y}\right|_{C_x^\lambda} + \|u_{m,n}\|_{C^\lambda} \le c_{m,n}\left\{\|f^*\zeta_{m,n}\|_{C^\lambda} + \right.$$
$$+ \left\|(2\alpha_{m,n} - r^*(x,y))\frac{\partial u_{m,n}}{\partial x}\right\|_{C_\lambda} + \|(q_{m,n} - q^*(x,y))u_{m,n}\|_{C_\lambda} + \varepsilon\left\|2\frac{\partial \zeta_{m,n}}{\partial x}\frac{\partial U^*}{\partial x} + 2\frac{\partial \zeta_{m,n}}{\partial y}\frac{\partial U^*}{\partial y} + \Delta\zeta_{m,n}U^*\right\|_{C^\lambda} + \quad (5.2)$$
$$\left. + \left\|r^*(x,y)\frac{\partial \zeta_{m,n}}{\partial x}U^*\right\|_{C^\lambda} + \varepsilon|g_n|_{C^{2,\lambda}} + \|g_n\|_{C^{1,\lambda}}\right\}, \quad \lambda \in (0,1).$$

В правой части полученного неравенства присутствуют слагаемые, зависящие от $u_{m,n}$, аналогичные тем, которые имеются и в левой части, и слагаемые, зависящие от U^*. Проведем сначала оценку величин, связанных с $u_{m,n}$. Начнем со второго слагаемого правой части. На основании определения нормы в C_λ и, принимая во внимание правила вычисления постоянной Гёльдера для произведения двух функций, найдем, что

$$\left\|(2\alpha_{m,n} - r^*(x,y))\frac{\partial u_{m,n}}{\partial x}\right\|_{C_\lambda} = \left\|(2\alpha_{m,n} - r^*(x,y))\frac{\partial u_{m,n}}{\partial x}\right\|_{C_\lambda(\Delta_{m,n})} \le$$
$$\le \sup_{(x,y)\in\Delta_{m,n}}|2\alpha_{m,n} - r^*(x,y)|\left(\left|\frac{\partial u_{m,n}}{\partial x}\right|_{C^\lambda} + \left\|\frac{\partial u_{m,n}}{\partial x}\right\|_C\right) + |r^*|_{C^\lambda}\left\|\frac{\partial u_{m,n}}{\partial x}\right\|_C. \quad (5.3)$$

Используя интерполяционное неравенство (см., например, [2], [5])

$$\left\|\frac{\partial v}{\partial x}\right\|_C \le \frac{t^\lambda}{1+\lambda}\left|\frac{\partial v}{\partial x}\right|_{C_x^\lambda} + \frac{2}{t}\|v\|_C, \quad t \in (0,\infty) - \text{любое}, \quad (5.4)$$

получим из (5.2), (5.3) оценку

$$c_{m,n}\left\|(2\alpha_{m,n} - r^*(x,y))\frac{\partial u_{m,n}}{\partial x}\right\|_{C_\lambda} \le c_{m,n}\left[c_{1,m,n}\sup_{(x,y)\in\Delta_{m,n}}|2\alpha_{m,n} - r^*(x,y)|\left|\frac{\partial u_{m,n}}{\partial x}\right|_{C^\lambda(\Delta_{m,n})} + \right.$$
$$\left. + c_{2,m,n}t^\lambda\left|\frac{\partial u_{m,n}}{\partial x}\right|_{C_x^\lambda(\Delta_{m,n})} + c_{3,m,n}\|u_{m,n}\|_C\right].$$

Выберем теперь $\alpha_{m,n}$ из условия

$$2\alpha_{m,n} = \left[\sup_{(x,y)\in\Delta_{m,n}} r^*(x,y) + \inf_{(x,y)\in\Delta_{m,n}} r^*(x,y)\right]/2 = r^*(x_m, y_n), \quad (x_m, y_n) \in \Delta_{m,n}.$$

Будем считать, что размеры $\Delta_{m,n}$ – носителя $\zeta_{m,n}(x,y)$, т.е. длина δ отрезка Δ_m – носителя $\xi_m(x)$ (и Δ_n – носителя $\zeta_n(y)$) столь малы (за счет величины N), что

$$c_{m,n}c_{1,m,n}\sup_{(x,y)\in\Delta_{m,n}}|r^*(x_m,y_n) - r^*(x,y)| \le \frac{1}{4},$$

и выберем t так, чтобы

$$c_{m,n}c_{2,m,n}t^\lambda = 1/4.$$

При оценке третьего слагаемого в правой части (5.2) поступим аналогично, т.е. в выражении

$$\|(q_{m,n} - q^*(x,y))u_{m,n}\|_{C_\lambda} = \|(q_{m,n} - q^*(x,y))\|_C\left[|u_{m,n}|_{C_x^\lambda} + |u_{m,n}|_{C_y^\lambda} + \|u_{m,n}\|_C\right] + |q^*|_{C^\lambda}\|u_{m,n}\|_C$$

выберем

$$q_{m,n} = \left[\sup_{(x,y)\in\Delta_{m,n}} q^*(x,y) + \inf_{(x,y)\in\Delta_{m,n}} q^*(x,y) \right]/2 = q^*(x_m, y_n), \quad (x_m, y_n) \in \Delta_{m,n}.$$

Снова за счет малости $\Delta_{m,n}$ получим оценку

$$c_{m,n} \sup_{(x,y)\in\Delta_{m,n}} |q^*(x_m, y_n) - q^*(x,y)| \le \frac{1}{2}.$$

Теперь, принимая во внимание вышесказанное, а также используя следующую оценку для пятого слагаемого из правой части (5.2) на $\Delta_{m,n}$

$$\left\| r^*(x,y)\frac{\partial \zeta_{m,n}}{\partial x} U^* \right\|_{C^\lambda} \le \left\| r^*(x,y)\frac{\partial \zeta_{m,n}}{\partial x} \right\|_C \|U^*\|_C + \left\| r^*(x,y)\frac{\partial \zeta_{m,n}}{\partial x} \right\|_C |U^*|_{C^\lambda} + \left| r^*\frac{\partial \zeta_{m,n}}{\partial x} \right|_{C^\lambda} \|U^*\|_C,$$

и очевидные оценки других слагаемых там же, будем иметь

$$\begin{aligned} \varepsilon |u_{m,n}|_{C^{2,\lambda}} + \left|\frac{\partial u_{m,n}}{\partial x}\right|_{C^\lambda} + D(u_{m,n}) + \|u_{m,n}\|_{C^\lambda} \le \overline{c}_{mn} \Bigg\{ \|f^*\|_{C^\lambda} + \varepsilon \left[\left|\frac{\partial U^*(x,y)}{\partial x}\right|_{C^\lambda} + \left|\frac{\partial U^*(x,y)}{\partial y}\right|_{C^\lambda} + |U^*(x,y)|_{C^\lambda} \right] + \\ + |U^*|_{C_x^\lambda} + |U^*|_{C_y^\lambda} + \|U^*\|_C + \varepsilon |g_n|_{C^{2,\lambda}} + \|g_n\|_{C^{1,\lambda}} \Bigg\}, \end{aligned} \tag{5.5}$$

где

$$D(u_{m,n}) = \sqrt{\varepsilon} \left|\frac{\partial u_{m,n}}{\partial y}\right|_{C_y^\lambda} + \varepsilon^{\frac{1-\lambda}{2}} \left|\frac{\partial u_{m,n}}{\partial y}\right|_{C_x^\lambda}.$$

Заметим еще, что, исходя из определения функций g_n, имеем следующие оценки для последних слагаемых в правой части (5.5):

$$|g_n|_{C^{2,\lambda}} \le c_{3,m,n} \left[|g^*|_{C^{2,\lambda}} + \|g^*\|_{C^{1,\lambda}} \right],$$
$$\|g_n\|_{C^{1,\lambda}} \le c_{4,m,n} \|g^*\|_{C^{1,\lambda}}.$$

Суммируя полученные для $u_{m,n}$ оценки (5.5), для

$$U^* = \sum_{m=0}^{2N} \sum_{n=-N}^{2N} u_{m,n}$$

получаем оценку

$$\begin{aligned} \varepsilon |U^*|_{C^{2,\lambda}} + \left|\frac{\partial U^*}{\partial x}\right|_{C^\lambda} + D(U^*) + \|U^*\|_{C^\lambda} \le c \Bigg\{ \|f^*\|_{C^\lambda} + \varepsilon \left[\left|\frac{\partial U^*}{\partial x}\right|_{C^\lambda} + \left|\frac{\partial U^*}{\partial y}\right|_{C_x^\lambda} + \left|\frac{\partial U^*}{\partial y}\right|_{C_y^\lambda} + \|U^*\|_{C^\lambda} \right] + \\ + |U^*|_{C_x^\lambda} + |U^*|_{C_y^\lambda} + \|U^*\|_C + \varepsilon |g^*|_{C^{2,\lambda}} + \|g^*\|_{C^{1,\lambda}} \Bigg\}. \end{aligned} \tag{5.6}$$

Пусть $\varepsilon_2 > 0$ таково, что при $\varepsilon \le \varepsilon_2 < 1$ выполняется (так как $\varepsilon < \sqrt{\varepsilon}$, $\varepsilon < \varepsilon^{\frac{1-\lambda}{2}}$ при $\varepsilon < 1$)

$$\varepsilon \left[\left|\frac{\partial U^*}{\partial x}\right|_{C^\lambda} + \left|\frac{\partial U^*}{\partial y}\right|_{C_x^\lambda} + \left|\frac{\partial U^*}{\partial y}\right|_{C_y^\lambda} + \|U^*\|_{C^\lambda} \right] \le \frac{1}{2} \left[\left|\frac{\partial U^*}{\partial x}\right|_{C^\lambda} + D(U^*) + \|U^*\|_{C^\lambda} \right].$$

Далее для оценки $|U^*|_{C_x^\lambda}$ воспользуемся неравенством Юнга (см., например, [9, гл. I, § 15])

$$ab \le \frac{1}{p}(\mu a)^p + \frac{1}{p'}\left(\frac{b}{\mu}\right)^{1/p'}, \quad a > 0, \quad b > 0, \quad \frac{1}{p} + \frac{1}{p'} = 1, \quad \mu > 0 - \text{любое},$$

и интерполяционными неравенствами

$$|u|_{C_x^\lambda} \le 2^{1-\lambda} \left\|\frac{\partial u}{\partial x}\right\|_C^\lambda \|u\|_C^{1-\lambda}$$

(см., например, [5]) и (5.4). Будем иметь

$$|U^*|_{C_x^\lambda} \le 2^{1-\lambda}\left[\lambda\left\|\frac{\partial U^*}{\partial x}\right\|_C + (1-\lambda)\|U^*\|_C\right] \le 2^{1-\lambda}\left\{\lambda\left[\frac{s^\lambda}{1+\lambda}\left|\frac{\partial U^*}{\partial x}\right|_{C_x^\lambda} + 2s^{-1}\|U^*\|_C\right] + (1-\lambda)\|U^*\|_C\right\}.$$

Выберем параметр s так, чтобы выполнялось (c – постоянная из (5.6))

$$c2^{1-\lambda}\frac{\lambda s^\lambda}{1+\lambda} \le \frac{1}{4}.$$

С учетом этого придем к оценке

$$\varepsilon|U^*|_{C^{2,\lambda}} + \left|\frac{\partial U^*}{\partial x}\right|_{C^\lambda} + D(U^*) + \|U^*\|_{C^\lambda} \le c\left\{\|f^*\|_{C^\lambda} + |U^*|_{C_y^\lambda} + \|U^*\|_C + \varepsilon|g^*|_{C^{2,\lambda}} + \|g^*\|_{C^{1,\lambda}}\right\}. \tag{5.7}$$

Замечание 1. Оценка (5.7) справедлива и при $\beta = 0$, однако оценить оставшиеся в правой части (5.7) величины $c\|U^*\|_C$ и $c|U^*|_{C_y^\lambda}$ так, чтобы их можно было исключить за счет присутствующих в левой части соответствующих величин, при $\beta = 0$ не удается, так как стоящий перед ними множитель невозможно сделать малым независимо от ε.

Оценим теперь величины $\|U^*\|_C$ и $|U^*|_{C_y^\lambda}$ в полуплоскости. Напомним, что $U^*(x,y)$ является решением задачи (4.5). Так как решение U^* на бесконечности стремится к нулю, то максимальное значение эта функция принимает в конечной (ограниченной) области, и на основании принципа сравнения (см. [8, Ch. 2, § 6])

$$\|U^*\|_C \le \frac{\|f^*\|_C}{\inf q^*} + \|g^*\|_C. \tag{5.8}$$

Для оценки $|U^*|_{C_y^\lambda}$ преобразуем задачу для U^* из (4.5) к виду

$$-\varepsilon\Delta U^* + 2\alpha\frac{\partial U^*}{\partial x} + QU^* = f^* + (2\alpha - r^*)\frac{\partial U^*}{\partial x} + (Q - q^*)U^* := F^*, \quad (x,y) \in \mathbb{R}^2_+,$$
$$U^*|_{x=0} = g^*(y), \quad -\infty < y < \infty, \quad \lim_{\sqrt{x^2+y^2}\to\infty} U^*(x,y) = 0,$$

где α и Q – некоторые постоянные, подлежащие выбору в дальнейшем. Теперь для оценки величины $|U^*|_{C_y^\lambda}$ воспользуемся результатами [2, лемма 5.1] и [4]. Согласно указанной лемме, при нулевой граничной функции оценка решения уравнения с постоянными коэффициентами имеет вид

$$|U^*|_{C_y^\lambda} \le \frac{1}{Q}\left[1 + \frac{\varepsilon Q}{4\alpha^2}\right]|F^*|_{C_y^\lambda},$$

а с учетом оценки решения с нулевой правой частью и ненулевой граничной функцией, из [2], [4] следует

$$|U^*|_{C_y^\lambda} \le \frac{1}{Q}\left[1 + \frac{\varepsilon Q}{4\alpha^2}\right]|F^*|_{C_y^\lambda} + c_5\|g^*\|_{C^\lambda}.$$

Отсюда вытекает следующая оценка:

$$|U^*|_{C_y^\lambda} \le \frac{1}{Q}\left[1 + \frac{\varepsilon Q}{4\alpha^2}\right]\left\{|f^*|_{C_y^\lambda} + \max|2\alpha - r^*|\left|\frac{\partial U^*}{\partial x}\right|_{C_y^\lambda} + \max|Q - q^*||U^*|_{C_y^\lambda} + \right.$$
$$\left. + |r^*|_{C_y^\lambda}\left\|\frac{\partial U^*}{\partial x}\right\|_C + |q^*|_{C_y^\lambda}|U^*|_C\right\} + c_6\|g^*\|_{C^\lambda}.$$

Пусть

$$Q = q^*_{\max}, \quad 2\alpha = \frac{r^*_{\max} + r^*_{\min}}{2}.$$

Тогда $(Q - q^*) > 0$. Рассмотрим коэффициент при $|U^*|_{C_y^\lambda}$. Заметим, что при

$$\varepsilon \le \frac{2\alpha^2 q^*_{\min}}{Q(Q - q^*_{\min})} = \varepsilon_3 = \frac{(r^*_{\max} + r^*_{\min})^2 q^*_{\min}}{8q^*_{\max}(q^*_{\max} - q^*_{\min})} \tag{5.9}$$

выполняется неравенство

$$\left[1 - \frac{Q - q^*_{\min}}{Q}\left(1 + \frac{\varepsilon Q}{4\alpha^2}\right)\right] \ge \frac{q^*_{\min}}{2Q},$$

а после применения интерполяционного неравенства (5.4) при $\varepsilon \le \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3, 1\}$ (см. (4.3), (5.9)) получим оценку $|U^*|_{C_y^\lambda}$:

$$\begin{aligned} |U^*|_{C_y^\lambda} \le \frac{2}{q^*_{\min}}\left[1 + \frac{\varepsilon q^*_{\max}}{4\alpha^2}\right]\left\{|f^*|_{C_y^\lambda} + \frac{(r^*_{\max} - r^*_{\min})}{2}\left|\frac{\partial U^*}{\partial x}\right|_{C_y^\lambda} + \right. \\ \left. + \frac{|r^*|_{C_y^\lambda} t^\lambda}{1+\lambda}\left|\frac{\partial U^*}{\partial x}\right|_{C_x^\lambda} + \left[\frac{2|r^*|_{C_y^\lambda}}{t} + |q^*|_{C_y^\lambda}\right]\|U^*\|_C + c_7\|g^*\|_{C^\lambda}\right\}. \end{aligned} \tag{5.10}$$

Выразим входящие в (5.10) величины, зависящие от коэффициентов уравнения (4.5), с учетом (4.2), (4.3) и будем считать, что $\varepsilon \le \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3, 1\}$. Тогда (5.10) будет иметь вид

$$\begin{aligned} |U^*|_{C_y^\lambda} \le \frac{c_8}{\beta}\left\{|f^*|_{C_y^\lambda} + \frac{(r^*_{\max} - r^*_{\min})}{2}\left|\frac{\partial U^*}{\partial x}\right|_{C_y^\lambda} + \right. \\ \left. + \frac{|r^*|_{C_y^\lambda} t^\lambda}{1+\lambda}\left|\frac{\partial U^*}{\partial x}\right|_{C_x^\lambda} + \left[\frac{2|r^*|_{C_y^\lambda}}{t} + |q^*|_{C_y^\lambda}\right]\|U^*\|_C + c_7\|g^*\|_{C^\lambda}\right\}. \end{aligned}$$

Выберем теперь β, а затем t так, чтобы

$$\frac{c_8}{\beta}\frac{(r^*_{\max} - r^*_{\min})}{2} \le \frac{1}{4}, \quad \frac{c_8}{\beta}|r^*|_{C_y^\lambda}\frac{t^\lambda}{1+\lambda} \le \frac{1}{4}.$$

Тогда после подстановки последней оценки в (5.7) получим оценку (3.3) для решения задачи (4.5) в полуплоскости. Из (3.3) очевидным образом следует оценка (2.5) при $k = 0$, так как $\varepsilon^{(1-\lambda)/2} > \sqrt{\varepsilon}$ при $0 < \lambda < 1$ и $0 < \varepsilon < 1$.

Теперь рассмотрим функцию $U(x, y)$, являющуюся сужением $U^*(x, y)e^{\beta x}$ на область $0 \le x \le 3/2$ и докажем для нее оценку (2.5). Тогда сужение $U(x, y)$ на Ω и будет искомой регулярной составляющей решения задачи (2.1).

Оценку (2.5) для $U(x, y)$ при $0 \le x \le 3/2$ докажем, подставив в левую часть (2.5) при $k = 0$ функцию $U(x, y) = U^* e^{\beta x}$. Далее, принимая во внимание оценки

$$|a(x,y)b(x,y)|_{C^\lambda} \le \|a(x,y)\|_C\,|b(x,y)|_{C^\lambda} + \|b(x,y)\|_C\,|a(x,y)|_{C^\lambda},$$

$$\left|e^{\beta x}\right|_{C_x^\lambda} \le \frac{\beta e^{3\beta}}{\lambda}, \quad \left|e^{-\beta x}\right|_{C_x^\lambda} \le \frac{\beta^\lambda}{\lambda}\left(\frac{1-\lambda}{e}\right)^{1-\lambda},$$

а также используя интерполяционное неравенство (5.4) в нужных местах, будем иметь

$$\begin{aligned} \varepsilon|U|_{C^{2,\lambda}} + \left|\frac{\partial U}{\partial x}\right|_{C^\lambda} + \sqrt{\varepsilon}\left|\frac{\partial U}{\partial y}\right|_{C^\lambda} + \|U\|_{C^\lambda} \le c_9\left\{\varepsilon|U^*|_{C^{2,\lambda}} + \left|\frac{\partial U^*}{\partial x}\right|_{C^\lambda} + \sqrt{\varepsilon}\left|\frac{\partial U^*}{\partial y}\right|_{C^\lambda} + \|U^*\|_{C^\lambda}\right\} \le \\ \le c\left(\|f\|_{C^\lambda} + \varepsilon|g|_{C^{2,\lambda}} + \|g\|_{C^{1,\lambda}}\right). \end{aligned}$$

Чтобы доказать оценку (2.5) при любом k, достаточно для решения U получить соответствующую оценку при $k = 1$. Для этого продифференцируем уравнение (2.4) для U по y и это же уравнение по x. Заметим, что полученные уравнения для новых функций $\bar{u}(x, y) = \partial U/\partial y$ и $\hat{u}(x, y) = \partial U/\partial x$ с граничными условиями $\bar{u}(0, y) = g'(y)$ и $\hat{u}(0, y) = (\partial U/\partial x)(0, y)$, соответственно, являются уже рассмотренными задачами (но с другими правыми частями и граничными условиями), и потому, воспользовавшись полученными ранее оценками для каждой из них и, где необходимо, интерполяционным неравенством (5.4), после сложения результатов, придем к оценке (2.5) при $k = 1$. Продолжая этот процесс далее, приходим к доказательству теоремы 1.

Изложим вышесказанное более подробно.

1. Задачу для производной функции U по y можно записать в виде

$$L\left(\frac{\partial U}{\partial y}\right) = \frac{\partial f}{\partial y} - \frac{\partial r}{\partial y}\frac{\partial U}{\partial x} - \frac{\partial q}{\partial y}U, \quad \left.\frac{\partial U}{\partial y}\right|_{x=0} = g', \tag{5.11}$$

где оператор L определен в (2.1).

Применим к решению этой задачи доказанную для $k = 0$ оценку (2.5). Будем иметь

$$\begin{aligned} &\varepsilon\left|\frac{\partial U}{\partial y}\right|_{C^{2,\lambda}} + \left|\frac{\partial}{\partial x}\frac{\partial U}{\partial y}\right|_{C^\lambda} + \sqrt{\varepsilon}\left|\frac{\partial}{\partial y}\frac{\partial U}{\partial y}\right|_{C^\lambda} + \left\|\frac{\partial U}{\partial y}\right\|_{C^\lambda} \le \\ &\le c\left(\left\|\frac{\partial f}{\partial y}\right\|_{C^\lambda} + \varepsilon|g'|_{C^{2,\lambda}} + \|g'\|_{C^{1,\lambda}} + \left\|\frac{\partial r}{\partial y}\frac{\partial U}{\partial x}\right\|_{C^\lambda} + \left\|\frac{\partial q}{\partial y}U\right\|_{C^\lambda}\right), \quad \lambda \in (0,1). \end{aligned} \tag{5.12}$$

Рассмотрим два последних слагаемых в правой части оценки (5.12). Так как по определению нормы в C^λ

$$\left\|\frac{\partial r}{\partial y}\frac{\partial U}{\partial x}\right\|_{C^\lambda} = \left|\frac{\partial r}{\partial y}\frac{\partial U}{\partial x}\right|_{C^\lambda} + \left\|\frac{\partial r}{\partial y}\frac{\partial U}{\partial x}\right\|_{C} \le \left\|\frac{\partial r}{\partial y}\right\|_{C}\left|\frac{\partial U}{\partial x}\right|_{C^\lambda} + \left|\frac{\partial r}{\partial y}\right|_{C^\lambda}\left\|\frac{\partial U}{\partial x}\right\|_{C} + \left\|\frac{\partial r}{\partial y}\right\|_{C}\left\|\frac{\partial U}{\partial x}\right\|_{C},$$

и $\left\|\frac{\partial U}{\partial x}\right\|_{C}$ оценивается при помощи интерполяционного неравенства (5.4) через $\left|\frac{\partial U}{\partial x}\right|_{C_x^\lambda}$ и $\|U\|_C$, причем эти величины уже были оценены в (2.5) при $k = 0$ через

$$c\left\{\|f\|_{C^\lambda} + \varepsilon|g|_{C^{2,\lambda}} + \|g\|_{C^{1,\lambda}}\right\},$$

а последнее слагаемое из (5.12) оценивается аналогично, то правая часть неравенства (5.12) оценивается величиной

$$c\left(\left\|\frac{\partial f}{\partial y}\right\|_{C^\lambda} + \|f\|_{C^\lambda} + \varepsilon|g'|_{C^{2,\lambda}} + \|g'\|_{C^{1,\lambda}} + \varepsilon|g|_{C^{2,\lambda}} + \|g\|_{C^{1,\lambda}}\right). \tag{5.13}$$

2. Теперь продифференцируем (2.4) по x. Для функции $\frac{\partial U}{\partial x}$ получим задачу, аналогичную (5.11):

$$L\left(\frac{\partial U}{\partial x}\right) = \frac{\partial f}{\partial x} - \frac{\partial r}{\partial x}\frac{\partial U}{\partial x} - \frac{\partial q}{\partial x}U, \quad \left.\frac{\partial U}{\partial x}\right|_{x=0} = \frac{\partial U}{\partial x}(0, y). \tag{5.14}$$

Функция $\frac{\partial U}{\partial x}$, как решение этой задачи, также оценивается при помощи (2.5), именно

$$\begin{aligned} &\varepsilon\left|\frac{\partial U}{\partial x}\right|_{C^{2,\lambda}} + \left|\frac{\partial}{\partial x}\frac{\partial U}{\partial x}\right\|_{C^\lambda} + \sqrt{\varepsilon}\left|\frac{\partial}{\partial y}\frac{\partial U}{\partial x}\right|_{C^\lambda} + \left\|\frac{\partial U}{\partial x}\right\|_{C^\lambda} \le \\ &\le c\left(\left\|\frac{\partial f}{\partial x}\right\|_{C^\lambda} + \varepsilon\left|\frac{\partial U}{\partial x}(0, y)\right|_{C^{2,\lambda}} + \left\|\frac{\partial U}{\partial x}(0, y)\right\|_{C^{1,\lambda}} + \left\|\frac{\partial r}{\partial x}\frac{\partial U}{\partial x}\right\|_{C^\lambda} + \left\|\frac{\partial q}{\partial x}U\right\|_{C^\lambda}\right), \quad \lambda \in (0,1). \end{aligned} \tag{5.15}$$

Заметим, что

$$\varepsilon\left|\frac{\partial U}{\partial x}(0, y)\right|_{C^{2,\lambda}} = \varepsilon\left|\frac{\partial^3 U}{\partial x \partial y^2}(0, y)\right|_{C^\lambda} \le \varepsilon\left|\frac{\partial^3 U}{\partial x \partial y^2}\right|_{C^\lambda},$$

а

$$\varepsilon\left|\frac{\partial U}{\partial y}\right|_{C^{2,\lambda}} = \varepsilon\left[\left|\frac{\partial^3 U}{\partial y^3}\right|_{C^\lambda} + \left|\frac{\partial^3 U}{\partial x \partial y^2}\right|_{C^\lambda} + \left|\frac{\partial^3 U}{\partial x^2 \partial y}\right|_{C^\lambda}\right],$$

и потому из (5.12), (5.13) следует оценка

$$\varepsilon\left|\frac{\partial U}{\partial x}(0,y)\right|_{C^{2,\lambda}} \le c\left\{\left\|\frac{\partial f}{\partial y}\right\|_{C^\lambda} + \varepsilon|g'|_{C^{2,\lambda}} + \|g'\|_{C^{1,\lambda}} + \|f\|_{C^\lambda} + \varepsilon|g|_{C^{2,\lambda}} + \|g\|_{C^{1,\lambda}}\right\}.$$

Заметим также, что

$$\left\|\frac{\partial U}{\partial x}(0,y)\right\|_{C^{1,\lambda}} = \left\|\frac{\partial^2 U}{\partial x \partial y}(0,y)\right\|_{C^\lambda} \le \left|\frac{\partial^2 U}{\partial x \partial y}\right|_{C^\lambda} + \left\|\frac{\partial^2 U}{\partial x \partial y}\right\|_C,$$

а на основании интерполяционного неравенства (5.4) имеем

$$\left\|\frac{\partial^2 U}{\partial x \partial y}\right\|_C \le \frac{t^\lambda}{1+\lambda}\left|\frac{\partial^2 U}{\partial x \partial y}\right|_{C_x^\lambda} + \frac{2}{t}\left\|\frac{\partial U}{\partial y}\right\|_C, \quad t \in n(0,\infty).$$

Входящие же в правые части двух последних неравенств величины оцениваются правой частью (5.12), т.е. величиной (5.13). Осталось оценить два последних слагаемых в правой части (5.15), а именно,

$$\left\|\frac{\partial r}{\partial x}\frac{\partial U}{\partial x}\right\|_{C^\lambda}, \quad \left\|\frac{\partial q}{\partial x}U\right\|_{C^\lambda}.$$

Снова используя правило вычисления коэффициента Гёльдера от произведения функций и интерполяционное неравенство (5.4), на основании доказанной для U оценки (2.5) при $k = 0$, оценим сумму указанных последних слагаемых. Складывая теперь оценки (5.12) и (5.15), с учетом (5.13) и очевидных неравенств, выражающих связь между младшими и старшими коэффициентами Гёльдера, получаем оценку (2.5) при $k = 1$. Продолжая этот процесс, т.е. дифференцируя каждое из уравнений для $\partial U/\partial x$, $\partial U/\partial y$ снова по x и эти же уравнения по y, после аналогичных рассуждений докажем оценку (2.5) для следующего k. И так далее. Тем самым, придем к установлению указанной оценки для U при всех k. Следовательно, оценка (2.5) установлена при $0 \le x \le 3/2$, а значит, и для $(x,y) \in \Omega$. Тем самым, теорема 1 полностью доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Шишкин Г.И.* Сеточные аппроксимации сингулярно возмущенных эллиптических и параболических уравнений. Екатеринбург: УрО РАН, 1992.
2. *Андреев В.Б.* Оценки в классах Гёльдера регулярной составляющей решения сингулярно возмущенного уравнения конвекции-диффузии // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 12. С. 1983–2020.
3. *Kellogg R.B., Stynes M.* Corner singularities and boundary layers in a simple convection-diffusion problem // J. Differ. Equat. 2005. V. 213. P. 81–120.
4. *Андреев В.Б., Белухина И.Г.* Оценки в классах Гёльдера решения неоднородной задачи Дирихле для сингулярно возмущенного однородного уравнения конвекции-диффузии // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 2. С. 264–276.
5. *Андреев В.Б.* К оценке гладкости регулярной составляющей решения одномерного сингулярно возмущенного уравнения конвекции-диффузии // Ж. вычисл. матем. и матем. физ. 2015. Т. 55. № 1. С. 22–33.
6. *Ладыженская О.А., Уральцева Н.Н.* Линейные и квазилинейные уравнения эллиптического типа. М.: Наука, 1971.
7. *Фихтенгольц Г.М.* Курс дифференциального и интегрального исчисления. Т. 1. М.: Наука, 1962.
8. *Protter M.H., Weinberger H.F.* Maximum principles in differential equations. 2nd ed. Berlin, Heidelberg: Springer, 1984.
9. *Беккенбах Э., Беллман Р.* Неравенства. М.: Мир, 1965.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 217–223

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.951

ОБ АППРОКСИМАЦИИ СЛАБЫХ РЕШЕНИЙ УРАВНЕНИЯ ЛАПЛАСА ГАРМОНИЧЕСКИМИ МНОГОЧЛЕНАМИ

© 2021 г. М. Е. Боговский[1,2]

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

[2] 141701 Долгопрудный, М.о., Институтский пер., 9, МФТИ, Россия

e-mail: bogovskii@ccas.ru, bogovskii.me@mipt.ru

Поступила в редакцию 16.06.2020 г.
Переработанный вариант 21.07.2020 г.
Принята к публикации 15.08.2020 г.

В статье дано новое, основанное на идеологии Ф. Браудера, доказательство теоремы об аппроксимации гармоническими многочленами в пространствах Лебега $L_p(\Omega)$ и Соболева $W_p^1(\Omega)$ слабых решений уравнения Лапласа в ограниченной области $\Omega \subset \mathbb{R}^n$, $n \geq 2$, со связной липшицевой границей. Библ. 9.

Ключевые слова: проблема аппроксимации, гармонические многочлены, ограниченная область в $\mathbb{R}^n$, липшицева граница, пространство Лебега $L_p(\Omega)$, пространство Соболева $W_p^1(\Omega)$, слабые решения уравнения Лапласа.

DOI: 10.31857/S0044466921010038

1. ВВЕДЕНИЕ

Ф. Браудером в [1], [2] было установлено, что если $\mathcal{A}$ – линейный эллиптический дифференциальный оператор, то решение уравнения

$$\mathcal{A}u = 0, \quad x \in \Omega, \tag{1.1}$$

в области $\Omega \in \mathbb{R}^n$, $n \geq 2$, может быть аппроксимировано в норме пространства Соболева $W_p^l(\Omega)$ решениями того же уравнения для какой-либо подходящей области $\widetilde{\Omega} \supset \overline{\Omega}$, содержащей замыкание $\overline{\Omega}$ области Ω, если решения для $\widetilde{\Omega}$ образуют аппроксимативный базис в $W_p^l(\widetilde{\Omega})$.

В настоящей работе рассматривается вопрос об аппроксимации гармоническими многочленами слабых решений уравнения Лапласа в пространствах Лебега $L_p(\Omega)$ и Соболева $W_p^1(\Omega)$ для ограниченной области $\Omega \in \mathbb{R}^n$, $n \geq 2$, со связной липшицевой границей при $p \in (1, \infty)$.

Необходимо отметить, что подход Ф. Браудера к аппроксимации решений эллиптических уравнений в норме W_p^l принципиально опирается на двойственность пространства $\overset{\circ}{W}{}_p^l$ и пространства функционалов $\overset{\circ}{W}{}_{p'}^{-l}$ с сопряженным показателем $p' = p/(p-1)$. Существенным недостатком такого подхода является чрезмерная сложность неизбежно возникающих при этом вспомогательных задач для оператора $\mathcal{A}$, когда $\mathcal{A}$ уже не является простейшим оператором Лапласа. Этого недостатка лишен представленный в настоящей статье новый подход к задачам аппроксимации решений.

2. АППРОКСИМАЦИЯ СЛАБЫХ РЕШЕНИЙ УРАВНЕНИЯ ЛАПЛАСА

В этом разделе рассматриваются вопросы аппроксимации гармоническими многочленами слабых решений уравнения Лапласа $\Delta u = 0$ в нормах $L_p(\Omega)$ и $W_p^1(\Omega)$ для области Ω, т.е. для открытого связного множества $\Omega \subset \mathbb{R}^n$, $n \geq 2$. Существенно, что на замкнутых подпространствах сла-

бых решений классов $L_p(\Omega)$ и $W_p^1(\Omega)$ устанавливается аппроксимативная базисность системы однородных гармонических многочленов, ортогональных на сфере S_R в $\mathbb{R}^n$ какого-либо радиуса $R > 0$, выбранного и зафиксированного так, чтобы сфера S_R содержала замыкание $\overline{\Omega}$ рассматриваемой области Ω. Отметим, что аппроксимативная базисность здесь понимается в строгом смысле определения 9.1 в [3, с. 275].

Как с более простого, начнем со случая аппроксимации в норме $L_p(\Omega)$. При этом функцию $u \in L_p(\Omega)$ будем называть слабым решением уравнения Лапласа в $\Omega \subset \mathbb{R}^n$, если выполнено тождество

$$\int_\Omega u(x)v(x)dx = 0 \quad \forall v \in \overset{\circ}{C}{}^\infty(\Omega),$$

где символом $\overset{\circ}{C}{}^\infty(\Omega)$ обозначено пространство всех бесконечно дифференцируемых в Ω функций с носителем, компактным в Ω. Справедлива следующая

Теорема 1. *Пусть $\Omega \subset \mathbb{R}^n$ – ограниченная область с липшицевой связной границей, $n \geq 2$, $1 < p < \infty$. Для всякого слабого решения уравнения Лапласа $u \in L_p(\Omega)$ найдется последовательность гармонических многочленов $\{h_m\}_{m=1}^\infty$ такая, что*

$$\lim_{m\to\infty} \|u - h_m\|_{L_p(\Omega)} = 0.$$

Доказательство. Через $\mathscr{H}_p(\Omega)$ обозначим замыкание в $L_p(\Omega)$ подпространства всех гармонических многочленов. И пусть

$$\widehat{\mathscr{H}}_p(\Omega) \overset{\text{def}}{=} \left\{ u \in L_p(\Omega) : \int_\Omega (u, \Delta\psi)dx = 0 \quad \forall \psi \in \overset{\circ}{C}{}^\infty(\Omega) \right\}$$

есть подпространство слабых решений уравнения Лапласа в области Ω. Очевидно, что $\mathscr{H}_p(\Omega) \subset \widehat{\mathscr{H}}_p(\Omega)$. Теорема будет доказана, если установить обратное включение $\widehat{\mathscr{H}}_p(\Omega) \subset \mathscr{H}_p(\Omega)$. Для этого введем обозначение

$$\overset{\circ}{\mathscr{D}}_p(\Omega) \overset{\text{def}}{=} \{f = \Delta v : v \in \overset{\circ}{W}{}_p^2(\Omega)\}, \tag{2.1}$$

где пространство Соболева $\overset{\circ}{W}{}_p^2(\Omega)$ определено как замыкание в $W_p^2(\Omega)$ его подпространства $\overset{\circ}{C}{}^\infty(\Omega)$, что для ограниченной области Ω обеспечивает замкнутость в $L_p(\Omega)$ подпространства $\overset{\circ}{\mathscr{D}}_p(\Omega)$, определенного в (2.1) как область значений оператора Лапласа $\Delta: \overset{\circ}{W}{}_p^2(\Omega) \to L_p(\Omega)$.

Нетрудно убедиться, что при $1 < p < \infty$ подпространство $\widehat{\mathscr{H}}_p(\Omega)$ замкнуто в $L_p(\Omega)$. Кроме того, имеем

$$\begin{aligned} &\widehat{\mathscr{H}}_p(\Omega)^\perp = \overset{\circ}{\mathscr{D}}_{p'}(\Omega), \\ &\overset{\circ}{\mathscr{D}}_p(\Omega)^\perp = \widehat{\mathscr{H}}_{p'}(\Omega), \end{aligned} \qquad p' = p/(p-1), \quad 1 < p < \infty,$$

где символом $X^\perp$ обозначен аннулятор подпространства $X \subset L_p(\Omega)$. Отметим (см. [4, разд. 4.5–6)], что при $1 < p < \infty$ аннулятор подпространства $X \subset L_p(\Omega)$ будет сильно замкнутым подпространством в $L_{p'}(\Omega)$ с сопряженным показателем $p' = p/(p-1)$.

По определению ограниченный сферой S_R открытый шар $B_R \overset{\text{def}}{=} \{x \in \mathbb{R}^n : |x| < R\}$ содержит замыкание $\overline{\Omega}$ области Ω. С помощью растяжений и усреднений легко проверить, что $\widehat{\mathscr{H}}_p(B_R)$ совпадает с замыканием в $L_p(B_R)$ его подпространства

$$\mathscr{H}^\infty(\overline{B}_R) = \{u \in C^\infty(\overline{B}_R) : \Delta u = 0\}.$$

Используя ортогональное разложение гармонических функций в $L_2(B_R)$ по сферическим гармоникам (см. [5], [6]), легко убедиться, что любой элемент подпространства $\mathscr{H}^\infty(\overline{B}_R)$ аппроксимируется гармоническими многочленами в норме пространства Соболева $W_2^l(B_R)$ для любого $l \geq n/2$, а значит, и в норме $L_p(B_R)$ при $p \in (1,\infty)$ в силу вложения $W_2^l(B_R)$ в $C(\overline{B}_R)$. Это означает совпадение подпространств $\widehat{\mathscr{H}}_p(B_R) = \mathscr{H}_p(B_R)$ при любом $p \in (1,\infty)$.

Теорема будет доказана, если убедиться, что замыкание в $L_p(\Omega)$ подпространства сужений на Ω функций из $\mathscr{H}^\infty(\overline{B}_R)$ совпадает с $\widehat{\mathscr{H}}_p(\Omega)$ при любом значении $p \in (1,\infty)$. Предположим противное. Тогда для некоторого $p \in (1,\infty)$ в силу теоремы Рисса об общем виде линейного непрерывного функционала на $L_p(\Omega)$ найдутся ненулевые элементы $f \in L_{p'}(\Omega)$ и $v \in \widehat{\mathscr{H}}_p(\Omega)$, удовлетворяющие условиям

$$\int_\Omega f(x)v(x)dx = 1, \quad \int_\Omega f(x)u(x)dx = 0 \quad \forall u \in \mathscr{H}^\infty(\overline{B}_R). \tag{2.2}$$

Доопределяя $f = f(x)$ нулем в точках $x \in B_R\backslash\Omega$, получаем функцию $f \in L_{p'}(B_R)$, удовлетворяющую условию

$$\int_{B_R} f(x)u(x)dx = 0 \quad \forall u \in \mathscr{H}^\infty(\overline{B}_R),$$

которое означает, что $f \in \mathscr{H}_p(B_R)^\perp = \mathring{\mathscr{D}}_{p'}(B_R)$ и найдется такая функция $w \in \mathring{W}^2_{p'}(B_R)$, что $f(x) = \Delta w(x)$ почти всюду в B_R. При этом для области $B_R\backslash\overline{\Omega}$ функция w оказывается решением однородной задачи Коши

$$\begin{aligned} \Delta w = 0, \quad x \in B_R\backslash\overline{\Omega}, \\ w|_{r=R} = \partial_r w|_{r=R} = 0, \end{aligned} \tag{2.3}$$

где ∂_r – производная по нормали к ∂B_R. А ввиду связности открытого множества $B_R\backslash\overline{\Omega}$, единственное решение однородной задачи Коши для оператора Лапласа может быть только нулевым на $B_R\backslash\overline{\Omega}$, т.е. функция $w \in \mathring{W}^2_{p'}(B_R)$ будет тождественным нулем $w = 0$ на $B_R\backslash\overline{\Omega}$.. Последнее означает, что $w \in \mathring{W}^2_{p'}(\Omega)$, так как граница $\partial\Omega$ липшицева (подробности см. в [7]). Но тогда $f = \Delta w \in \mathring{\mathscr{D}}_{p'}(\Omega) = \widehat{\mathscr{H}}_p(\Omega)^\perp$, что противоречит левому из двух равенств (2.2), так как $v \in \widehat{\mathscr{H}}_p(\Omega)$. Полученное противоречие завершает доказательство теоремы 1.

Функцию $u \in W_p^1(\Omega)$ будем называть слабым решением уравнения Лапласа в области Ω, если

$$\int_\Omega \nabla u \cdot \nabla\psi dx = 0 \quad \forall \psi \in \mathring{C}^\infty(\Omega),$$

где точка означает скалярное произведение векторов в $\mathbb{R}^n$. Для таких функций справедлива следующая теорема об аппроксимации гармоническими многочленами в норме пространства Соболева $W_p^1(\Omega)$.

Теорема 2. *Пусть $\Omega \subset \mathbb{R}^n$ – ограниченная область с липшицевой связной границей, $n \geq 2$, $1 < p < \infty$. Если $u \in W_p^1(\Omega)$ – слабое решение уравнения Лапласа в Ω, то найдется последовательность гармонических многочленов $\{h_m\}_{m=1}^\infty$ такая, что*

$$\lim_{m\to\infty} \|u - h_m\|_{W_p^1(\Omega)} = 0.$$

Доказательство теоремы существенно облегчила бы формальная отсылка к известному $\mathbf{L}_p$-разложению Гельмгольца–Вейля–Соболева. Но в рассматриваемом общем случае связной липшицевой границы $\partial\Omega$ это разложение известно только для достаточно малой окрестности значений показателя $p = 2$, расширение которой требует, вообще говоря, значительных допол-

нительных ограничений на $\partial\Omega$, заведомо лишних для доказываемой теоремы. Избежать лишних ограничений на связную липшицеву $\partial\Omega$ позволит использование лишь двух из трех компонент, участвующих в $\mathbf{L}_p$-разложении Гельмгольца–Вейля–Соболева, что потребует введения дополнительных обозначений.

Сначала введем вспомогательную ограниченную область $\Omega \in \mathbb{R}^n$ с липшицевой связной границей. Через $\mathbf{L}_p(\Omega)$ обозначим пространство Лебега векторных полей $\mathbf{v}\colon \Omega \to \mathbb{R}^n$, и пусть $\overset{\circ}{\mathbf{G}}{}^{\infty}(\Omega)$ и $\overset{\circ}{\mathbf{J}}{}^{\infty}(\Omega)$ – подпространства бесконечно дифференцируемых и финитных в Ω потенциальных и соленоидальных векторных полей соответственно, т.е.

$$\overset{\circ}{\mathbf{G}}{}^{\infty}(\Omega) = \{\nabla u\colon u \in \overset{\circ}{C}{}^{\infty}(\Omega)\}, \quad \overset{\circ}{\mathbf{J}}{}^{\infty}(\Omega) = \{\mathbf{v} \in \overset{\circ}{\mathbf{C}}{}^{\infty}(\Omega)\colon \operatorname{div}\mathbf{v} = 0\}.$$

Замыкания подпространств $\overset{\circ}{\mathbf{G}}{}^{\infty}(\Omega)$ и $\overset{\circ}{\mathbf{J}}{}^{\infty}(\Omega)$ в пространстве Лебега $\mathbf{L}_p(\Omega)$ обозначим через $\overset{\circ}{\mathbf{G}}_p(\Omega)$ и $\overset{\circ}{\mathbf{J}}_p(\Omega)$ соответственно. При этом для ограниченной области Ω с липшицевой связной границей $\partial\Omega$ имеем эквивалентное определение

$$\overset{\circ}{\mathbf{G}}_p(\Omega) = \{\nabla u\colon u \in \overset{\circ}{W}{}^1_p(\Omega)\},$$

где пространство Соболева $W^1_p(\Omega)$ определено как замыкание в $W^1_p(\Omega)$ его подпространства $\overset{\circ}{C}{}^{\infty}(\Omega)$. В общем случае обычно вводится еще одно, не всегда эквивалентное, определение подпространства $\overset{\circ}{\mathbf{J}}_p(\Omega)$ (см. [8]), но в рассматриваемом здесь простейшем случае ограниченной области $\Omega \in \mathbb{R}^n$ со связной липшицевой границей в этом нет необходимости.

Вышеупомянутая пара компонент $\mathbf{L}_p$-разложения Гельмгольца–Вейля–Соболева – это замкнутые в $\mathbf{L}_p(\Omega)$ подпространства $\overset{\circ}{\mathbf{G}}_p(\Omega)$ и $\overset{\circ}{\mathbf{J}}_p(\Omega)$. Сразу же заметим, что их алгебраическая сумма $\overset{\circ}{\mathbf{G}}_p(\Omega) + \overset{\circ}{\mathbf{J}}_p(\Omega)$ будет прямой, т.е. их пересечение тривиально: $\overset{\circ}{\mathbf{G}}_p(\Omega) \cap \overset{\circ}{\mathbf{J}}_p(\Omega) = \{0\}$, что в случае ограниченной Ω почти очевидно, так как открытое множество $\Omega' \overset{\text{def}}{=} \mathbb{R}^n \backslash \overline{\Omega}$ не пусто, а всякое $\mathbf{v} \in \overset{\circ}{\mathbf{G}}_p(\Omega) \cap \overset{\circ}{\mathbf{J}}_p(\Omega)$ представимо в виде $\mathbf{v} = \nabla u$ с некоторой $u \in \overset{\circ}{W}{}^1_p(\Omega)$.

Доопределяя $u \in \overset{\circ}{W}{}^1_p(\Omega)$ нулем на все $\mathbb{R}^n$ с сохранением класса W^1_p и не меняя обозначений, будем без ограничения общности считать, что $u \in W^1_p(\mathbb{R}^n)$, где $u \subset \overline{\Omega}$. Принадлежность $\mathbf{v} \in \overset{\circ}{\mathbf{J}}_p(\Omega)$ означает интегральное тождество

$$\int_{\Omega} \mathbf{v}\cdot\nabla\psi dx = 0 = \int_{\mathbb{R}^n} \nabla u\cdot\nabla\psi dx \qquad \psi \in \overset{\circ}{C}{}^{\infty}(\mathbb{R}^n),$$

в силу которого гармоническая в $\mathbb{R}^n$ функция $u \in C^{\infty}(\mathbb{R}^n)$, тождественно равняясь нулю вне Ω, может быть только тождественным нулем в $\mathbb{R}^n$. Таким образом, $\mathbf{v} \equiv 0$ и пересечение подпространств $\overset{\circ}{\mathbf{G}}_p(\Omega)$ и $\overset{\circ}{\mathbf{J}}_p(\Omega)$ тривиально при любом значении показателя $p \in (1,\infty)$.

Проверим теперь, что прямая алгебраическая сумма $\overset{\circ}{\mathbf{G}}_p(\Omega) + \overset{\circ}{\mathbf{J}}_p(\Omega)$ замкнутых подпространств $\overset{\circ}{\mathbf{G}}_p(\Omega)$ и $\overset{\circ}{\mathbf{J}}_p(\Omega)$ будет замкнутым подпространством в $\mathbf{L}_p(\Omega)$. Для этого достаточно установить существование такой постоянной $C > 0$, что

$$\|\nabla u\|_{\mathbf{L}_p(\Omega)} + \|\mathbf{v}\|_{\mathbf{L}_p(\Omega)} \le C\|\nabla u + \mathbf{v}\|_{\mathbf{L}_p(\Omega)} \qquad \forall u \in \overset{\circ}{W}{}^1_p(\Omega), \quad \forall \mathbf{v} \in \overset{\circ}{\mathbf{J}}_p(\Omega). \tag{2.4}$$

С этой целью, полагая $\mathbf{f} \overset{\text{def}}{=} \nabla u + \mathbf{v}$ и сохраняя прежние обозначения, доопределим $u \in \overset{\circ}{W}{}^1_p(\Omega)$, $\mathbf{v} \in (\Omega)$ и $\mathbf{f} \in \mathbf{L}_p(\Omega)$ нулем вне ограниченной области Ω. При этом получим $\nabla u \in \overset{\circ}{\mathbf{G}}_p(\mathbb{R}^n)$ и $\mathbf{v} \in \overset{\circ}{\mathbf{J}}_p(\mathbb{R}^n)$ в силу определения $\overset{\circ}{\mathbf{G}}_p(\Omega)$ и $\overset{\circ}{\mathbf{J}}_p(\Omega)$ как замыканий в $\mathbf{L}_p(\Omega)$ подпространств $\overset{\circ}{\mathbf{G}}{}^{\infty}(\Omega)$ и $\overset{\circ}{\mathbf{J}}{}^{\infty}(\Omega)$, тогда как носители элементов u, $\mathbf{v}$, $\mathbf{f}$ окажутся подмножествами замыкания $\overline{\Omega}$. В таком

случае почти всюду в $\mathbb{R}^n$ будет выполняться равенство $\nabla u(x) + \mathbf{v}(x) = \mathbf{f}$, преобразование Фурье которого с учетом принадлежности $\mathbf{v} \in \mathring{\mathbf{J}}_p(\mathbb{R}^n)$ приводит к равенствам

$$i\xi\hat{u}(\xi) + \hat{\mathbf{v}}(\xi) = \hat{\mathbf{f}}(\xi), \quad \xi \cdot \hat{\mathbf{v}}(\xi) = 0 \quad \forall \xi \in \mathbb{R}^n, \tag{2.5}$$

где крышечка означает преобразование Фурье, т.е.

$$\hat{u}(\xi) = F[u(x)] = \int\limits_{\mathbb{R}^n} u(x)e^{-i\xi\cdot x}dx, \quad \xi \in \mathbb{R}^n.$$

При этом компактность носителей элементов u, $\mathbf{v}$, $\mathbf{f}$ означает непрерывность на всем $\mathbb{R}^n$ их образов Фурье $\hat{u}$, $\hat{\mathbf{v}}$, $\hat{\mathbf{f}}$. Из (2.5) легко находим

$$i\xi_j\hat{u}(\xi) = \xi_j \frac{\xi \cdot \hat{\mathbf{f}}(\xi)}{|\xi|^2}, \quad v_j(\xi) = f_j(\xi) - \xi_j \frac{\xi \cdot \hat{\mathbf{f}}(\xi)}{|\xi|^2} \quad \forall \xi \in \mathbb{R}^n, \quad j = 1, \ldots, n,$$

где $\hat{\mathbf{f}} = (\hat{f}_1, \ldots, \hat{f}_n)$, откуда следует представление

$$\nabla u = \mathbf{R}(\mathbf{R} \cdot \mathbf{f}), \quad \mathbf{v} = \mathbf{f} - \mathbf{R}(\mathbf{R} \cdot \mathbf{f}) \tag{2.6}$$

слагаемых ∇u и $\mathbf{v}$ через их сумму $\mathbf{f}$ с помощью вектор-оператора

$$\mathbf{R}\colon L_p(\mathbb{R}^n) \to \mathbf{L}_p(\mathbb{R}^n), \quad \mathbf{R} = (R_1, \ldots, R_n), \quad 1 < p < \infty,$$

компонентами которого $R_j\colon L_p(\mathbb{R}^n) \to L_p(\mathbb{R}^n)$ служат ограниченные на $L_p(\mathbb{R}^n)$ преобразования Рисса

$$R_j w = F^{-1}\left[\frac{\xi_j}{|\xi|}\hat{w}\right], \quad \widehat{R_j w} = \frac{\xi_j}{|\xi|}\hat{w}, \quad j = 1, \ldots, n,$$

подробно описанные в [9]. Из представления (2.6) и ограниченности вектор-оператора $\mathbf{R}\colon L_p(\mathbb{R}^n) \to \mathbf{L}_p(\mathbb{R}^n)$ при $1 < p < \infty$ вытекает справедливость неравенства (2.4) с некоторой постоянной $C > 0$, зависящей только от n и p. А тогда алгебраическая сумма $\mathring{\mathbf{G}}_p(\Omega) + \mathring{\mathbf{J}}_p(\Omega)$ будет еще и прямой топологической суммой $\mathring{\mathbf{G}}_p(\Omega) \oplus \mathring{\mathbf{J}}_p(\Omega)$ замкнутых подпространств $\mathring{\mathbf{G}}_p(\Omega)$ и $\mathring{\mathbf{J}}_p(\Omega)$.

Подпространство градиентов слабо гармонических функций в пространстве Лебега $\mathbf{L}_p(\Omega)$ обозначим через $\mathbf{I}_p(\Omega)$. Поскольку ограниченная область Ω имеет липшицеву границу, можно без ограничения общности полагать, что

$$\mathbf{I}_p(\Omega) = \left\{\nabla u\colon u \in W_p^1(\Omega), \int\limits_\Omega u(x)dx = 0, \int\limits_\Omega \nabla u \cdot \nabla\psi dx = 0 \quad \forall \psi \in \mathring{C}^\infty(\Omega)\right\}.$$

Очевидно, определенное таким образом подпространство $\mathbf{I}_p(\Omega)$ будет замкнуто в пространстве Лебега $\mathbf{L}_p(\Omega)$.

Для завершения доказательства теоремы нам понадобятся еще два замкнутых в $\mathbf{L}_p(\Omega)$ подпространства потенциальных и соленоидальных векторных полей

$$\begin{gathered}\mathbf{G}_p(\Omega) = \{\nabla u\colon u \in W_p^1(\Omega)\},\\ \mathbf{J}_p(\Omega) = \left\{\mathbf{v} \in \mathbf{L}_p(\Omega)\colon \int\limits_\Omega \mathbf{v} \cdot \nabla\psi dx = 0 \quad \forall \psi \in \mathring{C}^\infty(\Omega)\right\},\end{gathered} \tag{2.7}$$

которые служат аннуляторами подпространств $\mathring{\mathbf{G}}_p(\Omega)$ и $\mathring{\mathbf{J}}_p(\Omega)$, т.е.

$$\mathring{\mathbf{G}}_p(\Omega)^\perp = \mathbf{J}_{p'}(\Omega), \quad \mathring{\mathbf{J}}_p(\Omega)^\perp = \mathbf{G}_{p'}(\Omega), \quad p' = p/(p-1), \quad 1 < p < \infty, \tag{2.8}$$

где верхний индекс $\perp$ в обозначении сильно замкнутого подпространства в рефлексивном $\mathbf{L}_p(\Omega)$ превращает это подпространство в его сильно замкнутый в $\mathbf{L}_{p'}(\Omega)$ аннулятор (см. [4, с. 108, 109]).

Отметим, что первое из двух равенств (2.9) вытекает непосредственно из теоремы Рисса об общем виде линейного непрерывного функционала на $\mathbf{L}_p(\Omega)$ и определения $\mathring{\mathbf{G}}_p(\Omega)$, тогда как второе принято считать очевидным следствием теоремы двойственности де Рама без каких бы то ни

было требований к области $\Omega \subset \mathbb{R}^n$. Однако для ограниченной области с липшицевой границей фундаментальная и необобщаемая теорема двойственности де Рама эквивалентна паре равенств, установленных в статье [8], а именно, следствию из теоремы 1 на с. 153 в совокупности с равенством (25) на с. 157.

Отметим также, что пересечение подпространств (2.7) легко вычисляется

$$\mathbf{G}_p(\Omega) \cap \mathbf{J}_p(\Omega) = \mathbf{I}_p(\Omega), \quad 1 < p < \infty, \tag{2.9}$$

так как принадлежность $\mathbf{v} \in \mathbf{G}_p(\Omega)$ означает, что $\mathbf{v} = \nabla u$ с некоторым $u \in W_p^1(\Omega)$, а принадлежность того же $\mathbf{v} \in \mathbf{J}_p(\Omega)$ означает выполнение условия

$$\int_\Omega \nabla u \cdot \nabla \psi dx = 0 \quad \forall \psi \in \overset{\circ}{C}{}^\infty(\Omega),$$

откуда сразу же следует принадлежность $\mathbf{v} \in \mathbf{I}_p(\Omega)$, т.е. $\mathbf{G}_p(\Omega) \cap \mathbf{J}_p(\Omega) \subset \mathbf{I}_p(\Omega)$. Обратное включение следует из определения $\mathbf{I}_p(\Omega)$ как подпространства градиентов всех слабых решений уравнения Лапласа класса $W_p^1(\Omega)$.

Вычислим теперь аннулятор подпространства $\mathbf{I}_p(\Omega)$. Для этого заметим сначала, что в силу (2.8) и (2.9) аннулятор алгебраической суммы

$$(\overset{\circ}{\mathbf{G}}_p(\Omega) + \overset{\circ}{\mathbf{J}}_p(\Omega))^\perp = \overset{\circ}{\mathbf{G}}_p(\Omega)^\perp \cap \overset{\circ}{\mathbf{J}}_p(\Omega)^\perp = \mathbf{G}_{p'}(\Omega) \cap \mathbf{J}O_{p'}(\Omega) = \mathbf{I}_{p'}(\Omega),$$

откуда ввиду уже установленной при всех значениях показателя $p \in (1,\infty)$ замкнутости в $\mathbf{L}_p(\Omega)$ его подпространства $\overset{\circ}{\mathbf{G}}_p(\Omega) + \overset{\circ}{\mathbf{J}}_p(\Omega)$ находим

$$\mathbf{I}_p(\Omega)^\perp = \overset{\circ}{\mathbf{G}}_{p'}(\Omega) + \overset{\circ}{\mathbf{J}}_{p'}(\Omega) = \overset{\circ}{\mathbf{G}}_{p'}(\Omega) \oplus \overset{\circ}{\mathbf{J}}_{p'}(\Omega) \tag{2.10}$$

с сопряженным показателем $p' = p/(p-1) \in (1,\infty)$.

Наконец все готово для завершения доказательства теоремы. Через $\mathscr{H}_p^1(\Omega)$ обозначим замыкание в $W_p^1(\Omega)$ подпространства гармонических многочленов, а через $\widehat{\mathscr{H}}_p^1(\Omega)$ – подпространство слабых решений уравнения Лапласа в области Ω класса $W_p^1(\Omega)$, т.е.

$$\widehat{\mathscr{H}}_p^1(\Omega) = \left\{ u \in W_p^1(\Omega) \colon \int_\Omega (\nabla u, \nabla \psi) dx = 0 \quad \forall \psi \in \overset{\circ}{C}{}^\infty(\Omega) \right\}.$$

Очевидно, что $\mathscr{H}_p^1(\Omega) \subset \widehat{\mathscr{H}}_p^1(\Omega)$. Осталось доказать обратное включение, т.е. $\widehat{\mathscr{H}}_p^1(\Omega) \subset \mathscr{H}_p^1\Omega)$.

Для рассматриваемой ограниченной области $\Omega \subset \mathbb{R}^n$ с липшицевой связной границей выберем и зафиксируем радиус $R > 0$ так, чтобы открытый шар $B_R = \{x \in \mathbb{R}^n : |x| < R\}$ содержал замыкание $\overline{\Omega}$. С помощью растяжений и усреднений легко проверить, что $\widehat{\mathscr{H}}_p^1(B_R)$ совпадает с замыканием в $W_p^1(B_R)$ его подпространства гармонических функций $\mathscr{H}^\infty(\overline{B}_R)$. А используя ортогональное разложение элементов $\mathscr{H}^\infty(\overline{B}_R)$ по сферическим гармоникам (см. [5], [6]), заключаем, что любой элемент подпространства $\mathscr{H}^\infty(\overline{B}_R)$ аппроксимируется гармоническими многочленами в норме пространства Соболева $W_2^l(B_R)$ для любого $l \geq 1$. Поэтому в силу вложения $W_2^l(B_R)$ в $W_p^1(B_R)$ при значениях $l \geq 1 + n/2$ любой элемент подпространства $\mathscr{H}^\infty(\overline{B}_R)$ аппроксимируется гармоническими многочленами в норме $W_p^1(B_R)$ с любым показателем $p \in (1,\infty)$. Последнее означает совпадение подпространств $\widehat{\mathscr{H}}_p^1(B_R) = \mathscr{H}_p^1(B_R)$ при любом $p \in (1,\infty)$.

Таким образом, для доказательства теоремы достаточно убедиться, что замыкание в $W_p^1(\Omega)$ подпространства сужений на Ω функций из $\mathscr{H}^\infty(\overline{B}_R)$ совпадает с $\widehat{\mathscr{H}}_p^1(\Omega)$ при любом $p \in (1,\infty)$, что

эквивалентно совпадению замыкания в $\mathbf{L}_p(\Omega)$ подпространства градиентов гармонических функций как элементов $\mathbf{I}_p(B_R)$ с подпространством $\mathbf{I}_p(\Omega)$. Предположим противное. Тогда для некоторого $p \in (1,\infty)$ по теореме Рисса об общем виде линейного непрерывного функционала на $\mathbf{L}_p(\Omega)$ найдутся ненулевые элементы $\mathbf{f} \in \mathbf{L}_{p'}(\Omega)$ и $\mathbf{v} \in \mathbf{I}_p(\Omega)$, удовлетворяющие условиям

$$\int_\Omega \mathbf{f}\cdot\mathbf{v}dx = 1, \quad \int_\Omega \mathbf{f}\cdot\mathbf{w}dx = 0 \quad \forall \mathbf{w} \in \mathbf{I}_p(B_R). \tag{2.11}$$

Доопределяя $\mathbf{f}$ нулем на $B_R\backslash\Omega$, получаем вектор-функцию $\mathbf{f} \in \mathbf{L}_{p'}(\Omega)$, удовлетворяющую условию

$$\int_{B_R} \mathbf{f}\cdot\mathbf{w}dx = 0 \quad \forall \mathbf{w} \in \mathbf{I}_p(B_R),$$

которое означает принадлежность

$$\mathbf{f} \in \mathbf{I}_p(B_R)^\perp = \mathring{\mathbf{G}}_{p'}(B_R) \oplus \mathring{\mathbf{J}}_{p'}(B_R).$$

А тогда доопределенная нулем вне Ω вектор-функция $\mathbf{f}$ должна иметь вид

$$\mathbf{f} = \nabla U + \mathbf{V}, \quad U \in \mathring{W}^l_{p'}(B_R), \quad \mathbf{V} \in \mathring{\mathbf{J}}_{p'}(B_R),$$

при условии $\mathbf{f} = 0$ на $B_R\backslash\Omega$, которое означает выполнение равенств $U = 0$ и $\mathbf{V} = 0$ вне области Ω.

Поскольку граница $\partial\Omega$ липшицева и $U \in \mathring{W}^l_{p'}(B_R)$, условие $U = 0$ вне Ω с липшицевой границей означает, что $U \in \mathring{W}^l_{p'}(\Omega)$ (см. [7]). В свою очередь, выполнение условия $\mathbf{V} = 0$ вне ограниченной области Ω с липшицевой границей для вектор-функции $\mathbf{V} \in \mathring{\mathbf{J}}_{p'}(B_R)$ означает, что $\mathbf{V} \in \mathring{\mathbf{J}}_{p'}(\Omega)$ (см. [8]). Таким образом, имеем

$$\mathbf{f} = \nabla U + \mathbf{V}, \quad U \in W^1(\Omega), \quad \mathbf{V} \in \mathring{J}_{p'}(\Omega),$$

что означает равенство нулю первого из двух интегралов (2.11), тогда как он равен единице. Полученное противоречие завершает доказательство теоремы.

Замечание. В доказательствах теорем 1 и 2 шар B_R можно заменить любой ограниченной областью $\widetilde{\Omega} \subset \mathbb{R}^n$ с кусочно-гладкой связной границей такой, что $\overline{\Omega} \subset \widetilde{\Omega}$, например параллелепипедом. При этом гармонические в $\widetilde{\Omega}$ функции могут дополнительно удовлетворять любым локально корректным краевым условиям (Дирихле, Неймана или третьему краевому) на одной или нескольких частях границы $\partial\widetilde{\Omega}$, но не на всей $\partial\widetilde{\Omega}$. Такой подход принципиально упрощает технику построения в явном виде, возможно, более подходящего для заданной области Ω аппроксимативного базиса из гармонических функций (не обязательно многочленов), который, учитывая форму области Ω, обеспечит более эффективную реализацию аналитико-численных методов решения корректно поставленных краевых задач для уравнения Лапласа в заданной области Ω.

СПИСОК ЛИТЕРАТУРЫ

1. *Browder F.E.* Approximation by solutions of partial differential equations // Am. J. Math. 1962. V. 84. P. 134–160.
2. *Browder F.E.* Functional analysis and partial differential equations. 2 // Math. Ann. 1962. V. 145. P. 81–226.
3. *Singer I.* Bases in Banach Spaces, vol. 2. Berlin–Heidelberg–New York: Springer, 1981.
4. *Рудин У.* Функциональнцй анализ. М.: Мир, 1975.
5. *Axler A., Bourdon P., Ramey W.* Harmonic functions theory (2nd ed.). New York: Springer, 2001.
6. *Efthimou C., Frye Ch.* Spherical Harmonics in p Dimensions. Hackensack, NJ: World Scientific Publ. Co., 2014.
7. *Буренков В.И.* О приближении функций из пространства $W^r_p(\Omega)$ финитными функциями для произвольного открытого множества Ω // Труды МИАН им. В.А. Стеклова. 1974. Т. 131. С. 51–63.
8. *Масленникова В.Н., Боговский М.Е.* Аппроксимация потенциальных и соленоидальных векторных полей // Сиб. матем. ж. 1983. Т. 24. № 5. С. 149–171.
9. *Стейн И.* Сингулярные интегралы и дифференциальные свойства функций М.: Мир, 1973.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 224–238

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 517.95

ЧИСЛЕННЫЙ АНАЛИЗ ТРЕХМЕРНЫХ ЗАДАЧ МАГНИТНОЙ МАСКИРОВКИ НА ОСНОВЕ ОПТИМИЗАЦИОННОГО МЕТОДА[1)]

© 2021 г. Г. В. Алексеев[1,*], Ю. Э. Спивак[1,2]

[1] *690041 Владивосток, ул. Радио, 7, Институт прикладной математики ДВО РАН, Россия*

[2] *690091 Владивосток, ул. Суханова, 8, Дальневосточный федеральный университет, Россия*

**e-mail: alekseev@iam.dvo.ru*

Поступила в редакцию 26.06.2020 г.
Переработанный вариант 26.06.2020 г.
Принята к публикации 16.09.2020 г.

Формулируются обратные задачи для трехмерной модели магнитостатики, возникающие при проектировании осесимметричных многослойных экранирующих и маскировочных устройств. В предположении, что проектируемое устройство состоит из конечного числа сферических слоев, каждый из которых заполнен однородной изотропной средой, предлагается численный алгоритм решения указанных задач, основанный на оптимизационном методе. С его помощью рассматриваемые обратные задачи сводятся к конечномерным экстремальным задачам, роль управлений в которых играют магнитные проницаемости каждого элементарного слоя. Для нахождения искомых управлений применяется метод роя частиц. На основе анализа проведенных вычислительных экспериментов показывается, что полученным оптимальным решениям отвечают маскировочные устройства, обладающие наивысшей эффективностью в рассматриваемом классе устройств и простотой технической реализации. Библ. 30. Фиг. 1. Табл. 7.

Ключевые слова: обратные задачи, метод оптимизации, метод роя частиц, магнитная маскировка.

DOI: 10.31857/S0044466921020034

1. ВВЕДЕНИЕ

В последние годы большое внимание уделяется разработке методов решения задач маскировки материальных тел. Начало указанному направлению было положено в работах [1]–[3], в которых авторы цитированных работ разработали оригинальный метод, получивший название “transfomation optics” (ТО-подход). Указанный метод основан на установленных в [1], [4] и ряде других работ свойстве инвариантности уравнений Максвелла относительно определенных преобразований координат. Позже этот метод был распространен на акустические [5] и статические (тепловые, электрические и магнитные) поля [6]–[9]. Близкие обратные задачи восстановления диэлектрической или магнитной проницаемости тела, помещенного в прямоугольный волновод, изучались в [10], [11].

Нужно отметить, что полученные в цитируемых работах решения задач маскировки обладают рядом недостатков. Главным недостатком является трудность технической реализации полученных решений. По этой причине в последние годы стала развиваться другая стратегия решения задач дизайна устройств невидимости. Она основана на использовании оптимизационного метода решения обратных задач. Начиная с фундаментальных работ А.Н. Тихонова [12], оптимизационный метод широко используется при решении прикладных обратных задач. Преимуществом оптимизационного метода по сравнению с другими является то, что его применение позволяет учесть априори многие из требований, касающиеся технической реализации искомых решений задач дизайна. Использованию оптимизационного метода для решения задач дизайна маскировки и других специальных функциональных устройств посвящен ряд работ. Отметим среди них статьи [13]–[19], в которых оптимизационный метод вместе с методом топологической оптимизации или методом роя частиц (в качестве метода численной оптимизации) приме-

[1)]Работа обоих авторов выполнена при поддержке государственного задания ИПМ ДВО РАН. Работа второго автора выполнена при финансовой поддержке РФФИ (проект № 19-31-90039).

няется для численного решения задач дизайна устройств маскировки и концентрирования статических полей. Упомянем также работы [20]–[24], посвященные качественному анализу задач электромагнитной или тепловой маскировки на основе оптимизационного метода.

Оптимизационный метод применяется и в настоящей работе для решения обратных задач магнитной маскировки в рамках трехмерной осесимметричной модели магнитостатики. Указанные задачи заключаются в нахождении материальных параметров неоднородной среды, заполняющей материальную (маскировочную либо экранирующую) оболочку в виде сферического слоя, исходя из выполнения условий маскировки либо экранирования. В предположении, что исходная оболочка состоит из конечного числа элементарных сферических слоев, заполненных однородными изотропными средами, мы сведем указанные задачи к решению соответствующих конечномерных экстремальных задач, в которых постоянные магнитные проницаемости каждого слоя играют роль управляющих параметров. Для нахождения решений указанных экстремальных задач мы предложим численный алгоритм их решения, основанный на методе роя частиц [25], по схеме, используемой в заметке [16], и обсудим результаты вычислительных экспериментов для случая однородного внешне приложенного магнитного поля. На основе проведенного анализа мы покажем, что применение разработанного алгоритма позволяет спроектировать маскировочные или экранирующие слоистые оболочки, обладающие высокой маскировочной эффективностью даже при малом количестве слоев и простотой технической реализации.

2. ПОСТАНОВКА ПРЯМОЙ И ОБРАТНОЙ ЗАДАЧ

Задача магнитной маскировки по своей структуре содержит три основные компоненты: область, где рассматривается физический процесс, внешне приложенное магнитное поле и материальную оболочку, служащую для маскировки материальных тел [26]. В трехмерных задачах магнитной маскировки роль основной области играет все пространство $\mathbb{R}^3$, исходное внешне приложенное магнитное поле создается компактно распределенными источниками, либо источниками, расположенными в бесконечности, тогда как роль маскировочной оболочки играет область в виде сферического слоя $\Omega = \{a < r = |\mathbf{x}| < b\}$, заполненная неоднородной изотропной либо анизотропной (в общем случае) средой с переменной магнитной проницаемостью μ. Обозначим через Ω_i и Ω_e^∞ внутренность и внешность области Ω соответственно. Будем предполагать, что эти области Ω_i и Ω_e^∞ заполнены однородной изотропной средой с постоянной магнитной проницаемостью μ_0 (см. фиг. 1).

Обозначим через B_R шар $|\mathbf{x}| < R$ радиуса R, содержащий внутри себя области Ω_i и Ω (см. фиг. 1), и предположим, что за пределами шара B_R находятся внешние источники, создающие в частном случае, когда $\mu = \mu_0$ в Ω и, следовательно, все пространство $\mathbb{R}^3$ заполнено однородной средой с постоянной магнитной проницаемостью μ_0, однородное поле $\mathbf{H}_a = -\operatorname{grad}\Phi_a$. Оно описывается магнитным потенциалом Φ_a, сужение которого на шар B_R удовлетворяет уравнению Лапласа $\Delta\Phi_a = 0$ в B_R.

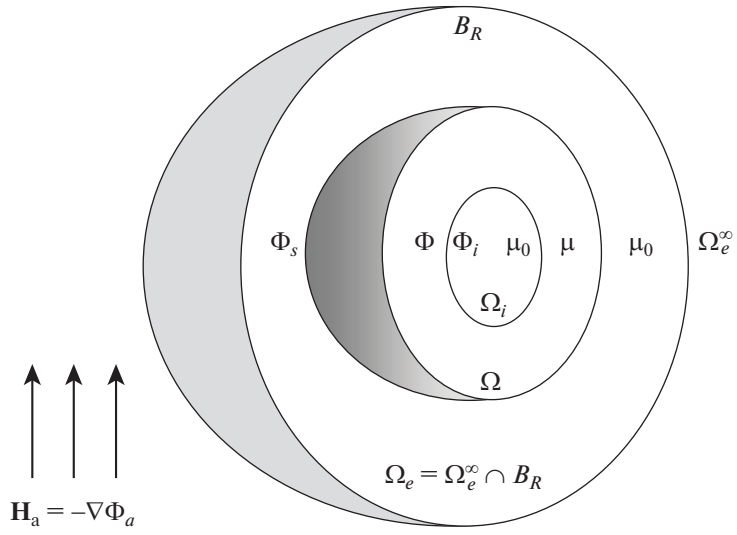
Наличие оболочки (Ω, μ) внутри B_R приводит к изменению поля Φ_a, в результате чего оно принимает вид $\tilde{\Phi} = \Phi_a + \tilde{\Phi}_s$. Здесь $\tilde{\Phi}_s$ – возмущение поля Φ_a, вызываемое внесением объекта (Ω, μ) в область B_R. Обозначим через Φ_i либо Φ сужения поля $\tilde{\Phi}$ на области Ω_i, либо Ω соответственно. Положим $\Phi_s = \tilde{\Phi}_s\big|_{\Omega_e^\infty}$. Введенные поля Φ_i в Ω_i, Φ в Ω и Φ_s в Ω_e^∞ являются решением следующей задачи сопряжения [23]:

$$\mu_0\Delta\Phi_i = 0 \quad \text{в} \quad \Omega_i, \qquad \operatorname{div}(\mu\nabla\Phi) = 0 \quad \text{в} \quad \Omega, \qquad \mu_0\Delta\Phi_s = 0 \quad \text{в} \quad \Omega_e^\infty, \tag{2.1}$$

$$\Phi_i = \Phi, \qquad \mu_0\frac{\partial\Phi_i}{\partial r} = \mu\frac{\partial\Phi}{\partial r} \qquad \text{при} \qquad r = a, \tag{2.2}$$

$$\Phi = \Phi_a + \Phi_s, \qquad \mu\frac{\partial\Phi}{\partial r} = \mu_0\frac{\partial(\Phi_a + \Phi_s)}{\partial r} \qquad \text{при} \qquad r = b, \tag{2.3}$$

$$\Phi_s(\mathbf{x}) = o(1) \qquad \text{при} \qquad r = |\mathbf{x}| \to \infty. \tag{2.4}$$



Фиг. 1. Схематическое изображение внешне приложенного магнитного поля и шара B_R, содержащего сферическую магнитную оболочку Ω.

Ниже будем ссылаться на (2.1)–(2.4) как на задачу магнитного рассеяния, а на Φ_s – как на рассеянное оболочкой (Ω, μ) поле.

В случае, когда μ – диагональный в сферических координатах r, θ, φ тензор, причем $\mu = \mathrm{diag}(\mu_r, \mu_\theta, \mu_\theta)$, где μ_r и μ_θ – заданные ограниченные положительные функции, задача (2.1)–(2.4) была исследована теоретически в работе [23]. В ней установлены достаточные условия на исходные данные, обеспечивающие существование единственного слабого решения. В рассматриваемом нами в данной статье случае, когда μ – скалярная функция, описывающая переменную изотропную среду, заполняющую область Ω, указанные условия, обеспечивающие корректность задачи (2.1)–(2.4), имеют вид

$$\mu \in L^\infty(\Omega), \quad \mu(\mathbf{x}) \geq \mu^0 = \mathrm{const} > 0 \ \text{ в } \ \Omega, \quad \Delta\Phi_a = 0 \ \text{ в } \ B_R. \tag{2.5}$$

Как уже указывалось, основное внимание в этой статье будет уделено численному анализу обратных задач для модели (2.1)–(2.4), связанных с проектированием устройств маскировки материальных тел. Указанные задачи состоят в нахождении неизвестной магнитной проницаемости μ, исходя из условий маскировки. Чтобы сформулировать их, обозначим через $\tilde{\Phi}^\mu \equiv (\Phi_i^\mu, \Phi^\mu, \Phi_s^\mu)$ решение задачи (2.1)–(2.4), отвечающее проницаемости μ в Ω и проницаемости μ_0 в Ω_i и Ω_e^∞. Ниже будем рассматривать общую и две частные обратные задачи маскировки. Общая обратная задача, называемая задачей полной магнитной маскировки, состоит в нахождении магнитной проницаемости μ среды, заполняющей область Ω, исходя из следующих двух независимых условий:

$$\nabla\Phi_i^\mu = 0 \quad (\text{т.е. } \Phi_i^\mu = \mathrm{const}) \ \text{ в } \ \Omega_i, \quad \Phi_s^\mu = 0 \ \text{ в } \ \Omega_e^\infty. \tag{2.6}$$

Первое условие в (2.6) относится к поведению внутреннего поля Φ_i^μ (т.е. поля $\tilde{\Phi}^\mu$ внутри области Ω_i), тогда как второе условие в (2.6) описывает поведение рассеянного поля Φ_s^μ во внешности Ω_e^∞. С учетом этого задачу нахождения проницаемости μ, исходя из выполнения первого условия в (2.6), принято называть задачей внутренней маскировки или задачей экранирования, тогда как задачу нахождения μ, исходя из выполнения второго условия в (2.6), называют задачей внешней маскировки [26]. На саму пару (Ω, μ), обеспечивающую выполнение первого условия (либо обоих условий) в (2.6), будем ссылаться как на экранирующую (либо маскировочную) оболочку.

Ниже мы будем предполагать, исходя из условий простоты технической реализации проектируемых с помощью оптимизационного метода оболочек, что искомая оболочка Ω является слоистой и состоит из M сферических слоев $\Omega_m = \{R_{m-1} < r = |\mathbf{x}| < R_m\}$, $m = 1, 2, \ldots, M$, $R_0 = a$, $R_M = b$

одной и той же толщины $d = (b-a)/M$. Каждый из них заполнен однородной изотропной средой, магнитные свойства которой характеризуются постоянной магнитной проницаемостью $\mu_m > 0$, $m = 1,2,\ldots,M$. Последнее означает, что искомую магнитную проницаемость μ искомой оболочки (Ω,μ) следует искать в виде

$$\mu(\mathbf{x}) = \sum_{m=1}^{M} \mu_m \chi_m(\mathbf{x}). \tag{2.7}$$

Здесь $\chi_m(\mathbf{x})$ – характеристическая функция слоя Ω_m, равная 1 в Ω_m и 0 вне Ω_m, а μ_m, $m = 1,2,\ldots,M$ – неизвестные положительные константы, которые находятся, исходя из точного или приближенного выполнения одного или двух условий в (2.6). Отметим, что функция (2.7) удовлетворяет при любых $\mu_m > 0$ первым двум условиям в (2.5).

3. СЛУЧАЙ ПОСТОЯННОГО ВНЕШНЕ ПРИЛОЖЕННОГО ПОЛЯ

Аналогично [8] будем рассматривать ниже важный частный случай, когда внешне приложенное магнитное поле $\mathbf{H}_a$ постоянно во всем пространстве и направлено вдоль оси z. В этом случае, означающем, что источники находятся в бесконечности, поле $\mathbf{H}_a$ представимо в виде

$$\mathbf{H}_a = -\nabla\Phi_a \quad \text{в} \quad \mathbb{R}^3, \qquad \Phi_a = -H_a r\cos\theta, \qquad H_a = |\mathbf{H}_a|. \tag{3.1}$$

Поскольку потенциал Φ_a, введенный в (3.1), и магнитная проницаемость μ, определенная в (2.7), удовлетворяют условиям (2.5) при любом $R > 0$, то точное решение $\tilde{\Phi} \equiv (\Phi_i, \Phi, \Phi_s)$ задачи (2.1)–(2.4), отвечающее упомянутой паре μ и Φ_a в (2.7) и (3.1) (мы опускаем для простоты верхний индекс μ в обозначении решения), существует и единственно. Более того, используя метод Фурье, указанное решение можно записать в явном виде. С этой целью обозначим через $\Phi_m \equiv \Phi|_{\Omega_m}$ сужение компоненты Φ решения $\tilde{\Phi}$ на подобласть $\Omega_m \subset \Omega$, $m = 1,2,\ldots,M$, и положим $\Phi_0 = \Phi_i$ в Ω_i, $\Phi_{M+1} = \Phi_a + \Phi_s$ в Ω_e^∞. Простой анализ показывает, что отдельные компоненты Φ_m (мы будем ссылаться на них как на поля), позволяющие определить решение $\tilde{\Phi} = (\Phi_i, \Phi_1, \ldots, \Phi_M, \Phi_s)$ во всем пространстве $\mathbb{R}^3$, представимы в виде

$$\Phi_0(r,\theta) = \alpha_0 r\cos\theta \quad \text{в} \quad \Omega_i, \qquad \Phi_{M+1}(r,\theta) = \Phi_a(r,\theta) + (\beta_{M+1}/r^2)\cos\theta \quad \text{в} \quad \Omega_e^\infty, \tag{3.2}$$

$$\Phi_1(r,\theta) = (\alpha_1 r + \beta_1/r^2)\cos\theta \quad \text{в} \quad \Omega_1, \qquad \Phi_2(r,\theta) = (\alpha_2 r + \beta_2/r^2)\cos\theta \quad \text{в} \quad \Omega_2, \tag{3.3}$$

$$\ldots$$

$$\Phi_M(r,\theta) = (\alpha_M r + \beta_M/r^2)\cos\theta \quad \text{в} \quad \Omega_M. \tag{3.4}$$

Здесь константы $\alpha_0, \alpha_m, \beta_m$, $m = 1,\ldots,M$, и β_{M+1} определяются из условий непрерывности смежных полей Φ_m и Φ_{m+1} на общих границах $r = R_m$, $m = 0,1,\ldots,M$, имеющих вид

$$\Phi_m = \Phi_{m+1}, \qquad \mu_m \frac{\partial\Phi_m}{\partial r} = \mu_{m+1}\frac{\partial\Phi_{m+1}}{\partial r} \qquad \text{при} \qquad r = R_m, \qquad m = 0,1,\ldots,M, \qquad (\mu_{M+1} = \mu_0). \tag{3.5}$$

Под Φ_m при $m = 0$ и $M+1$ в (3.5) мы понимаем, соответственно, Φ_i и $\Phi_a + \Phi_s$.

Подставляя (3.2)–(3.4) в (3.5), приходим к следующей системе $2M+2$ линейных алгебраических уравнений относительно коэффициентов α_0, α_m, β_m, $m = 1,\ldots,M$, и β_{M+1}:

$$-R_0^3\alpha_0 + R_0^3\alpha_1 + \beta_1 = 0, \qquad -\mu_0 R_0^3\alpha_0 + \mu_1 R_0^3\alpha_1 - 2\mu_1\beta_1 = 0, \tag{3.6}$$

$$R_1^3\alpha_1 - R_1^3\alpha_2 + \beta_1 - \beta_2 = 0, \qquad \mu_1 R_1^3\alpha_1 - \mu_2 R_1^3\alpha_2 - 2\mu_1\beta_1 + 2\mu_2\beta_2 = 0, \tag{3.7}$$

$$\ldots$$

$$R_m^3\alpha_m - R_m^3\alpha_{m+1} + \beta_m - \beta_{m+1} = 0, \tag{3.8}$$

$$\mu_m R_m^3\alpha_m - \mu_{m+1} R_m^3\alpha_{m+1} - 2\mu_m\beta_m + 2\mu_{m+1}\beta_{m+1} = 0, \qquad m = 2,\ldots,M-1, \tag{3.9}$$

$$\ldots$$

$$\begin{gathered} -R_M^3\alpha_M - \beta_M + \beta_{M+1} = H_a R_M^3, \\ -\mu_M R_M^3\alpha_M + 2\mu_M\beta_M - 2\mu_0\beta_{M+1} = \mu_0 H_a R_M^3. \end{gathered} \tag{3.10}$$

Решив систему (3.6)–(3.10) относительно неизвестных коэффициентов α_0, α_m, β_m, $m = 1,\ldots,M$, и β_{M+1} и подставив найденные значения в (3.2)–(3.4), мы получим соответствующие поля $\Phi_i = \Phi_0$ в Ω_i, Φ_m в Ω_m, $m = 1,2,\ldots,M$, и $\Phi_s = \Phi_{M+1} - \Phi_a = (\beta_{M+1}/r^2)\cos\theta$ в Ω_e^∞, образующие искомое решение задачи (2.1)–(2.4). Ясно, что так полученные поля Φ_i, Φ_m и Φ_s зависят от постоянных проницаемостей $\mu_1,\mu_2,\ldots,\mu_M$ слоев $\Omega_1,\Omega_2,\ldots,\Omega_M$. Поэтому рассматриваемые нами обратные задачи магнитной маскировки сводятся при выполнении условия (3.1) к поиску таких проницаемостей $\mu_1,\ldots,\mu_M$, что соответствующее им решение $(\Phi_i,\Phi_1,\ldots,\Phi_M,\Phi_s)$, определяемое формулами (3.2)–(3.4), удовлетворяет точно или приближенно одному или обоим условиям в (2.6).

В простейшем случае, когда число слоев M равно 2, соответствующее точное решение $(\Phi_i,\Phi_1,\Phi_2,\Phi_s)$ определяется формулами (3.2), (3.3), в которых следует положить $M = 2$. Здесь неизвестные коэффициенты α_0, α_1, α_2, β_1, β_2 и β_3 являются решением системы шести линейных алгебраических уравнений, состоящей из (3.6), (3.7) и (3.10), где следует положить $M = 2$. Из (3.2) вытекает, что первое (либо второе) условие в (2.6) выполняется тогда и только тогда, когда $\alpha_0 = 0$ (либо $\beta_3 = 0$). Используя аналитический модуль пакета Wolfram Mathematica, точные выражения для коэффициентов α_0 и β_3 можно записать в виде

$$\alpha_0(\mu_1,\mu_2) = \frac{\Delta_{\alpha_0}}{\Delta} = -\frac{27H_a\mu_0\mu_1\mu_2R_0^3R_1^3R_2^3}{\Delta}, \tag{3.11}$$

$$\begin{aligned}\beta_3(\mu_1,\mu_2) = \frac{\Delta_{\beta_3}}{\Delta} = \frac{H_aR_0^3R_2^3(\mu_0(3\mu_1\mu_2R_1^3(R_0^3-R_2^3)-2\mu_1^2(R_0^3-R_1^3)(R_1^3-R_2^3)}{\Delta} + \\ + \frac{2\mu_2^2(R_0^3-R_1^3)(R_1^3-R_2^3))+2\mu_1\mu_2(-\mu_2(R_0^3+2R_1^3)(R_1^3-R_2^3)-\mu_1(R_0^3-R_1^3)(2R_1^3+R_2^3))}{\Delta} + \\ + \frac{\mu_0^2(\mu_1(2R_0^3+R_1^3)(R_1^3-R_2^3)+\mu_2(R_0^3-R_1^3)(R_1^3+2R_2^3)))}{\Delta},\end{aligned} \tag{3.12}$$

где определитель Δ системы (3.6), (3.7), (3.10) при $M = 2$ определяется формулой

$$\begin{aligned}\Delta = R_0^3(2\mu_1\mu_2(-\mu_2(R_0^3+2R_1^3)(R_1^3-R_2^3)-\mu_1(R_0^3-R_1^3)(2R_1^3+R_2^3))- \\ - 2\mu_0^2(\mu_1(2R_0^3+R_1^3)(R_1^3-R_2^3)+\mu_2(R_0^3-R_1^3)(R_1^3+2R_2^3))+\mu_0(4\mu_1^2(R_0^3-R_1^3)(R_1^3-R_2^3)+ \\ + 2\mu_2^2(R_0^3-R_1^3)(R_1^3-R_2^3)+3\mu_1\mu_2(2R_1^6+3R_1^3R_2^3+2R_0^3(R_1^3+R_2^3)))).\end{aligned} \tag{3.13}$$

Из (3.11) следует в предположении $\mu_0 > 0$, что $\alpha_0 = 0$ тогда и только тогда, когда $\mu_1 = 0$ (либо $\mu_2 = 0$). Подставляя $\mu_1 = 0$ в (3.12), выводим, что

$$\beta_3(0,\mu_2) = \beta_3 = \frac{H_aR_2^3(2\mu_2(R_1^3-R_2^3)+\mu_0(R_1^3+2R_2^3))}{2\mu_2(R_1^3-R_2^3)-2\mu_0(R_1^3+2R_2^3)}. \tag{3.14}$$

Полагая $\beta_3 = 0$ в (3.14), получаем линейное уравнение относительно неизвестной проницаемости μ_2. Его решение, обеспечивающее выполнение условия $\beta_3 = 0$, имеет вид

$$\mu_2 = \mu_2^0 \equiv \mu_0\frac{2R_2^3+R_1^3}{2(R_2^3-R_1^3)}. \tag{3.15}$$

Из построения следует, что пара $(\mu_1,\mu_2) = (0,\mu_2^0)$ является искомым решением обратной задачи полной маскировки. Впервые указанное решение было приведено в [8]. Хотя построенное точное решение описывается простой формулой (3.15), его техническая реализация не представляется возможной (вследствие условия $\mu_1 = 0$) из-за отсутствия природных либо инженерных материалов с нулевой магнитной проницаемостью. Ввиду этого указанное решение $(0,\mu_2^0)$ называют сингулярным. Один из способов преодоления трудностей с технической реализацией решений задач магнитной маскировки состоит в замене исходных обратных задач соответствующими приближенными задачами и в применении для решения последних задач оптимизационных методов, позволяющих учесть требования, связанные с технической реализацией отыскиваемых решений. Этим мы займемся в следующем разделе.

4. ПРИМЕНЕНИЕ ОПТИМИЗАЦИОННОГО МЕТОДА. ФОРМУЛИРОВКА ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

Для решения сформулированных выше обратных задач мы применим оптимизационный метод [12]. В соответствии с этим методом обратные задачи магнитной маскировки заменяются соответствующими экстремальными задачами, которые адекватно отвечают рассматриваемым обратным задачам. Указанные задачи заключаются в минимизации определенных функционалов качества, зависящих от магнитных проницаемостей $\mu_1, \ldots, \mu_M$ отдельных слоев $\Omega_1, \ldots, \Omega_M$. Чтобы сформулировать их, введем в рассмотрение M-мерный вектор $\mathbf{m} = (\mu_1, \mu_2, \ldots, \mu_M)$, на который мы будем ссылаться как на вектор магнитных проницаемостей, и определим ограниченное множество K в пространстве $\mathbb{R}^M$ формулой

$$K = \{\mathbf{m} \equiv (\mu_1, \mu_2, \ldots, \mu_M) : 0 < \mu_{\min} \le \mu_j \le \mu_{\max},\ j = 1, 2, \ldots, M\}. \tag{4.1}$$

На введенное множество K, где заданные положительные константы $\mu_{\min}$ и $\mu_{\max}$ определяют его нижнюю и верхнюю границы, будем ссылаться как на множество управлений. Напомним, что введение множества K по формуле (4.1) отвечает так называемой схеме коробки или схеме простых ограничений.

Введем переобозначение $\Phi[\mathbf{m}] = (\Phi_i[\mathbf{m}], \Phi[\mathbf{m}], \Phi_s[\mathbf{m}])$ для решения $\tilde{\Phi}^{\mu} \equiv (\Phi_i^{\mu}, \Phi^{\mu}, \Phi_s^{\mu})$ задачи (2.1)–(2.4), отвечающего магнитной проницаемости μ в Ω, связанной с вектором $\mathbf{m} = (\mu_1, \mu_2, \ldots, \mu_M) \in K$ формулой (2.7). Кроме того, в аналогичной ситуации будем использовать обозначение $(\Omega, \mathbf{m})$ для маскировочной оболочки вместо обозначения (Ω, μ). Положим $\Omega_e = \Omega_e^{\infty} \cap B_R$.

Предполагая ниже, что вектор $\mathbf{m} = (\mu_1, \mu_2, \ldots, \mu_M)$ принадлежит множеству управлений K, сформулируем следующие три экстремальные задачи:

$$J_i(\mathbf{m}) \to \inf, \quad \mathbf{m} \in K, \tag{4.2}$$

$$J_e(\mathbf{m}) \to \inf, \quad \mathbf{m} \in K, \tag{4.3}$$

$$J(\mathbf{m}) = 0.5[J_i(\mathbf{m}) + J_e(\mathbf{m})] \to \inf, \quad \mathbf{m} \in K. \tag{4.4}$$

Здесь функционалы качества $J_i(\mathbf{m})$, $J_e(\mathbf{m})$ и $J(\mathbf{m})$ определяются формулами

$$J_i(\mathbf{m}) = \frac{\|\nabla \Phi_i[\mathbf{m}]\|_{L^2(\Omega_i)}}{\|\nabla \Phi_a\|_{L^2(\Omega_i)}}, \quad J_e(\mathbf{m}) = \frac{\|\Phi_s[\mathbf{m}]\|_{L^2(\Omega_e)}}{\|\Phi_a\|_{L^2(\Omega_e)}}, \quad J(\mathbf{m}) = 0.5[J_i(\mathbf{m}) + J_e(\mathbf{m})]. \tag{4.5}$$

Напомним, в частности, что $\Phi_a = -H_a r \cos\theta$ – потенциал заданного внешне приложенного поля, а L^2 – нормы, входящие в (4.5), определяются формулами

$$\begin{gathered} \|\Phi_a\|^2_{L^2(\Omega_e)} = \int\limits_{\Omega_e} |\Phi_a|^2 \, d\mathbf{x}, \quad \|\nabla \Phi_a\|^2_{L^2(\Omega_i)} = \int\limits_{\Omega_i} |\nabla \Phi_a|^2 \, d\mathbf{x}, \\ \|\Phi_s[\mathbf{m}]\|^2_{L^2(\Omega_e)} = \int\limits_{\Omega_e} |\Phi_s[\mathbf{m}]|^2 \, d\mathbf{x}, \quad \|\nabla \Phi_i[\mathbf{m}]\|^2_{L^2(\Omega_i)} = \int\limits_{\Omega_i} |\nabla \Phi_i[\mathbf{m}]|^2 \, d\mathbf{x}. \end{gathered} \tag{4.6}$$

Из вида функционала $J_i(\mathbf{m})$ следует, что условие $J_i(\mathbf{m}^*) = 0$ на некотором векторе $\mathbf{m}^* = (\mu_1^*, \mu_2^*, \ldots, \mu_M^*) \in K$ эквивалентно тому, что $\nabla \Phi_i[\mathbf{m}^*] = 0$ в Ω_i. Это означает, что отвечающая вектору $\mathbf{m}^*$ в силу формулы (2.7) проницаемость

$$\mu^*(\mathbf{x}) = \sum_{m=1}^{M} \mu_m^* \chi_m(\mathbf{x})$$

является решением задачи экранирования. Аналогично, условие $J_e(\mathbf{m}^*) = 0$ эквивалентно тому, что рассеянное поле $\Phi_s^{\mu^*} \equiv \Phi_s[\mathbf{m}^*]$ обращается в нуль всюду в подобласти $\Omega_e \equiv \Omega_e^{\infty} \cap B_R$ области Ω_e^{∞}. Отсюда вытекает в силу принципа единственного продолжения гармонической в Ω_e^{∞} функции $\Phi_s^{\mu^*}$, что $\Phi_s^{\mu^*} = 0$ всюду в Ω_e^{∞}. Последнее эквивалентно тому, что проницаемость μ^* является

решением задачи внешней маскировки. Наконец, условие $J(\mathbf{m}^*) \equiv 0.5[J_i(\mathbf{m}^*) + J_e(\mathbf{m}^*)] = 0$ означает, что проницаемость μ^* является решением задачи полной маскировки.

Способность проектируемой оболочки $(\Omega, \mathbf{m})$ экранировать либо маскировать материальные объекты характеризуется так называемой экранирующей либо маскировочной эффективностью. Количественно указанные характеристики описывают точность, с которой выполняются оба условия в (2.6), либо одно из этих условий. Простой анализ формул (4.5), (4.6) показывает, что введенные выше функционалы $J_i(\mathbf{m})$, $J_e(\mathbf{m})$ и $J(\mathbf{m})$ имеют наглядный смысл среднеквадратичных интегральных ошибок выполнения первого, второго или обоих условий маскировки в (2.6) на векторе $\mathbf{m} = (\mu_1, \mu_2, ..., \mu_M)$.

Из сказанного вытекает, что для оценки экранирующей либо маскировочной эффективности проектируемой оболочки $(\Omega, \mathbf{m})$ следует использовать именно значения $J_i(\mathbf{m})$, $J_e(\mathbf{m})$ и $J(\mathbf{m})$ функционалов J_i, J_e и J. Так, маскировочную эффективность оболочки $(\Omega, \mathbf{m})$ следует оценивать при помощи значения $J(\mathbf{m})$, которое связано с ней обратной зависимостью: чем меньше значение $J(\mathbf{m})$, тем выше маскировочная эффективность оболочки $(\Omega, \mathbf{m})$, и наоборот. В частности, условие $J(\mathbf{m}^*) = 0$ для некоторого $\mathbf{m}^* \in K$, математически эквивалентное тому, что $\mathbf{m}^*$ является точным решением задачи полной маскировки, физически означает, что соответствующая маскировочная оболочка обладает наивысшей маскировочной эффективностью.

Но нужно отметить, что введенный в (4.5) функционал $J \equiv 0.5(J_i + J_e)$ необходимо удовлетворяет условию $J(\mathbf{m}) > 0$ для всех $\mathbf{m} \in K$, где K – любое ограниченное множество, введенное в (4.1). Фактически это вытекает из того известного в теории маскировки факта (см., например, [9], [26]), что проницаемость μ, обеспечивающая точный маскировочный эффект, и, следовательно, отвечающая в силу (2.7) вектору $\mathbf{m} \in K$, для которого $J(\mathbf{m}) = 0$, необходимо должна принимать сингулярные (например, нулевые) значения в некоторых точках множества $\bar{\Omega}$. В то же время проницаемость $\mu(\mathbf{x})$, определенная формулой (2.7), является регулярной положительной в $\bar{\Omega}$ функцией для любого вектора $\mathbf{m} = (\mu_1, \mu_2, ..., \mu_M)$ с положительными μ_m. Поэтому наша цель при решении, например, задачи (4.4) при заданном множестве K будет заключаться в том, чтобы найти вектор проницаемостей (оптимальное решение задачи (4.4)) $\mathbf{m}^{\text{opt}} \in K$, на котором функционал J принимает минимальное на множестве K значение $J^{\text{opt}} = J(\mathbf{m}^{\text{opt}})$, а следовательно, спроектированная оболочка $(\Omega, \mathbf{m}^{\text{opt}})$ обладает максимальной (на множестве K) маскировочной эффективностью. На аналогичные цели направлены задачи (4.2) и (4.3).

Для реализации указанных целей мы применим численный алгоритм, основанный на методе роя частиц (МРЧ) [25]. Напомним, что МРЧ был предложен в 1995 г. в работе [27]. Он не использует значений производных от минимизируемого функционала, является достаточно универсальным и простым при численной реализации. Поэтому в последнее время этот метод широко применяется при решении большого класса обратных и экстремальных задач в различных областях науки и техники. Указанный метод применялся, в частности, в статьях [15]–[18] при численном решении двумерной и трехмерной задач статической маскировки при помощи оболочек, состоящих из однородных изотропных (либо анизотропных, в общем случае) материалов. Мы будем использовать МРЧ по схеме, подробно описанной в заметке [16]. Из нее следует, что основную роль в описанном выше алгоритме играет вычисление значений $J(\mathbf{m})$ минимизируемого функционала J для конкретного вектора $\mathbf{m} \in K$, моделирующего положения конкретных частиц, составляющих используемый рой. С учетом особенностей предложенного выше метода решения прямой задачи магнитной маскировки указанная процедура состоит из двух этапов.

На первом этапе мы находим компоненты α_0 и β_{M+1} решения системы (3.6)–(3.10) для конкретного вектора $\mathbf{m} = (\mu_1, \mu_2, ...\mu_M)$ с использованием пакета Matlab R2019a и, подставив найденные значения α_0 и β_{M+1} в формулы (3.2), содержащие поля $\Phi_i \equiv \alpha_0 r\cos\theta$ и $\Phi_s \equiv (\beta_{M+1}/r^2)\cos\theta$, находим внутреннее и рассеянное поля $\Phi_i[\mathbf{m}] = \alpha_0 r\cos\theta$ и $\Phi_s[\mathbf{m}] = (\beta_{M+1}/r^2)\cos\theta$, отвечающие конкретному вектору $\mathbf{m}$. Далее мы подставляем найденные выражения $\Phi_i[\mathbf{m}]$ и $\Phi_s[\mathbf{m}]$ в (4.5) и вы-

числяем соответствующие интегралы, определяющие нормы, входящие в (4.6), с помощью следующих аналитических формул:

$$\|\nabla\Phi_a\|^2_{L^2(\Omega_i)} \int\limits_{\Omega_i} |\nabla\Phi_a|^2 d\mathbf{x} = \int\limits_0^{R_0}\int\limits_0^{2\pi}\int\limits_0^{\pi} |\nabla\Phi_a|^2 r^2 \sin\theta\, dr d\theta d\varphi = \frac{4}{3}\pi R_0^3 H_a^2,$$

$$\|\nabla\Phi_i[\mathbf{m}]\|^2_{L^2(\Omega_i)} = \int\limits_{\Omega_i} |\nabla\Phi_i[\mathbf{m}]|^2 d\mathbf{x} = \int\limits_0^{R_0}\int\limits_0^{2\pi}\int\limits_0^{\pi} |\nabla\Phi_i[\mathbf{m}]|^2 r^2 \sin\theta\, dr d\theta d\varphi = \frac{4}{3}\pi\alpha_0^2 R_0^3,$$

$$\|\Phi_a\|^2_{L^2(\Omega_e)} = \int\limits_{\Omega_e} \Phi_a^2 d\mathbf{x} = H_a^2 \int\limits_{R_M}^{R_{M+1}}\int\limits_0^{2\pi}\int\limits_0^{\pi} r^4 \cos^2\theta \sin\theta\, dr d\theta d\varphi = H_a^2 \pi \frac{4(R_{M+1}^5 - R_M^5)}{15}, \tag{4.7}$$

$$\|\Phi_s[\mathbf{m}]\|^2_{L^2(\Omega_e)} = \int\limits_{R_M}^{R_{M+1}}\int\limits_0^{2\pi}\int\limits_0^{\pi} \left[(\beta_{M+1})^2/r^2\right] \cos^2\sin\theta\, dr d\theta d\varphi = \frac{4\pi\beta_{M+1}^2 (R_{M+1} - R_M)}{3R_M R_{M+1}},$$

$$J_i(\mathbf{m}) = \sqrt{\frac{\|\nabla\Phi_i[\mathbf{m}]\|^2_{L^2(\Omega_i)}}{\|\nabla\Phi_a\|^2_{L^2(\Omega_i)}}} = \frac{\alpha_0}{H_a},$$

$$J_e(\mathbf{m}) = \sqrt{\frac{\|\Phi_s[\mathbf{m}]\|^2_{L^2(\Omega_s)}}{\|\Phi_a\|^2_{L^2(\Omega_e)}}} = \frac{\beta_{M+1}}{H_a}\sqrt{\frac{5(R_{M+1} - R_M)}{R_M R_{M+1}(R_{M+1}^5 - R_M^5)}}, \tag{4.8}$$

$$J(\mathbf{m}) = \frac{1}{2}\left[J_i(\mathbf{m}) + J_e(\mathbf{m})\right] = \frac{\alpha_0}{2H_a} + \frac{\beta_{M+1}}{2H_a}\sqrt{\frac{5(R_{M+1} - R_M)}{R_M R_{M+1}(R_{M+1}^5 - R_M^5)}}. \tag{4.9}$$

Здесь $R_0, R_1, \ldots, R_M$ – величины (радиусы сфер), введенные в (3.5), $R_{M+1} = R$.

Важно отметить, что все приведенные выше интегралы вычисляются точно, поэтому этот этап не вносит дополнительной ошибки в процедуру нахождения решения. Это важно в вычислительном плане, поскольку рассматриваемые обратные задачи относятся к классу некорректных задач. Тем не менее с учетом плохой обусловленности системы (3.6)–(3.10) в общем случае задание ее коэффициентов, нахождение решения и все другие расчеты производились с достаточно высокой точностью, обеспечиваемой правилами пакета Matlab R2019a.

5. АНАЛИЗ РЕЗУЛЬТАТОВ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Обсудим здесь некоторые результаты по численному решению рассматриваемых задач магнитной маскировки с использованием метода роя частиц (МРЧ). Численное моделирование проводилось для следующих исходных данных:

$$a = 0.035 \text{ м}, \quad b = 0.05 \text{ м}, \quad \mu_0 = 1, \quad R = 0.7 \text{ м}. \tag{5.1}$$

Внешне приложенное магнитное поле имеет вид (3.1). Основное внимание мы уделим анализу вычислительных экспериментов по решению задачи экранирования (4.2) либо задачи полной маскировки (4.4).

Напомним, что параметр R входит в виде R_{M+1} в формулы (4.8), (4.9) для вычисления $J_e(\mathbf{m})$ и $J(\mathbf{m})$. Анализ этих формул в сравнении с (2.6) показывает, что чем больше R, тем больше информации об условии $\Phi_s = 0$ в Ω_e^∞, входящем в (2.6), учитывается в выражениях (4.8), (4.9). Отсюда следует, что точность решения задач внешней и полной маскировки должна увеличиваться с увеличением R. Этот же факт подтвердили численные эксперименты, проводимые (в случае, когда $a = 0.035$ м, $b = 0.05$ м) при разных значениях R, равных 0.07, 0.1, 0.15 и 0.7 м, причем дальнейшее увеличение R не приводило к сколь-нибудь заметным различиям в результатах. По этой причине приводимые ниже результаты расчетов относятся именно к значению $R = 0.7$ м в (5.1).

Анализ большого количества проведенных вычислительных экспериментов по решению задачи (4.2) позволил выявить весьма интересную тенденцию в поведении компонент μ_m^{opt} ее опти-

мального решения $\mathbf{m}^{\text{opt}} = (\mu_1^{\text{opt}}, \mu_2^{\text{opt}}, ..., \mu_M^{\text{opt}})$. Оказалось (при определенных условиях на исходные данные задачи (4.2)), что оптимальные значения μ_m^{opt} всех параметров μ_m с нечетными индексами $m = 1, 3, 5, ..., M-1$ совпадают с одной из границ $\mu_{\min}, \mu_{\max}$ множества управлений K, тогда как оптимальные значения $\mu_2^{\text{opt}}, \mu_4^{\text{opt}}, ..., \mu_M^{\text{opt}}$ остальных параметров (с четными индексами) совпадают с другой границей. Таким образом, выполняются соотношения:

$$\mu_1^{\text{opt}} = \mu_3^{\text{opt}} = ... = \mu_{M-1}^{\text{opt}} = \mu_{\min}, \quad \mu_2^{\text{opt}} = \mu_4^{\text{opt}} = ... = \mu_M^{\text{opt}} = \mu_{\max}, \tag{5.2}$$

либо

$$\mu_1^{\text{opt}} = \mu_3^{\text{opt}} = ... = \mu_{M-1}^{\text{opt}} = \mu_{\max}, \quad \mu_2^{\text{opt}} = \mu_4^{\text{opt}} = ... = \mu_M^{\text{opt}} = \mu_{\min}. \tag{5.3}$$

Следуя [17], [18], мы будем ссылаться на (5.2) (либо на (5.3)) как на соотношения чередующегося дизайна 1-го (либо 2-го) типа. Напомним, что под термином “чередующийся дизайн” в теории маскировки понимают решение в виде слоистой оболочки, состоящей из конечного числа слоев, заполненных чередующимися (по слоям) материалами с большим (определенным) отношением параметров указанных материалов [6]. Указанное отношение принято называть их контрастом. Одним из условий, обеспечивающих выполнение условия (5.2) (либо (5.3)), является условие

$$\mu_0^2 - \mu_{\min}\mu_{\max} \geq 0. \tag{5.4}$$

Из (5.2) (либо (5.3)) следует, что для соответствующего оптимального решения $\mathbf{m}^{\text{opt}} \equiv (\mu_1^{\text{opt}}, \mu_2^{\text{opt}}, ..., \mu_M^{\text{opt}})$ задачи (4.2) справедлив аналог так называемого свойства bang-bang. Согласно этому свойству каждая компонента μ_m^{opt} оптимального решения принимает одно из двух значений $\mu_{\min}, \mu_{\max}$, являющихся границами множества K [28]. Как мы увидим ниже, именно это свойство играет основополагающую роль для получения легко реализуемых решений рассматриваемых задач.

Обсудим теперь результаты решения задачи экранирования (4.2) для трех конкретных тестов, отвечающих следующим трем парам $(\mu_{\min}, \mu_{\max})$:

$$1)\ (0.025, 40), \quad 2)\ (0.0045, 40) \quad \text{и} \quad 3)\ (0.0045, 70). \tag{5.5}$$

Отметим, что все введенные в (5.5) значения, кроме $\mu_{\min} = 0.025$, отвечают магнитным проницаемостям известных материалов. Например, значение $\mu = 40$ описывает относительную магнитную проницаемость закаленной нержавеющей стали, значение $\mu = 70$ отвечает магнитной проницаемости кобальта, тогда как значение $\mu = 0.0045$ отвечает известному метаматериалу под названием “сверхпроводник SuperPower SCS12050” [29], [30], который широко используется в приложениях. Подчеркнем, что все пары в (5.5) удовлетворяют условию (5.4).

Результаты решения с помощью МРЧ задачи (4.2) для первой пары $(\mu_{\min}, \mu_{\max})$ в (5.5) с контрастом $\mu_{\max}/\mu_{\min} = 40/0.025 = 1600$ для четных значений M, изменяющихся от 2 до 16, приведены в табл. 1. Она содержит оптимальные значения μ_1^{opt}, μ_M^{opt} магнитных проницаемостей первого и последнего слоев, совпадающие в силу (5.3) с $\mu_{\max}$ и $\mu_{\min}$ соответственно, и оптимальные значения $J_i(\mathbf{m}^{\text{opt}})$ функционала $J_i(\mathbf{m})$, где $\mathbf{m}^{\text{opt}} = \mathbf{m}^{\text{alt}} \equiv (\mu_{\max}, \mu_{\min}, \mu_{\max}, ..., \mu_{\min})$, вместе со значениями $J_e(\mathbf{m}^{\text{opt}})$ и $J(\mathbf{m}^{\text{opt}})$ (для сравнения) для четных значений M, изменяющихся от 2 до 16. Остальные значения управлений μ_m^{opt}, $m = 2, 3, ..., M-1$, определяются из соотношений (5.3), где следует положить $\mu_{\min} = 0.025$, $\mu_{\max} = 40$. Анализ табл. 1 показывает, что при изменении M от 2 до 16 значения $J_i(\mathbf{m}^{\text{opt}})$ изменяются в пределах от 2.16×10^{-2} до значения 1.28×10^{-4}, которое соответствует невысокой экранирующей эффективности.

Из общих соображений (см. детали в [18]) следует, что для повышения экранирующей эффективности следует увеличить контраст $\mu_{\max}/\mu_{\min}$. Этого можно добиться как за счет уменьшения $\mu_{\min}$, так и за счет увеличения $\mu_{\max}$. Полагая в соответствии со вторым сценарием в (5.5) $\mu_{\min} = 0.0045$, $\mu_{\max} = 40$, чему соответствует контраст $\mu_{\max}/\mu_{\min} = 8888.9$, и применяя МРЧ к задаче (4.2), мы получаем результаты, приведенные в табл. 2, которая является аналогом табл. 1 для

Таблица 1. Задача экранирования: $\mu_{\min} = 0.025$, $\mu_{\max} = 40$, $\mathbf{m}^{\text{opt}} = \mathbf{m}^{\text{alt}}$

M	μ_1^{opt}	μ_M^{opt}	$J_i(\mathbf{m}^{\text{opt}})$	$J(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
2	40	0.025	2.16×10^{-2}	1.14×10^{-2}	1.18×10^{-3}
4	40	0.025	2.77×10^{-3}	1.83×10^{-3}	9.02×10^{-4}
6	40	0.025	8.46×10^{-4}	7.68×10^{-4}	6.91×10^{-4}
8	40	0.025	4.08×10^{-4}	4.69×10^{-4}	5.29×10^{-4}
10	40	0.025	2.55×10^{-4}	3.30×10^{-4}	4.06×10^{-4}
12	40	0.025	1.86×10^{-4}	2.48×10^{-4}	3.10×10^{-4}
14	40	0.025	1.50×10^{-4}	1.92×10^{-4}	2.34×10^{-4}
16	40	0.025	1.28×10^{-4}	1.51×10^{-4}	1.74×10^{-4}

Таблица 2. Задача экранирования: $\mu_{\min} = 0.0045$, $\mu_{\max} = 40$, $\mathbf{m}^{\text{opt}} = \mathbf{m}^{\text{alt}}$

M	μ_1^{opt}	μ_M^{opt}	$J_i(\mathbf{m}^{\text{opt}})$	$J(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
2	40	0.0045	4.14×10^{-3}	2.78×10^{-3}	1.41×10^{-3}
4	40	0.0045	1.10×10^{-4}	7.30×10^{-4}	1.35×10^{-3}
6	40	0.0045	7.97×10^{-6}	6.50×10^{-4}	1.29×10^{-3}
8	40	0.0045	1.08×10^{-6}	6.19×10^{-4}	1.24×10^{-3}
10	40	0.0045	2.25×10^{-7}	5.93×10^{-4}	1.19×10^{-3}
12	40	0.0045	6.46×10^{-8}	5.69×10^{-4}	1.14×10^{-3}
14	40	0.0045	2.36×10^{-8}	5.48×10^{-4}	1.10×10^{-3}
16	40	0.0045	1.04×10^{-8}	5.28×10^{-4}	1.06×10^{-3}

новой пары $(\mu_{\min}, \mu_{\max}) = (0.0045, 40)$. Она содержит оптимальные значения μ_1^{opt}, μ_M^{opt} магнитных проницаемостей первого и последнего слоев, совпадающие в силу (5.3) с $\mu_{\max}$ и $\mu_{\min}$ соответственно, и оптимальные значения $J_i(\mathbf{m}^{\text{opt}})$ функционала $J_i(\mathbf{m})$, где $\mathbf{m}^{\text{opt}} = \mathbf{m}^{\text{alt}}$, вместе со значениями $J_e(\mathbf{m}^{\text{opt}})$ и $J(\mathbf{m}^{\text{opt}})$. Анализ табл. 2 показывает, что значения $J_i(\mathbf{m}^{\text{opt}})$ изменяются в пределах от 4.14×10^{-3} до 1.04×10^{-8} при увеличении M от 2 до 16. В то же время значения $J_e(\mathbf{m}^{\text{opt}})$ и $J(\mathbf{m}^{\text{opt}})$, приведенные в двух последних столбцах табл. 2, достаточно велики. Это, естественно, связано с тем, что мы минимизируем именно функционал $J_i(\mathbf{m})$.

Последнее значение $J_i(\mathbf{m}^{\text{opt}}) = 1.04\times10^{-8}$ (при $M = 16$) соответствует достаточно высокой экранирующей эффективности оптимальной оболочки $(\Omega, \mathbf{m}^{\text{opt}})$, причем спроектированная в рамках теста 2 оболочка $(\Omega, \mathbf{m}^{\text{alt}})$ допускает простую техническую реализацию, поскольку она состоит из слоев, заполненных чередующимися распространенными материалами. Первым из них является распространенный инженерный материал – закаленная нержавеющая сталь с относительной магнитной проницаемостью $\mu = 40$, вторым является известный метаматериал SuperPower SCS12050 с относительной магнитной проницаемостью $\mu = 0.0045$.

Наконец, выбрав в качестве следующего теста третий сценарий в (5.5), когда $\mu_{\min} = 0.0045$, $\mu_{\max} = 70$ и применяя МРЧ, мы получаем результаты, приведенные в табл. 3, являющейся аналогом табл. 1 для третьей пары $(\mu_{\min}, \mu_{\max})$ в (5.5). Видно, что при увеличении M от 2 до 16 значения $J_i(\mathbf{m}^{\text{opt}})$ уменьшаются от 2.43×10^{-3} до значения 2.45×10^{-10}, которое отвечает очень высокой экранирующей эффективности.

Обсудим теперь результаты вычислительных экспериментов по решению задачи полной маскировки (4.4). Предварительно напомним, что в физической литературе общепринято в качестве

Таблица 3. Задача экранирования: $\mu_{\min} = 0.0045$, $\mu_{\max} = 70$, $\mathbf{m}^{\text{opt}} = \mathbf{m}^{\text{alt}}$

M	μ_1^{opt}	μ_M^{opt}	$J_i(\mathbf{m}^{\text{opt}})$	$J(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
2	70	0.0045	2.43×10^{-3}	1.92×10^{-3}	1.41×10^{-3}
4	70	0.0045	3.80×10^{-5}	6.94×10^{-4}	1.35×10^{-3}
6	70	0.0045	1.64×10^{-6}	6.46×10^{-4}	1.29×10^{-3}
8	70	0.0045	1.35×10^{-7}	6.17×10^{-4}	1.23×10^{-3}
10	70	0.0045	1.77×10^{-8}	5.90×10^{-4}	1.18×10^{-3}
12	70	0.0045	3.28×10^{-9}	5.64×10^{-4}	1.13×10^{-3}
14	70	0.0045	8.03×10^{-10}	5.41×10^{-4}	1.08×10^{-3}
16	70	0.0045	2.45×10^{-10}	5.18×10^{-4}	1.04×10^{-3}

решения задачи маскировки использовать решение $\mathbf{m}^{\text{alt}}$, отвечающее чередующемуся дизайну [6], [26]. Это связано с тем, что оболочку, состоящую из чередующихся слоев с большим контрастом, принято считать хорошей аппроксимацией анизотропной оболочки, которая получается в результате применения для решения задач дизайна устройств маскировки метода, основанного на ТО подходе [2]. Однако, как показывает анализ табл. 1–3, оболочки $(\Omega, \mathbf{m}^{\text{alt}})$, отвечающие схеме чередующегося дизайна, обладают низкой маскировочной эффективностью. Действительно, в то время как значения $J_i(\mathbf{m}^{\text{opt}})$, где $\mathbf{m}^{\text{opt}} = \mathbf{m}^{\text{alt}}$, приведенные в табл. 2 и 3, достаточно малы при M, близких к 16, отвечая высокой экранирующей эффективности оболочки $(\Omega, \mathbf{m}^{\text{alt}})$, значения $J(\mathbf{m}^{\text{opt}})$, наоборот, относительно велики, имея порядок 10^{-3}–10^{-4} даже при M, близких к 16, что отвечает невысокой маскировочной эффективности. Для того, чтобы получить решение задачи маскировки, обладающее высокой маскировочной эффективностью, необходимо решить именно экстремальную задачу (4.4), отвечающую задаче полной маскировки.

Мы начнем наш анализ результатов решения задачи (4.4) с анализа простейшего случая двухслойной оболочки ($M = 2$). Напомним, что в этом случае существует точное решение $(0, \mu_2^0)$ задачи полной маскировки, где μ_2^0 определяется формулой (3.15). Простые вычисления с учетом соотношений (5.1), согласно которым $R_0 = a = 0.035$ м, $R_1 = (a+b)/2 = 0.0425$ м, $R_2 = b = 0.05$ м, показывают, что $\mu_2^0 = \mu_0(2R_2^3 + R_1^3)/(2(R_2^3 - R_1^3)) = 3.38726919339$. (Последнее значение записано с 11-ю верными цифрами после запятой). С учетом этого мы зафиксируем верхнюю границу $\mu_{\max} = 10$, а в качестве нижней границы $\mu_{\min}$ выберем убывающую последовательность $\mu_1^n = 10^{-n}$. Результаты решения задачи (4.4) в виде оптимальных проницаемостей μ_1^{opt}, μ_2^{opt} и оптимальных значений $J(\mathbf{m}^{\text{opt}})$ функционала $J(\mathbf{m})$, где $\mathbf{m}^{\text{opt}} = (\mu_1^{\text{opt}}, \mu_2^{\text{opt}})$, представлены вместе со значениями $J_e(\mathbf{m}^{\text{opt}})$ и $J_i(\mathbf{m}^{\text{opt}})$ (для сравнения) в табл. 4 для ряда значений μ_1^n, изменяющихся от 10^{-1} до 10^{-12}.

Анализ табл. 4 показывает, что с уменьшением нижней границы $\mu_{\min}$ множества K от 10^{-1} до 10^{-12}, чему соответствует увеличение контраста $\mu_{\max}/\mu_{\min}$ от 100 до 10^{13}, оптимальное решение $(\mu_1^{\text{opt}}, \mu_2^{\text{opt}})$ задачи полной маскировки (4.4), найденное с помощью МРЧ, при стремлении $\mu_{\min}$ к нулю приближается к точному (сингулярному) решению $(0, \mu_2^0)$ задачи полной маскировки. При этом значение $J(\mathbf{m}^{\text{opt}})$, где $\mathbf{m}^{\text{opt}} = (\mu_1^{\text{opt}}, \mu_2^{\text{opt}})$, изменяется от 1.10×10^{-1} при $\mu_{\min} = 10^{-1}$ до значения 3.90×10^{-12} при $\mu_{\min} = 10^{-12}$, отвечающего очень высокой маскировочной эффективности оболочки $(\Omega, \mathbf{m}^{\text{opt}})$. Указанные факты подтверждают высокую точность используемого нами оптимизационного метода для решения обратной задачи полной маскировки. Хотя, как уже отмечалось выше, в практическом плане полученные с высокой точностью решения, отвечающие малым значениям μ_1^{opt}, мало перспективны в виду их сингулярности.

Таблица 4. Задача маскировки: $M = 2$, $\mu_{\min}^{n} = 10^{-n}$, $n = 1,...,12$, $\mu_{\max} = 10$

$\mu_{\min}$	$\mu_{\max}$	μ_1^{opt}	μ_2^{opt}	$J(\mathbf{m}^{\text{opt}})$	$J_i(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
10^{-1}	10.0	10	0.10000000000	1.10×10^{-1}	2.20×10^{-1}	5.86×10^{-4}
10^{-2}	10.0	10	0.01000000000	1.60×10^{-2}	3.06×10^{-2}	1.35×10^{-3}
10^{-4}	10.0	10^{-4}	3.38595617966	3.90×10^{-4}	7.80×10^{-4}	0.0
10^{-6}	10.0	10^{-6}	3.38725605599	3.90×10^{-6}	7.80×10^{-6}	0.0
10^{-8}	10.0	10^{-8}	3.38726906202	3.90×10^{-8}	7.80×10^{-8}	0.0
10^{-10}	10.0	10^{-10}	3.38726919208	3.90×10^{-10}	7.80×10^{-10}	0.0
10^{-12}	10.0	10^{-12}	3.38726919338	3.90×10^{-12}	7.80×10^{-12}	0.0

Из табл. 4 следует, что $\mu_1^{\text{opt}} = \mu_{\min}$ для всех $n = 4,...,12$ (кроме $n = 1,2$, когда $\mu_1^{\text{opt}} = \mu_{\max}$ и $\mu_2^{\text{opt}} = \mu_{\min}$), тогда как μ_2^{opt} принимает некоторое промежуточное значение между $\mu_{\min}$ и $\mu_{\max}$. Это является проявлением общей тенденции в поведении отдельных компонент оптимального решения задачи (4.4) для всех значений M. Указанная тенденция состоит в том, что соотношения (5.2) (либо (5.3)) для оптимальных проницаемостей, отвечающие схеме чередующегося дизайна 1-го (либо 2-го) типа, выполняются при выполнении условия (5.4) для всех управлений μ_m^{opt}, кроме, быть может, последнего μ_M^{opt}, которое принимает некоторое промежуточное значение между $\mu_{\min}$ и $\mu_{\max}$. Другими словами, для задачи (4.4) вместо (5.2) (либо (5.3)) выполняются следующие соотношения почти чередующегося дизайна:

$$\mu_1^{\text{opt}} = \mu_3^{\text{opt}} = ... = \mu_{M-1}^{\text{opt}} = \mu_{\min}, \quad \mu_2^{\text{opt}} = ... = \mu_{M-2}^{\text{opt}} = \mu_{\max}, \quad \mu_{\min} \le \mu_M^{\text{opt}} \le \mu_{\max} \tag{5.6}$$

либо

$$\mu_1^{\text{opt}} = \mu_3^{\text{opt}} = ... = \mu_{M-1}^{\text{opt}} = \mu_{\max}, \quad \mu_2^{\text{opt}} = ... = \mu_{M-2}^{\text{opt}} = \mu_{\min}, \quad \mu_{\min} \le \mu_M^{\text{opt}} \le \mu_{\max}. \tag{5.7}$$

Приведем теперь результаты решения задачи маскировки для нескольких конкретных тестов. Первый тест отвечает первой выбранной ранее паре $(\mu_{\min}, \mu_{\max}) = (0.025, 40)$ в (5.5). Соответствующие результаты решения задачи (4.4) в виде оптимальных управлений μ_1^{opt}, μ_M^{opt} и оптимальных значений $J(\mathbf{m}^{\text{opt}})$ функционала $J(\mathbf{m})$ представлены вместе со значениями $J_i(\mathbf{m}^{\text{opt}})$ и $J_e(\mathbf{m}^{\text{opt}})$ (для сравнения) в табл. 5. Остальные управления $\mu_2^{\text{opt}},...,\mu_{M-1}^{\text{opt}}$ при $M \ge 4$ определяются из соотношений (5.7), отвечающих схеме почти чередующегося дизайна 2-го типа. Из табл. 5 видно, что последнее управление μ_M^{opt} совпадает с $\mu_{\min} = 0.025$ при $M = 2, 4$ и 6, при $M \ge 8$ оно принимает промежуточные между $\mu_{\min}$ и $\mu_{\max}$ значения, убывающие от 0.0546 при $M = 8$ до 0.0339 при $M = 16$, тогда как значение $J(\mathbf{m}^{\text{opt}})$ убывает от 1.14×10^{-2} при $M = 2$ до значения 7.28×10^{-5} при $M = 16$, отвечающего невысокой маскировочной эффективности оболочки $(\Omega, \mathbf{m}^{\text{opt}})$.

Следующий тест отвечает второй выбранной ранее паре $(\mu_{\min}, \mu_{\max}) = (0.0045, 40)$ в (5.5). Соответствующие результаты решения задачи (4.4) в виде управлений μ_1^{opt}, μ_M^{opt} и оптимальных значений $J(\mathbf{m}^{\text{opt}})$ функционала $J(\mathbf{m})$ представлены вместе со значениями $J_i(\mathbf{m}^{\text{opt}})$ и $J_e(\mathbf{m}^{\text{opt}})$ в табл. 6, которая является аналогом табл. 5. Остальные значения $\mu_2^{\text{opt}},...,\mu_{M-1}^{\text{opt}}$ при $M \ge 4$ определяются из соотношений (5.6), отвечающих схеме почти чередующегося дизайна 1-го типа. Из табл. 6 видно, что последнее управление μ_M^{opt} совпадает с $\mu_{\min} = 0.0045$ при $M = 2$, а при изменении M от 4 до 16 μ_M^{opt} возрастает от значения 6.3817 до 21.3467, тогда как значение $J(\mathbf{m}^{\text{opt}})$ убывает от 2.78×10^{-3} при $M = 2$ до значения 8.09×10^{-9} при $M = 16$, отвечающего достаточно высокой маскировочной эффективности оболочки $(\Omega, \mathbf{m}^{\text{opt}})$.

Таблица 5. Задача маскировки: $\mu_{\min} = 0.025$, $\mu_{\max} = 40$

M	μ_1^{opt}	μ_M^{opt}	$J(\mathbf{m}^{\text{opt}})$	$J_i(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
2	40	0.025	1.14×10^{-2}	2.16×10^{-3}	1.18×10^{-3}
4	40	0.025	1.83×10^{-3}	2.77×10^{-3}	9.02×10^{-4}
6	40	0.025	7.68×10^{-4}	8.46×10^{-3}	6.91×10^{-4}
8	40	0.0546	3.19×10^{-4}	6.37×10^{-4}	4.78×10^{-19}
10	40	0.0465	1.76×10^{-4}	3.52×10^{-4}	1.59×10^{-19}
12	40	0.0411	1.18×10^{-4}	2.36×10^{-4}	7.96×10^{-20}
14	40	0.0370	8.91×10^{-5}	1.78×10^{-4}	3.19×10^{-19}
16	40	0.0339	7.28×10^{-5}	1.46×10^{-4}	0.0

Таблица 6. Задача маскировки: $\mu_{\min} = 0.0045$, $\mu_{\max} = 40$

M	μ_1^{opt}	μ_M^{opt}	$J(\mathbf{m}^{\text{opt}})$	$J_i(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
2	40	0.0045	2.78×10^{-3}	4.14×10^{-3}	1.41×10^{-3}
4	0.0045	6.3817	2.43×10^{-4}	4.85×10^{-4}	0.0
6	0.0045	9.2633	1.24×10^{-5}	2.48×10^{-5}	0.0
8	0.0045	11.9708	1.33×10^{-6}	2.66×10^{-6}	0.0
10	0.0045	14.5166	2.35×10^{-7}	4.71×10^{-7}	0.0
12	0.0045	16.9168	5.97×10^{-8}	1.19×10^{-7}	0.0
14	0.0045	19.1880	1.98×10^{-8}	3.96×10^{-8}	0.0
16	0.0045	21.3467	8.09×10^{-9}	1.62×10^{-8}	0.0

Таблица 7. Задача маскировки: $\mu_{\min} = 0.0045$, $\mu_{\max} = 70$

M	μ_1^{opt}	μ_M^{opt}	$J(\mathbf{m}^{\text{opt}})$	$J_i(\mathbf{m}^{\text{opt}})$	$J_e(\mathbf{m}^{\text{opt}})$
2	70	0.0045	1.92×10^{-3}	2.43×10^{-3}	1.41×10^{-3}
4	0.0045	6.3806	1.39×10^{-4}	2.79×10^{-4}	0.0
6	0.0045	9.2573	4.15×10^{-6}	8.31×10^{-6}	0.0
8	0.0045	11.9512	2.66×10^{-7}	5.31×10^{-7}	0.0
10	0.0045	14.4691	2.88×10^{-8}	5.75×10^{-8}	0.0
12	0.0045	16.8218	4.61×10^{-9}	9.22×10^{-9}	0.0
14	0.0045	19.0213	1.00×10^{-9}	2.00×10^{-9}	0.0
16	0.0045	21.0804	2.77×10^{-10}	5.54×10^{-10}	0.0

Достигнутую эффективность можно сделать еще выше, если увеличить контраст $\mu_{\max}/\mu_{\min}$ за счет увеличения $\mu_{\max}$. В этом можно убедиться из табл. 7, которая является аналогом табл. 6 для случая $\mu_{\max} = 70$, отвечающего магнитной проницаемости кобальта. Видно, что указанное изменение $\mu_{\max}$ до 70 привело к изменению $J(\mathbf{m}^{\text{opt}})$ от 1.92×10^{-3} при $M = 2$ до значения 2.77×10^{-10} при $M = 16$, которое отвечает очень высокой маскировочной эффективности оболочки $(\Omega, \mathbf{m}^{\text{opt}})$. Отметим еще одну особенность последних двух тестов. Она состоит в том, что построенное оптимальное решение $\mathbf{m}^{\text{opt}}$ задачи полной маскировки является при $M \geq 4$ одновременно точным решением задачи внешней маскировки. Это вытекает из соотношений $J_e(\mathbf{m}^{\text{opt}}) = 0$, приведенных в последнем столбце каждой из табл. 6 и 7 при $M > 2$.

На наш взгляд, представляет интерес сравнить табл. 7 и 3 (а также 6 и 2), отвечающие одним и тем же данным ($\mu_{\min}, \mu_{\max}$). Указанное сравнение позволяет сделать вывод об очень сильном влиянии на качество (т.е. точность) решения задачи маскировки именно последнего управления μ_M^{opt}. Действительно, ошибка решения задачи маскировки, определяемая значением $J(\mathbf{m}^{\text{alt}})$, где $\mathbf{m}^{\text{alt}}$ – чередующееся решение, равна 5.18×10^{-4} в табл. 3 (при $M=16$) и равна $J(\mathbf{m}^{\text{opt}}) = 2.77\times10^{-10}$ в табл. 7 на оптимальном решении $\mathbf{m}^{\text{opt}}$ задачи (4.4), отличающемся от $\mathbf{m}^{\text{alt}}$ лишь последней компонентой μ_M^{opt}. Такое существенное уменьшение значения $J(\mathbf{m}^{\text{opt}})$ на шесть порядков обусловлено изменением последнего управления μ_M^{opt}.

Таким образом, применение численного алгоритма решения обратных задач экранирования и маскировки, основанного на оптимизационном методе, позволило получить решения, обладающие наивысшей эффективностью в рассматриваемом классе. Для задачи экранирования полученные решения обладают также простотой технической реализации при определенном выборе множества управлений K в (4.1), поскольку отвечающие им экранирующие оболочки состоят из чередующихся слоев, заполненных распространенными материалами. Для задачи маскировки материал последнего слоя может не принадлежать классу природных или инженерных материалов. В связи с этим может возникнуть техническая трудность с реализацией полученных решений. Однако эта трудность не является принципиальной ввиду больших успехов, достигнутых в последнее время в создании метаматериалов с заданными магнитными свойствами.

Полученные в статье результаты относятся к случаю однородного внешне приложенного магнитного поля, имеющего вид (3.1). На наш взгляд, большой интерес представляет перенесение полученных выше результатов на случай неоднородного внешне приложенного поля, создаваемого компактно распределенными источниками. Исследованию этой проблемы авторы собираются посвятить будущую работу.

СПИСОК ЛИТЕРАТУРЫ

1. *Долин Л.С.* О возможности сопоставления трехмерных электромагнитных систем с неоднородным анизотропным заполнением // Изв. вузов. Радиофизика. 1961. Т. 4. № 4. С. 964–967.
2. *Pendry J.B., Schurig D., Smith D.R.* Controlling electromagnetic fields // Science. 2006. V. 312. P. 1780–1782.
3. *Leonhardt U.* Optical conformal mapping // Science. 2006. V. 312. P. 1777–1780.
4. *Ward A.J., Pendry J.B.* Refraction and geometry in Maxwell's equations // J. of Modern Optics. 1996. V. 43. P. 773–793.
5. *Chen H., Chan C.T.* Acoustic cloaking in three dimensions using acoustic metamaterial // Appl. Phys. Lett. 2007. V. 91. № 183518. P. 1–3.
6. *Han T., Qiu C.-W.* Transformation Laplacian metamaterials: recent advances in manipulating thermal and dc fields // J. Opt. 2016. V. 18. № 044003. P. 1–13.
7. *Yang F., Zhong Mei Z.L., Jin T.Y., Cui T.J.* DC electric invisibility cloak // Phys. Rev. Lett. 2012. V. 109. № 053902. P. 1–5.
8. *Gomory F., Solovyov M., Souc J., Navau C., Prat-Camps J., Sanchez A.* Experimental realization of a magnetic cloak // Science. 2012. V. 335. P. 1466–1468.
9. *Guenneau S., Amra C., Veynante D.* Transformation thermodynamics: cloaking and concentrating heat flux // Opt. Express. 2012. V. 20. P. 8207–8218.
10. *Shestopalov Yu.V., Smirnov Yu.G.* Determination of permittivity of an inhomogeneous dielectric body in a waveguide // Inverse Problems. 2011. V. 27. № 9. P. 095010.
11. *Smirnov Y.G., Medvedik M.Y., Moskaleva M.A.* Two-step method for permittivity determination of an inhomogeneous body placed in a rectangular waveguide // Lobachevskii J. Math. 2018. V. 39. P. 1114–1147.
12. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. М.: Наука, 1986. 288 с.
13. *Peralta I., Fachinotti V.D.* Optimization-based design of heat flux manipulation devices with emphasis on fabricability // Sci. Rep. 2017. V. 7. № 6261. P. 1–8.
14. *Fachinotti V.D., Ciarbonetti A.A., Peralta I., Rintoul I.* Optimization-based design to easy-to-make devices for heat flux manipulation // Int. J. Therm. Sci. 2018. V. 128. P. 38–48.
15. *Алексеев Г.В., Левин В.А., Терешко Д.А.* Оптимизационный анализ задачи тепловой маскировки цилиндрического тела // Докл. АН. 2017. Т. 472. № 4. С. 398–402.
16. *Алексеев Г.В., Левин В.А., Терешко Д.А.* Оптимизационный метод в задачах дизайна сферических слоистых тепловых оболочек // Докл. АН. 2017. Т. 476. № 5. С. 512–517.

17. *Алексеев Г.В., Терешко Д.А.* Оптимизационный метод в осесимметричных задачах электрической маскировки материальных тел // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 2. С. 217–234.
18. *Alekseev G.V., Tereshko D.A.* Particle swarm optimization-based algorithms for solving inverse problems of designing thermal cloaking and shielding devices // Int. J. Heat Mass Transf. 2019. V. 135. P. 1269–1277.
19. *Lobanov A.V., Spivak Yu.E.* Numerical analysis of problem of designing magnetic bilayer cloak // Progress In Electromagnetics Research Symposium – Spring (PIERS). 2017. P. 1362–1366.
20. *Alekseev G.V., Tereshko D.A., Shestopalov Yu.V.* Optimization approach for axisymmetric electric field cloaking and shielding // Inverse Probl. Sci. Eng. 2020. V. 28. P. 1–16. Published online: 02.06.2020
21. *Алексеев Г.В.* Оценки устойчивости в задаче маскировки материальных тел для уравнений Максвелла // Ж. вычисл. матем. и матем. физ. 2014. Т. 54. № 12. С. 1863–1878.
22. *Алексеев Г.В.* Анализ и оптимизация в задачах маскировки материальных тел для уравнений Максвелла // Дифференц. ур-ния. 2016. Т. 52. № 3. С. 366–377.
23. *Алексеев Г.В., Спивак Ю.Э.* Теоретический анализ задачи магнитной маскировки на основе оптимизационного метода // Дифференц. ур-ния. 2018. Т. 54. № 9. С. 1155–1166.
24. *Спивак Ю.Э.* Оптимизационный метод в двумерных задачах магнитной маскировки // Сиб. электрон. матем. изв. 2019. Т. 16. С. 812–825.
25. *Poli R., Kennedy J., Blackwell T.* Particle swarm optimization: an overview // Swarm Intel. 2007. V. 1. P. 33–57.
26. *Алексеев Г.В.* Проблема невидимости в акустике, оптике и теплопереносе. Владивосток: Дальнаука, 2016. 224 с.
27. *Kennedy J., Eberhart R.* Particle swarm optimization // Proceedings of IEEE International Conference on Neural Networks. IV. 1995. P. 1942–1948.
28. *Chiang A.C.* Elements of Dynamic Optimization. New York: McGraw-Hill, 1992. 327 p.
29. *Solovyov M., Gomory F., Souc J., Mikulasova E., Usakova M., Usak E.* Force acting on a magnetic cloak placed in magnetic field // The 13th biennial European Conference on Applied Superconductivity. 2017. Poster № 3LP4-03.
30. SuperPower Inc., http://www.superpower-inc.com/

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 239–255

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.634

АНАЛИТИЧЕСКОЕ ИССЛЕДОВАНИЕ ХАОТИЧЕСКОЙ ДИНАМИКИ ДВУМЕРНОЙ МОДЕЛИ ЛОТКИ–ВОЛЬТЕРРА С СЕЗОННОСТЬЮ

© 2021 г. Ю. В. Бибик

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

e-mail: yvbibik@ccas.ru

Поступила в редакцию 01.01.2020 г.
Переработанный вариант 01.01.2020 г.
Принята к публикации 15.08.2020 г.

Аналитически исследована динамика классической биологической системы Лотки–Вольтерра, к которой добавлен фактор сезонности. Исходная модель описана простым гамильтонианом. В ходе исследования для выявления хаотического поведения в системе гамильтониан представлен в виде суммы не зависящего от времени гамильтониана и отдельных резонансов. Исследование взаимодействия выделенных резонансов с помощью метода перекрытия резонансов Чирикова позволило определить аналитический критерий в терминах критических значений амплитуд сезонности, при которых в исходной системе происходит переход к хаосу. Результаты исследования показывают, что при наличии периодического возмущения (в данном случае сезонности) в системе с двумя зависимыми переменными возникает хаотическое поведение. Библ. 36.

Ключевые слова: хаос, система Лотки–Вольтерра, сезонность, метод перекрытия резонансов Чирикова.

10.31857/S0044466921010026

1. ВВЕДЕНИЕ

В работе исследуются особенности хаотического поведения обобщения биологической системы Лотки–Вольтерра [1]–[4] с сезонностью. Система описывается простым гамильтонианом. Для определения характера динамики системы, включая выявление в ней хаотического поведения, и для определения параметров, при которых хаотическое поведение возникает, использован метод перекрытия резонансов Чирикова [5].

Теория хаотического поведения динамических систем является важнейшим научным достижением прошлого и текущего столетий. По значимости, которую эта теория внесла в понимание фундаментальных законов мироздания, она стоит в одном ряду с такими важнейшими теориями, как теория относительности и квантовая теория. Отличительной чертой сегодняшней теории хаоса является то, что она имеет междисциплинарное значение. Известные на сегодня основные характеристики хаоса, такие как нелинейность, детерминизм, чувствительность к начальным условиям, невозможность долгосрочного прогноза, устойчивые нарушения в поведении системы, наличие окон периодичности, оказались очень информативными и эффективными для понимания и объяснения природы землетрясений, лесных пожаров, эпидемий, колебаний в экономических и финансовых системах. Эта теория не имеет узко научной направленности, на ее основе формируются методы и подходы, позволяющие предотвратить катастрофические явления в экологии, биологии, медицине.

До описания деталей исследования представляется необходимым вкратце отметить наиболее значимые, основополагающие достижения и открытия, составляющие сегодняшний фундамент этой теории. Работы выдающихся математиков Анри Пуанкаре [6] 1892 г., [7] 1905 г. и Жака Адамара [8] 1898 г. для математической теории хаоса являются базовыми. Как известно, Пуанкаре впервые отметил, что на орбитах трех или более взаимодействующих небесных тел может возникать нестабильность и непредсказуемость поведения, а также то, что небольшие различия в начальных условиях дают в финале очень большие изменения, и предсказание поведения системы становится невозможным. Жак Адамар описал первую динамическую структуру, оказавшуюся

хаотической. Работы Пуанкаре и Адамара были известны, но долгое время не находили широкого отклика среди исследователей.

В 1970-х годах появился целый ряд исследований, выдвинувших теорию хаоса в число важнейших научных достижений. Основные результаты важнейших исследований были получены после появления компьютерной техники, по мере накопления экспериментальных, теоретических данных, развития численных методов. Прорывные исследования этого периода начинаются с работ метеоролога Лоренца. Занимаясь исследованиями в области прогнозирования погоды, он исследовал компьютерную модель, содержащую уравнения, учитывающие зависимость между температурой, атмосферным давлением и скоростью ветра. Результаты исследований он опубликовал в знаменитых статьях "Deterministic Nonperiodic flow" ("Детерминированный непериодический поток") [9] 1963 г. и "Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?" ("О возможности предсказаний: может ли взмах крыльев бабочки в Бразилии вызвать торнадо в Техасе?") [10] 1972 г. Результаты его исследований были оценены исследователями-современниками как "революционные". Они обогатили теорию хаоса пониманием того, что небольшие изменения, происходящие в атмосфере, приводят к радикальным и неожиданным последствиям, что сложные динамические системы имеют порог предсказуемости, поэтому прогнозировать динамическое поведение сложных динамических моделей на длительный период невозможно. Лоренц также открыл первый "странный аттрактор" – подмножество фазового пространства, для которого все траектории, стартующие недалеко от него, стремятся к нему с течением времени.

К прорывным работам относятся работы Митчелла Фейгенбаума 1978–1981 гг., показавшего, что хаос может возникать через различные последовательности бифуркаций. Он обратил внимание на универсальные закономерности перехода к хаосу при удвоении периода. Используя метод ренормализационной группы, Фейгенбаум создал теорию, объясняющую универсальность удвоений периода [11], [12], [13]. Важнейшими для понимания природы хаоса стали также исследования этого периода Бенуа Мальденброта, обратившего внимание на то, что абсолютно случайные процессы имеют элементы подобия. Он впервые ввел понятие "фрактал", одно из глобальных понятий в физике хаоса [14].

Следующий важнейший вклад в развитие теории хаоса был сделан в 1990 г., когда были опубликованы две выдающиеся новаторские работы, предложившие метод контроля хаоса и метод его синхронизации.

В первой их них (статья Эдварда Отта, Селсо Гребоги и Джеймса Йорке "Контроль хаоса" (Edward Ott, Celso Grebogi and James A. Yorke "Controlling Chaos")) [15] был изложен универсальный метод контроля хаоса. Сегодня этот метод известен как метод OGY (по первым буквам фамилий авторов). Суть метода состоит в использовании аттрактора хаотической системы для выбора нестабильной орбиты, которая давала бы улучшенную производительность, и ее стабилизации с применением небольших возмущений системы. Предложенный метод позволяет использовать исключительную чувствительность хаотических систем к крошечным возмущениям для контроля хаоса и стабилизации системы. Указанный метод является универсальным и используется, например, в медицине, для исследования и выработки стратегии контроля за поведенческим возбуждением, эпилептическими припадками, сердечными приступами.

Во второй работе "Синхронизация в хаотических системах" (Louis M. Pecora and T.L. Carroll "Synchronization in chaotic systems") [16], авторы Луис Пекора и Т.Л. Кэрролл предложили первую рабочую схему синхронизации хаоса. Техника синхронизации позволяет использовать хаос в шифровании, при передаче информации по широкополосным каналам, более устойчивым к шуму и помехам. Фактически авторы этих работ дали в руки исследователям инструменты управления хаосом.

В 2000-х годах были опубликованы значимые работы, прояснившие строение пространственно-временных хаотических структур. Особого внимания заслуживает работа Мэтью Фишмана и Дэвида Эгольфа "Выявление строительных блоков пространственно-временного хаоса: отклонения от экстенсивности" (Matthew P. Fishman, David A. Egolf "Revealing the building block of spatiotemporal chaos: deviations from extensivity") [17] 2006. Авторы выполнили высокоточные вычислительные исследования фрактальной размерности как функции длины системы для пространственно-временных хаотических состояний одномерного комплексного уравнения Гинзбурга–Ландау. Отклонения от экстенсивности по шкале длин показали, что эта пространственно-временная хаотическая система состоит из слабо взаимодействующих строительных блоков, каждый из которых содержит около 2-х степеней свободы. Результаты исследования дали объяснение "окнам периодичности", обнаруженным в пространственно-временных структурах умерен-

ных размеров. Выводы, полученные в работе [17] путем компьютерного моделирования уравнения Гинзбурга–Ландау, были подтверждены 2008 г. экспериментально. В работе [18], при исследовании жидкокристаллической электроконвекции с мягкими граничными условиями, было подтверждено существование четко определенных строительных блоков, управляющих динамикой пространственно-временного хаоса.

В 2015 г. опубликована работа, впервые доказавшая наличие хаоса в природе (Статья Элизы Бенинки и соавт. "Колебания видов, поддерживаемые циклической преемственностью на грани хаоса" (Elisa Benincà, Bill Ballantine, Stephen P. Ellner, and Jef Huisman. "Species fluctuations sustained by a cyclic succession at the edge of chaos")) [19]. Работа основана на двадцатилетних наблюдениях в природных условиях за динамикой популяций в многовидовом сообществе (моллюсков, бурых водорослей и мидий). Учеными впервые подтверждено наличие в природных условиях хаотического поведения указанного сообщества в циклической сукцессии (голая порода → → ракушки и водоросли коры → мидии → голая порода) на Северном острове Новой Зеландии. Указанная работа обогатила теорию хаоса углубленным пониманием особенностей хаотического поведения экосистем в природных условиях.

В последнее десятилетие, вплоть до настоящего времени, проводятся теоретические, численные и экспериментальные исследования исключительной важности в области анализа и прогнозирования поведения хаотических систем. Эта проблема носит междисциплинарный характер и является ключевой в физике, экологии, биологии, медицине. Для прогнозирования возможных катастрофических сдвигов в физических и экологических системах проводятся исследования по выявлению и использованию в прогностических целях универсальных сигналов-индикаторов раннего предупреждения (Early Warning Signals (EWS)), которые бы заранее указывали на возможные сдвиги в процессах, которые могут изменить структуру и функции системы. Общая оценка состояния вопроса, перечень рисков, которые планируется выявить и предупредить с помощью сигналов-индикаторов, методы выявления самих сигналов, а также перечень проблем, требующих решения, перечислены в работах [20], [21]. Экспериментальное исследование по выявлению и использованию сигналов раннего предупреждения в термоакустической системе [22] подтверждает эффективность их использования. Исследования сигналов-индикаторов в работе [23] показали, что катастрофический коллапс в экологической системе может произойти без наличия сигналов раннего предупреждения, и для того, чтобы оценить, можно ли считать данный сигнал действительно эффективным сигналом раннего предупреждения, необходимы дополнительные исследования и проверка результатов.

Успешными и перспективными являются исследования на стыке теории хаоса с другими научными направлениями. Примером эффективного сотрудничества являются исследования в области теории лагранжевых когерентных структур (LCSs), использующих идеи теории хаоса и динамики жидкости. Теория позволяет не только понять, как потоки жидкости изолированы друг от друга, но и предсказать направление транспортировки в сложных потоках жидкости и в атмосферных потоках [24]. Углубленное исследование LCSs позволит улучшить на практике мониторинг переноса загрязненных потоков жидкости на поверхности океана и/или загрязненных потоков в атмосфере, вызванных как природными, так и техногенными катастрофами, а также улучшить мониторинг промышленных и биологических потоков. Исследуется использование идей LCSs в турбулентной жидкости [25] для описания транспортных процессов в плазме [26], в экологии, в работе [27] отмечено, что LCSs отслеживаются птицами для нахождения добычи, а использование метода LCSs "открывает интересные перспективы в управлении экосистемами и рыболовством".

В настоящее время основной проблемой хаотических систем является разработка универсальной методики их анализа в условиях сильной нелинейности взаимосвязей и случайных возмущений.

Данное аналитическое исследование выполняется в рамках указанной проблемы. Исследуемая в работе система Лотки–Вольтерра представляет собой задачу, традиционно связанную с математической биологией. Однако в последнее время она находит применение при исследовании нелинейной динамики плазмы [28], для исследования хаотической динамики вселенной [29], для анализа и прогнозирования хаотической динамики криптовалют [30].

Как известно, основной сложностью исследования хаотической динамики нелинейных систем является то, что они трудно поддаются прямому аналитическому исследованию.

Большинство исследований в области хаотического поведения различных обобщений системы Лотки–Вольтерра выполнено с использованием численных методов. В приведенных ниже

численных работах выполнен достаточно подробный анализ различных типов поведения обобщенных систем Лотки–Вольтерра при разных параметрах, методах и подходах.

В работе [31] численно исследована возможность существования странных аттракторов в уравнениях Вольтерра для более чем трех видов с конкуренцией. В работе [32] численно исследовано возникновение странных аттракторов в трехмерных уравнениях Вольтерра. В работе [33] численно исследован механизм возникновения бифуркационного хаоса в четырехмерной системе Лотки–Вольтерра с конкуренцией. В работе [34] методом грубой силы численно изучено возникновение хаоса в базовых моделях Лотки–Вольтерра с четырьмя конкурирующими видами. По мнению авторов, изученная в работе система “является самым простым примером хаоса в реалистичной конкурентной модели Лотки–Вольтерра”.

Аналитических работ по исследованию хаоса в обобщенных системах Лотки–Вольтерра мало. Остановимся вкратце на результатах аналитических исследований.

В работе [35] изучена система Лотки–Вольтерра с N видами и n ресурсами. Аналитически доказано существование в системе многих гиперболических динамик на некоторых инвариантных множествах. Показано, что в зависимости от выбора параметров система генерирует все типы гиперболической динамики, включая хаотическую. В статье приведен пример системы Лотки–Вольтерра (для 10-ти видов с 3-мя ресурсами), которая имеет динамику Лоренца. Показано, что система Лоренца может быть реализована системой Лотки–Вольтерра. При этом “система Лотки–Вольтерра не имеет хаотического аттрактора, но имеет семейство параметров хаотических инвариантных множеств (эквивалентно аттракторам Лоренца)”. В статье отмечено, что существуют системы Лотки–Вольтерра для N видов и n ресурсов, которые сильно устойчивы и в то же время демонстрируют хаотичность поведения.

В работе [36] с помощью метода Мельникова аналитически исследована трехмерная система Лотки–Вольтерра с периодическими по времени возмущениями. Метод Мельникова использован для установления существования в системе поперечного гетероклинического цикла, подтверждающего наличие в трехмерной системе хаотического поведения.

Настоящая статья организована следующим образом:

В разд. 2 дано описание исследуемой двумерной модели системы Лотки–Вольтерра с сезонностью. В подразделах 2.1, 2.2 и 2.3 приведены основные понятия, определения и преобразования, выполненные для того, чтобы представить систему Лотки–Вольтерра с сезонностью в виде, который позволил бы применить к ней метод перекрытия резонансов Чирикова для выявления в системе хаотического поведения.

В разд. 3 предложен аналитический подход исследования динамики двумерной системы Лотки–Вольтерра с сезонностью с использованием метода перекрытия резонансов Чирикова.

В разд. 4 выполнен анализ полученных результатов. Показано, что исследованная двумерная система Лотки–Вольтерра с сезонностью имеет сложную динамику и при определенных условиях в системе происходит переход к хаотическому поведению.

Перейдем к описанию исходной модели в следующем разделе.

2. ОПИСАНИЕ ДВУМЕРНОЙ СИСТЕМЫ ЛОТКИ–ВОЛЬТЕРРА С СЕЗОННОСТЬЮ

Исследуемая в работе двумерная система Лотки–Вольтерра с сезонностью является обобщением классической системы Лотки–Вольтерра. Как известно, классическая система представляет собой нелинейный осциллятор, моделирующий колебания численности видов в системах типа хищник–жертва. Она представляет собой гамильтонову систему с одной степенью свободы и поэтому полностью интегрируема. Ее динамика является регулярной, хаос отсутствует. Она имеет следующий вид:

$$\frac{dx}{dt} = x - xy, \tag{2.1}$$

$$\frac{dy}{dt} = -y + xy. \tag{2.2}$$

Здесь x – численность жертв, y – численность хищников.

Однако реальные системы развиваются в среде с периодически меняющимися условиями. Такие системы могут быть описаны системой Лотки–Вольтерра с сезонностью. Уравнение Лотки–Вольтерра с учетом фактора сезонности имеет вид

$$\frac{dx}{dt} = (1 + \alpha\cos\Omega t)x - xy, \tag{2.3}$$

$$\frac{dy}{dt} = -(1 - \alpha\cos\Omega t)y + xy, \tag{2.4}$$

где $1 + \alpha\cos\Omega t$ – коэффициент рождаемости жертв, $1 - \alpha\cos\Omega t$ – коэффициент смертности хищников, α – амплитуда сезонности.

В уравнениях (2.3), (2.4) коэффициенты рождаемости жертв и смертности хищников периодически зависят от времени. Такая периодическая зависимость представляет собой математическое выражение фактора сезонности. После добавления к классической системе Лотки–Вольтерра фактора сезонности гамильтонова система Лотки–Вольтерра с одной степенью свободы превращается в гамильтонову систему с 3/2 степнями свободы. В работе исследуется, допускает ли такая система хаотическую динамику. Существуют разные методы исследования динамики и выявления хаотического поведения в таких системах. Для рассматриваемого случая используется метод перекрытия резонансов Чирикова, позволяющий определить условия перехода исследуемой системы к хаотическому поведению.

2.1 Преобразование исходной системы к гамильтонову виду, выражение не зависящего от времени гамильтониана в промежуточных переменных

Приведем уравнения (2.3), (2.4) к гамильтоновому виду с помощью следующей замены переменных:

$$q = \ln x, \quad p = \ln y. \tag{2.1.1}$$

В новых переменных эти уравнения примут следующий вид:

$$\frac{dq}{dt} = (1 + \alpha\cos\Omega t) - e^p, \tag{2.1.2}$$

$$\frac{dp}{dt} = -(1 - \alpha\cos\Omega t) + e^q. \tag{2.1.3}$$

Уравнения (2.1.2), (2.1.3) гамильтоновы с гамильтонианом H:

$$H = e^q - (1 - \alpha\cos\Omega t)q + e^p - (1 + \alpha\cos\Omega t)p. \tag{2.1.4}$$

Преобразуем гамильтониан (2.1.4), представив его в виде суммы, не зависящего от времени гамильтониана $\bar{H}$ и зависящей от времени добавки

$$H = \bar{H} + \alpha\cos\Omega t = (q - p)e^q - q + e^p - p + \alpha\cos\Omega t(q - p). \tag{2.1.5}$$

В формуле (2.1.5) не зависящая от времени часть гамильтониана записана в терминах канонически сопряженных переменных p и q. Поэтому первым шагом на пути исследования будет переход к переменным типа действие–угол для не зависящей от времени части гамильтониана (2.1.5).

Вначале преобразуем не зависящий от времени гамильтониан $\bar{H}$. Представим его в следующем виде:

$$\bar{H} = e^q - q + e^p - p \cong \frac{q^2}{2} + \frac{p^2}{2} + \frac{q^3}{6} + \frac{p^3}{6}. \tag{2.1.6}$$

В правой части формулы (2.1.6) оставим полиномиальное приближение гамильтониана $\bar{H}$ с точностью до членов третьего порядка. Оно может быть представлено в виде следующей суммы:

$$\bar{H} = H_0 + \frac{q^3}{6} + \frac{p^3}{6} = H_0 + \varepsilon H_1. \tag{2.1.7}$$

Здесь H_0 – гамильтониан гармонического осциллятора, а H_1 – возмущающий гамильтониан.

Гамильтонианы H_0 и H_1 имеют вид

$$H_0 = \frac{q^2}{2} + \frac{p^2}{2}, \tag{2.1.8}$$

$$H_1 = q^3 + p^3. \tag{2.1.9}$$

Малый параметр ε определен следующим образом:

$$\varepsilon = \frac{1}{6}. \tag{2.1.10}$$

Перейдем к переменным типа действие–угол для гамильтониана $\bar{H}$. Поскольку гамильтониан $\bar{H}$ представлен в виде малого возмущения гармонического осциллятора (2.1.7), удобно перейти к переменным типа действие–угол в два этапа. На первом этапе будем использовать стандартные переменные типа действие–угол для гамильтониана гармонического осциллятора H_0. На втором этапе, в пп. 2.2, выполним переход от этих переменных к переменным типа действие–угол для гамильтониана $\bar{H}$.

Приступим к выполнению первого этапа и выполним переход от переменных p и q к переменным типа действие–угол для гармонического осциллятора H_0 (основной части гамильтониана $\bar{H}$)

$$(p,q) \to (I,\theta). \tag{2.1.11}$$

Старые переменные p и q выражаются через новые переменные I и θ по следующим формулам:

$$p = \sqrt{2I}\cos\theta, \tag{2.1.12}$$

$$q = \sqrt{2I}\sin\theta. \tag{2.1.13}$$

Гамильтониан гармонического осциллятора H_0 в новых переменных равен просто переменной действия I:

$$H_0 = \frac{q^2}{2} + \frac{p^2}{2} = I. \tag{2.1.14}$$

На этом преобразование гамильтониана гармонического осциллятора H_0 в новых переменных завершено. Тогда не зависящий от времени гамильтониан $\bar{H}$ примет вид

$$\bar{H} = H_0 + \varepsilon H_1 = I + \varepsilon 2^{3/2} I^{3/2}(\cos^3\theta + \sin^3\theta). \tag{2.1.15}$$

Гамильтониан $\bar{H}$ записан в промежуточных переменных I и θ. Они представляют собой переменные типа действие–угол для гамильтониана H_0, но не являются таковыми для гамильтониана $\bar{H}$. Поэтому дальнейшая задача заключается в построении переменных типа действие–угол для гамильтониана $\bar{H}$. Для этого в следующем разделе будет развит метод Гамильтона–Якоби, основанный на разложении в ряд теории возмущений по параметру ε.

2.2. Построение переменных типа действие–угол для не зависящего от времени гамильтониана $\bar{H}$ (2.1.15) с использованием метода Гамильтона–Якоби

Перейдем к построению переменных типа действие–угол для гамильтониана $\bar{H}$ (2.1.15)

$$(I,\theta) \to (J,\varphi). \tag{2.2.1}$$

Переход к переменным J, φ даст возможность преобразовать гамильтониан (2.1.5) к виду, который позволит применить метод Чирикова к исследуемой системе. Для перехода к этим переменным воспользуемся методом Гамильтона–Якоби, использующим производящую функцию канонических преобразований S. Она позволит выразить старые переменные через новые по следующим формулам:

$$\varphi = \frac{\partial S(J,\theta)}{\partial J}, \tag{2.2.2a}$$

$$I = \frac{\partial S(J,\theta)}{\partial \theta}. \tag{2.2.2б}$$

Далее, используя формулу (2.2.2а), можно будет свести процедуру по определению новых переменных к исследованию дифференциального уравнения первого порядка – к уравнению Гамильтона–Якоби. Уравнение Гамильтона–Якоби для производящей функции S с использованием гамильтониана (2.1.15) имеет следующий вид:

$$\bar{H}(I,\theta) = \bar{H}\left(\frac{\partial S}{\partial \theta},\theta\right) = K(J). \tag{2.2.3}$$

Здесь $K(J)$ – не зависящий от времени гамильтониан (2.1.15) в терминах новой переменной J.

Решить уравнение Гамильтона–Якоби в общем случае довольно сложно. Для его решения будем использовать метод возмущения. При этом используем наличие в гамильтониане (2.1.15) малого параметра ε. С учетом разложения гамильтонианов $\bar{H}$ и $K(J)$ в ряд по малому параметру ε уравнение (2.2.3) примет вид

$$\bar{H}\left(\frac{\partial S}{\partial \theta},\theta\right) = H_0\left(\frac{\partial S}{\partial \theta}\right) + \varepsilon H_1\left(\frac{\partial S}{\partial \theta},\theta\right) = \frac{\partial S}{\partial \theta} + \varepsilon 2^{3/2}\left(\frac{\partial S}{\partial \theta}\right)^{3/2}(\cos^3\theta + \sin^3\theta) = K_0(J) + \varepsilon K_1(J) + \ldots. \tag{2.2.4}$$

Для применения теории возмущений производящая функция S также разлагается в ряд по ε:

$$S = J\theta + \varepsilon S_1 + \varepsilon^2 S_2 + \ldots. \tag{2.2.5}$$

На первом этапе решения уравнения Гамильтона–Якоби приступим к определению первых трех членов $K_0(J)$, $K_1(J)$, $K_2(J)$ ряда теории возмущений для нового гамильтониана $K(J)$.

После подстановки формулы (2.2.5) в формулу (2.2.4) получим следующее уравнение:

$$J + \frac{\varepsilon\partial S_1}{\partial\theta} + \frac{\varepsilon^2\partial S_2}{\partial\theta} + \ldots + \varepsilon 2^{3/2}\left(J + \varepsilon\frac{\partial S_1}{\partial\theta} + \ldots\right)^{3/2}\left(\cos^3\theta + \sin^3\theta\right) = K_0(J) + \varepsilon K_1(J) + \ldots. \tag{2.2.6}$$

В нулевом порядке по ε из уравнения (2.2.6) следует уравнение для $K_0(J)$:

$$K_0(J) = J. \tag{2.2.7}$$

В первом порядке по ε из формулы (2.2.6) вытекает следующее уравнение для K_1:

$$\varepsilon 2\pi K_1(J) = \int_0^{2\pi} d\theta\left[\frac{\varepsilon\partial S_1}{\partial\theta} + \varepsilon 2^{3/2}J^{3/2}(\cos^3\theta + \sin^3\theta)\right] = 0. \tag{2.2.8}$$

Из формулы (2.2.8) видно, что $K_1 = 0$. Для производящей функции в первом порядке по ε из формулы (2.2.6) следует уравнение:

$$\varepsilon\frac{\partial S_1}{\partial\theta} = -\varepsilon 2^{3/2}J^{3/2}(\cos^3\theta + \sin^3\theta). \tag{2.2.9}$$

Интегрируя уравнение (2.2.9), получаем следующее выражение для функции S_1:

$$\begin{aligned} S_1 &= C - 2^{3/2}J^{3/2}\left(\frac{1}{3}\cos^3\theta - \cos\theta + \sin\theta - \frac{1}{3}\sin^3\theta\right) = \\ &= C - 2^{3/2}J^{3/2}\left(\frac{1}{12}\cos 3\theta - \frac{3}{4}\cos\theta + \frac{3}{4}\cos\theta + \frac{1}{12}\sin 3\theta\right). \end{aligned} \tag{2.2.10}$$

Во втором порядке по ε из уравнения (2.2.6) получим следующее уравнение для функции $K_2(J)$:

$$\begin{aligned} \varepsilon^2 2\pi K_2(J) &= \int_0^{2\pi} d\theta\left[\frac{\varepsilon^2\partial S_2}{\partial\theta} + \varepsilon^2 2^{3/2}\frac{3}{2}J^{1/2}\frac{\partial S_1}{\partial\theta}(\cos^3\theta + \sin^3\theta)\right] = \\ &= \varepsilon^2\int_0^{2\pi} d\theta 2^{3/2}\frac{3}{2}J^{1/2}(-1)2^{3/2}J^{3/2}(\cos^3\theta + \sin^3\theta)^2 = \\ &= -12\varepsilon^2 J^2\int_0^{2\pi} d\theta(\cos^3\theta + \sin^3\theta)^2 = -12\varepsilon^2 J^2\frac{5}{4}\pi = -15\varepsilon^2 J^2\pi. \end{aligned} \tag{2.2.11}$$

Величина $K_2(J)$ имеет вид

$$K_2(J) = -\frac{15}{2}J^2. \tag{2.2.12}$$

Таким образом, первый этап завершен. Получены первые три члена $K_0(J)$, $K_1(J)$, $K_2(J)$ ряда теории возмущений для нового гамильтониана $K(J)$. Кроме того, производящая функция канонических преобразований $S = J\theta + \varepsilon S_1$ найдена с точностью до первого порядка по ε.

На втором этапе используем производящую функцию канонических преобразований S_1 для перехода к новым переменным J и φ с точностью до первого порядка по ε. Уравнения (2.2.2) с точностью до первого порядка по ε, примут следующий вид:

$$I = \frac{\partial S}{\partial \theta} = J + \frac{\varepsilon \partial S_1}{\partial \theta} = J - \varepsilon J^{3/2} 2^{3/2} (\cos^3\theta + \sin^3\theta). \tag{2.2.13}$$

$$\varphi = \frac{\partial S}{\partial \theta} = \theta + \frac{\varepsilon \partial S_1}{\partial J} = \theta + \varepsilon(-2^{3/2})\frac{3}{2}J^{1/2}\left(\frac{1}{3}\cos^3\theta - \cos\theta + \sin\theta - \frac{1}{3}\sin^3\theta\right). \tag{2.2.14}$$

Для решения уравнений (2.2.13) и (2.2.14) разложим функцию θ в ряд по степеням ε:

$$\theta = \theta_0 + \varepsilon\theta_1 + \varepsilon^2\theta_2 + \ldots. \tag{2.2.15}$$

Тогда формула для переменой φ будет иметь вид

$$\varphi = \theta_0 + \varepsilon\theta_1 - \varepsilon 2^{3/2}\frac{3}{2}J^{1/2}\left(\frac{1}{3}\cos^3(\theta_0 + \varepsilon\theta_1) - \cos(\theta_0 + \varepsilon\theta_1) + \sin(\theta_0 + \varepsilon\theta_1) - \frac{1}{3}\sin^3(\theta_0 + \varepsilon\theta_1)\right). \tag{2.2.16}$$

В нулевом порядке по ε получим следующее уравнение:

$$\theta_0 = \varphi. \tag{2.2.17}$$

В первом порядке по ε получим уравнение

$$\theta_1 = 2^{3/2}\frac{3}{2}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right). \tag{2.2.18}$$

Объединив формулы (2.2.16) , (2.2.17) и (2.2.18), получим с точностью до первого порядка по ε выражение для переменной θ через переменные J и φ:

$$\theta = \varphi + \varepsilon 2^{3/2}\frac{3}{2}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^{-3}\varphi\right). \tag{2.2.19}$$

Таким образом, в данном подразделе построены переменные типа действие–угол для гамильтониана (2.1.15). Старые переменные I и θ выражены через новые переменные J и φ. Это позволит применить метод Чирикова к исследованию динамики исходной системы. В следующем разделе представим полный гамильтониан (2.1.15) в виде гамильтониана, представляющего собой сумму не зависящего от времени гамильтониана в переменных J и φ и отдельных резонансов.

2.3. Преобразование полного гамильтониана системы Лотки–Вольтерра с сезонностью (2.1.5) *в виде гамильтониана, представляющего собой сумму не зависящего от времени гамильтониана и отдельных резонансов*

Для того чтобы исследовать динамику двумерной системы Лотки–Вольтерра с сезонностью методом Чирикова, преобразуем полный гамильтониан таким образом, чтобы в его состав входили в качестве слагаемых отдельные резонансы. Для этого выполним следующие преобразования.

Вначале выразим полный гамильтониан (2.1.5) через не зависящий от времени гамильтониан $\bar{H}$ (2.1.15), записанный в переменных J и φ, а также через зависящий от времени дополнительный член, пропорциональный α и выраженный в переменных I и θ.

С использованием формул (2.2.7), (2.2.8), (2.2.12) гамильтониан будет иметь вид

$$H = \bar{H} + \alpha(q - p)\cos\Omega t = J - \varepsilon^2\frac{15}{2}J^2 + \alpha\cos\Omega t\sqrt{2I}(\sin\theta - \cos\theta). \tag{2.3.1}$$

В зависящем от времени члене формулы (2.3.1) выразим сомножитель, содержащий I через переменные J и θ. Получим следующую формулу:

$$\sqrt{2I} = 2^{1/2}(J - \varepsilon J^{3/2}2^{3/2}(\cos^3\theta + \sin^3\theta))^{1/2} = 2^{1/2}J^{1/2}\left(1 - \frac{1}{2}\varepsilon J^{1/2}2^{3/2}(\cos^3\theta + \sin^3\theta)\right). \tag{2.3.2}$$

Осталось выразить переменную θ через переменные J и φ. Как следует из формулы (2.2.19), выражения для $\sin\theta$ и $\cos\theta$ будут иметь следующий вид:

$$\begin{aligned}\sin\theta &= \sin\left(\varphi + \varepsilon 2^{3/2}\frac{2}{3}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\right) = \\ &= \sin\varphi\cos\left(\varepsilon 2^{3/2}\frac{2}{3}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\right) + \\ &+ \cos\varphi\sin\left(\varepsilon 2^{3/2}\frac{2}{3}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\right) = \\ &= \sin\varphi + \varepsilon 2^{3/2}\frac{2}{3}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\cos\varphi.\end{aligned} \tag{2.3.3}$$

$$\begin{aligned}\cos\theta &= \cos\left(\varphi + \varepsilon^{3/2}\frac{2}{3}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\right) = \\ &= \cos\varphi - \sin\varphi\varepsilon 2^{3/2}\frac{2}{3}J^{1/2}\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right).\end{aligned} \tag{2.3.4}$$

Используя формулы (2.3.2), (2.3.3) и (2.3.4), выразим гамильтониан, представленный в формуле (2.3.1) через переменные J и φ:

$$\begin{aligned}H = J - \varepsilon^2\frac{15}{2}J^2 + \alpha\cos\Omega t\Big(&\sin\varphi - \cos\varphi + \varepsilon 2^{3/2}\frac{2}{3}J^{1/2}\times \\ &\times\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\cos\varphi + \\ + \varepsilon 2^{3/2}\frac{2}{3}J^{1/2}&\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right)\sin\varphi\Big)\times \\ \times 2^{1/2}&\left(J^{1/2} - \frac{1}{2}\varepsilon J 2^{3/2}\left(\cos^3\varphi + \sin^3\varphi\right)\right) = J - \varepsilon^2\frac{15}{2}J^2 + \\ + \alpha\cos\Omega t 2^{1/2}J^{1/2}(\sin\varphi - \cos\varphi) &+ \varepsilon\alpha\cos\Omega t 2^{1/2}J^{1/2}2^{3/2}\frac{2}{3}J^{1/2}\times \\ \times(\cos\varphi + \sin\varphi)&\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right) + \\ + \varepsilon\alpha\cos\Omega t 2^{1/2}&\left(-\frac{1}{2}\right)J 2^{3/2}(\cos^3\varphi + \sin^3\varphi)(\sin\varphi - \cos\varphi) + \ldots .\end{aligned} \tag{2.3.5}$$

Преобразуем формулу (2.3.5) к более удобному виду. Для удобства работы введем коэффициенты A, B и C:

$$A = \alpha\cos\Omega t 2^{1/2}J^{1/2}, \tag{2.3.6}$$

$$B = \varepsilon\cos\Omega t 2^{1/2}\frac{8}{3}J, \tag{2.3.7}$$

$$C = -2\varepsilon\cos\Omega t J. \tag{2.3.8}$$

После введения указанных коэффициентов формула (2.3.5) примет вид

$$\begin{aligned}H = J - \varepsilon^2\frac{15}{2}J^2 &+ A(\sin\varphi - \cos\varphi) + \\ + B(\cos\varphi + \sin\varphi)&\left(\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi\right) + \\ + C(\cos^3\varphi + \sin^3\varphi)&(\sin\varphi - \cos\varphi) + \ldots .\end{aligned} \tag{2.3.9}$$

Формулу (2.3.9) можно упростить, если произвести сдвиг фазы φ, выполнив следующую замену переменных:

$$\varphi \to \varphi - \frac{\pi}{4}. \tag{2.3.10}$$

После сдвига фазы члены формулы (2.3.9), содержащие $\sin\varphi$, $\sin^3\varphi$, $\cos\varphi$, $\cos^3\varphi$ примут следующий вид:

$$\sin\varphi - \cos\varphi \to -\sqrt{2}\cos\left(\varphi + \frac{\pi}{4}\right), \tag{2.3.11}$$

$$\cos\varphi + \sin\varphi \to \sqrt{2}\sin\left(\varphi + \frac{\pi}{4}\right), \tag{2.3.12}$$

$$\begin{aligned}\frac{1}{3}\cos^3\varphi - \cos\varphi + \sin\varphi - \frac{1}{3}\sin^3\varphi = \frac{1}{12}\cos 3\varphi - \frac{3}{4}\cos\varphi + \frac{1}{12}\sin 3\varphi + \frac{3}{4}\sin\varphi = \\ = \frac{3}{4}(-\sqrt{2})\cos\left(\varphi + \frac{\pi}{4}\right) + \frac{1}{12}\sqrt{2}\sin\left(3\varphi + \frac{\pi}{4}\right) = \frac{3}{4}(-\sqrt{2})\cos\left(\varphi + \frac{\pi}{4}\right) + \frac{1}{12}\sqrt{2}\sin\left(3\varphi + \frac{3}{4}\pi - \frac{\pi}{2}\right) = \\ = \frac{3}{4}\left(-\sqrt{2}\right)\cos\left(\varphi + \frac{\pi}{4}\right) + \frac{1}{12}\sqrt{2}\cos\left(3\left(\varphi + \frac{\pi}{4}\right)\right),\end{aligned} \tag{2.3.13}$$

$$\begin{aligned}\cos^3\varphi + \sin^3\varphi = \frac{3}{4}\sin\varphi - \frac{1}{4}\sin 3\varphi + \frac{3}{4}\cos\varphi + \frac{1}{4}\cos 3\varphi = \\ = \frac{3}{4}\sqrt{2}\sin\left(\varphi + \frac{\pi}{4}\right) + \frac{1}{4}\sqrt{2}\cos\left(3\varphi + \frac{\pi}{4}\right) = \frac{3}{4}\sqrt{2}\sin\left(\varphi + \frac{\pi}{4}\right) + \frac{1}{4}\sqrt{2}\sin\left(3\varphi + \frac{\pi}{4}\right).\end{aligned} \tag{2.3.14}$$

После сдвига фазы формулу (2.3.9) можно представить в следующем виде:

$$\begin{aligned}H = J - \varepsilon^2\frac{15}{2}J^2 + \tilde{A}\cos\varphi + \tilde{B}_1\sin\varphi\cos\varphi + \\ + \tilde{B}_2\sin\varphi\cos 3\varphi + \tilde{C}_1\sin\varphi\cos\varphi + \tilde{C}_2\cos\varphi\sin 3\varphi.\end{aligned} \tag{2.3.15}$$

В формуле (2.3.15) коэффициенты $\tilde{A}$, $\tilde{B}_1$, $\tilde{B}_2$, $\tilde{C}_1$, $\tilde{C}_2$ пропорциональны функции $\cos\Omega t$. После выполненных выше преобразований члены, содержащие эти коэффициенты, очень похожи на резонансы. Разница состоит в том, что некоторые из этих членов содержат произведения тригонометрических функций вместо самих тригонометрических функций. Поэтому далее выполним их замену на одиночные тригонометрические функции. Коэффициенты $\tilde{A}$, $\tilde{B}_1$, $\tilde{B}_2$, $\tilde{C}_1$, $\tilde{C}_2$ в формуле (2.3.15) имеют следующий вид:

$$\tilde{A} = \left(-\sqrt{2}\right)A = -2\alpha J^{1/2}\cos\Omega t, \tag{2.3.16}$$

$$\tilde{B}_1 = \frac{3}{2}B = -\varepsilon\alpha\cos\Omega t 4J, \tag{2.3.17}$$

$$\tilde{B}_2 = \frac{1}{6}B = \varepsilon\alpha J^{1/2}\cos\Omega t\frac{4}{9}J, \tag{2.3.18}$$

$$\tilde{C}_1 = \frac{3}{2}C = 3\varepsilon\alpha\cos\Omega t J, \tag{2.3.19}$$

$$\tilde{C}_2 = -\frac{1}{2}C = \varepsilon\alpha\cos\Omega t J. \tag{2.3.20}$$

Произведем дальнейшее преобразование тригонометрических функций по следующим формулам:

$$\sin\varphi\cos\varphi = \frac{1}{2}\sin 2\varphi, \tag{2.3.21}$$

$$\sin\varphi\cos 3\varphi = \frac{1}{2}(\sin(\varphi + 3\varphi) + \sin(\varphi - 3\varphi)), \tag{2.3.22}$$

$$\sin 3\varphi\cos\varphi = \frac{1}{2}(\sin(\varphi + 3\varphi) + \sin(3\varphi - \varphi)). \tag{2.3.23}$$

Воспользовавшись формулами (2.3.21)–(2.3.23), приведем выражение (2.3.15) к следующему виду:

$$H = J - \varepsilon^2\frac{15}{2}J^2 + \tilde{A}\cos\varphi + \left[\left(\frac{\tilde{B}_1 + \tilde{C}_1}{2}\right) - \frac{\tilde{B}_2}{2} + \frac{\tilde{C}_2}{3}\right]\sin 2\varphi + \left[\frac{\tilde{B}_2}{2} + \frac{\tilde{C}_2}{2}\right]\sin 4\varphi. \tag{2.3.24}$$

Подставив коэффициенты из формул (2.3.16)–(2.3.20) в формулу (2.3.24), получим окончательную формулу для полного гамильтониана с тремя выделенными резонансами

$$H = J - \varepsilon^2\frac{15}{2}J^2 - 2\alpha J^{1/2}\cos\Omega t\cos\varphi - \frac{2}{9}\varepsilon\alpha J\cos\Omega t\sin 2\varphi + \frac{13}{2}\varepsilon\alpha J\cos\Omega t\sin 4\varphi. \tag{2.3.25}$$

Формула (2.3.25) содержит резонансные члены, выраженные в виде произведений тригонометрических функций, содержащих параметры Ωt и φ. Выполним последнее преобразование и разложим каждое их этих произведений в виде суммы из двух слагаемых. Окончательная формула преобразованного гамильтониана с выделенными резонансами будет иметь вид

$$\begin{aligned} H = J - \varepsilon^2\frac{15}{2}J^2 - \alpha J^{1/2}[\cos(\varphi - \Omega t) + \cos(\varphi + \Omega t)] - \varepsilon\alpha\left(\frac{1}{9}\right)J[\sin(2\varphi - \Omega t) + \sin(2\varphi + \Omega t)] + \\ + \varepsilon\alpha\left(\frac{13}{36}\right)J\left[\sin(4\varphi - \Omega t) + \sin(4\varphi + \Omega t)\right]. \end{aligned} \tag{2.3.26}$$

Таким образом, после выполненных выше преобразований полный гамильтониан системы Лотки–Вольтерра с сезонностью (2.1.5) удалось записать приближенно в виде гамильтониана (2.3.26), содержащего три резонанса: $\cos(\varphi - \Omega t)$, $\sin(2\varphi - \Omega t)$, $\sin(4\varphi - \Omega t)$, с помощью которых в следующем разделе будет исследована динамика исходной системы.

3. АНАЛИТИЧЕСКОЕ ИССЛЕДОВАНИЕ ДИНАМИКИ ИСХОДНОЙ СИСТЕМЫ С ПОМОЩЬЮ МЕТОДА ПЕРЕКРЫТИЯ РЕЗОНАНСОВ ЧИРИКОВА

Используемый в работе метод Чирикова основан на выделении из множества резонансов исследуемого гамильтониана нескольких наиболее значимых резонансов. Для них определяется момент, когда эти резонансы начинают перекрываться и взаимодействовать. При этом каждый выделенный резонанс в отдельности описывается гамильтонианом нелинейного маятника. С помощью таких гамильтонианов для каждого резонанса определяется его частотная ширина. Выделенные резонансы разделены определенным частотным интервалом. Каждый из них занимает область на оси частот, определяемую его шириной. Резонансы начинают взаимодействовать, когда эти области перекрываются. Это происходит, когда сумма полуширин этих резонансов становится равной частотному расстоянию между резонансами. Количественная оценка взаимодействия резонансов определяется с помощью критерия, определяющего степень перекрытия резонансов. Критерий был предложен Чириковым для характеристики динамики гамильтоновых систем [5].

Задача исследования, выполняемая в этом разделе, состоит в том, чтобы выяснить, при какой амплитуде сезонности в исходной системе Лотки–Вольтерра происходит переход к хаотическому поведению. Динамика системы будет исследоваться поэтапно, с использованием трех резонансов полного гамильтониана системы (2.3.26). Каждый из трех резонансов гамильтониана (2.3.26) описывается с помощью гамильтониана нелинейного маятника. Это позволяет оценить размеры резонансов как по оси переменой действия J, так и по оси частот ω. Зная размеры резонансов, можно определить амплитуду сезонности, при которой возникает хаотическое поведение в системе.

Исследование выполняется в три этапа.

На первом этапе, для удобства исследования, гамильтониан (2.3.26) представлен в общем виде (3.1), (3.2). С помощью уравнений динамики для данного гамильтониана выделяется один из трех резонансов. Он преобразовывается в гамильтониан хорошо изученного нелинейного осциллятора, для которого по известным формулам можно определить ширину резонанса. Остальные резонансы исследуются аналогичным образом.

На втором этапе исследования оцениваются размеры резонансов с помощью полученного гамильтониана нелинейного осциллятора.

На третьем этапе, после получения размеров выделенных резонансов, определяются условия, при которых полуширины резонансов перекрывают расстояния между ними. В этот момент происходит перекрытие резонансов, которое приводит к возникновению хаотического поведения в системе. С помощью критерия Чирикова определяется конкретное значение амплитуды сезонности, при котором исследуемая система переходит к хаосу.

Перейдем к реализации первого этапа. Представим гамильтониан (2.3.26) в общем виде:

$$H = \bar{H}(J) + \alpha V(J, \varphi, t). \tag{3.1}$$

Здесь зависящее от времени слагаемое представлено следующим образом:

$$V(J, \varphi, t) = \sum_{k,l} V_{k,l}(J) \cos(k\varphi - l\Omega t + \varphi_{k,l}). \tag{3.2}$$

Гамильтониан (3.1), (3.2) порождает уравнения движения, которые описывают динамику взаимодействующих резонансов. Уравнения динамики для данного гамильтониана имеют вид

$$\frac{dJ}{dt} = -\alpha \frac{\partial V}{\partial \varphi} = \alpha \sum_{k,l} k V_{k,l}(J) \sin(k\varphi - l\Omega t + \varphi_{k,l}), \tag{3.3}$$

$$\frac{d\varphi}{dt} = -\alpha \frac{\partial H}{\partial J} = \frac{d\bar{H}(J)}{dJ} + \alpha \frac{\partial \bar{H}(J, \varphi, t)}{\partial J} = w(J) + \alpha \sum_{k,l} \frac{\partial V}{\partial J} \cos(k\varphi - l\Omega t + \varphi_{k,l}). \tag{3.4}$$

Вначале выделим отдельные резонансы из множества всех резонансов. До тех пор пока частотные полуширины резонансов не перекрывают частотного расстояния между ними, такое выделение можно осуществить. При достаточно малом параметре α в возмущающем члене динамика системы представляет собой динамику невзаимодействующих резонансов. Резонансы возникают, когда одна из фаз, заданных парой k, l в уравнениях (3.3), (3.4), начинает меняться медленно. При этом другие фазы меняются значительно быстрее, а их средний вклад незначителен. Исходя из этих соображений, можно выделить из множества резонансов, влияющих на динамику системы (3.3), (3.4), один главный.

Приступим к выделению отдельного резонанса.

Резонанс проявляется наиболее сильно, когда соответствующая фаза перестает меняться. При этом ее производная по времени становится равной нулю:

$$k\omega - l\Omega = 0. \tag{3.5}$$

Здесь ω – собственная частота системы, Ω – частота внешнего воздействия, k и l – целые числа, определяющие кратности вхождения частот в фазу.

Когда выполняется условие резонанса (3.5), соответствующая фаза в уравнениях (3.3), (3.4) становится медленнее по сравнению с остальными и вносит главный вклад в динамику. Поэтому оставим в уравнениях (3.3), (3.4) только один определенный резонанс. Ему будет соответствовать действие J_0, удовлетворяющее уравнению

$$k_0\omega(J_0) - l_0\Omega = 0. \tag{3.6}$$

Переобозначим входящие в формулы (3.3), (3.4) коэффициенты следующим образом:

$$V_{k_0,l_0} = V_0. \tag{3.7}$$

Тогда упрощенная система уравнений (3.3), (3.4) примет следующий вид:

$$\frac{dJ}{dt} = \alpha k_0 V_0 \sin(k_0\varphi - l_0\Omega t + \varphi), \tag{3.8}$$

$$\frac{d\phi}{dt} = \omega(J) + \alpha \frac{\partial V_0}{\partial J} \cos(k_0\varphi - l_0\Omega t + \varphi). \tag{3.9}$$

Система (3.8), (3.9) представляет собой редукцию системы (3.3), (3.4), выполненную для описания выделенного резонанса, заданного парой k_0, l_0. Для упрощения этой системы вначале введем переменную

$$\Delta J = J - J_0. \tag{3.10}$$

Эта переменная будет представлять собой одну из пары канонически сопряженных переменных, в которых можно будет выразить динамику системы (3.8), (3.9). Как будет показано ниже (фор-

мула (3.16)), другой канонически сопряженной переменной будет фаза, от которой зависят тригонометрические функции в системе (3.8), (3.9).

Далее выполним следующие шаги аппроксимации.

1. В правых частях системы (3.8), (3.9) коэффициент V_0 заменим на константу $V_0 = V_0(J_0)$. Это приближение позволит избавиться от второго слагаемого в правой части формулы (3.9). Формула (3.9) примет следующий вид:

$$\frac{d\varphi}{dt} = \omega(J). \tag{3.11}$$

2. Оставшееся первое слагаемое в формуле (3.11) допускает следующее преобразование:

$$\omega(J) = \omega_0 + \omega'\Delta J, \tag{3.12}$$

где

$$\omega_0 = \omega(J_0), \quad \omega' = \frac{d\omega(J_0)}{dJ}. \tag{3.13}$$

После завершения аппроксимации подставим формулы (3.12), (3.13) в формулу (3.11) и получим формулу

$$\frac{d\varphi}{dt} = \omega_0 + \omega'\Delta J. \tag{3.14}$$

Формула (3.14) допускает упрощения. Умножим уравнения (3.14) на константу k_0. Получим следующее уравнение:

$$\frac{dk_0\varphi}{dt} = k_0\omega_0 + k_0\omega'\Delta J = l_0\Omega + k_0\omega'\Delta J = \frac{d}{dt}l_0\Omega t + k_0\omega'\Delta J. \tag{3.15}$$

Перенесем слагаемое $\frac{d}{dt}l_0\Omega t$ в левую часть формулы (3.15). В итоге в левой части появится производная по времени от новой фазы ψ:

$$\frac{d}{dt}(k_0\varphi - l_0\Omega t + \pi) = \frac{d}{dt}\psi = k_0\omega'\Delta J. \tag{3.16}$$

Здесь ψ – новая фаза:

$$\psi = k_0\varphi - l_0\Omega t + \varphi - \pi. \tag{3.17}$$

В результате выполненных упрощений и преобразований система (3.8), (3.9) приводится к следующему виду:

$$\frac{d}{dt}(\Delta J) = -\alpha k_0 V_0 \sin\psi, \tag{3.18}$$

$$\frac{d}{dt}\psi = k_0\omega'\Delta J. \tag{3.19}$$

Представим уравнения (3.18), (3.19) в гамильтоновой форме

$$\frac{d}{dt}(\Delta J) = \frac{\partial\tilde{H}}{\partial\psi}, \quad \frac{d}{dt}\psi = \frac{\partial\tilde{H}}{\partial(\Delta J)}, \tag{3.20}$$

где гамильтониан $\tilde{H}$ имеет вид

$$\tilde{H} = \frac{1}{2}k_0\omega'(\Delta J)^2 - \varepsilon k_0 V_0 \cos\psi. \tag{3.21}$$

Гамильтониан (3.21) представляет собой гамильтониан нелинейного маятника. Он описывается уравнением

$$\frac{d^2\psi}{dt^2} + \Omega_0^2 \sin\psi = 0, \tag{3.22}$$

где частота колебаний маятника имеет вид

$$\Omega_0 = (\alpha k_0^2 V_0 |\omega'|)^{1/2}. \tag{3.23}$$

Таким образом, первый этап завершен. Динамику одного выделенного резонанса удалось описать в терминах колебаний нелинейного маятника. Аналогичным образом описывается динамика остальных резонансов.

Перейдем ко второму этапу исследования и определим ширину выделенного резонанса. Для определения его ширины используем формулу для сепаратрисы нелинейного маятника. Сепаратриса нелинейного маятника, заданного гамильтонианом (3.21), определяется уравнением

$$\Delta J_s = \pm\sqrt{\frac{4\alpha V_0}{\omega'}}\cos\frac{\psi}{2}. \tag{3.24}$$

С помощью формулы (3.24) определим полуширину резонанса в терминах переменной действия ΔJ:

$$\Delta J = 2\sqrt{\frac{\alpha V_0}{\omega'}}. \tag{3.25}$$

Далее, через полуширину в терминах переменной действия, выразим частотную полуширину выделенного резонанса

$$\Delta\omega = \omega'\Delta J = 2\sqrt{\alpha\omega' V_0}. \tag{3.26}$$

На втором этапе, с помощью полученного гамильтониана нелинейного осциллятора, определены размеры выделенного резонанса. Таким же образом определяются размеры остальных резонансов.

После получения размеров выделенных резонансов перейдем к третьему этапу и исследуем систему с учетом взаимодействия этих резонансов. Рассмотрим пару, образованную первым и вторым резонансом. Формулой (3.26) определены их полуширины. Располагая двумя полуширинами этих резонансов, критерий Чирикова можно выразить в виде

$$(\Delta\omega)_1 + (\Delta\omega)_2 = \Delta\Omega. \tag{3.27}$$

Здесь $(\Delta\omega)_1$ – полуширина первого резонанса, $(\Delta\omega)_2$ – полуширина второго резонанса, $\Delta\Omega$ – расстояние между резонансами.

Согласно критерию Чирикова появление хаотического поведения следует ожидать, когда сумма полуширин соседних резонансов становится равной расстоянию между резонансами. Приведенные выше формулы (3.26) и (3.27) справедливы для любого резонанса и любой пары резонансов. Располагая этими формулами, можно перейти к исследованию конкретных резонансов, заданных формулой (2.3.26). Из формулы (2.3.26) следует конкретное выражение для частоты $\omega(J)$ исследуемой системы Лотки–Вольтерра с сезонностью:

$$\omega(J) = 1 - 15\varepsilon^2 J. \tag{3.28}$$

Теперь с учетом формулы (3.28) формула (3.6) для первого, второго и третьего резонанса примет вид

$$1 - 15\varepsilon^2 J_1 = \Omega, \tag{3.29}$$

$$1 - 15\varepsilon^2 J_2 = \frac{\Omega}{2}, \tag{3.30}$$

$$1 - 15\varepsilon^2 J_3 = \frac{\Omega}{4}. \tag{3.31}$$

Первая производная от частоты по параметру действия одинакова для всех трех резонансов и равна

$$\omega' = -15\varepsilon^2. \tag{3.32}$$

Осталось определить входящий в формулу (3.26) параметр V_0. Значения параметра V_0 для первого, второго и третьего резонанса равны

$$V_{01} = J_1^{1/2} = \left(\frac{1-\Omega}{15\varepsilon^2}\right)^{1/2} = \left(\frac{12}{5}(1-\Omega)\right)^{1/2}, \tag{3.33}$$

$$V_{02} = \frac{1}{9}\varepsilon J_2 = \frac{\varepsilon}{9}\frac{\left(1-\frac{\Omega}{2}\right)}{15\varepsilon^2} = \frac{2}{45}\left(1-\frac{\Omega}{2}\right), \tag{3.34}$$

$$V_{03} = \frac{13}{36}\varepsilon J_3 = \frac{13}{36}\varepsilon\frac{\left(1-\frac{\Omega}{4}\right)}{15\varepsilon^2} = \frac{13}{90}\left(1-\frac{\Omega}{4}\right). \tag{3.35}$$

Подставим формулы (3.32), (3.33), (3.34), (3.35) в формулы (3.26), (3.27). В итоге получим формулы для определения частотных полуширин для первого, второго и третьего резонансов

$$(\Delta\omega)_1 = 2\alpha^{1/2}\sqrt{\frac{5}{12}}\left(\frac{12}{15}(1-\Omega)\right)^{1/4}, \tag{3.36}$$

$$(\Delta\omega)_2 = 2\alpha^{1/2}\sqrt{\frac{5}{12}}\left(\frac{2}{45}\left(1-\frac{\Omega}{2}\right)\right)^{1/2}, \tag{3.37}$$

$$(\Delta\omega)_3 = 2\alpha^{1/2}\sqrt{\frac{5}{12}}\left(\frac{13}{90}\left(1-\frac{\Omega}{4}\right)\right)^{1/2}. \tag{3.38}$$

Располагая конкретными полуширинами этих резонансов, можно использовать критерий Чирикова для определения условий перехода системы к хаосу. Когда частотные полуширины известны, критерий Чирикова формулируется следующим образом:

$$K_{ch} = \frac{(\Delta\omega)_1 + (\Delta\omega)_2}{\Delta\Omega} \geq 1. \tag{3.39}$$

Если умножить левую и правую части формулы (3.39) на $\Delta\Omega$, то получим более удобную для дальнейших расчетов форму критерия Чирикова – формулу (3.27). Подставив в формулу (3.27) частотные полуширины соседних резонансов из формул (3.36), (3.38), получим окончательные формулы критерия Чирикова, из которых можно определить критическую амплитуду сезонности α, при которой в исследуемой системе происходит переход к хаотическому поведению

$$(\Delta\omega)_1 + (\Delta\omega)_2 = 2\alpha^{1/2}\sqrt{\frac{5}{12}}\left[\left(\frac{12}{5}(1-\Omega)\right)^{1/4} + \left(\frac{2}{45}\left(1-\frac{\Omega}{2}\right)\right)^{1/2}\right] = \frac{\Omega}{2}, \tag{3.40}$$

$$(\Delta\omega)_2 + (\Delta\omega)_3 = 2\alpha^{1/2}\sqrt{\frac{5}{12}}\left[\left(\frac{2}{45}\left(1-\frac{\Omega}{2}\right)\right)^{1/2} + \left(\frac{13}{90}\left(1-\frac{\Omega}{4}\right)\right)^{1/2}\right] = \frac{\Omega}{4}. \tag{3.41}$$

Формула (3.40) определяет критерий Чирикова для первого и второго резонансов, формула (3.41) – для второго и третьего резонансов. Для того чтобы оценить порядок величины $\alpha_{1,2}$, при котором перекрываются первый и второй резонансы и в исследуемой системе происходит переход к хаосу, положим для определенности $\Omega = \frac{1}{2}$. Тогда из формулы (3.40) следует, что $\alpha_{1,2} = 0.024$. Второй расчет проведем для второго и третьего резонансов, приняв значение $\Omega = \frac{3}{2}$. В этом случае из формулы (3.41) следует, что значение $\alpha_{2,3}$, при котором перекрываются второй и третий резонансы и происходит переход системы к хаосу, составляет $\alpha_{2,3} = 0.515$. С помощью критерия Чирикова определено значение критической амплитуды сезонности α, при которой в результате взаимодействия выделенных резонансов (первого и второго, второго и третьего) происходит переход к хаотической динамике в системе Лотки–Вольтерра с сезонностью. При этом размеры резонансов прямо пропорциональны $\alpha^{1/2}$. Таким образом, при наличии периодического возмущения (в данном случае сезонности), при определенном параметре амплитуды сезонности, в исходной системе с двумя зависимыми переменными возникает хаотическое поведение.

4. ЗАКЛЮЧЕНИЕ

В работе исследована динамика системы Лотки–Вольтерра с сезонностью с применением эффективного и достаточно простого метода перекрытия резонансов Чирикова [5]. Каких-либо сложностей с использованием этого метода для выявления хаотического поведения в исходной системе не возникало. Метод Чирикова широко известен, хорошо изучен и позволяет с помощью параметра K, определяющего степень перекрытия резонансов, охарактеризовать динамику гамильтоновых систем. В соответствии с теорией Чирикова, при малом взаимодействии резонансов и $K \ll 1$ динамика системы характеризуется как регулярная. При сильном перекрытии резонансов и $K \geq 1$ динамика системы оказывается очень сложной, в ней возникает нерегулярное движение, сопровождающееся переходом системы к хаосу.

Основная сложность исследования заключалась в поэтапном преобразовании системы Лотки–Вольтерра с сезонностью таким образом, чтобы к ней можно было применить метод Чирикова. Необходимое преобразование выполнено в три этапа в пп. 2.1, 2.2 и 2.3. В результате преобразований удалось выразить гамильтониан исходной системы в виде суммы не зависящего от времени гамильтониана в переменных действие–угол и нескольких резонансов. Выполненное в разд. 3 исследование взаимодействия выделенных резонансов с использованием метода Чирикова позволило определить аналитический критерий возникновения хаотического поведения в исследуемой системе. Аналитический критерий основан на определении критических величин амплитуд сезонности, при которых в исследуемой системе Лотки–Вольтерра с сезонностью возникает хаотическое поведение.

СПИСОК ЛИТЕРАТУРЫ

1. *Volterra V.* Fluctuations in the abundance of a species considered mathematically // Nature. 1926. V. 118. P. 558–560. https://doi.org/10.1038/118558a0
2. *Volterra V.* Variazioni e fluttuazioni dei numero d'individui in specie animali conviventi; Memorie della Regia Accademia Nazionale dei Lincei. 1926. V. 2. P. 31–113. English translation in Chapman, R.N., Animal Ecology, New York: McGraw–Hill, 1931.
3. *Volterra V.* Lessons on the mathematical theory of struggle for life. (Original: Leçons sur la théorie mathématique de la lutte pour la vie). Paris: Gauthier-Villars, 1931. P. 214.
4. *Lotka A.J.* Elements of physical biology. Baltimore: Williams and Wilkins, 1925. P. 495.
5. *Chirikov B.V.* A universal instability of many-dimensional oscillator systems // Phys. Rep. 1979. V. 52. P. 263–379.
6. *Poincaré H.* Les méthodes nouvelles de la mécanique céleste. V. 1–3, Paris: Gauthier-Villars, 1892. P. 412.
7. *Poincaré H.* Leçons de mécanique céleste. Paris: Gauthier-Villars, 1905. P. 365.
8. *Hadamard J.S.* Sur le billard non-Euclidien // Soc. Sci. Bordeaux, Procès Verbaux 1898. 147 p.
9. *Lorenz E.N.* Deterministic nonperiodic flow // Journal of the Atmospheric Sciences. 1963.V. 20. № 2. P. 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2
10. *Lorenz E.N.* Predictability: Does the flap of a butterfly's wings in Brazil set off a Tornado in Texas? // American Association for the Advancement of Science. 1972. http://gymportalen.dk/sites/lru.dk/files/lru/132_kap6_lorenz_artikel_the_butterfly_effect.pdf
11. *Feigenbaum M.J.* Quantitative universality for a class of nonlinear transformations // J. Stat. Phys. 1978. V. 19. № 1. P. 25–52. https://doi.org/10.1007/BF01020332
12. *Feigenbaum M.J.* The universal metric properties of nonlinear transformations // J. Stat. Phys. 1979. V. 21. № 6. P. 669–706. https://doi.org/10.1007/BF01107909
13. *Feigenbaum M.J., Greene J.M., MacKay R.S., Vivaldi F.* Universal behaviour in families of area-preserving maps // Physica D: Nonlinear Phenomena. 1981. V. 3. № 3. P. 468–486. https://doi.org/10.1016/0167-2789(81)90034-8
14. *Mandelbrot B.B.* The Fractal Geometry of Nature. New York: W. H. Freeman, 1982. P. 468.
15. *Ott E., Grebogi C., Yorke J.A.* Controlling chaos // Phys. Rev. Lett. 1990. V. 64. № 11. P. 1196–1199. https://doi.org/10.1103/PhysRevLett.64.1196
16. *Pecora L.M., Carroll T.L.* Synchronization in chaotic systems // Phys. Rev. Lett. 1990. V. 64. № 8. P. 821–824. https://doi.org/10.1103/PhysRevLett.64.821

17. *Fishman M.P., Egolf D.A.* Revealing the building block of spatiotemporal chaos: deviations from extensivity // Phys. Rev. Lett. 2006. V. 96. № 5. Article ID 054103. P. 4.
https://doi.org/10.1103/PhysRevLett.96.054103

18. *Spiegel D.R., Johnson E.R.* Experimental investigation of the transition to spatiotemporal chaos with a system-size control parameter // Research Letters in Physics. V. 2008. Article ID 891324. P. 5.
https://doi.org/10.1155/2008/891324

19. *Benincà E., Ballantine B., Ellner S.P., Huisman J.* Species fluctuations sustained by a cyclic succession at the edge of chaos // PNAS (Proceeding of the National Academy of Sciences of the United States of America). 2015. V. 112. № 20. P. 6389–6394.
https://doi.org/10.1073/pnas.1421968112

20. *Clements C.F., Ozgul A.* Indicators of transitions in biological systems // Ecol. Lett. 2018. V. 21. № 6. P. 905–919.
https://doi.org/10.1111/ele.12948

21. *Scheffer M., Bascompte J., Brock W. et al.* Early-warning signals for critical transitions // Nature. 2009. V. 461. P. 53–59.
https://doi.org/10.1038/nature08227

22. *Gopalakrishnan E. et al.* Early warning signals for critical transitions a thermoacoustic system // Sci. Rep. 2016. V.6. Article number: 35310. P. 1-10.
https://doi.org/10.1038/srep35310

23. *Boerlijst M.C., Oudman T., de Roos A.M.* Catastrophic collapse can occur without early warning: examples of silent catastrophes in structured ecological models // PLOS ONE. 2013. V. 8. № 4. P. 1–6.
https://doi.org/10.1371/journal.pone.0062033

24. *Peacock T., Haller G.* Lagrangian coherent structures. The hidden skeleton of fluid flows. // Physics Today. 2013. P. 41.

25. *Mathur M., Haller G., Peacock T., Ruppert-Felsot J.E., Swinney H.L.* Uncovering the Lagrangian skeleton of turbulence // Phys. Rev. Lett. 2007. V. 98. № 14. P. 144502-1–144502-4.
https://doi.org/10.1103/PhysRevLett.98.144502

26. *Falessi M.V., Pegoraro F., Schep T.J.* Lagrangian coherent structures and plasma transport processes // J. of Plasma Physics. 2015. V. 81. № 5.
https://doi.org/10.1017/S0022377815000690

27. *Kai E.T., Rossi V., Sudre J., Weimerskirch H., Lopez C., Hernandez-Garcia E., Marsac F., Garcon V.* Top marine predators track Lagrangian coherent structures // PNAS (Proceeding of the National Academy of Sciences of the United States of America). 2009. V. 106. № 20. P. 8245–8250.

28. *Zhu H., Chapman S.C., Dendy R.O.* Robustness of predator-prey models for confinement regime transitions in fusion plasmas // Phys. Plasmas. 2013. V. 20. № 4. P. 042302-1–042302-11.
https://doi.org/10.1063/1.4800009

29. *Aydiner E.* Chaotic universe model: Lotka–Volterra dynamics of the universe evolution // arXiv: 1610.07338v3 [gr-qc]. 2017. P. 9.

30. *Gatabazi P., Mba J.C., Pindza E.* Fractional gray Lotka–Volterra models with application to cryptocurrencies adoption // Chaos. 2019. V. 29. № 7. P. 10.
https://doi.org/10.1063/1.5096836

31. *Arneodo A., Coullet P., Peyraud J. et al.* Strange attractors in Volterra equations for species in competition // J. Math. Biology. 1982. V. 14. № 2. P. 153–157.
https://doi.org/10.1007/BF01832841

32. *Arneodo A., Coullet P., Tresser C.* Occurrence of strange attractors in three-dimensional Volterra equations // Phys. Lett. A . 1980. V. 79A. № 4. P. 259–263.
https://doi.org/10.1016/0375-9601(80)90342-4

33. *Wang R., Xiao D.* Bifurcations and chaotic dynamics in a 4-dimensional competitive Lotka–Volterra system // Nonlinear Dyn. 2010. V. 59. № 3. P. 411–422.
https://doi.org/10.1007/s11071-009-9547-3

34. *Vano J.A., Wildenberg J.C., Anderson M.B., Noel J.K., Sprott J.C.* Chaos in low-dimensional Lotka–Volterra models of competition // Nonlinearity. 2006. V. 19. № 10. P. 2391–2404.
https://doi.org/10.1088/0951-7715/19/10/006

35. *Kozlov V., Vakulenko S.* On chaos in Lotka–Volterra systems: an analytical approach // Nonlinearity. 2013. V. 26. № 8. P. 2299–2314.
https://doi.org/10.1088/0951-7715/26/8/2299

36. *Christie J.R., Gopalsamy K., Li J.* Chaos in perturbed Lotka–Volterra systems // Anziam J. (Australian and New Zealand Industrial and Applied Mathematics Journal). 2001. V. 42. № 3. P. 399–412.
https://doi.org/10.1017/S1446181100012025

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 256–267

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 517.956.4

УГЛОВОЙ ПОГРАНИЧНЫЙ СЛОЙ В КРАЕВЫХ ЗАДАЧАХ ДЛЯ СИНГУЛЯРНО ВОЗМУЩЕННЫХ ПАРАБОЛИЧЕСКИХ УРАВНЕНИЙ С КУБИЧЕСКИМИ НЕЛИНЕЙНОСТЯМИ

© 2021 г. И. В. Денисов

300026 Тула, пр-т Ленина, 125, Тульский государственный педагогический университет им. Л.Н. Толстого, Россия

e-mail: den@tspu@mail:ru

Поступила в редакцию 04.06.2000 г.
Переработанный вариант 23.07.2000 г.
Принята к публикации 16.09.2020 г.

Для сингулярно возмущенного параболического уравнения

$$\epsilon^2\left(a^2\frac{\partial^2 u}{\partial x^2}-\frac{\partial u}{\partial t}\right)=F(u,x,t,\epsilon)$$

в прямоугольнике рассматривается задача с краевыми условиями I рода. Предполагается, что в угловых точках прямоугольника функция F относительно переменной u является кубической. Строится полное асимптотическое разложение решения при $\varepsilon \to 0$ и обосновывается его равномерность в замкнутом прямоугольнике. Библ. 10.

Ключевые слова: пограничный слой, асимптотическое приближение, сингулярно возмущенное уравнение.

DOI: 10.31857/S0044466921020071

ВВЕДЕНИЕ

В работе рассматривается начально-краевая задача вида

$$\epsilon^2\left(a^2\frac{\partial^2 u}{\partial x^2}-\frac{\partial u}{\partial t}\right)=F(u,x,t,\epsilon),\quad (x,t)\in\Omega, \tag{0.1}$$

$$u(x,0,\epsilon)=\phi(x),\quad 0\le x\le 1, \tag{0.2}$$

$$u(0,t,\epsilon)=\psi_1(t),\quad u(1,t,\epsilon)=\psi_2(t),\quad 0\le t\le T. \tag{0.3}$$

Через Ω обозначен прямоугольник $\{(x,t)|0<x<1,0<t<T\}$.

Ранее в работах [1]– [5] проведены исследования в предположении, что в угловых точках прямоугольника $(0,0)$ и $(1,0)$ функция F является квадратичной относительно переменной u на промежутке от корня вырожденного уравнения до граничного значения. Построено полное асимптотическое приближение решения задачи (0.1)–(0.3) при $\epsilon\to 0$ и обоснована равномерность этого приближения в замкнутом прямоугольнике с точностью любого порядка.

В работе [6] была предпринята попытка распространить полученные результаты на произвольную монотонную нелинейность. Для доказательства существования подходящих решений нелинейных задач использовался метод верхних и нижних решений. Однако выполнение необходимых неравенств в общем случае доказать не удалось и пришлось их выполнение постулировать. Кроме рассмотренных ранее квадратичных нелинейностей этим неравенствам удовлетворяли и некоторые другие нелинейные функции, что, несомненно, являлось продвижением в исследовании общей задачи. Однако класс подходящих нелинейных функций не был полностью изучен. В данной статье рассматриваются кубические нелинейности. Для предлагаемых барьерных функций необходимые неравенства доказываются. Строится полное асимптотическое при-

ближение решения задачи (0.1)–(0.3) при $\epsilon \to 0$ и обосновывается равномерность этого приближения в замкнутом прямоугольнике с точностью любого порядка.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть, как и в предыдущих работах, выполнены следующие условия.

Условие 1. Функции $F(u,x,t,\epsilon)$, $\phi(x)$, $\psi_1(t)$ и $\psi_2(t)$ являются достаточно гладкими и в угловых точках прямоугольника Ω выполняются условия согласованности начально-краевых значений:

$$\phi(0) = \psi_1(0), \quad \phi(1) = \psi_2(0).$$

Условие 2. Вырожденное уравнение $F(u,x,t,0) = 0$ в замкнутом прямоугольнике $\bar{\Omega}$ имеет решение, которое обозначается как $u = \bar{u}_0(x,t)$.

Заметим, что в силу нелинейности это уравнение может иметь и другие решения.

Условие 3. Производная $F_u'(\bar{u}_0(x,t),x,t,0) > 0$ в замкнутом прямоугольнике $\bar{\Omega}$.

Условие 4. Начальная задача

$$\frac{d\Pi_0}{d\tau} = -F(\bar{u}_0(x,0) + \Pi_0, x, 0, 0), \quad \Pi_0(x,0) = \phi(x) - \bar{u}_0(x,0), \tag{1.1}$$

имеет решение $\Pi_0(x,\tau)$ при $\tau \geq 0$, удовлетворяющее условию $\Pi_0(x,\infty) = 0$ (здесь параметр $x \in [0,1]$).

Условие 5. Для систем

$$\frac{dz_1}{dy} = z_2, \quad a^2\frac{dz_2}{dy} = F(\bar{u}_0(k,t) + z_1, k, t, 0), \tag{1.2}$$

прямые $z_1 = \psi_{1+k}(t) - \bar{u}_0(k,t)$ пересекают сепаратрисы, входящие в точку покоя $(z_1,z_2) = (0,0)$ при $y \to \infty$ (здесь t – параметр, $k = 0$ или 1).

Условий 1–5 недостаточно, чтобы гарантировать существование решения задачи (0.1)–(0.3) для произвольной функции $F(u,x,t,\epsilon)$. Поэтому требуются дополнительные условия, гарантирующие возможность построения асимптотики решения. Эти условия будут заключаться в выборе определенного класса функций. Решение задачи (0.1)–(0.3) строится согласно методу угловых пограничных функций (см. [7]) в виде суммы

$$u(x,t,\epsilon) = \bar{u} + (\Pi + Q + Q^*) + (P + P^*), \tag{1.3}$$

где $\bar{u}$ обозначает функцию, называемую регулярной частью асимптотики. Эта функция представляет решение задачи во внутренней части прямоугольника Ω без учета граничных условий. Пограничные функции Π, Q и Q^* осуществляют гладкий переход от регулярной части к граничным условиям на сторонах прямоугольника Ω: $t = 0$, $x = 0$ и $x = 1$ соответственно. Угловые пограничные функции P и P^* сглаживают невязки, вносимые пограничными функциями вблизи вершин прямоугольника Ω: $(0,0)$ и $(1,0)$ соответственно.

2. РЕГУЛЯРНАЯ И ПОГРАНСЛОЙНАЯ ЧАСТИ АСИМПТОТИКИ

Формальная процедура построения регулярной части асимптотики и пограничных функций хорошо отработана (см. [7]), однако, чтобы не обращаться к другим источникам, повторим ее схематично. В уравнении (0.1) функция F заменяется выражением, аналогичным (1.3):

$$F(u,x,t,\epsilon) = \bar{F} + (\Pi F + QF + Q^*F) + (PF + P^*F). \tag{2.1}$$

Выражения (1.3) и (2.1) подставляются в уравнение (0.1), которое разделяется на шесть частей: регулярную, три пограничных и две угловых. Регулярная часть асимптотики $\bar{u}$ строится в виде ряда по степеням ϵ:

$$\bar{u}(x,t,\epsilon) = \sum_{k=0}^{\infty} \epsilon^k \bar{u}_k(x,t). \tag{2.2}$$

Коэффициент $\bar{u}_0 = \bar{u}_0(x,t)$ выбирается в соответствии с условиями 2 и 3, а последующие функции $\bar{u}_k, k \geq 1$, строятся рекуррентно.

Пусть $\partial\Omega$ обозначает границу прямоугольника Ω без стороны $t = T$. Регулярная часть $\bar{u}(x,t,\epsilon)$ асимптотики дает решение задачи (0.1)–(0.3) внутри прямоугольника Ω, но на границе $\partial\Omega$ функция $\bar{u}(x,t,\epsilon)$, вообще говоря, не совпадает с начальным и граничными значениями. В связи с этим возникает так называемая “невязка”.

Погранслойная часть асимптотики вводится для устранения невязок регулярной части с начальным и граничными условиями. Погранслойные функции Π, Q и Q^* определяются из уравнений, в которых переходят к растянутым переменным

$$\xi = \frac{x}{\epsilon}, \quad \xi_* = \frac{1-x}{\epsilon}, \quad \tau = \frac{t}{\epsilon^2}.$$

Функции Π, Q и Q^* ищутся в виде рядов

$$\Pi(x,\tau,\epsilon) = \sum_{k=0}^{\infty} \epsilon^k \Pi_k(x,\tau), \tag{2.3}$$

$$Q(\xi,t,\epsilon) = \sum_{k=0}^{\infty} \epsilon^k Q_k(\xi,t), \tag{2.4}$$

$$Q^*(\xi_*,t,\epsilon) = \sum_{k=0}^{\infty} \epsilon^k Q_k^*(\xi_*,t). \tag{2.5}$$

На стороне $t = 0$ невязки в начальном условии (0.2) призвана устранить функция $\Pi = \Pi(x,\tau,\epsilon)$. При переходе от переменных (x,t) к переменным (x,τ) прямоугольник Ω при $\epsilon \to 0$ растягивается до полуполосы $0 \le x \le 1$, $0 \le \tau < \infty$.

Для $\Pi_0 = \Pi_0(x,\tau)$ получается задача

$$-\frac{\partial \Pi_0}{\partial \tau} = F\left(\bar{u}_0(x,0) + \Pi_0, x, 0, 0\right), \quad \Pi_0(x,0) = \phi(x) - \bar{u}_0(x,0). \tag{2.6}$$

Здесь x играет роль параметра. В силу условия 4 эта задача имеет решение, для которого в силу условия 3 справедлива экспоненциальная оценка убывания вида

$$\left|\Pi_0(x,\tau)\right| \le C\exp(-\kappa\tau), \tag{2.7}$$

где C и κ – некоторые положительные числа.

Задачи для определения функций $\Pi_k = \Pi_k(x,\tau)$, $k \ge 1$, получаются линейными:

$$-\frac{\partial \Pi_k}{\partial \tau} = F_u'\left(\bar{u}_0(x,0) + \Pi_0, x, 0, 0\right)\Pi_k + \tilde{\pi}_k, \quad \Pi_k(x,0) = -\bar{u}_k(x,0). \tag{2.8}$$

Функции $\tilde{\pi}_k$ рекуррентно выражаются через функции Π_j, $j < k$, и их производные. Поэтому, если для функций Π_j, $j < k$, справедливы оценки вида (2.7), то для функций $\tilde{\pi}_k$ справедливы оценки того же вида.

Если величина $\phi(x) - \bar{u}_0(x,0)$ не равна тождественно нулю, то решения задач (2.8) имеют вид

$$\Pi_k(x,\tau) = -U(x,\tau)\bar{u}_k(x,0) - U(x,\tau)\int_0^{\tau} \left(U(x,\sigma)\right)^{-1}\tilde{\pi}_k(x,\sigma)d\sigma,$$

где

$$U(x,\tau) = \exp\left(-\int_0^{\tau} F_u'\left(\bar{u}_0(x,0) + \Pi_0(x,\lambda), x, 0, 0\right)d\lambda\right)$$

есть фундаментальное решение ($U(x,0) = 1$) соответствующего однородного уравнения, и справедлива оценка вида

$$\left|U(x,\tau)\left(U(x,\sigma)\right)^{-1}\right| \le C\exp(-\kappa(\tau-\sigma)),$$

где переменные $0 \le x \le 1$, $0 \le \sigma \le \tau$, а постоянные C и κ – положительные числа. Эта оценка позволяет для функции $\Pi_k(x,\tau)$ получить оценку вида (2.7).

Если величина $\phi(x) - \bar{u}_0(x,0) \equiv 0$, то $\Pi_0(x,\tau) \equiv 0$. Коэффициенты при Π_k в задачах (2.8) оказываются постоянными и положительными, т.е. задачи упрощаются.

Таким образом определяются коэффициенты ряда (2.3), и функция $\Pi(x,\tau,\epsilon)$ устраняет невязки с начальным условием (0.2) на стороне $t = 0$.

Построенная регулярная часть асимптотики вносит невязки и в граничные условия. На стороне $x = 0$ невязки в граничных условиях призвана устранить функция $Q = Q(\xi,t,\epsilon)$, где $\xi = x/\epsilon$ – растянутая переменная. При переходе от переменных (x,t) к переменным (ξ,t) прямоугольник Ω при $\epsilon \to 0$ растягивается до полуполосы $0 \le \xi < \infty$, $0 \le t \le T$.

Задача для $Q_0 = Q_0(\xi,t)$ имеет вид

$$a^2 \frac{\partial^2 Q_0}{\partial \xi^2} = F\left(\bar{u}_0(0,t) + Q_0, 0, t, 0\right), \quad Q_0(0,t) = \psi_1(t) - \bar{u}_0(0,t), \quad Q_0(\infty,t) = 0, \tag{2.9}$$

где t играет роль параметра. Уравнение (2.9) эквивалентно системе (1.2), в которой следует положить $z_1 = Q_0(\xi,t)$, $k = 0$, $y = \xi$. Условия затухания выделяют решения уравнения (2.9), для которых справедливы экспоненциальные оценки убывания вида

$$|Q_0(\xi,t)| \le C\exp(-\kappa\xi), \tag{2.10}$$

где C и κ – положительные числа. Так как возможен переход с сепаратрисы на сепаратрису, то решение задачи (2.9) не единственно. Однако такие случаи мы исключаем и рассматриваем только монотонные решения.

Задачи для определения функций $Q_k(\xi,t)$, $k \ge 1$, линейны:

$$a^2 \frac{\partial^2 Q_k}{\partial \xi^2} = F_u'(\bar{u}_0(0,t) + Q_0, 0, t, 0)Q_k + \tilde{q}_k, \quad Q_k(0,t) = -\bar{u}_k(0,t), \quad Q_k(\infty,t) = 0. \tag{2.11}$$

Функции $\tilde{q}_k$ рекуррентно выражаются через функции Q_j, $j < k$, и их производные. Поэтому, если для функций Q_j, $j < k$, справедливы экспоненциальные оценки вида (2.10), то для функций $\tilde{q}_k$ справедливы оценки того же вида.

Если величина $\psi_1(t) - \bar{u}_0(0,t)$ не равна тождественно нулю, то решение задачи (2.11) имеет вид

$$Q_k(\xi,t) = -\frac{\Phi(\xi,t)}{\Phi(0,t)}\bar{u}_k(0,t) - \frac{\Phi(\xi,t)}{a(0,t)}\int_0^\xi \frac{d\lambda}{\Phi^2(\lambda,t)}\int_\lambda^\infty \Phi(\sigma,t)\tilde{q}_k(\sigma,t)d\sigma, \quad \Phi(\xi,t) = \frac{\partial Q_0(\xi,t)}{\partial \xi},$$

и для него справедлива оценка вида (2.10).

Если величина $\psi_1(t) - \bar{u}_0(0,t) \equiv 0$, то $Q_0(\xi,t) \equiv 0$, а коэффициент при Q_k в уравнениях (2.11) оказывается постоянным и положительным, т.е. задача упрощается.

Таким образом определяются коэффициенты ряда (2.4), и функция $Q(\xi,t,\epsilon)$ устраняет невязки в граничном условии на стороне $x = 0$.

Регулярная часть асимптотики вносит невязки в граничные условия и на стороне $x = 1$. Эти невязки устраняет функция $Q^* = Q^*(\xi_*,t,\epsilon)$, $\xi_* = (1-x)/\epsilon$, которая строится в виде ряда (2.5). Коэффициенты этого ряда Q_k^* определяются аналогично коэффициентам ряда (2.4) и для них справедливы экспоненциальные оценки убывания вида

$$|Q_k^*(\xi_*,t)| \le C\exp(-\kappa\xi_*),$$

где C и κ – положительные числа.

Таким образом, погранслойная часть асимптотики определяется полностью. Однако каждая в отдельности погранслойная функция, устраняя невязки на соответствующей стороне, в свою очередь вносит невязки на примыкающие стороны прямоугольника. Так, погранслойные функции $\Pi_k(x,\tau)$, устраняя невязки в начальном условии на стороне $t = 0$, вносят дополнительные невязки в граничные условия на сторонах $x = 0$ и $x = 1$. Эти невязки существенны только вблизи угловых точек (0,0) и (1,0), а далее, с ростом t, они экспоненциально затухают. Аналогичное влияние функции $Q_k(\xi,t)$ и $Q_k^*(\xi_*,t)$ оказывают на начальное условие на стороне $x = 0$.

3. УГЛОВАЯ ЧАСТЬ АСИМПТОТИКИ. ОСНОВНЫЕ ПРОБЛЕМЫ

С целью устранения невязок с начальным и граничными условиями вблизи угловых точек $(0,0)$ и $(1,0)$ прямоугольника Ω вводятся угловые пограничные функции $P(\xi,\tau,\epsilon)$ и $P^*(\xi_*,\tau,\epsilon)$. Ввиду того, что не существует универсального метода их нахождения, задачи для определения функций $P(\xi,\tau,\epsilon)$ и $P^*(\xi_*,\tau,\epsilon)$ доставляют основные трудности. Практически каждая такая задача требует разработки новых методов решения.

Угловые пограничные функции будем искать в виде рядов

$$P(\xi,\tau,\epsilon)=\sum_{k=0}^{\infty}\epsilon^k P_k(\xi,\tau),\qquad P^*(\xi_*,\tau,\epsilon)=\sum_{k=0}^{\infty}\epsilon^k P_k^*(\xi_*,\tau).$$

Рассмотрим угловую точку $(0,0)$. В окрестности этой точки невязки в условия (0.2), (0.3) вносят функции $\Pi(x,\tau,\epsilon)$ и $Q(\xi,t,\epsilon)$. Для устранения этих невязок служит функция $P(\xi,\tau,\epsilon)$, которая определяется из уравнения

$$\epsilon^2\left(a^2\frac{\partial^2 P}{\partial x^2}-\frac{\partial P}{\partial t}\right)\Bigg|_{\substack{x=\epsilon\xi\\ t=\epsilon^2\tau}}=PF.$$

Задача для определения $P_0(\xi,\tau)$ ставится в первой четверти

$$\mathbb{R}^2_+:=\{(\xi,\tau)|\xi>0,\ \tau>0\}$$

плоскости переменных (ξ,τ) и имеет вид

$$a^2\frac{\partial^2 P_0}{\partial\xi^2}-\frac{\partial P_0}{\partial\tau}=F\left(\overline{u}_0(0,0)+\Pi_0(0,\tau)+Q_0(\xi,0)+P_0(\xi,\tau),0,0,0\right)-$$
$$-F\left(\overline{u}_0(0,0)+\Pi_0(0,\tau),0,0,0\right)-F\left(\overline{u}_0(0,0)+Q_0(\xi,0),0,0,0\right),\quad \xi>0,\quad \tau>0, \tag{3.1}$$

$$P_0(0,\tau)=-\Pi_0(0,\tau),\quad P_0(\xi,0)=-Q_0(\xi,0),\quad P_0(\xi,\tau)\to 0\quad \text{при}\quad \xi+\tau\to\infty. \tag{3.2}$$

Для функций $P_k(\xi,\tau)$, $k\geq 1$, в области $\mathbb{R}^2_+$ получаются линейные задачи

$$a^2\frac{\partial^2 P_k}{\partial\xi^2}-\frac{\partial P_k}{\partial\tau}=F_u'\left(\overline{u}_0(0,0)+\Pi_0(0,\tau)+Q_0(\xi,0)+P_0(\xi,\tau),0,0,0\right)P_k+h_k, \tag{3.3}$$

$$P_k(0,\tau)=-\Pi_k(0,\tau),\quad P_k(\xi,0)=-Q_k(\xi,0),\quad P_k(\xi,\tau)\to 0\quad \text{при}\quad \xi+\tau\to\infty, \tag{3.4}$$

где неоднородности h_k удовлетворяют экспоненциальным оценкам убывания вида

$$|h_k(\xi,\tau)|\leq C\exp(-\kappa(\xi+\tau)), \tag{3.5}$$

если такого же вида оценкам удовлетворяют функции $P_0,\ldots,P_{k-1}$. Здесь C и κ – некоторые положительные числа.

Если число $\phi(0,0)-\overline{u}_0(0,0)=0$, то решением задачи (2.6) при $x=0$ будет функция $\Pi_0(0,\tau)\equiv 0$, решением задачи (2.8) при $t=0$ будет функция $Q_0(\xi,0)\equiv 0$. Решением задачи (3.1), (3.2) будет функция $P_0(\xi,\tau)\equiv 0$, а коэффициент в задачах (3.3), (3.4) будет постоянным и положительным:

$$F_u'\left(\overline{u}_0(0,0)+\Pi_0(0,\tau)+Q_0(\xi,0)+P_0(\xi,\tau),0,0,0\right)=F_u'\left(\overline{u}_0(0,0),0,0,0\right)>0.$$

В этом случае решения задач (3.3), (3.4) выписываются в явном виде и для них получаются экспоненциальные оценки вида (3.5).

Если число $\varphi(0,0)-\overline{u}_0(0,0)\neq 0$, то, вообще говоря, неизвестно, имеет или нет задача (3.1), (3.2) решение и удовлетворяет ли решение в случае существования экспоненциальной оценке вида (3.5). Кроме этого, в задачах (3.3), (3.4) коэффициент

$$F_u':=F_u'\left(\overline{u}_0(0,0)+\Pi_0(0,\tau)+Q_0(\xi,0)+P_0(\xi,\tau),0,0,0\right)$$

может в зависимости от вида функции F и величины $\phi(0,0)$ принимать как положительные, так и отрицательные значения.

Даже в случае разрешимости задач (3.1)–(3.4) доказательство того, что задача (0.1)–(0.3) имеет решение, все равно остается проблемой. Это связано с тем, что, не зная явного вида функции $P_0(\xi,\tau)$,

мы не можем знать явного вида коэффициента F_u', который может оказаться как положительным, так и отрицательным. Если не накладывать дополнительных условий, то это обстоятельство, вообще говоря, не позволит обосновать построенную асимптотику решения.

Таким образом, в результате реализации метода угловых пограничных функций для исследования задачи (0.1)–(0.3) мы перешли к задачам (3.1)–(3.4). Дальнейшие исследования предполагают разрешение, по крайней мере, трех основных проблем.

1. Имеет ли задача (3.1), (3.2) решение, удовлетворяющее экспоненциальной оценке убывания вида (3.5)?

2. Если задача (3.1), (3.2) имеет решение, удовлетворяющее экспоненциальной оценке вида (3.5), то имеют ли задачи (3.3), (3.4) решения, удовлетворяющие подобным оценкам?

3. Если задачи (3.1)–(3.4) разрешимы, т.е. если ряд (1.3) может быть построен, то имеет ли задача (0.1)–(0.3) решение, для которого этот ряд будет асимптотическим представлением при $\epsilon \to 0$ в замкнутом прямоугольнике $\bar{\Omega}$?

Задачи для угловых пограничных функций $P_k^*(\xi_*, \tau)$, ставятся аналогично.

4. КУБИЧЕСКАЯ НЕЛИНЕЙНОСТЬ

Будем рассматривать угловую точку $(0,0)$ прямоугольника Ω и определим оператор L:

$$L(P) := a^2 \frac{\partial^2 P}{\partial \xi^2} - \frac{\partial P}{\partial \tau} - F(\bar{u}_0 + \Pi_0 + Q_0 + P) + F(\bar{u}_0 + \Pi_0) + F(\bar{u}_0 + Q_0),$$

где

$$P = P(\xi, \tau), \quad F(u) = F(u, 0, 0, 0), \quad \bar{u}_0 = \bar{u}_0(0,0), \quad \Pi_0 = \Pi_0(0, \tau), \quad Q_0 = Q_0(\xi, 0).$$

Задачу (3.1), (3.2) можно переписать в операторной форме:

$$L(P_0) = 0 \quad \text{в области} \quad \mathbb{R}^2_+, \tag{4.1}$$

$$P_0(0, \tau) = -\Pi_0(0, \tau), \quad P_0(\xi, 0) = -Q_0(\xi, 0), \quad P_0(\xi, \tau) \to 0 \quad \text{при} \quad \xi + \tau \to \infty. \tag{4.2}$$

Для доказательства существования решения задачи (4.1), (4.2) используется метод верхних и нижних решений (см. [8]– [10]), который заключается в том, что задача

$$L(Z) = 0 \quad \text{в области} \quad \Omega, \quad Z = h \quad \text{на границе} \quad \partial\Omega$$

имеет решение Z в промежутке между барьерными функциями $Z_- \le Z \le Z_+$, если в области Ω выполняются неравенства

$$L(Z_+) \le 0, \quad L(Z_-) \ge 0, \quad Z_- \le Z_+,$$

а на ее границе

$$Z_- \le h \le Z_+.$$

Барьерные функции для задачи (4.1), (4.2) желательно строить с расчетом, чтобы коэффициент F_u' в задачах (3.3), (3.4) был положительным. Это обеспечит существование решений $P_k(\xi, \tau)$ с оценками вида (3.5). В качестве возможного верхнего решения задачи (4.1), (4.2) рассмотрим функцию

$$Z_+(\xi, \tau) = 0.$$

Для определенности будем считать, что $\phi(0,0) > \bar{u}_0(0,0)$, т.е. граничное значение в точке $(0,0)$ лежит правее корня вырожденного уравнения. Тогда значения Π_0 и Q_0 принадлежат промежутку $(0, \phi - \bar{u}_0]$, где $\phi - \bar{u}_0 > 0$, и на границе области $\mathbb{R}^2_+$ выполняются необходимые для верхнего решения неравенства:

$$Z_+(0, \tau) = 0 \ge -\Pi_0, \quad Z_+(\xi, 0) = 0 \ge -Q_0.$$

Внутри области $\mathbb{R}^2_+$ нужно доказать неравенство

$$L(Z_+) = -F(\bar{u}_0 + \Pi_0 + Q_0) + F(\bar{u}_0 + \Pi_0) + F(\bar{u}_0 + Q_0) \le 0. \tag{4.3}$$

Для краткости обозначим $\Pi_0 = y$, $Q_0 = z$, $L(Z_+) = L$, где $L = L(y, z)$. Тогда неравенство (4.3) примет вид

$$L = -F(\overline{u}_0 + y + z) + F(\overline{u}_0 + y) + F(\overline{u}_0 + z) \le 0, \quad y, z \in (0, \phi - \overline{u}_0].$$

Необходимые условия экстремума

$$\frac{\partial L}{\partial y} = -F'(\overline{u}_0 + y + z) + F'(\overline{u}_0 + y) = 0,$$

$$\frac{\partial L}{\partial z} = -F'(\overline{u}_0 + y + z) + F'(\overline{u}_0 + z) = 0,$$

приводят к равенствам

$$F'(\overline{u}_0 + y + z) = F'(\overline{u}_0 + y) = F'(\overline{u}_0 + z). \tag{4.4}$$

На границе области из-за симметрии функции $L(y, z)$ можно рассмотреть только две из четырех сторон квадрата $(0, \phi - \overline{u}_0] \times (0, \phi - \overline{u}_0]$.

1. При $y = 0$ значения

$$L(0, z) = -F(\overline{u}_0 + z) + F(\overline{u}_0) + F(\overline{u}_0 + z) = 0.$$

2. При $y = \phi - \overline{u}_0$ значения

$$L(\phi - \overline{u}_0, z) = -F(\phi + z) + F(\phi) + F(\overline{u}_0 + z).$$

Необходимое неравенство $L(\phi - \overline{u}_0, z) \le 0$ выполняется, если

$$F(\overline{u}_0 + u) + F(\phi) \le F(\phi + u), \quad u \in (0, \phi - \overline{u}_0]. \tag{4.5}$$

Для проверки условия (4.5) нужно рассмотреть два куска графика функции $F(u)$: один – при $u \in (\overline{u}_0, \phi]$, а другой – при $u \in (\phi, \phi + \phi - \overline{u}_0]$. Первый кусок нужно сдвинуть параллельно вертикальной оси на величину $F(\phi)$, в связи с чем появляется функция

$$F_1(u) = F(\overline{u}_0 + u) + F(\phi), \quad u \in (0, \phi - \overline{u}_0].$$

Второй кусок нужно сдвинуть влево на величину $\phi - \overline{u}_0$, в связи с чем появляется функция

$$F_2(u) = F(\phi + u), \quad u \in (0, \phi - \overline{u}_0].$$

Графики обеих функций выходят из точки $(0, F(\phi))$. Условие (4.5) оказывается эквивалентным простому условию

$$F_1(u) \le F_2(u), \quad u \in (0, \phi - \overline{u}_0].$$

В качестве возможного нижнего решения задачи (4.1), (4.2) рассмотрим функцию

$$Z_-(\xi, \tau) = -2\sqrt{\Pi_0 Q_0}.$$

Так как значения Π_0 и Q_0 принадлежат промежутку $(0, \phi - \overline{u}_0]$, то необходимые для нижнего решения неравенства выполняются на границе области $\mathbb{R}^2_+$:

$$Z_-(0, \tau) = -2\sqrt{(\phi - \overline{u}_0)\Pi_0} \le -\Pi_0, \quad Z_-(\xi, 0) = -2\sqrt{(\phi - \overline{u}_0)Q_0} \le -Q_0.$$

Внутри области $\mathbb{R}^2_+$ нужно доказать неравенство $L(Z_-) \ge 0$. Имеем

$$L(Z_-) = a^2 \frac{\partial^2 Z_-}{\partial \xi^2} - \frac{\partial Z_-}{\partial \tau} - F(\overline{u}_0 + \Pi_0 + Q_0 - 2\sqrt{\Pi_0 Q_0}) + F(\overline{u}_0 + \Pi_0) + F(\overline{u}_0 + Q_0).$$

Производная

$$\frac{\partial^2 Z_-}{\partial \xi^2} = -2\sqrt{\Pi_0}\frac{d^2\sqrt{Q_0}}{d\xi^2} = -2\sqrt{\Pi_0}\frac{d}{d\xi}\left(\frac{1}{2\sqrt{Q_0}}\frac{dQ_0}{d\xi}\right) = -2\sqrt{\Pi_0}\left(-\frac{1}{4Q_0\sqrt{Q_0}}\left(\frac{dQ_0}{d\xi}\right)^2 + \frac{1}{2\sqrt{Q_0}}\frac{d^2Q_0}{d\xi^2}\right).$$

Задача для определения функции $Q_0 = Q_0(\xi, 0)$ получается из (2.9) при $t = 0$:

$$a^2 \frac{d^2 Q_0}{d\xi^2} = F(\overline{u}_0 + Q_0), \quad Q_0(0,0) = \phi - \overline{u}_0, \quad Q_0(\infty, 0) = 0.$$

Отсюда

$$\frac{d^2 Q_0}{d\xi^2} = \frac{1}{a^2} F(\overline{u}_0 + Q_0), \quad \left(\frac{dQ_0}{d\xi}\right)^2 = \frac{2}{a^2} \int_0^{Q_0} F(\overline{u}_0 + u)du.$$

Поэтому

$$\frac{\partial^2 Z_-}{\partial \xi^2} = -2\sqrt{\Pi_0}\left[-\frac{1}{4Q_0\sqrt{Q_0}}\frac{2}{a^2}\int_0^{Q_0} F(\overline{u}_0 + u)du + \frac{1}{2\sqrt{Q_0}}\frac{1}{a^2}F(\overline{u}_0 + Q_0)\right] =$$
$$= \frac{1}{a^2}\sqrt{\Pi_0 Q_0}\left(\frac{1}{Q_0^2}\int_0^{Q_0} F(\overline{u}_0 + u)du - \frac{1}{Q_0}F(\overline{u}_0 + Q_0)\right).$$

Производная

$$\frac{\partial Z_-}{\partial \tau} = -2\sqrt{Q_0}\frac{d\sqrt{\Pi_0}}{d\tau} = -\frac{\sqrt{Q_0}}{\sqrt{\Pi_0}}\frac{d\Pi_0}{d\tau}.$$

Задача для определения функции $\Pi_0 = \Pi_0(0, \tau)$ получается из (2.6) при $y = 0$:

$$-\frac{d\Pi_0}{d\tau} = F(\overline{u}_0 + \Pi_0), \quad \Pi_0(0,0) = \phi - \overline{u}_0.$$

Поэтому

$$\frac{\partial Z_-}{\partial \tau} = \frac{\sqrt{Q_0}}{\sqrt{\Pi_0}}F(\overline{u}_0 + \Pi_0) = \sqrt{\Pi_0 Q_0}\frac{1}{\Pi_0}F(\overline{u}_0 + \Pi_0).$$

В результате для $L(Z_-)$ получается выражение

$$L(Z_-) = \sqrt{\Pi_0 Q_0}\left(\frac{1}{Q_0^2}\int_0^{Q_0} F(\overline{u}_0 + u)du - \frac{1}{Q_0}F(\overline{u}_0 + Q_0) - \frac{1}{\Pi_0}F(\overline{u}_0 + \Pi_0)\right) -$$
$$- F(\overline{u}_0 + \Pi_0 + Q_0 - 2\sqrt{\Pi_0 Q_0}) + F(\overline{u}_0 + \Pi_0) + F(\overline{u}_0 + Q_0). \tag{4.6}$$

Теорема 4.1. *Пусть функция $F(u)$ имеет вид $F(u) = u^3 - b^3$, где $b > 0$, и граничное значение $\varphi(0,0) > \overline{u}_0(0,0)$. Тогда задача* (4.1), (4.2) *имеет решение $P_0(\xi, \tau)$, удовлетворяющее экспоненциальной оценке убывания вида* (3.5).

Доказательство. Считаем, что $\overline{u}_0 = b$. С помощью метода верхних и нижних решений покажем, что существует решение задачи (4.1), (4.2), заключенное между барьерными функциями следующего вида:

$$Z_-(\xi, \tau) = -2\sqrt{\Pi_0 Q_0} \le P_0(\xi, \tau) \le 0 = Z_+(\xi, \tau).$$

Для доказательства того, что функция $Z_+(\xi, \tau) = 0$ является верхним решением задачи, нужно проверить выполнение условия (4.4), которое для рассматриваемой функции имеет вид

$$(b + y + z)^2 = (b + y)^2 = (b + z)^2,$$

так как $\overline{u}_0(0,0) = b$. Это условие выполняется в единственной точке $y = z = 0$, которая не является внутренней для области. Поэтому точек экстремума нет. Условие (4.5), очевидно, выполняется. Таким образом, функция $Z_+(\xi, \tau) = 0$ является верхним решением задачи.

Для доказательства того, что функция $Z_-(\xi,\tau) = -2\sqrt{\Pi_0 Q_0}$ является нижним решением задачи, нужно проверить выполнение неравенства $L(Z_-) \geq 0$. С этой целью запишем (4.6) с учетом замены $\bar{u}_0 \to b$, $\Pi_0 \to y$, $Q_0 \to z$, $L(Z_+) \to L$ в виде

$$L = \sqrt{yz}\left(\frac{1}{z^2}\int_0^z F(b+u)du - \frac{1}{z}F(b+z) - \frac{1}{y}F(b+y)\right) - F(b+y+z-2\sqrt{yz}) + F(b+y) + F(b+z). \quad (4.7)$$

Доказательство неравенства $L \geq 0$ значительно упрощается, если отбросить первое слагаемое и рассмотреть более сильное неравенство

$$-\sqrt{yz}\left(\frac{1}{z}F(b+z) + \frac{1}{y}F(b+y)\right) - F(b+y+z-2\sqrt{yz}) + F(b+y) + F(b+z) \geq 0. \quad (4.8)$$

В отличие от L левая часть (4.8) является функцией, симметричной относительно y и z. Ее можно привести к зависимости от переменных

$$s = \sqrt{yz} \in (0, \phi - b], \quad t = \frac{y+z}{2} \in (0, \phi - b]. \quad (4.9)$$

Имеем

$$F(b+u) = (b+u)^3 - b^3 = u[(b+u)^2 + b(b+u) + b^2] = u(u^2 + 3bu + 3b^2),$$

$$F(b+y) + F(b+z) = (y^3 + 3by^2 + 3b^2y) + (z^3 + 3bz^2 + 3b^2z) = (y+z)^3 -$$

$$- 3yz(y+z) + 3b(y+z)^2 - 6byz + 3b^2(y+z) = 8t^3 - 6s^2t + 12bt^2 - 6bs^2 + 6b^2t,$$

$$\frac{1}{z}F(b+z) + \frac{1}{y}F(b+y) = (z^2 + 3bz + 3b^2) + (y^2 + 3by + 3b^2) =$$

$$= (y+z)^2 - 2yz + 3b(y+z) + 6b^2 = 4t^2 - 2s^2 + 6bt + 6b^2,$$

$$F(b+y+z-2\sqrt{yz}) = F(b+2t-2s).$$

В результате замены (4.9) неравенство (4.8) примет вид

$$-s(4t^2 - 2s^2 + 6bt + 6b^2) - F(b+2t-2s) + 8t^3 - 6s^2t + 12bt^2 - 6bs^2 + 6b^2t \geq 0,$$

или

$$H := 8t^3 - 6s^2t + 12bt^2 - 6bs^2 + 6b^2t - 4st^2 + 2s^3 - 6bst - 6b^2s - F(b+2t-2s) \geq 0. \quad (4.10)$$

Найдем образ квадрата $G : (0, \phi - b] \times (0, \phi - b]$ при отображении

$$(y,z) \to \left(\sqrt{yz}, \frac{y+z}{2}\right).$$

Так как

$$\frac{y+z}{2} \geq \sqrt{yz},$$

то оба треугольника $y < z$ и $z < y$, составляющие квадрат G, отображаются в угол $t > s$. Две стороны квадрата G:

$$y = 0, \quad z \in (0, \phi - b] \quad \text{и} \quad z = 0, \quad y \in (0, \phi - b],$$

переходят в один и тот же отрезок

$$s = 0, \quad t \in \left(0, \frac{\phi - b}{2}\right]. \quad (4.11)$$

Диагональ квадрата G: $z = y$, $y \in (0, \phi - b]$, переходит в отрезок биссектрисы

$$t = s, \quad s \in (0, \phi - b]. \quad (4.12)$$

И, наконец, оставшиеся две стороны квадрата G:

$$y = \phi - b, \quad z \in (0, \phi - b] \quad \text{и} \quad z = \phi - b, \quad y \in (0, \phi - b],$$

переходят в один и тот же кусок параболы

$$t = \frac{s^2}{2(\phi - b)} + \frac{\phi - b}{2}, \quad s \in \left(0, \frac{\phi - b}{2}\right]. \tag{4.13}$$

Итак, образом квадрата G является криволинейный треугольник Γ, ограниченный линиями (4.11)–(4.13) и накрываемый дважды.

Исследуем функцию H, определяемую в (4.10), на экстремумы в треугольнике Γ. Необходимые условия экстремума

$$\frac{\partial H}{\partial s} = -12st - 12bs - 4t^2 + 6s^2 - 6bt - 6b^2 + 2F'(b + 2t - 2s) = 0,$$
$$\frac{\partial H}{\partial t} = 24t^2 - 6s^2 + 24bt + 6b^2 - 8st - 6bs - 2F'(b + 2t - 2s) = 0$$

приводят к соотношениям

$$F'(b + 2t - 2s) = 6st + 6bs + 2t^2 - 3s^2 + 3bt + 3b^2 = 12t^2 - 3s^2 + 12bt + 3b^2 - 4st - 3bs,$$

которые выполняются только при $t = s$. Поэтому внутри треугольника Γ точек экстремума нет. Остается проверить выполнение неравенства (4.10) на границе треугольника Γ.

1. На стороне

$$s = 0, \quad t \in \left(0, \frac{\phi - b}{2}\right]$$

значения

$$H(0,t) = 8t^3 + 12bt^2 + 6b^2t - F(b + 2t) = F(b + 2t) - F(b + 2t) = 0 \geq 0.$$

2. На стороне

$$t = s, \quad s \in (0, \phi - b]$$

значения

$$H(s,s) = 8s^3 - 6s^3 + 12bs^2 - 6bs^2 + 6b^2s - 4s^3 + 2s^3 - 6bs^2 - 6b^2s - F(b) = 0 \geq 0.$$

3. Для проверки неравенства (4.10) на стороне

$$t = \frac{s^2}{2(\phi - b)} + \frac{\phi - b}{2}, \quad s \in \left(0, \frac{\phi - b}{2}\right]$$

удобнее вернуться к переменным (y, z) и рассмотреть одну из двух сторон квадрата G:

$$y = \phi - b, \quad z \in (0, \varphi - b] \quad \text{или} \quad z = \phi - b, \quad y \in (0, \phi - b].$$

Для определенности будем рассматривать сторону $z = \phi - b$, $y \in (0, \phi - b]$. Нужно доказать неравенство (4.8) при $z = \phi - b$. Запишем левую часть этого неравенства в виде

$$\left(1 - \sqrt{\frac{y}{z}}\right) F(b + z) + \left(1 - \sqrt{\frac{z}{y}}\right) F(b + y) - F((b + (\sqrt{z} - \sqrt{y})^2) \geq 0,$$

и сделаем замены

$$t = \sqrt{\frac{y}{z}}, \quad 0 < t < 1, \quad a = z, \quad a > 0. \tag{4.14}$$

В результате получим неравенство

$$(1 - t)F(b + a) - \frac{1 - t}{t} F(b + at^2) - F(b + a(1 - t)^2) \geq 0.$$

Здесь

$$F(b + u) = uh(u), \quad \text{где} \quad h(u) = u^2 + 3bu + 3b^2.$$

Соответственно имеем неравенство

$$(1-t)ah(a) - (1-t)ath(at^2) - a(1-t)^2 h(a(1-t)^2) \geq 0,$$

которое после деления на $(1-t)a$ принимает вид

$$h(a) - th(at^2) - (1-t)h(a(1-t)^2) \geq 0. \tag{4.15}$$

Преобразуем левую часть этого неравенства:

$$\begin{aligned} h(a) - th(at^2) - (1-t)h(a(1-t)^2) &= (a^2 + 3ba + 3b^2) - t(a^2t^4 + 3bat^2 + 3b^2) - \\ &- [a^2(1-t)^4 + 3ba(1-t)^2 + 3b^2] + t[a^2(1-t)^4 + 3ba(1-t)^2 + 3b^2] = \\ &= a[a + 3b - a(1-t)^4 - 3b(1-t)^2 + a(4t^4 - 6t^3 + 4t^2 - t) + 3b(2t^2 - t)]. \end{aligned}$$

Поэтому неравенство (4.15) эквивалентно неравенству

$$g := a + 3b - a(1-t)^4 - 3b(1-t)^2 + a(4t^4 - 6t^3 + 4t^2 - t) + 3b(2t^2 - t) \geq 0. \tag{4.16}$$

Изучим зависимость значений g от параметра a на промежутке $a > 0$. Производная

$$\frac{\partial g}{\partial a} = 1 - (1-t)^4 + 4t^4 - 6t^3 + 4t^2 - t = t(3t^3 - 2t^2 - 2t + 3) > 0,$$

поэтому g возрастает на промежутке $a > 0$ и ее величина

$$g(a) > g(0) = 3b - 3b(1-t)^2 + 3b(2t^2 - t) = 3b(t^2 + t) > 0.$$

Таким образом, (4.16), а вместе с ним и (4.8) при $z = \phi - b$, являются верными неравенствами. Это завершает доказательство неравенства (4.10), и можно утверждать, что функция $Z_-(\xi,\tau) = -2\sqrt{\Pi_0 Q_0}$ является нижним решением задачи (4.1), (4.2).

Обе барьерные функции $Z_-(\xi,\tau)$ и $Z_+(\xi,\tau)$ удовлетворяют экспоненциальным оценкам убывания вида (3.5), поэтому решение $P_0(\xi,\tau)$ задачи (4.1), (4.2) также удовлетворяет оценке того же вида. Теорема доказана.

Теорема 4.2. *Пусть функция $F(u)$ имеет вид $F(u) = u^3 - b^3$, где $b > 0$, и граничное значение $\phi(0,0) > \overline{u}_0(0,0)$. Тогда задачи* (3.3), (3.4) *имеют решения $P_k(\xi,\tau)$, удовлетворяющие экспоненциальным оценкам убывания вида* (3.5).

Доказательство. Исследуем знак производной $F'(\overline{u}_0 + \Pi_0 + Q_0 + P_0)$. Так как

$$P_0 \geq -2\sqrt{\Pi_0 Q_0},$$

то

$$\Pi_0 + Q_0 + P_0 \geq \Pi_0 + Q_0 - 2\sqrt{\Pi_0 Q_0} = (\sqrt{\Pi_0} - \sqrt{Q_0})^2 \geq 0.$$

Поэтому величина

$$\overline{u}_0 + \Pi_0 + Q_0 + P_0 = b + \Pi_0 + Q_0 + P_0 \geq b,$$

вследствие чего производная $F'(\overline{u}_0 + \Pi_0 + Q_0 + P_0) > 0$ и задачи (3.3), (3.4) имеют решения с оценками вида (3.5). Теорема доказана.

Отметим, что функции $P_k^*(\xi_*,\tau)$, $k \geq 0$, определяются аналогично.

5. ОЦЕНКА ОСТАТОЧНОГО ЧЛЕНА

Итак, ряд (1.3) оказывается полностью построенным при дополнительном условии 6:

Условие 6. В угловых точках $(k,0)$, $k = 0,1$, прямоугольника Ω функция $F(u)$ имеет вид $F(u) = u^3 - b^3$, где положительное число b не обязательно одно и то же, и граничные значения $\phi(k,0) > \overline{u}_0(k,0)$.

Теорема 5.1. *Пусть выполнены условия* 1–6. *Тогда для достаточно малых* ε *задача* (0.1)–(0.3) *имеет решение* $u(x,t,\varepsilon)$, *для которого ряд*

$$\sum_{k=0}^{\infty}\varepsilon^k(\overline{u}_k(x,t)+\Pi_k(x,\tau)+Q_k(\xi,t)+Q_k^*(\xi_*,t)+P_k(\xi,\tau)+P_k^*(\xi_*,\tau))$$

является асимптотическим представлением при $\varepsilon\to 0$ *в замкнутом прямоугольнике* $\overline{\Omega}$.

Доказательство основано на разрешимости задач для пограничных функций Π_k, Q_k, Q_k^*, P_k и P_k^* при $k\geq 1$ и повторяет доказательство соответствующей теоремы из статьи [3].

Выражаю искреннюю благодарность В.Ф. Бутузову за плодотворное обсуждение полученных результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Денисов И.В.* Первая краевая задача для квазилинейного сингулярно возмущенного параболического уравнения в прямоугольнике // Ж. вычисл. матем. и матем. физ. 1996. Т. 36. № 10. С. 56–72.
2. *Денисов И.В.* Оценка остаточного члена в асимптотике решения краевой задачи // Ж. вычисл. матем. и матем. физ. 1996. Т. 36. № 12. С. 64–67.
3. *Денисов И.В.* Угловой пограничный слой в краевых задачах для сингулярно возмущенных параболических уравнений с квадратичной нелинейностью // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 2. С. 255–274.
4. *Денисов А.И., Денисов И.В.* Угловой пограничный слой в краевых задачах для сингулярно возмущенных параболических уравнений с нелинейностями // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 1. С. 102–117.
5. *Денисов А.И., Денисов И.В.* Угловой пограничный слой в краевых задачах для сингулярно возмущенных параболических уравнений с немонотонными нелинейностями // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 9. С. 1581–1590.
6. *Денисов И.В.* Угловой пограничный слой в краевых задачах для сингулярно возмущенных параболических уравнений с монотонной нелинейностью // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 4. С. 1–11.
7. *Васильева А.Б., Бутузов В.Ф.* Асимптотические методы в теории сингулярных возмущений. М.: Высш. школа, 1990.
8. *Amann H.* On the existence of positive solutions of nonlinear elliptic boundary value problems // Indiana Univ. Math. J. 1971. V. 21. № 2. P. 125–146.
9. *Sattinger D.H.* Monotone methods in nonlinear elliptic and parabolic boundary value problems // Indiana Univ. Math. J. 1972. V. 21. № 11. P. 979–1000.
10. *Amann H.* // Nonlinear Analysis: coll. of papers in honor of E.H. Rothe / Ed. by L. Cesari et al. New York etc: Acad press, cop. 1978. XIII. P. 1–29.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 268–280

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.634

МОДЕЛИРОВАНИЕ ЛАМИНАРНО-ТУРБУЛЕНТНОГО ПЕРЕХОДА С ПРИМЕНЕНИЕМ ДИССИПАТИВНЫХ ЧИСЛЕННЫХ СХЕМ[1)]

© 2021 г. И. В. Егоров[1,2], Н. К. Нгуен[1,*], Т. Ш. Нгуен[3], П. В. Чувахов[1,2]

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, МФТИ, Россия

[2] 140180 Жуковский, М.о., ул. Жуковского, 1, Центральный аэрогидродинамический институт им. Жуковского, Россия

[3] Ханой, район Бак Ту Лием, Улица Кау Дьен, № 298, Ханойский индустриальный университет, Вьетнам

**e-mail: nguyennhucan528@gmail.com*

Поступила в редакцию 15.03.2020 г.
Переработанный вариант 18.08.2020 г.
Принята к публикации 16.09.2020 г.

Продемонстрирована возможность применения монотонной схемы сквозного счета с низким порядком аппроксимации для моделирования процесса ламинарно-турбулентного перехода. Рассмотрена задача моделирования ламинарно-турбулентного перехода в сверхзвуковом пограничном слое над плоской пластиной, число Маха 3. Результаты расчетов сопоставлены с результатами других авторов, которые получены с применением низкодиссипативных схем. Сравниваются спектральные характеристики возмущений в области их линейного и нелинейного развития, а также структура переходного течения и характеристики осредненного пограничного слоя. Библ. 6. Фиг. 13.

Ключевые слова: монотонная схема, прямое численное моделирование, возмущение, ламинарно-турбулентный переход, сверхзвуковой пограничный слой, нелинейный распад.

DOI: 10.31857/S0044466921020083

1. ВВЕДЕНИЕ

Разработка перспективных летательных аппаратов опирается на результаты аэродинамических исследований. Экспериментальная часть обычно сопряжена с большими финансовыми затратами на постановку и проведение эксперимента в аэродинамических трубах, а получаемые в эксперименте данные ограничены. В отличие от эксперимента численное моделирование нестационарного течения сжимаемого газа позволяет изучать обтекание тел произвольной конфигурации, выявлять тонкую структуру наблюдаемых явлений и получать результаты, которые затруднительно получить экспериментальным путем. На базе результатов моделирования рассчитываются аэротермодинамические коэффициенты – коэффициенты давления, трения и теплообмена. Последний становится критически важным в случае больших сверхзвуковых и гиперзвуковых скоростей потока, в особенности, когда течение в пограничном слое турбулизуется и коэффициенты трения и теплопередачи к поверхности возрастают в разы.

Трубные данные по положению ламинарно-турбулентного перехода (ЛТП) непостоянны, так как зависят от фона возмущений в потоке конкретной аэродинамической установки. Поэтому особую ценность представляют результаты прямого численного моделирования течений на режиме ЛТП, когда такой фон можно строго контролировать. К сожалению, для практически важных режимов пространственно-временные затраты на проведение такого моделирования велики. Использование высокопроизводительных многопроцессорных вычислительных кластеров (суперЭВМ) позволяет проводить лишь отдельные расчеты, целью которых является исследование линейных и нелинейных механизмов, лежащих в основе ЛТП. Понизить затраты можно, используя численные схемы невысокого порядка аппроксимации.

[1)]Работа выполнена при финансовой поддержке РНФ (код проекта 19-79-10132) в МФТИ с использованием оборудования центра коллективного пользования “Комплекс моделирования и обработки данных исследовательских установок мега-класса” НИЦ “Курчатовский институт”, http://ckp.nrcki.ru/.

Монотонные схемы сквозного счета диссипативны. Это позволяет устойчиво рассчитывать течения с ударными волнами, отрывными областями, пограничными слоями и другими особенностями с учетом их взаимодействия. Избыточная диссипация приводит к численному затуханию малых возмущений. Однако темпы роста возмущений в неустойчивых пограничных слоях могут заметно превосходить эффекты, связанные с численной диссипацией. Настоящая работа демонстрирует успешное применение монотонной схемы второго порядка точности по пространству и времени для моделирования ламинарно-турбулентного перехода сверхзвукового пограничного слоя на плоской пластине, число Маха 3. Результаты моделирования сопоставляются с аналогичными результатами, полученными с применением значительно менее диссипативного метода [1], который основан на схеме четвертого порядка точности по продольному и нормальному к поверхности направлениям, и спектрального метода в боковом направлении; интегрирование по времени проводится по методу Рунге–Кутты четвертого порядка аппроксимации. Начало настоящей работы положено в [2]. В отличие от [2] далее детально сравниваются спектральные характеристики возмущений в области их линейного и нелинейного развития, а также интегральные характеристики течения.

2. ПОСТАНОВКА ЗАДАЧИ

2.1. Краткое описание численного метода

В настоящей работе использован пакет программ HSFlow++ [3], с помощью которого проводится прямое численное моделирование течений в рамках уравнений Навье–Стокса. Дифференциальные уравнения решаются в криволинейной системе координат (ξ, η, ζ) в безразмерном дивергентном виде:

$$\frac{\partial \mathbf{Q}}{\partial t} + \frac{\partial \mathbf{E}}{\partial \xi} + \frac{\partial \mathbf{G}}{\partial \eta} + \frac{\partial \mathbf{F}}{\partial \zeta} = 0.$$

Для численного решения используются безразмерные значения. Декартовы координаты $x^* = xL$, $y^* = yL$, $z^* = zL$ отнесены к характерному масштабу длины L, время $t^* = tL/V_\infty$ – к характерному масштабу времени L/V_∞, компоненты вектора скорости $u^* = uV_\infty$, $v^* = vV_\infty$, $w^* = wV_\infty$ – к модулю вектора скорости набегающего потока V_∞, давление $p^* = p(\rho_\infty V_\infty^2)$ – к удвоенному скоростному напору набегающего потока, остальные газодинамические переменные – к их значениям в набегающем потоке. Звездочка в верхнем индексе обозначает размерную величину; если звездочка отсутствует, переменная предполагается безразмерной. Символ “∞” обозначает значение переменной в набегающем потоке.

Для аппроксимации дифференциальных уравнений используются полностью неявный метод конечного объема и схема второго порядка аппроксимации по времени:

$$\frac{3\mathbf{Q}^{n+1}_{i,j,k} - 4\mathbf{Q}^{n}_{i,j,k} + \mathbf{Q}^{n-1}_{i,j,k}}{\Delta t} + \frac{\mathbf{E}^{n+1}_{i+\frac{1}{2},j,k} - \mathbf{E}^{n+1}_{i-\frac{1}{2},j,k}}{h_\xi} + \frac{\mathbf{G}^{n+1}_{i,j+\frac{1}{2},k} - \mathbf{G}^{n+1}_{i,j-\frac{1}{2},k}}{h_\eta} + \frac{\mathbf{F}^{n+1}_{i,j,k+\frac{1}{2}} - \mathbf{F}^{n+1}_{i,j,k-\frac{1}{2}}}{h_\zeta} = 0.$$

При определении конвективных потоковых величин на гранях ячейки сетки, например, $\mathbf{E}^{n+1}_{i+1/2,j,k}$, потоки расщепляются по направлениям. Для каждого направления определяется матрица Якоби ($A = \partial\mathbf{E}/\partial\mathbf{Q}$ для направления ξ), которая диагонализируется в виде $A = B\Lambda B^{-1}$, где B – матрица, составленная из правых собственных векторов, а Λ – диагональная матрица собственных значений матрицы A**.** Конвективная составляющая потоковых величин $\mathbf{E}$, $\mathbf{G}$, $\mathbf{F}$ на грани ячейки аппроксимируется с использованием монотонной схемы типа Годунова:

$$\mathbf{E}_{i+\frac{1}{2}} = \frac{1}{2}[\mathbf{E}(\mathbf{Q}_L) + \mathbf{E}(\mathbf{Q}_R) - B_{LR}\Lambda(\varphi(\lambda_{LRi}))B^{-1}_{LR}(\mathbf{Q}_R - \mathbf{Q}_L)],$$

где нижними индексами L и R отмечены величины, рассчитанные справа и слева на рассматриваемой грани с использованием восстановленных с помощью процедуры реконструкции значений газодинамических переменных. Например, для грани $i + 1/2$ индекс L соответствует ячейке i, а индекс R – ячейке $i + 1$. Нижним индексом LR отмечены величины, вычисляемые с помощью приближенного метода Роу решения задачи Римана о распаде произвольного разрыва. Модификация собственных значений $\varphi(\lambda)$ обеспечивает физически корректное изменение энтропии на разрывах. В работе использована процедура реконструкции по методу WENO-3.

Для аппроксимации диффузионной составляющей векторов $\mathbf{E}$, $\mathbf{G}$, $\mathbf{F}$ на грани элементарной ячейки применена разностная схема типа центральных разностей второго порядка точности:

$$\left.\frac{\partial q}{\partial \xi}\right|_{i+\frac{1}{2},j,k} = \frac{1}{h_\xi}(q_{i+1,j,k} - q_{i,j,k}),$$

$$\left.\frac{\partial q}{\partial \eta}\right|_{i+\frac{1}{2},j,k} = \frac{1}{4h_\eta}(q_{i+1,j+1,k} + q_{i,j+1,k} - q_{i+1,j-1,k} - q_{i,j-1,k}),$$

$$\left.\frac{\partial q}{\partial \zeta}\right|_{i+\frac{1}{2},j,k} = \frac{1}{4h_\zeta}(q_{i+1,j,k+1} + q_{i,j,k+1} - q_{i+1,j,k-1} - q_{i,j,k-1}),$$

где q – любая из неконсервативных (“примитивных”) зависимых переменных задачи, газодинамическая функция u, v, w, p или T.

После аппроксимаций уравнений Навье–Стокса и граничных условий интегрирование исходных уравнений в частных производных сводится к решению системы нелинейных алгебраических уравнений $R(\mathbf{U}) = 0$, где R – оператор дискретизации, вычисляющий вектор невязки согласно аппроксимации уравнений, $\mathbf{U}$ – вектор искомых неконсервативных переменных (u, v, w, p, T) (компоненты скорости, давление, температура) во всех узлах расчетной сетки. Длина вектора $\mathbf{U}$ составляет n_qN, где N – полное число узлов расчетной сетки, включая граничные, n_q – число неизвестных в каждом узле ($n_q = 5(u, v, w, p, T)$ в трехмерной постановке и $n_q = 4(u, v, p, T)$ в двухмерной). Система сеточных уравнений $R(\mathbf{U}) = 0$ решается с помощью модифицированного метода Ньютона–Рафсона

$$\mathbf{U}^{[k+1]} = \mathbf{U}^{[k]} - \tau^{[k+1]}(J^{[k_0]})^{-1}R(\mathbf{U}^{[k]}),$$

где k, k_0 – номера итераций по нелинейности, $k_0 \le k$, $J^{[k_0]} = (\partial R/\partial \mathbf{U})^{[k_0]}$ – матрица Якоби системы нелинейных уравнений, $R(\mathbf{U}^{[k]})$ – вектор невязки, τ – параметр регуляризации. Здесь выражение $(J^{[k_0]})^{-1}R(\mathbf{U}^{[k]}) \equiv Y^{[k]}$ является решением линейной системы уравнений

$$(J^{[k_0]})Y^{[k]} = R(\mathbf{U}^{[k]}).$$

Параметр регуляризации метода Ньютона относительно начального приближения $\tau^{[k]}$ определяется по формуле

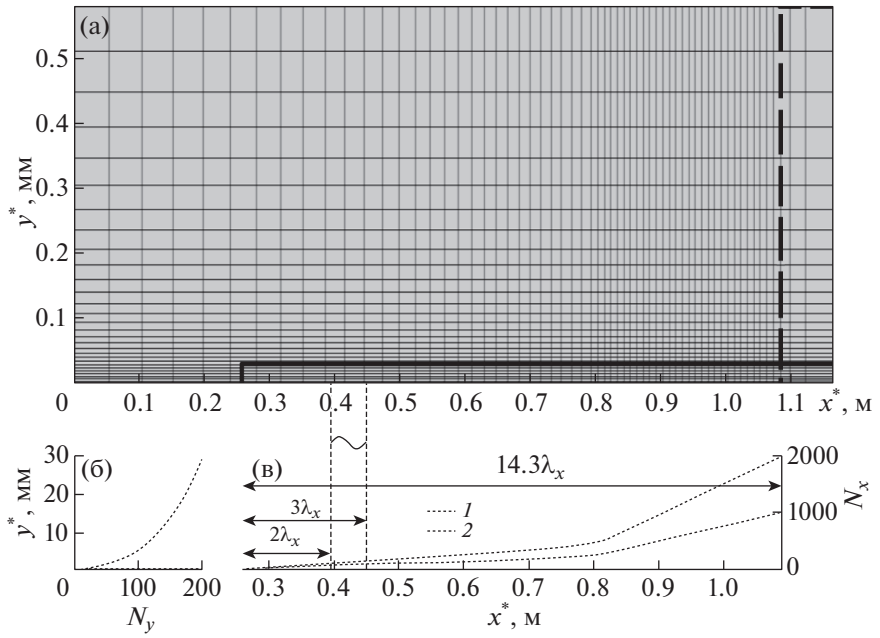
$$\tau^{[k+1]} = \frac{(Y^{[k]} - Y^{[k-1]})Y^{[k]}}{(Y^{[k]} - Y^{[k-1]})^2}.$$

По мере сходимости итерационного процесса $\tau^{[k]} \to 1$, а скорость сходимости теоретически стремится к квадратичной.

Получение аналитического вида матрицы Якоби J для рассматриваемой численной схемы, включающей решение задачи Римана о распаде разрыва, представляется весьма трудоемким. В настоящей программе применяется универсальный метод формирования матрицы $J^{[k_0]} = (\partial R/\partial \mathbf{U})^{[k_0]}$ на итерации по нелинейности k_0 с помощью процедуры конечных приращений вектора невязки R по вектору искомых переменных $\mathbf{U}$. При этом m-й столбец матрицы $J^{[k_0]}$ вычисляется как в виде

$$J_m^{[k_0]} = \frac{R(\mathbf{U}^{[k_0]} + \varepsilon \mathbf{e}_m) - R(\mathbf{U}^{[k_0]})}{\varepsilon}, \quad \varepsilon = 10^{-8}, \quad m = 1, \ldots, n_qN,$$

где $\mathbf{e}_m$ – единичный вектор длины n_qN, полностью состоящий из 0, кроме единственной 1 на позиции m. Такая методика вычисления Якобиана применима к произвольной системе сеточных уравнений.



Фиг. 1. Постановка задачи: (а) – расчетная область и сетка (показана каждая 10-я сеточная линия); (б) – сгущение сетки по нормали к стенке; (в) – сгущение сетки в продольном направлении: *1* – сетка 80 миллионов узлов, *2* – сетка 20 миллионов узлов.

2.2. Параметры потока и постановка расчетной задачи

Рассматривается номинально двухмерное течение над заостренной плоской пластиной при числе Маха набегающего потока $M_\infty = 3$, температуре набегающего потока $T_\infty = 103.6$ K. Расчет эволюции возмущений проводится в подобласти; процедура расчета аналогична процедуре, описанной в [4]. Число Рейнольдса составляет $\mathrm{Re}_{1,\infty} = 2.181\times10^6$ м$^{-1}$. Число Прандтля принимается постоянным: $\mathrm{Pr} = \mu c_p/\lambda = 0.71$. Уравнения Навье–Стокса замыкаются уравнением состояния $\gamma M_\infty^2 p = \rho T$, где $\gamma = 1.4$ – показатель адиабаты. Динамический коэффициент молекулярной вязкости рассчитывается по формуле Сазерленда: $\mu = (1+T_\mu)/(T+T_\mu)T^{3/2}$, где $T_\mu = T_\mu^*/T_\infty^* = 110.4\ \text{K}/103.6\ \text{K} \approx 1.07$.

Численное интегрирование проводится в прямоугольной области, показанной на фиг. 1. На входной и верхней границах фиксируются безразмерные параметры набегающего потока: $(u,v,w,p,T) = (1,0,0,1/\gamma M_\infty^2,1)$. Для стационарных расчетов стенка принимается теплоизолированной с условиями прилипания на ней. Выходной границе предшествует буферная зона с укрупненными ячейками по x и y для демпфирования выходящих через границу возмущений. На выходной границе накладываются мягкие условия как линейная экстраполяция примитивных переменных из расчетной области. На боковых границах $z_{\min}$ и $z_{\max}$ ставятся условия симметрии.

Расчет проводится следующим образом. Во-первых, двухмерное стационарное невозмущенное течение над плоской пластиной рассчитывается до достижения невязкой величины 10^{-8}. Во-вторых, из полученного решения вырезается подобласть, в которой далее будет моделироваться развитие возмущений; на новых входных границах подобласти фиксируются газодинамические величины из расчета на первом шаге; стационарное поле устанавливается дополнительно до полной сходимости (величина невязки не превосходит 10^{-8}). В-третьих, полученное в подобласти стационарное поле дублируется в третьем поперечном направлении z; распределение температуры по поверхности *фиксируется*; в пограничный слой вводятся возмущения типа “вдув–отсос” по процедуре, описанной ниже. Нестационарный расчет проводится до установления квазистационарного режима течения. При таком подходе поверхность пластины является адиабатической, но пульсации температуры на поверхности отсутствуют.

2.3. Генератор возмущений

Генератор возмущений моделируется при $x^* \in [x_1^*, x_2^*] = [0.394, 0.452]$ м в соответствии с работой [1]. В этом диапазоне нормальная составляющая вектора скорости имеет вид

$$v(x, y = 0, t) = A(t)v_p(x_p)\cos(\beta_0 z)\cos(-\omega_0 t),$$

где

$$v_p = \begin{cases} 1.5^4(1 + x_p)^3(3(1 + x_p)^2 - 7(1 + x_p) + 4), & -1 \le x_p \le 0, \\ -1.5^4(1 - x_p)^3(3(1 - x_p)^2 - 7(1 - x_p) + 4), & 0 \le x_p \le 1, \end{cases} \quad x_p = \frac{2x - (x_2 + x_1)}{x_2 - x_1},$$

$$A(t) = \varepsilon \begin{cases} 0, & t < 0, \\ 0.1^{((T-t)/(0.9T))^2}, & 0 \le t \le T, \\ 1, & t > T, \end{cases}$$

где $T = 2\pi/\omega$, $A(t)$ – амплитуда, $\varepsilon = 0.00573$. Остальные параметры потока в области генератора вычисляются как для случая стенки без генератора. Возмущение с частотой $\omega_0^*/2\pi = f_0^* = 6.36$ кГц и волновым числом $\beta_0^* = 211.52$ м$^{-1}$ будем называть фундаментальным.

Как будет показано далее, результаты настоящих расчетов хорошо согласуются с результатами работы [1] как качественно, так и количественно. Однако амплитуда возмущений ε в настоящей работе отличается от амплитуды из работы [1] практически вдвое (в работе [1] $\varepsilon = 0.003$) и подобрана для совпадения положения начала ЛТП. Как показано ниже, численная диссипация не влияет на положение ЛТП в настоящей работе. Поэтому, вероятно, в работе [1] амплитуда возмущений указана ошибочно.

2.4. Расчетная сетка

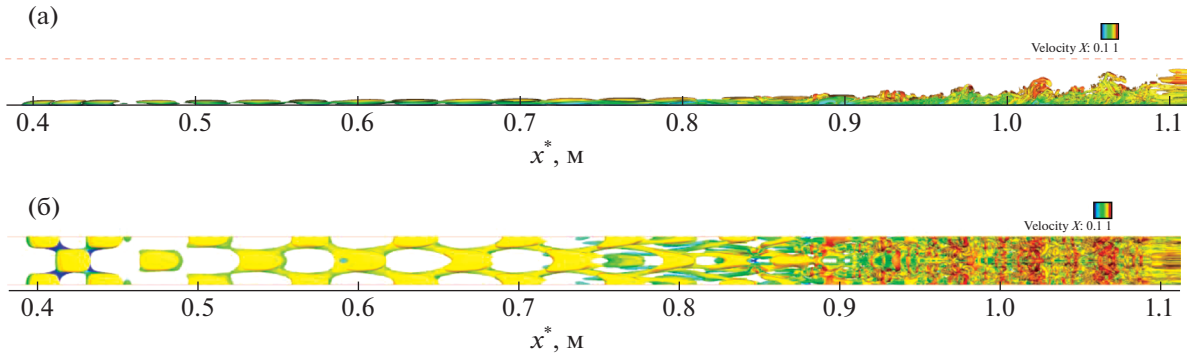
Характерный масштаб длины выбран как $L = 0.7239$ м. Продольный размер буферной зоны, которая ограничена пунктирным прямоугольником на фиг. 1, составляет полторы длины волны фундаментального возмущения, или $1.5\lambda_x$, где $\lambda_x = x_2^* - x_1^*$.

На фиг. 1 показаны расчетная область и расчетная сетка (вид сбоку). Подобласть для проведения основных нестационарных расчетов ограничена сплошным прямоугольником и начинается на расстоянии $x_0^* = 0.258$ м от передней кромки пластины. Длина подобласти в 14.3 раза больше продольной длины волны фундаментального возмущения. Высота подобласти выбрана $y_H^* = 0.03$ м, что составляет не менее пяти местных толщин пограничного слоя на выходной границе. Размер подобласти в боковом направлении составляет одну длину волны λ_z в поперечном направлении, где $\lambda_z = 2\pi/\beta_0^* \approx 0.0297$ м.

Расчетная сетка представлена на фиг. 1а, соответствующие сеточные сгущения – на фиг. 1б и 1в. Основные расчеты настоящей работы проведены на сетке 80 миллионов узлов (подробная сетка). Эта сетка соответствует сетке из работы [1] в плоскости $x0y$. Сеточные линии распределены равномерно в поперечном направлении. На подробной сетке количество точек вдоль оси z составляет 201. Грубая сетка имеет вдвое меньше узлов по x и по z, чем подробная сетка. В вертикальном направлении количество точек для обеих сеток одинаково; поперек пограничного слоя приходится не менее 100 точек. На подробной сетке возмущение разрешено в боковом направлении (по z) 201 точкой на длину волны по z, а в продольном направлении (по x) – 320 точками. Стоит отметить, что для распространения монохроматической акустической волны в равномерном потоке требуется около 40 точек на длину волны, чтобы добиться близкого к естественному уровня вязкого затухания волны для используемого численного метода. Поэтому на построенных сетках численная диссипация фундаментального возмущения незначительна.

2.5. Анализируемые величины

Данные для обработки и сравнения собираются начиная с момента безразмерного времени $t = 2.261$ после ввода возмущений в пограничный слой, когда установился квазипериодический режим течения. В следующем разделе анализируются свойства переходного течения.



Фиг. 2. Визуализация вихревых структур пограничного слоя с помощью изоповерхностей Q-критерия, $Q = 5$: (а) – вид сбоку со стороны $+z$; (б) – вид сверху со стороны $+y$. Окраска соответствует величине продольной компоненты вектора скорости. Буферная зона начинается при $x^* \approx 1.09$ м.

Анализ проводится для спектрального состава возмущений, где сопоставляются амплитуды отдельных гармоник Фурье или максимумы этих амплитуд по нормали к поверхности в рассматриваемом сечении $x = \text{const}$. Фурье-анализ и обработка нестационарных результатов выполнены с помощью возможностей языка программирования Python (библиотека numpy). Результат работы процедур быстрого преобразования Фурье по времени и координате z нормирован на $N_t \times N_z/4$, где N_t и N_z – количество точек анализируемого сигнала по времени и по координате z соответственно. В настоящей работе исследованы амплитуды гармоник пульсаций продольной компоненты скорости, давления, температуры, а также максимумы этих величин по нормали к поверхности.

Вихревая структура полей течения визуализируется с помощью Q-критерия: $Q = 0.5(\Omega_{ij}\Omega_{ij} - S_{ij}S_{ij})$, $S_{ij} = 0.5(\partial_j u_i + \partial_i u_j)$, $\Omega_{ij} = 0.5(\partial_j u_i - \partial_i u_j)$, u_i – компоненты вектора скорости (для записи использованы тензорные обозначения; предполагается соглашение о суммировании по повторяющемуся в произведении индексу).

Также сопоставляются поля мгновенных значений продольной и поперечной компонент вектора завихренности и средние параметры течения (коэффициент трения, профили продольной компоненты скорости, осредненные по Фавру и по Рейнольдсу).

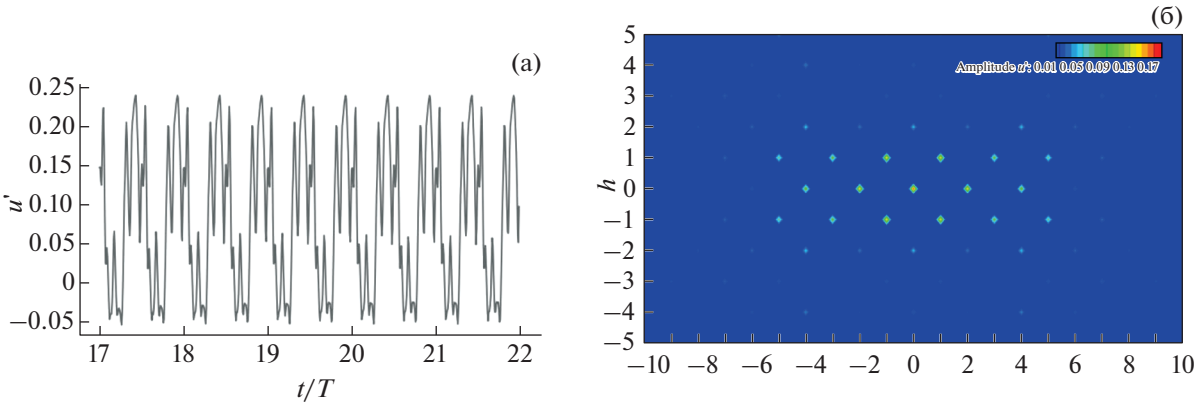
3. АНАЛИЗ РЕЗУЛЬТАТОВ

3.1. Структуры течения и спектральный состав возмущений

Мгновенная структура возмущенного течения представлена на фиг. 2 с помощью Q-критерия. Сразу за источником возмущений расположена область линейного развития возмущений, где они формируют X-образные структуры, усиливающиеся вниз по потоку. Вблизи $x^* \approx 0.6$ мм появляются первые признаки нелинейного взаимодействия – возмущения начинают искажаться. При $x^* \in (0.75, 0.85)$ м наблюдается интенсивный нелинейный распад возмущений. За этой областью формируется зона молодой турбулентности с ростом мелкомасштабных вихрей. Эта зона развивается вниз по потоку.

Так как возбуждающие возмущения носят периодический характер с выделенной частотой, а нелинейные взаимодействия приводят к порождению кратных гармоник, то отклик пограничного слоя на такие возмущения также должен быть периодическим (квазистационарный режим течения). Для изучения спектральных свойств процесса ЛТП нестационарное течение сначала устанавливается до квазистационарного режима, а затем набирается статистика в течение пяти периодов квазигармонического режима. Пример квазистационарного сигнала в некоторой точке внутри пограничного слоя показан на фиг. 3а.

В каждом поперечном сечении $x^* = \text{const}$ возмущение можно представить через сумму гармонических колебаний путем преобразования Фурье. Для рассматриваемого течения вдоль плоской пластины разумно делать двухмерное преобразование Фурье по времени и поперечной координате для каждой линии $x^* = \text{const}$, $y^* = \text{const}$. Результат такого двухмерного преобразования можно представить в виде амплитуды гармоники $(f^*, \beta^*) = (hf_0^*, k\beta_0^*)$. Таким образом, результат



Фиг. 3. (а) – Квазистационарные пульсации продольной компоненты скорости, $u'(t, x_0^*, y_0^*, z_0^*)$, в точке $(x_0^*, y_0^*, z_0^*) = (0.9201,\ 0.0035,\ -0.0076)$; (б) – двухмерное преобразование Фурье поля $u'(t, x_0^*, y_0^*, z)$ на линии $(x_0^*, y_0^*) = (0.9201,\ 0.0035)$ м.

двухмерного преобразования Фурье можно представить в виде амплитуд двухмерных гармоник $\hat{u}_{hk}$. Пример представлен на фиг. 3б для начала области молодой турбулентности.

В описанной постановке фурье-спектр является симметричным, поэтому интерес представляет только часть спектра при $h \geq 0$, $k \geq 0$. Следует отметить, что пики спектра расположены в шахматном порядке, что объясняется квадратичным (нелинейным) взаимодействием возмущений друг с другом. Например, для стационарного возмущения и других четных частот $h = 0, 2, 4, \ldots$ наблюдаются только максимумы на четных волновых числах $k = 2, 4, 6, \ldots$, а для нечетных частот $h = 1, 3, 5, \ldots$ – на нечетных волновых числах $k = 1, 3, 5, \ldots$. Такая картина свойственна механизму наклонного распада, когда нелинейно (квадратично) взаимодействуют две гармоники с одинаковыми частотами, но противоположными по знаку волновыми числами. При этом частота удваивается, а волновое число обнуляется: $[1, 1] + [1, -1] \to [2, 0]$. Чем ближе гармоника к фундаментальной гармонике, тем выше ее амплитуда. Это связано как с тем, что нелинейный распад продвигается в область высоких частот постепенно, так и с тем, что имеется численная пространственно-временная диссипация используемого численного метода. Расчеты настоящей работы выполнены на двух разных сетках, одна из которых вдвое мельче в продольном и боковом направлении. Как показано ниже, для обеих сеток результаты оказываются близкими.

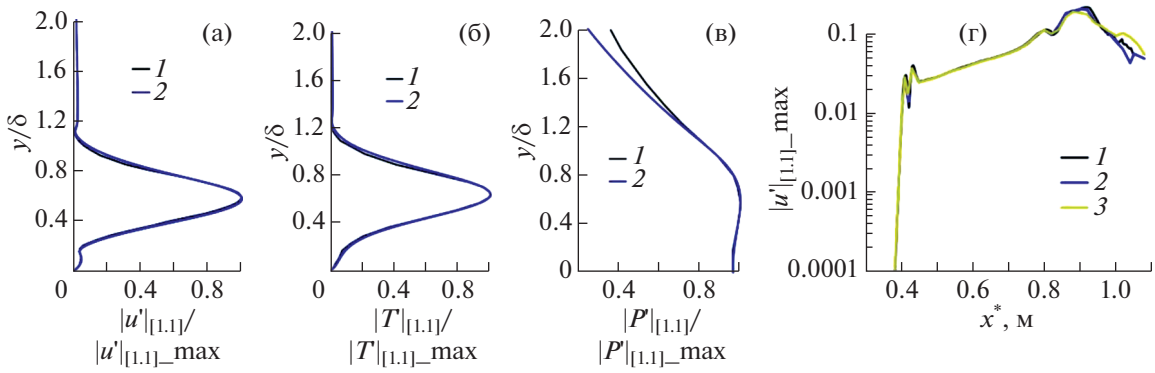
Ниже полученные результаты сопоставляются с результатами работы [1].

3.2. Линейный режим

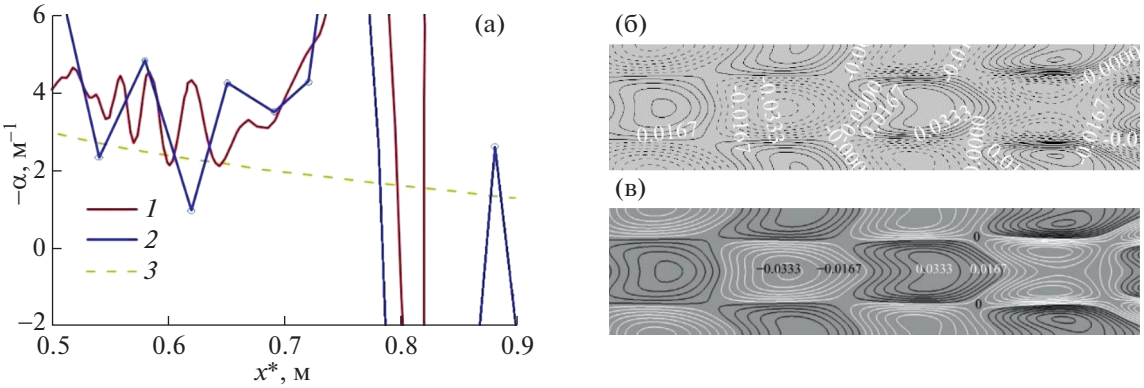
Рассмотрим линейный режим развития возмущений, который наблюдается примерно от $x^* = 0.4$ м до $x^* = 0.6$ м. На фиг. 4 амплитуды фундаментальной моды [1,1] пульсаций продольной компоненты вектора скорости u' в сечении $x^* = 0.5$ м, полученные в настоящей работе, сопоставляются с результатами работы [1]. Для данного случая и для других линий $x^* = \text{const}$, $y^* = \text{const}$ наблюдается хорошее согласование (фиг. 4а–4в).

Среди всех возможных линий $y^* = \text{const}$ для данного сечения $x^* = \text{const}$ можно выделить линию y_0^*, на которой амплитуда рассматриваемой гармоники максимальна. Для пульсаций u' или T' эта линия будет находиться в критическом слое пограничного слоя, $y_0/\delta \approx 0.65$. Эволюция такого максимума вниз по потоку хорошо согласуется с результатами [1] (фиг. 4г) даже на грубой сетке.

Инкремент роста возмущений является чувствительной к структуре стационарного решения характеристикой неустойчивого пограничного слоя. Рассмотрим эволюцию продольного инкремента роста возмущения продольной компоненты вектора скорости, $\alpha_i = -\dfrac{d}{dx}[\ln(u'_{\max}(x))]$. Здесь нижний индекс max обозначает, что рассматриваются максимальная по y фурье-амплитуда гармоники. Фиг. 5а демонстрирует хорошее согласование уровня инкрементов в се-



Фиг. 4. Амплитуды гармоники [1, 1]: (а)–(в) в сечении $x^* = 0.5$ м: (а) – для $|u'_{[1,1]}|$, (б) – для $|T'_{[1,1]}|$, (в) – для $|p'_{[1,1]}|$; (г) – максимальная по y^* амплитуда $|u'_{[1,1]}|$ в зависимости от x^*. *1* – работа [1], *2* – подробная сетка (HSFlow++), *3* – грубая сетка (HSFlow++).



Фиг. 5. (а) – Продольный инкремент усиления величины $|u'_{[1,1]}|$, *1* – работа [1], *2* – настоящая работа (80 миллионов узлов), *3* – результаты, полученные по линейной теории устойчивости (код Мэка); (б), (в) – мгновенный контур пульсации u' для сечения $x^* = 0.546$–0.67 м, $y^* = 2.3$ мм: (б) – работа [1], (в) – подробная расчетная сетка.

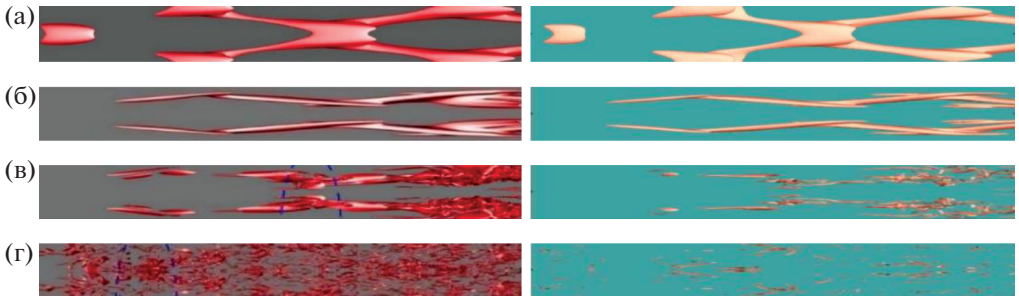
чениях $x = \text{const}$. Начиная с $x^* \approx 0.65$ м, величина инкремента сильно растет. Этот момент соответствует началу нелинейной стадии развития возмущений.

Также следует отметить хорошее согласование в структуре возмущений внутри пограничного слоя для различных сечений $y^* = \text{const}$, продемонстрированное на фиг. 5б, 5в.

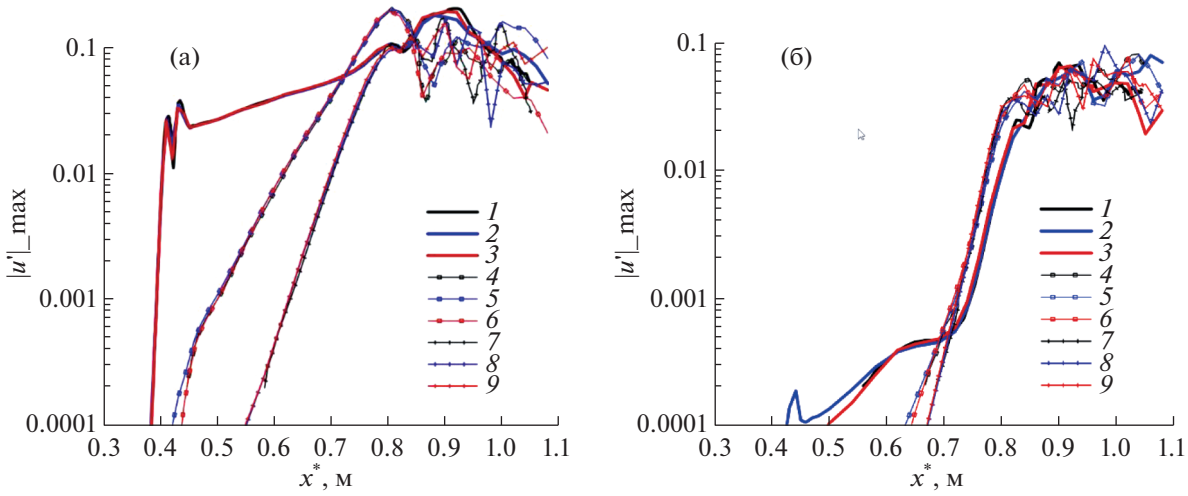
3.3. Нелинейный режим

Рассмотрим нелинейную стадию развития возмущений. Момент проявления нелинейного взаимодействия можно отметить по картинам Q-критерия, на которых заметно усиление “канатообразных” структур (фиг. 6). Для малых значений Q (15 и 100) настоящие результаты и результаты [1] хорошо согласуются. Однако в области молодой турбулентности, где появляются мелкомасштабные структуры и максимальная величина Q растет (10000 и 40000), диссипативная схема не позволяет идеально воспроизвести результаты низкодиссипативной схемы [1]. Наиболее вероятными причинами этого расхождения являются применение в [1] спектрального метода в боковом направлении и использование высокочастотных гармоник, которые располагаются вблизи частоты (волнового числа) Найквиста для используемой подробной расчетной сетки и поэтому плохо разрешаются.

Рассмотрим далее процесс нелинейного распада с помощью эволюции максимальных по y амплитуд гармоник возмущений. Механизм наклонного резонанса последовательный. Изначально растет наиболее неустойчивая фундаментальная наклонная волна [1, ±1], что обусловле-



Фиг. 6. Мгновенные изоповерхности Q-критерия, вид сверху; слева – работа [1], справа – настоящая работа, подробная сетка: (а) – для $Q = 15$, $x^* = 0.546$–0.670 м; (б) – для $Q = 100$, $x^* = 0.670$–0.798 м; (в) – для $Q = 10000$, $x^* = 0.798$–0.924 м; (г) – для $Q = 40\,000$, $x^* = 0.924$–1.051 м.

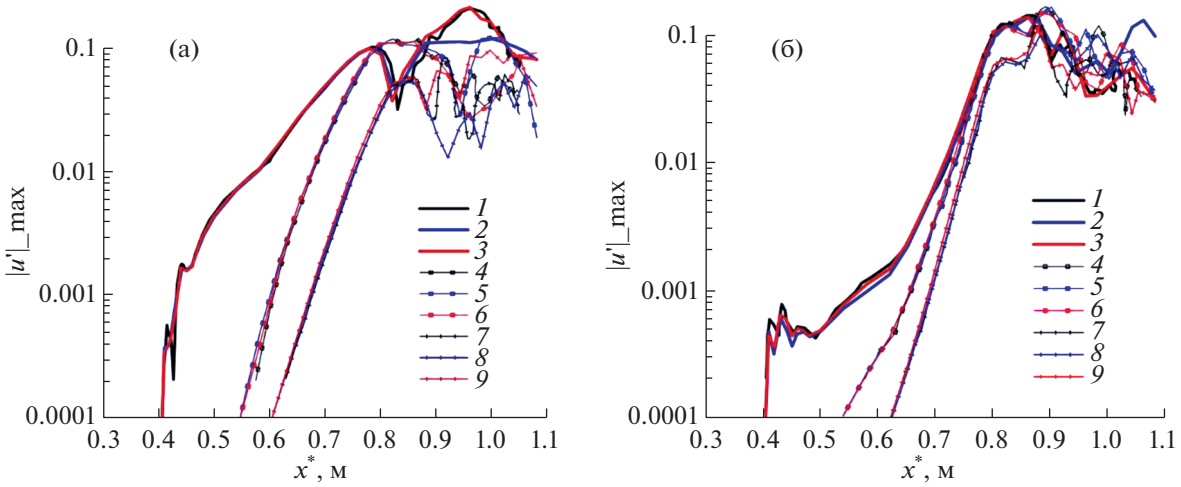


Фиг. 7. Эволюция максимальной по y амплитуды фурье-гармоники для нечетных частот: (а) – для $h = 1$, (б) – для $h = 3$: *1*, *4*, *7* – работа [1]; *2*, *5*, *8* – настоящая работа, грубая сетка; *3*, *6*, *9* – настоящая работа, подробная сетка; *1*, *2*, *3* – $k = 1$; *4*, *5*, *6* – $k = 3$; *7*, *8*, *9* – $k = 5$.

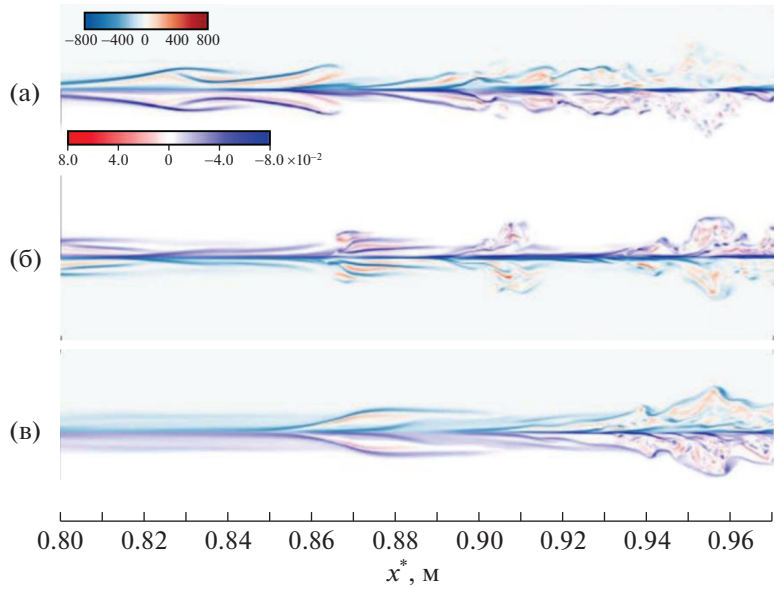
но чисто неустойчивостью пограничного слоя. При достижении некоторой критической амплитуды она начинает нелинейно взаимодействовать с собой, порождая кратные гармоники: $h = 0$ и $h = 2$, $k = 0$, $k = 2$, которые начинают расти благодаря нелинейному взаимодействию фундаментальных гармоник. Когда кратные гармоники достигают достаточных амплитуд, они начинают нелинейно взаимодействовать друг с другом и с фундаментальными гармониками, порождая все больше кратных гармоник. Такой процесс и его временная последовательность прослеживаются на фиг. 7 и фиг. 8, на которых изображена эволюция амплитуд гармоник вниз по потоку. Описанный механизм объясняет шахматную структуру спектра, приведенного на фиг. 3б.

Следует отметить хорошее совпадение в эволюции гармоник с работой [1] в областях линейного и слабонелинейного развития возмущений. Однако в области молодой турбулентности результаты, полученные на грубой сетке, начинают отличаться от результатов на подробной сетке. Последние продолжают хорошо согласовываться с результатами работы [1] для основных энергосодержащих частот $h = 0, 1$; $k = 1, 2$, и с ростом h и k начинает появляться небольшое рассогласование. При этом во всех случаях Фурье-амплитуды остаются на одном уровне и лишь разбегаются по фазе. Это может быть следствием накапливающейся ошибки более диссипативного метода. Однако в целом описанное сравнение результатов подтверждает надежность и применимость диссипативных схем для моделирования процесса ламинарно-турбулентного перехода.

На фиг. 9 приведены мгновенные структуры поперечного вектора завихренности в различных сечениях $z^* = \text{const}$ и сравнение полученных результатов (верхняя часть картины) с результатами работы [1] (нижняя часть картины после зеркального отражения). Видно, что вблизи $x^* = 0.865$ м



Фиг. 8. То же, что на фиг. 7, но для четных частот: (а) – для $h = 0$, (б) – для $h = 2$: *1*, *2*, *3* – $k = 2$; *4*, *5*, *6* – $k = 4$; *7*, *8*, *9* – $k = 6$.

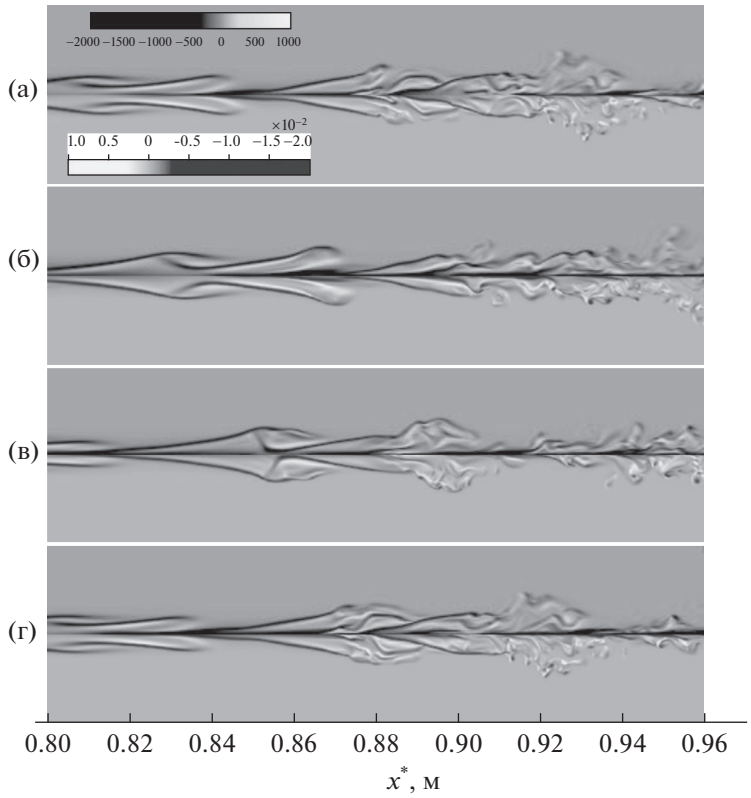


Фиг. 9. Мгновенное поле поперечного вектора завихренности в момент времени $t = 2.82891$: (а) – для $z^* = -0.0092$ м, (б) – для $z^* = -0.0047$ м, (в) – для $z^* = -0.0017$ м. Сверху – настоящая работа, подробная сетка; внизу – [1].

появляются мелкомасштабные структуры, свойственные развитой турбулентности. Детальное сравнение показывает, что вихри, полученные в настоящей работе, менее интенсивны по сравнению с вихрями из работы [1], и содержат меньше мелкомасштабных вихревых структур, что обсуждалось выше. Основные крупномасштабные структуры хорошо совпадают с [1].

Рассмотрим временную визуализацию процесса развития мелкомасштабных структур в конкретном сечении $z = \text{const}$. На фиг. 10 приведены мгновенные структуры поперечного вектора завихренности в различные моменты времени. На фиг. 10 дано сопоставление с результатами работы [5], в которой параметры задачи совпадают с параметрами работы [1], но расчеты проводились на измельченной сетке 211 млн. узлов. По результатам фиг. 10 можно сделать вывод о том, что скорость распространения вихрей внутри пограничного слоя составляет ≈0.7.

Фиг. 10 также демонстрирует развитие мелкомасштабных структур. Вихрь в точке $x^* = 0.84$ м (фиг. 10а) передвигается через точку $x^* = 0.87$ м (фиг. 10б) и доходит до точки $x^* = 0.9$ м



Фиг. 10. Мгновенное поле поперечного вектора завихренности в сечении $z^* = -0.0087$ м в различные моменты времени: (а) – для $t = t_0$, (б) – для $t = t_0 + 6T/20$, (в) – для $t = t_0 + 12T/20$, (г) – для $t = t_0 + 18T/20$. Сверху – настоящая работа, подробная сетка; внизу – работа [5].

(фиг. 10в), где активно делится на мелкие вихри, которые далее сносятся потоком до $x^* = 0.92$ м (фиг. 10г). К этому моменту времени в сечение $x^* = 0.84$ м подходит новый вихрь, и процесс повторяется (квазистационарный режим).

Снова можно отметить, что диссипативная схема HSFlow++ хорошо воспроизводит крупномасштабные структуры, сохраняя их в качественном и количественном отношении. Однако мелкомасштабные структуры воспроизводятся недостаточно, чтобы полностью соответствовать результатам работы [5].

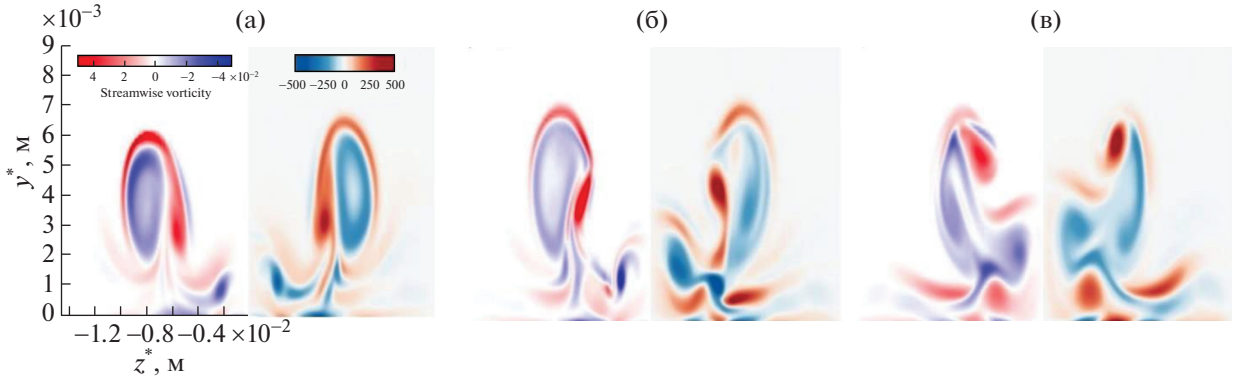
На фиг. 11 показаны мгновенные сечения продольной компоненты вектора завихренности, которые сопоставлены с результатами работы [1]. Структура течения симметрична относительно плоскости $z^* = 0$; поэтому приводится только половина области по z^*. В сечении $x^* = 0.862$ м (фиг. 11а) расположен большой вихрь. С ростом x растет и начинает распадаться при $x^* = 0.866$ м (фиг. 11б), потом в точке $x^* = 0.870$ м (фиг. 11в) образуется много мелких вихрей. Результаты, полученные на диссипативной численной схеме (настоящая работа), хорошо согласуются с результатами работы [1].

Рассмотрим зависимость осредненной величины коэффициента трения c_f от продольной координаты x. Локальный коэффициент трения вычисляется по формуле

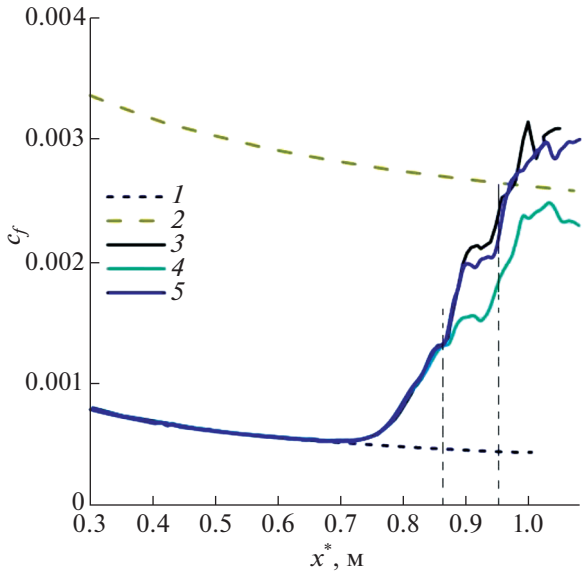
$$c_f = \frac{2}{\mathrm{Re}_\infty} \times \mu \Big| \times \frac{\partial u}{\partial y}\Big|_{y=0}.$$

В настоящей работе рассматривается усреднение по времени в течение пяти периодов фундаментальной гармоники $\Delta t = 10\pi/\omega_0$ и по размаху z всей расчетной области:

$$\overline{\varphi} = \frac{1}{\lambda_z}\frac{1}{\Delta t}\int_0^{\lambda_z}\int_{t_0}^{t_0+\Delta t} \varphi(t,z)\,dt\,dz.$$



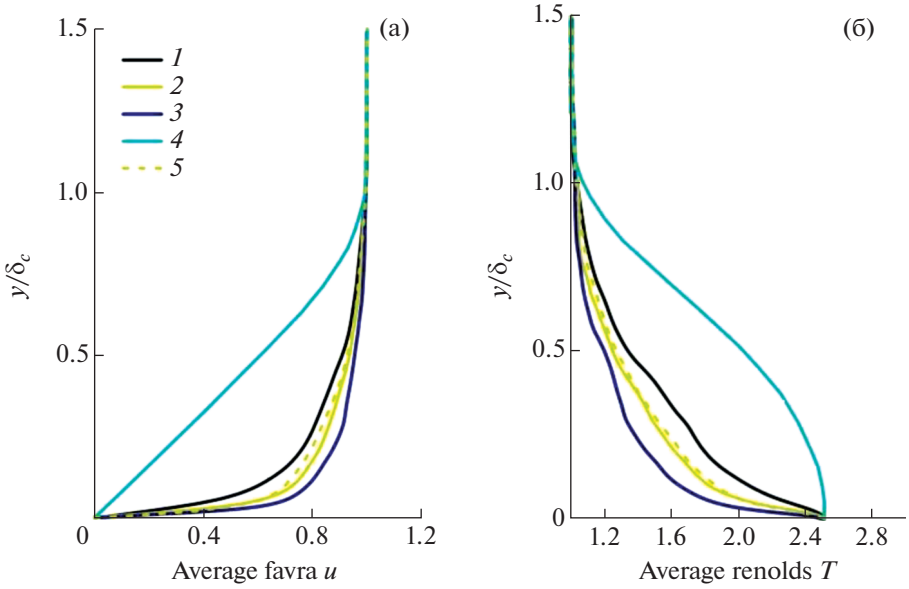
Фиг. 11. Мгновенный продольный вектор завихренности в моменте времени $t = 2.82891$: (а) – для $x^* = 0.862$ м, (б) – для $x^* = 0.866$ м, (в) – для $x^* = 0.870$ м. Слева – результаты работы [1]; справа – настоящая работа, подробная сетка.



Фиг. 12. Осредненный по пространству и времени коэффициент трения: *1* – ламинарная ветвь, *2* – теоретическая турбулентная ветвь [6], *3* – работа [1], *4* – настоящий расчет на грубой сетке, *5* – настоящий расчет на подробной сетке.

Начиная с $x^* = 0.72$ м величина c_f начинает резко усиливаться в окрестности точки $x^* = 0.86$ м. В этом диапазоне результаты диссипативной схемы совпадают с результатами работы [1] даже на грубой сетке. Ниже по потоку результаты на грубой сетке оказываются ниже по сравнению с результатами на подробной сетке. Последние удовлетворительно согласуются с результатами работы [1] в области молодой турбулентности при $x^* \geq 0.9$ м.

На фиг. 13 показаны средние профили газодинамических переменных. Чем сильнее эффект нелинейных взаимодействий в рассматриваемом сечении, тем более наполнены средние профили. Сечение $x^* = 0.996$ м вновь демонстрирует хорошее согласование результатов, полученных на диссипативной схеме, с результатами работы [1]. Фиг. 13 подтверждает, что несмотря на недостаточно детальную картину мелкомасштабных вихрей, полученную с помощью диссипативного метода, интегральные характеристики течения (средние профили газодинамических переменных, средний коэффициент трения) оказываются достаточно близки к результатам, полученным на низкодиссипативных схемах [1]. Этот вывод важен для прикладных задач, где нет необходимости подробно разрешать все структуры развитого турбулентного движения, но требуется получить надежные интегральные характеристики течения.



Фиг. 13. Осредненная (а) по Фавру величина продольной компоненты скорости и (б) осредненная по Рейнольдсу величина температуры на подробной расчетной сетке: *1* – для $x^* = 0.942$ м; *2* – для $x^* = 0.996$ м; *3* – для $x^* = 1.051$ м; *4* – ламинарный пограничный слой; *5* – работа [1], $x^* = 0.996$ м.

4. ВЫВОДЫ

Диссипативные численные схемы пригодны для моделирования процесса ламинарно-турбулентного перехода и надежного воспроизведения интегральных характеристик течения, таких как коэффициенты трения и средние профили газодинамических переменных. Это подтверждено путем детального сравнения настоящих результатов с результатами, полученными с применением низкодиссипативных схем [1].

Положение точки начала перехода практически не зависит от количества узлов сетки и порядка точности схемы, когда основная фундаментальная гармоника и ее ближайшие кратные гармоники достаточно разрешены. При этом сеточное разрешение гармоник более высокого порядка, по-видимому, играет второстепенную роль при моделировании начала ЛТП и интегральных характеристик течения.

В линейном режиме результаты, полученные с помощью диссипативной схемы, хорошо согласуются с результатами [1]. В развитом нелинейном режиме избыточная диссипативность схемы может приводить к недостаточно детальному воспроизведению мелкомасштабных структур, что можно компенсировать путем измельчения расчетной сетки. Для более аккуратного моделирования ламинарно-турбулентного перехода можно понижать диссипативность схемы там, где она не требуется (например, в пограничном слое). Это является предметом дальнейшей работы авторов.

СПИСОК ЛИТЕРАТУРЫ

1. *Mayer C.S.J., Terzi D.A.V., Fasel H.F.* DNS of Complete Transition to Turbulence Via Oblique Breakdown at Mach 3 // AIAA 2008-4398, 2008.
2. *Егоров И.В., Федоров А.В., Динь К.Х.* Прямое численное моделирование ламинарно-турбулентного перехода при сверхзвуковом обтекании острой пластины // Уч. зап. ЦАГИ. 2018. Т. 49. № 5. С. 17–25.
3. *Егоров И.В., Новиков А.В.* Прямое численное моделирование ламинарно-турбулентного обтекания плоской пластины при гиперзвуковых скоростях потока // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 6. С. 145–162.
4. *Chuvakhov P.V., Fedorov A.V., Obraz A.O.*, Numerical simulation of turbulent spots generated by unstable wave packets in a hypersonic boundary layer. Computers & Fluids. 2018. V. 162. P. 26–38. Available at: https://doi.org/10.1016/j.compfluid.2017.12.001
5. *Mayer C.S.J., Terzi D.A.V., Fasel H.F.* Direct numerical simulation of complete transition to turbulence via oblique breakdown at Mach 3 // J. Fluid Mech. 2011. V. 674. P. 5–42.
6. *White F.M.* Viscous Fluid Flow. New York: McGraw-Hill, 1991.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 281–302

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.634

ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ НЕСТАЦИОНАРНЫХ ДОЗВУКОВЫХ ТЕЧЕНИЙ ВЯЗКОГО ГАЗА НА ОСНОВЕ СОСТАВНЫХ КОМПАКТНЫХ СХЕМ ВЫСОКОГО ПОРЯДКА

© 2021 г. А. Д. Савельев

119991 Москва, ул. Вавилова, 40, Вычислительный центр им. А.А. Дородницына РАН Федерального исследовательского центра "Информатика и управление" РАН, Россия

e-mail: savel-cc09@yandex.ru

Поступила в редакцию 20.05.2020 г.
Переработанный вариант 28.08.2020 г.
Принята к публикации 16.09.2020 г.

Рассматривается семейство мультиоператорных компактных схем высокого порядка для расчетов течений вязкого газа на криволинейных сетках. В зависимости от количества используемых операторов порядок схемы для аппроксимации конвективных членов уравнений может составлять от 6-го до 22-го. С таким же порядком аппроксимируются вязкие члены исходных уравнений и метрические коэффициенты обобщенной криволинейной системы координат. На примере трех схем, в том числе самой трудоемкой из описываемых пятиоператорной, рассматривается их структура. Исследуется изменение аппроксимационных и диссипативных свойств схем, сопутствующее повышению их порядка. Выполняются сравнительные расчеты данными схемами дозвуковых течений газа на основе уравнений Эйлера и Навье–Стокса. Приводятся результаты расчетов вязкого дозвукового обтекания аэродинамического профиля в широком диапазоне изменения угла атаки и непроницаемого парашютного купола с использованием схемы 22-го порядка. Библ. 34. Табл. 9. Фиг. 13.

Ключевые слова: компактные разностные схемы, 22-й порядок аппроксимации, дозвуковые течения вязкого газа, отрыв пограничного слоя, аэродинамический профиль, купол парашюта.

DOI: 10.31857/S0044466921020113

1. ВВЕДЕНИЕ

В настоящее время все чаще для решения прикладных задач широко используются методы численного моделирования. При этом область расчета покрывается сеточными узлами с необходимой плотностью их распределения и решаются разностные аналоги уравнений сплошной среды в виде системы уравнений в частных производных на основе тех или иных дифференциальных операторов. При этом на конечный результат влияет как полнота описания выбранной модели течения за счет уточнения ее свойств, так и используемые разностные схемы. Дело в том, что конечно-разностное представление производных отличается от их исходных дифференциальных аналогов. Так, разложение в ряд Тейлора разностного аналога первой производной от некой гладкой функции f по координате x на сетке $\omega_h = \{x_j = jh,\ h = \text{const}\}$ имеет вид

$$h^{-1}\Delta_1^{(n)} f = \frac{\partial f}{\partial x} + \sum_{k=n}^{\infty} \alpha_k \frac{\partial^{k+1} f}{\partial x^{k+1}} h^k, \tag{1}$$

где $\Delta_1^{(n)}$ – сеточная функция, n – порядок аппроксимации производной, а коэффициенты $|\alpha_k| \to 0$ при $k \to \infty$. У схем n-го порядка аппроксимации члены под знаком суммы со значениями $k < n$ отсутствуют. Поскольку суммарное влияние членов, остающихся под знаком суммы, снижается, это приводит к тому, что разностное решение на их основе в сравнении со схемами более низкого порядка становится ближе к решению исходных дифференциальных уравнений.

Примеры разностных схем с порядком аппроксимации выше второго для расчетов задач газовой динамики известны с конца 60-х годов. Примером может служить известная схема Русанова (см. [1]). Чуть позже были предложены несимметричные компактные схемы третьего порядка,

учитывающие направление потока (см. [2]). В дальнейшем данный подход получил развитие и для схем более высокого порядка аппроксимации (см., например, [3]). В то же время за рубежом большее признание получили симметричные компактные аппроксимации, использующие пятиточечный шаблон (см. [4], [5]). Они, как правило, применяются для моделирования нестационарных течений, акустического излучения аэродинамических поверхности и струй (см. [6], [8]). Помимо компактных разностей высокого порядка, большое распространение получили схемы ENO и WENO, использующие некомпактную запись разностных операторов (см. [9], [10]), в том числе и на неструктурированных сетках (см. [11], [13]). В основном, они применяются при моделировании нестационарных невязких течений с контактными разрывами и скачками уплотнения. В целом можно отметить общее стремление к использованию при решении задач газовой динамики схем с порядком аппроксимации выше двух.

Составные компактные схемы (см. [14]) были построены на основе опыта, полученного при использовании разностных аппроксимаций типа (см. [2], [3]), при намерении получить более простое и эффективное управление их свойствами. Они представляют собой комбинацию симметричных компактных разностей высокого, изначально 6-го и 8-го порядков, и некомпактных четных разностей диффузного типа, ориентированных в соответствии с направлением характеристик. Позднее, когда были разработаны нецентрированные мультиоператорные схемы для уравнений газовой динамики (см. [15], [16]), мультиоператорный подход был использован и применительно к составным компактным схемам для описания конвективных и диффузных членов исходных уравнений. В [17] представлены схемы 6-го, 10-го и 14-го порядков, в [18] – 18-го и 22-го.

При высокой аппроксимации составных компактных схем непосредственное их применение в задачах ударного взаимодействия затруднительно. Это связано с диссипативными свойствами используемых диффузных операторов, которые с ростом их порядка сильно снижаются. Между тем, существует малоисследованная область дозвуковых течений, моделирование которых представляет большой интерес. Численное исследование вихреобразования, отрыва ламинарного и турбулентного течений на гладкой изогнутой поверхности, генерация и распространение звуковых волн требуют использования разностных схем с высокой разрешающей способностью. При этом сравнительный анализ результатов, полученных на основе схем разного порядка аппроксимации, должен способствовать получению верных выводов о рассматриваемом течении.

2. ИСХОДНЫЕ УРАВНЕНИЯ И ПОСТАНОВКА ЗАДАЧИ

Будем рассматривать течения вязкого теплопроводного газа с постоянным отношением удельных теплоемкостей. В общем пространственном случае они описываются трехмерными уравнениями. Однако подобный подход требует несоизмеримо большего времени расчетов, чем решение двумерных уравнений. Если учитывать тот объем полезной информации, которую можно получить по двумерной технологии за время одного трехмерного расчета, напрашивается вывод о необходимости ее использования, по крайней мере, на начальном этапе решения задачи. На данном этапе будем использовать двумерные нестационарные осредненные уравнения Навье–Стокса, дополненные уравнениями SST-модели турбулентности (см. [19]). Обезразмеренные по параметрам набегающего потока и характерному линейному размеру, они имеют следующий вид в системе координат (x, y):

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = \mathrm{Re}^{-1}\left(\frac{\partial \mathbf{F}_\mu}{\partial x} + \frac{\partial \mathbf{G}_\mu}{\partial y}\right) + C_l\mathbf{H},$$

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \\ \rho k \\ \rho\omega \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \rho u \\ \rho u^2 + p + p_t \\ \rho uv \\ (e + p + p_t)u \\ \rho uk \\ \rho u\omega \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p + p_t \\ (e + p + p_t)v \\ \rho vk \\ \rho v\omega \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ H_k \\ H_\omega \end{bmatrix}, \quad \boldsymbol{\varphi} = \begin{bmatrix} \rho \\ u \\ v \\ h \\ k \\ \omega \end{bmatrix},$$

$$\mathbf{F}_\mu = \begin{bmatrix} 0 \\ \mu\left(2u_x - \frac{2}{3}\nabla U\right) \\ \mu(u_y + v_x) \\ u\sigma_{xx} + v\tau_{xy} + \mu\,\mathrm{Pr}^{-1} h_x \\ \mu_k\,\mathrm{Pr}_k^{-1} k_x \\ \mu_\omega\,\mathrm{Pr}_\omega^{-1}\omega_x \end{bmatrix}, \quad \mathbf{G}_\mu = \begin{bmatrix} 0 \\ \mu(u_y + v_x) \\ \mu\left(2v_y - \frac{2}{3}\nabla U\right) \\ u\tau_{yx} + v\sigma_{yy} + \mu\,\mathrm{Pr}^{-1} h_y \\ \mu_k\,\mathrm{Pr}_k^{-1} k_y \\ \mu_\omega\,\mathrm{Pr}_\omega^{-1}\omega_y \end{bmatrix},$$

$$\sigma_{xx} = \mu\left(2u_x - \frac{2}{3}\nabla U\right), \quad \nabla U = u_x + v_y,$$

$$\tau_{xy} = \tau_{yx} = \mu(u_y + v_x), \quad p = (\gamma - 1)\gamma^{-1}\rho h,$$

$$\sigma_{yy} = \mu\left(2v_y - \frac{2}{3}\nabla U\right), \quad e = \rho[\gamma^{-1}h + 0.5(u^2 + v^2) + k],$$

$$p_t = \frac{2}{3}\rho k, \quad \mu\mathrm{Pr}^{-1} = \mu_l\mathrm{Pr}_l^{-1} + \mu_t\mathrm{Pr}_t^{-1},$$

$$\mu = \mu_l + \mu_t, \quad \mu_k\,\mathrm{Pr}_k^{-1} = \mu_l + \mathrm{Pr}_k^{-1}\mu_t,$$

$$\mathrm{Re} = \rho_\infty u_\infty c\mu_{l\infty}^{-1}, \quad \mu_\omega\,\mathrm{Pr}_\omega^{-1} = \mu_l + \mathrm{Pr}_\omega^{-1}\mu_t,$$

$$H_k = \rho k\left(\frac{\Omega}{F_\mu} - \beta^*\omega\right),$$

$$H_\omega = \rho\omega\left(\chi\frac{\Omega}{F_\mu} - \beta\omega\right) + 2(1 - F_1)\frac{\sigma_{\omega 2}\rho}{\omega}\left(\frac{\partial k}{\partial x}\frac{\partial\omega}{\partial x} + \frac{\partial k}{\partial y}\frac{\partial\omega}{\partial y}\right).$$

Здесь t и x, y – время и декартовы координаты, ρ, u и v – плотность и компоненты вектора скорости соответственно, h – энтальпия, $\gamma = 1.4$ – отношение удельных теплоемкостей, $\boldsymbol{\varphi}$ – вектор элементарных переменных, μ и μ_t – коэффициенты молекулярной и турбулентной вязкостей, $\mathrm{Pr} = 0.72$ и $\mathrm{Pr}_t = 0.9$ – молекулярное и турбулентное числа Прандтля, C_{lt} – коэффициент подключения источниковых членов модели турбулентности. Для определения величины молекулярной вязкости μ_l используется формула Саттерлэнда (см. [20]). Коэффициент турбулентной вязкости определяется по формуле

$$\mu_t = \mathrm{Re}\frac{\rho k}{F_\mu\omega}.$$

Необходимые константы и функции модели турбулентности (см. [19])

$$F_\mu = \max\left(1; \Omega\frac{F_2}{a_1\omega}\right), \quad F_2 = \tanh\left\{\left[\max\left(\frac{2k^{1/2}}{C_\mu\omega d}; \frac{500\mu_l}{\mathrm{Re}\,\rho\omega d^2}\right)\right]^2\right\},$$

$$\Omega^2 = 2\Omega_{xy}\Omega_{xy}, \quad \Omega_{xy} = 0.5\left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\right), \quad \beta^* = C_\mu = 0.09, \quad a_1 = 0.31,$$

d – расстояние до твердой поверхности. Поскольку модель представляет собой суперпозицию известных $k - \varepsilon$ и $k - \omega$ моделей, в ней используются две серии констант. Переход от констант внутреннего слоя к константам внешнего осуществляется с помощью функции перемешивания F_1:

$$\phi = F_1\phi_1 + (1 - F_1)\phi_2, \quad \phi = (\mathrm{Pr}_k, \mathrm{Pr}_\omega, \beta, \chi),$$

где

$$F_1 = \tanh\left(\left\{\min\left[\max\left(\frac{k^{1/2}}{C_\mu\omega d}; \frac{500\mu_l}{\mathrm{Re}\,\rho\omega d^2}\right); \frac{4\sigma_{\omega 2}\rho k}{CD_{k\omega}d^2}\right]\right\}^4\right),$$

$$CD_{k\omega} = \max\left(\frac{2\sigma_{\omega 2}\rho}{\omega}\left(\frac{\partial k}{\partial x}\frac{\partial \omega}{\partial x} + \frac{\partial k}{\partial y}\frac{\partial \omega}{\partial y}\right); 10^{-20}\right).$$

Наборы констант для внутреннего и внешнего решений (1 и 2 соответственно) имеют значения

$$\mathrm{Pr}_{k1} = 0.85, \quad \mathrm{Pr}_{\omega 1} = 0.5, \quad \beta_1 = 0.075, \quad \chi_1 = 0.533,$$

$$\mathrm{Pr}_{k2} = 1.0, \quad \mathrm{Pr}_{\omega 2} = 0.856, \quad \beta_2 = 0.0828, \quad \chi_2 = 0.44.$$

На поверхности тела задаются условия прилипания для скорости, температура поверхности или условие равенства нулю нормального градиента температуры. Плотность газа на поверхности определяется путем решения уравнения неразрывности или из условия нулевого нормального градиента давления. Так же на твердой поверхности задаются нулевые значения кинетической энергии k_w и турбулентной вязкости μ_{tw}, а частота турбулентности в соответствии с рекомендациями из [19] определяется как

$$\omega_w = 10\frac{6\mu_w}{\mathrm{Re}\,\rho_w\beta_1 d_w^2},$$

где d_w – расстояние до ближайшего узла разностной сетки.

На входной границе фиксируются параметры набегающего потока на бесконечности, а на выходной – условия свободного вытекания. При этом необходимо учитывать то, что возмущения, генерируемые обтекаемым телом, распространяются вверх против потока тем дальше, чем ниже число Маха набегающего потока. При достаточно близких внешних границах области расчета целесообразно использовать граничные условия на основе характеристик. В набегающем потоке задаются полные давление и температура, угол между горизонтальной и вертикальной компонентами скорости и решается уравнение на основе выходной характеристики. Напротив, на выходной границе задается давление, другие параметры экстраполируются. Кинетическая энергия турбулентности на бесконечности задается равной $k_\infty = 10^{-6}$. При этом полагается, что турбулентная вязкость лежит в диапазоне $0.1 \le \mu_{t\infty} \le 1$, откуда и следует значение частоты турбулентности в набегающем потоке.

В представленном виде уравнения представляют собой законы сохранения осредненных по массе параметров потока. Если же полагать равенству нулю кинетической энергии турбулентности, вихревой вязкости и источниковых членов в уравнениях для k и ω, то полученные таким образом уравнения позволяют получать характеристики ламинарного течения вязкого газа. Вопреки названию и устоявшимся представлениям, такие течения далеко не всегда являются стационарными. Возникающие при высоких числах Рейнольдса отрывные и пульсационные эффекты провоцируют переход от ламинарного режима течения газа к турбулентному.

Далее производится переход к обобщенной криволинейной системе координат для возможности расчета обтекания тел с искривленными и изломанными границами. Данная процедура подробно описана в [21]. Согласно ей, область течения в физической плоскости (x, y) отображается на единичный квадрат расчетной плоскости (ξ, η) с помощью преобразования общего вида $\xi = \xi(x, y)$, $\eta = \eta(x, y)$, $0 \le \xi \le 1$, $0 \le \eta \le 1$, а производные по декартовым координатам некой функции f уравнениях приобретают вид

$$\frac{\partial f}{\partial x} = \frac{1}{J^{-1}}\left(\frac{\partial f y_\eta}{\partial \xi} - \frac{\partial f y_\xi}{\partial \eta}\right), \quad \frac{\partial f}{\partial y} = \frac{1}{J^{-1}}\left(\frac{\partial f x_\xi}{\partial \eta} - \frac{\partial f x_\eta}{\partial \xi}\right),$$

где $J^{-1} = x_\xi y_\eta - x_\eta y_\xi$ – якобиан, а x_ξ, y_ξ, x_η и y_η – метрические коэффициенты преобразования. Уравнения, использующиеся для проведения расчетов, приобретают следующий вид:

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \frac{\partial \hat{\mathbf{F}}}{\partial \xi} + \frac{\partial \hat{\mathbf{G}}}{\partial \eta} = \mathrm{Re}^{-1}\left(\frac{\partial \hat{\mathbf{F}}_\mu}{\partial \xi} + \frac{\partial \hat{\mathbf{G}}_\mu}{\partial \eta}\right) + C_{lt}\hat{\mathbf{H}},$$

$$\hat{\mathbf{U}} = J^{-1}\mathbf{U}, \quad \hat{\mathbf{H}} = J^{-1}\mathbf{H}, \hat{\mathbf{F}} = \mathbf{F}y_\eta - \mathbf{G}y_\xi, \quad \hat{\mathbf{G}} = \mathbf{G}x_\xi - \mathbf{F}x_\eta,$$

$$\hat{\mathbf{F}}_\mu = \mathbf{F}_\mu y_\eta - \mathbf{G}_\mu y_\xi, \quad \hat{\mathbf{G}}_\mu = \mathbf{G}_\mu x_\xi - \mathbf{F}_\mu x_\eta,$$

а векторы $\hat{\mathbf{F}}_\mu$ и $\hat{\mathbf{G}}_\mu$ также содержат производные переменных и метрические коэффициенты. Таким образом, расчеты осуществляются на равномерной сетке, позволяющей применять компактные схемы высокого порядка.

3. РАЗНОСТНЫЕ ОПЕРАТОРЫ И ИХ СВОЙСТВА

Составные компактные разности были предложены в [14] на основе опыта практического применения несимметричных компактных аппроксимаций (см. [3]). Они основаны на идее разделения единого разностного оператора на части, каждая из которых отвечает за одно из его характерных свойств. Разностный аналог производной $\partial f/\partial x$, предназначенный для описания конвективных членов уравнений, на сетке $\omega_h = \{x_j = jh,\ h = \text{const}\}$ имеет вид

$$f_j' = \left[\Delta_1^{(2l)} + (-1)^m s\frac{\Delta_2^m}{c_{2m}}\right]\frac{f_j}{2h}, \tag{1}$$

где $l, m = 1, 2\ldots$ – коэффициенты при обозначении порядка центральной и повторной разностей, $s = \pm 1$ – параметр, учитывающий направление потока, а c_{2m} – нормирующая константа, своя для каждого значения m. При этом операторами $\Delta_1^{(2l)}$ и Δ_2^m обозначены первая и $2m$-я разности, описываемые соответственно с помощью симметричного компактного и обычного оператора диффузного типа. Первый отвечает за аппроксимацию производной, второй способствует получению устойчивого решения задачи. Для их представления используются операторы на основе простейших разностей $\Delta_1 = T_1 - T_{-1}$, $\Delta_2 = T_1 - 2T_0 + T_{-1}$, где T – оператор сдвига $T_n f_j = f(x_j + nh)$. Основу составных компактных схем представляет аппроксимирующий оператор, имеющий вид симметричной пятиточечной компактной разности

$$f_j' = \Delta_1^{(2l)}\frac{f_j}{2h} = (1 + a\Delta_2)^{-1}(1 + b\Delta_2)\Delta_1\frac{f_j}{2h}. \tag{2}$$

Данная разность при произвольных значениях a и b формально имеет второй порядок аппроксимации и реализуется трехточечными прогонками. В случае $a = 1/5$ и $b = 1/30$ она представляет собой классическую Паде-разность 6-го порядка (см. [4]).

Повышение порядка аппроксимации обычно сопровождается укрупнением сеточного шаблона. Альтернативным направлением получения разностных схем с высокой разрешающей способностью является мультиоператорный подход, описанный в [15], [16]. Однако для составных компактных схем высокий порядок достигается не путем подбора весов компонентов мультиоператора, как в [15], [16], а подбором значений коэффициентов a и b этих разностей при заданных весах. Веса всех разностных операторов равны и обратно пропорциональны их количеству. Мультиоператорное представление разности $\Delta_1^{(2l)}$ имеет вид

$$\Delta_1^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(1 + a_p\Delta_2)^{-1}(1 + b_p\Delta_2)\Delta_1, \tag{3}$$

где n – количество используемых операторов, a_p и b_p – коэффициенты, свои для каждого оператора, а k – общий порядок разностной аппроксимации. В зависимости от количества операторов порядок разностной схемы может меняться от 6 до 22 (см. [17], [18]). В данной работе используются разности 6-го, 14-го и 22-го порядков. Их коэффициенты представлены в табл. 1.

Поскольку разностные операторы (3) реализуются трехточечными прогонками, они требуют постановки соответствующих граничных условий для каждого внутреннего оператора. На каждой границе задаются по два граничных условия. Их вид представлен формулой (4), а необходимые коэффициенты сведены в табл. 2 и 3:

$$\Delta_{1,b}^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(a_p^0T_0 + a_p^1T_{+1} + a_p^2T_{+2})^{-1}(b_p^0T_0 + b_p^1T_{+1} + b_p^2T_{+2} + b_p^3T_{+3} + b_p^4T_{+4}). \tag{4}$$

Табл. 2 соответствует разности, содержащей две точки в левой части оператора. Остальные коэффициенты этого граничного оператора определяются из соотношений $a_p^1 = 1 - a_p^0$, $a_p^2 = 0$,

Таблица 1

n	k	p	a_p	b_p
1	6	1	0.2	0.0333333333333
3	14	1	0.239048459346	−0.0567444588747
		2	0.1624727238229	−0.0379667519389
		3	0.0600172783691	0.0220795956071
5	22	1	0.245338324132	−0.0875922106964
		2	0.149866943731	−0.0093729917038
		3	0.026791341338	0.01450958533096
		4	0.210093756803	−0.0538634205423
		5	0.0821953482816	0.0172714185639

Таблица 2

N	k	p	a_p^0	b_p^0	b_p^1
1	5	1	0.2	−0.61666666667	0.1333333333333
3	5	1	1.3707470112474	−2.392347386564	4.4383149363567
		2	0.223948025612	−0.661256249069	0.2514243312644
		3	0.0376713398493	−0.517006992215	0.0117873019992
5	5	1	140.32598921687	−233.9337759912	583.253061457344
		2	0.2135067189865	−0.6541547386747	0.24209209266732
		3	0.0120683273273	−0.5055658003922	0.00437413925759
		4	0.7353423046579	−1.3806499641782	1.94820127241312
		5	0.0673734960562	−0.53298694541299	0.028437971229197

Таблица 3

n	k	p	a_p^0	a_p^1	b_p^0	b_p^1
1	6	1	0.06666666666	0.53333333333	−0.23888888889	−0.44444444444
3	6	1	−0.0308111106	0.49717689768	0.10516977549	−1.24740708804
		2	0.18684519996	0.70131442416	−0.55589510643	0.090497388666
		3	0.02370623266	0.65008260032	−0.3388766994	−0.17650437274
5	6	1	−0.00901353453	0.49965746749	0.09951983608	−1.27858087272
		2	0.21224996404	0.78362809119	−0.65043676019	0.236544983249
		3	0.00734736459	0.62242727035	−0.31561619167	−0.18828384108
		4	−0.13495595651	0.44873160152	0.292203419413	−1.53813007655
		5	0.0458509465	0.6872125923	−0.37520150835	−0.14849271155

$b_p^2 = -2.5 - a_p^0 - 6b_p^0 - 3b_p^1$, $b_p^3 = 4 + 2a_p^0 + 8b_p^0 + 3b_p^1$, $b_p^4 = -1.5 - a_p^0 - 3b_p^0 - b_p^1$, где p меняется от 1 до n, n – количество операторов в схеме, а k – порядок аппроксимации.

Другое граничное условие содержит три узла сетки в левой части оператора. Коэффициенты a_p^0, a_p^1, b_p^0 и b_p^1 для него содержатся в табл. 3.

Остальные коэффициенты разностной формулы определяются соотношениями $a_p^2 = 1 - a_p^0 - a_p^1$, $b_p^2 = -1.5 - 2a_p^0 - a_p^1 - 6b_p^0 - 3b_p^1$, $b_p^3 = 2 + 4a_p^0 + 2a_p^1 - 3b_p^0 - b_p^1$, $b_p^4 = -0.5 - 2a_p^0 - a_p^1 - 3b_p^0 - b_p^1$.

Граничные условия, необходимые для получения решения с помощью прогонок, имеют порядки 5 и 6 не зависимо от количества операторов. Дело в том, что высокий порядок мультиоператорной схемы получается после суммирования отдельных операторов, каждый из которых имеет лишь второй порядок. При проведении прогонок граничные условия для каждого оператора формируются таким образом, чтобы остаточные члены их разложений приближались бы к остаточным членам соответствующих операторов. Использование традиционных односторонних разностей у границы снижает их порядок, хотя всего лишь в двух узлах.

Мультиоператорная формула (3) является бездиссипативной составляющей разности (1) и отвечает за аппроксимацию конвективных членов исходных уравнений. На ее основе также проводится вычисление смешанных производных вязких членов, компонент источникового вектора $\hat{\mathbf{H}}$ и метрических коэффициентов преобразования координат. Последнее особенно важно, так как обеспечивает выполнение условий

$$\frac{\partial y_\eta}{\partial \xi} = \frac{\partial y_\xi}{\partial \eta}, \quad \frac{\partial x_\xi}{\partial \eta} = \frac{\partial x_\eta}{\partial \xi},$$

необходимых для отсутствия наведенных искажений решения при расчетах на криволинейных сетках. Мультиоператорные схемы годятся и для трехмерных расчетов. В этом случае требуется несколько иное определение метрических коэффициентов преобразования координат. Примеры проведения трехмерных расчетов с помощью компактных разностей на криволинейных сетках можно найти, например, в [5].

Аппроксимирующей составляющей разности (2) может оказаться вполне достаточно для расчетов участков течения с существенным влиянием вязких сил, например, в зонах отрыва пограничного слоя у препятствий на твердой поверхности. В общем же случае потоки необходимо корректировать. Для этого используются стабилизирующие добавки в виде четных разностей высокого порядка, которые на сетке $\omega_h = \{x_j = jh,\ h = \text{const}\}$ выглядят следующим образом:

$$\sum_{n=1}^{3} (-1)^m c_{2m}^{-1} s_n \Delta_2^{2m-2} (\Delta_- T_{1/2} \mathbf{M}_n \Delta_+ \boldsymbol{\varphi}), \tag{5}$$

где $\Delta_- = T_0 - T_{-1}$, $\Delta_+ = T_1 - T_0$, $\mathbf{M}$ – якобиевы матрицы расщепления векторов конвективных членов уравнений, полученные в соответствии с количеством собственных значений λ_n^ξ и λ_n^η, $n = 1$–3,

$$\mathbf{M}_n^\xi = \frac{\partial \hat{\mathbf{f}}_n}{\partial \boldsymbol{\varphi}}, \quad \mathbf{M}_n^\eta = \frac{\partial \hat{\mathbf{g}}_n}{\partial \boldsymbol{\varphi}},$$

s_n – знаки собственных значений, c_{2m} – коэффициенты. Процедура расщепления потоковых векторов $\hat{\mathbf{F}}$ и $\hat{\mathbf{G}}$ подробно описана в [22]. В данном случае они имеют вид

$$\hat{\mathbf{F}} = \sum_{n=1}^{3} \hat{\mathbf{f}}_n, \quad \hat{\mathbf{G}} = \sum_{n=1}^{3} \hat{\mathbf{g}}_n,$$

где

$$\hat{\mathbf{f}}_1 = \frac{\rho \lambda_1^\xi (\gamma - 1)}{\gamma} \begin{bmatrix} 1 \\ u \\ v \\ (u^2 + v^2)/2 + k \\ \gamma k (\gamma - 1) \\ \gamma \omega (\gamma - 1) \end{bmatrix}, \quad \hat{\mathbf{f}}_2 = \frac{\rho \lambda_2^\xi}{2\gamma} \begin{bmatrix} 1 \\ u - \hat{a}_\xi y_\eta \\ v + \hat{a}_\xi x_\eta \\ h + (u^2 + v^2)/2 + k - w_\xi \hat{a}_\xi \\ 0 \\ 0 \end{bmatrix},$$

$$\hat{\mathbf{f}}_3 = \frac{\rho\lambda_3^\xi}{2\gamma}\begin{bmatrix} 1 \\ u + \hat{a}_\xi y_\eta \\ v - \hat{a}_\xi x_\eta \\ h + (u^2+v^2)/2 + k + w_\xi \hat{a}_\xi \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\mathbf{g}}_1 = \frac{\rho\lambda_1^\eta(\gamma-1)}{\gamma}\begin{bmatrix} 1 \\ u \\ v \\ (u^2+v^2)/2 + k \\ \gamma k(\gamma-1) \\ \gamma\omega(\gamma-1) \end{bmatrix},$$

$$\hat{\mathbf{g}}_2 = \frac{\rho\lambda_2^\eta}{2\gamma}\begin{bmatrix} 1 \\ u + \hat{a}_\eta y_\xi \\ v - \hat{a}_\eta x_\xi \\ h + (u^2+v^2)/2 + k - w_\eta \hat{a}_\eta \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\mathbf{g}}_3 = \frac{\rho\lambda_3^\eta}{2\gamma}\begin{bmatrix} 1 \\ u - \hat{a}_\eta y_\xi \\ v + \hat{a}_\eta x_\xi \\ h + (u^2+v^2)/2 + k + w_\eta \hat{a}_\eta \\ 0 \\ 0 \end{bmatrix},$$

$$w_\xi = uy_\eta - vx_\eta, \quad a_\xi = a(x_\eta^2 + y_\eta^2)^{1/2}, \quad \hat{a}_\xi = a_\xi(x_\eta^2 + y_\eta^2)^{-1},$$

$$\lambda_1^\xi = w_\xi, \quad \lambda_2^\xi = w_\xi - a_\xi, \quad \lambda_3^\xi = w_\xi + a_\xi, \quad s_n^\xi = \operatorname{sign}(\lambda_n^\xi),$$

$$w_\eta = -uy_\xi + vx_\xi, \quad a_\eta = a(x_\xi^2 + y_\xi^2)^{1/2}, \quad \hat{a}_\eta = a_\eta(x_\xi^2 + y_\xi^2)^{-1},$$

$$\lambda_1^\eta = w_\eta, \quad \lambda_2^\eta = w_\eta - a_\eta, \quad \lambda_3^\eta = w_\eta + a_\eta, \quad s_n^\eta = \operatorname{sign}(\lambda_n^\eta),$$

$$a = [\gamma(p + p_t)\rho^{-1}]^{1/2}.$$

Помимо конвективных членов, разностное представление которых можно осуществлять рассмотренным способом, в уравнениях динамики вязкого газа присутствуют диссипативные члены вида $\partial(\mu\partial f/\partial x)/\partial x$. Обычно они описываются стандартным образом операторами второго порядка $\Delta_- T_{1/2}\mu\Delta_+ f_j$. В данном случае будем применять разностные аппроксимации вида

$$\delta^{(k)} T_{1/2}(I^{(k)}\mu)\delta^{(k)} f_j, \tag{6}$$

где $\delta^{(k)}$ и $I^{(k)}$ – формулы компактных разности и интерполяции. Их мультиоператорное представление выглядит следующим образом:

$$\delta^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(1 + a_p^\delta\Delta_2)^{-1}[(1 - 3b_p^\delta)(T_{1/2} - T_{-1/2}) + b_p^\delta(T_{3/2} - T_{-3/2})], \tag{7}$$

$$I^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(1 + a_p^I\Delta_2)^{-1}[(1 - b_p^I)(T_{1/2} + T_{-1/2}) + b_p^I(T_{3/2} + T_{-3/2})]. \tag{8}$$

Здесь a_p^δ, b_p^δ, a_p^I и b_p^I – коэффициенты свои для каждого оператора, их значения представлены в табл. 4 и 5. Видно, что с увеличением количества операторов растет и порядок аппроксимации.

Расчет диссипативных членов уравнений проводится в два этапа. Сначала определяются производные параметров и интерполированные коэффициенты вязких членов в полуцелых узлах разностной сетки. При прогонках используются граничные условия следующего вида:

$$\delta_{1,b}^{(k)} = \frac{1}{n}\sum_{p=1}^{n}[b_{1,p}^{\delta,-1/2}T_{-1/2} + b_{1,p}^{\delta,1/2}T_{1/2} + b_{1,p}^{\delta,3/2}T_{3/2} + b_{1,p}^{\delta,5/2}T_{5/2}],$$

$$I_b^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(a_p^{I,0}T_0 + a_p^{I,1}T_{+1} + a_p^{I,2}T_{+2})^{-1}(b_p^{I,-1/2}T_{-1/2} + b_p^{I,1/2}T_{1/2} + b_p^{I,3/2}T_{3/2} + b_p^{I,3/2}T_{5/2}).$$

Основные коэффициенты данных операторов и содержатся в табл. 6 и 7.

Таблица 4

N	k	p	a_p^{δ}	b_p^{δ}
1	6	1	0.1125	0.070833333333333
3	14	1	0.2109472736338	0.183939274770512
		2	0.1186395423325546	0.057924127866374
		3	0.031172041017277	−0.006104545653217
5	22	1	0.0141849346249354	−0.009836597930706
		2	0.2324819204002052	0.218256088277739
		3	0.0580630497952101	−0.000139903346171
		4	0.120744725801449	0.0540743151222133
		5	0.1849034349504165	0.1396908301158074

Таблица 5

n	k	p	a_p^{I}	b_p^{I}
1	6	1	0.1875	0.03125
3	14	1	0.2376211084878	0.033945872641115
		2	0.152815116744539	0.0218307309635056
		3	0.047063774767658	0.0067233963953797
5	22	1	0.0730731233747809	0.0066430112158892
		2	0.2449366217017918	0.0222669656092538
		3	0.2068575917431287	0.0188052356130117
		4	0.0198433083961389	0.0018039371269217
		5	0.1427893547841596	0.0129808504349236

Таблица 6

n	k	p	$b_{1,p}^{\delta,-1/2}$
1	2	1	−1
3	3	1	−0.97299200113667
		2	−0.93928458553382
		3	−0.96272341332951
5	3	1	−0.97597846744436
		2	−0.98577416787753
		3	−0.94179704685862
		4	−0.93332958932076
		5	−0.95478739516539

Остальные коэффициенты граничного условия для разности $\delta_{1,b}^{(k)}$ определяются из значений коэффициентов $b_{1,p}^{\delta,-1/2}$ табл. 4 с помощью следующих соотношений: $b_{1,p}^{\delta,1/2} = -2 - 3b_{1,p}^{\delta,-1/2}$, $b_{1,p}^{3/2} = 3 + 3b_{1,p}^{\delta,1/2}$, $b_{1,p}^{\delta,5/2} = -1 - b_{1,p}^{\delta,1/2}$.

Коэффициенты граничного условия для компактной интерполяции $I_b^{(k)}$ рассчитываются на основе данных табл. 7:

$$a_p^{I,1} = 1 - a_p^{I,0}, \quad a_p^{I,2} = 0, \quad b_p^{I,3/2} = 1.5 + a_p^{I,0} - 3b_p^{I,-1/2} - 2b_p^{I,1/2},$$
$$b_p^{I,5/2} = -0.5 - a_p^{I,0} + 2b_p^{I,-1/2} + b_p^{I,1/2}.$$

Таблица 7

n	k	p	$a_p^{I,0}$	$b_p^{I,-1/2}$	$b_p^{I,1/2}$
1	5	1	0.375	0.078125	0.703125
3	6	1	0.4752422169756	0.03898841993765	0.89222282980377
		2	0.30563023348908	0.047290135444342	0.68559055811956
		3	0.09412754953532	0.023096444618009	0.53156161207667
5	6	1	0.146146246749562	0.030149555732161	0.562340590768969
		2	0.489873243403584	0.0245218870179106	0.938574547959106
		3	0.413715183486257	0.0350312986566758	0.827426523129242
		4	0.039686616792278	0.0101077011467953	0.511167450478814
		5	0.285578709568319	0.040814557446493	0.676115887663763

Таблица 8

n	k	p	$a_{2,p}^{\delta,0}$	$b_{2,p}^{\delta,1/2}$
1	4	1	0.043478260869565	−1.0452898550724637
3	4	1	0.02775767820473	−1.028507357546134
		2	0.064640062661807	−1.068564710857438
		3	0.0387199336324187	−1.040163280594343
5	4	1	0.02461276898715095	−1.02520400541866
		2	0.01443112691124347	−1.0146364217
		3	0.0617998891964227	−1.06539682525146
		4	0.071432869419425	−1.0761953281596
		5	0.04735358370203	−1.04949456256945

Таблица 9

n	k	p	$a_{3,p}^{\delta,0}$	$a_{3,p}^{\delta,1}$	$b_{3,p}^{\delta,1/2}$
1	5	1	0.1125	2.475	−2.7708333333333
3	5	1	0.2109472736338	7.388653531518	−7.9944873535566
		2	0.1186395423326	1.716747637802	−2.0119508503331
		3	0.0311720410173	0.773892381696	−0.8301319180773
5	5	1	0.0141849346249	0.562139274664	−0.5806725459833
		2	0.2324819204002	15.87727318258	−16.560493111661
		3	0.0580630497952	0.881470184814	−0.9974563810583
		4	0.1207447258015	1.569579725432	−1.8651434921575
		5	0.1849034349504	3.719836618387	−4.2293343184042

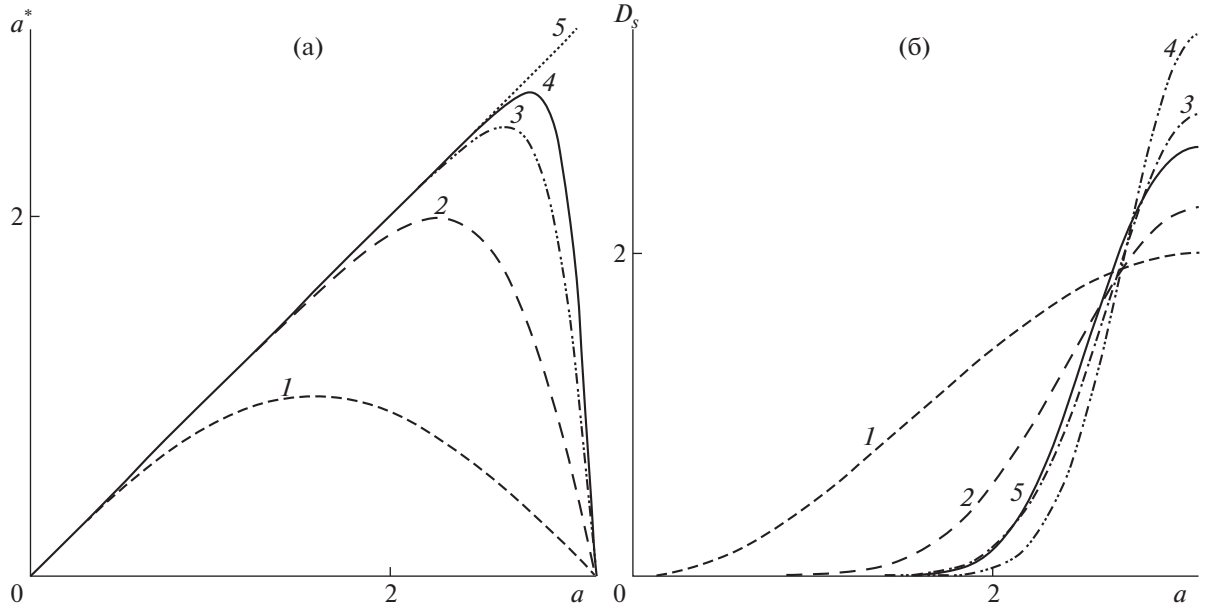
Далее следует расчет собственно вязких членов уравнений по формуле (7) с использованием граничных условий

$$\delta_{2,b}^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(a_{2,p}^{\delta,0}T_0 + a_{2,p}^{\delta,1}T_1)^{-1}(b_{2,p}^{\delta,1/2}T_{1/2} + b_{2,p}^{\delta,3/2}T_{3/2} + b_{2,p}^{\delta,5/2}T_{5/2} + b_{2,p}^{\delta,7/2}T_{7/2})$$

и

$$\delta_{3,b}^{(k)} = \frac{1}{n}\sum_{p=1}^{n}(a_{3,p}^{\delta,0}T_0 + a_{3,p}^{\delta,1}T_1 + a_{3,p}^{\delta,2}T_2)^{-1}(b_{3,p}^{\delta,1/2}T_{1/2} + b_{3,p}^{\delta,3/2}T_{3/2} + b_{3,p}^{\delta,5/2}T_{5/2} + b_{4,p}^{\delta,7/2}T_{7/2}),$$

базовые коэффициенты которых содержатся в табл. 8 и 9.



Фиг. 1. Дисперсионные и диссипативные свойства разностных операторов.

Остальные коэффициенты оператора $\delta_{2,b}^{(k)}$ определяются зависимостями $a_{2,p}^{\delta,1} = 1 - a_{2,p}^{\delta,0}$, $a_{2,p}^{\delta,2} = 0$, $b_{2,p}^{\delta,3/2} = -2 - a_{2,p}^{\delta,0} - 3b_{2,p}^{\delta,1/2}$, $b_{2,p}^{\delta,5/2} = 3 + 2a_{2,p}^{\delta,0} + 3b_{2,p}^{\delta,1/2}$, $b_{2,p}^{\delta,7/2} = -1 - a_{2,p}^{\delta,0} - b_{2,p}^{\delta,1/2}$ с использованием данных табл. 8.

С помощью приведенных в табл. 9 коэффициентов все остальные коэффициенты разности $\delta_{3,b}^{(k)}$ находятся из соотношений $a_{3,p}^{\delta,2} = 1 - a_{3,p}^{\delta,0} - a_{3,p}^{\delta,1}$, $b_{3,p}^{\delta,3/2} = -1 - 2a_{3,p}^{\delta,0} - a_{3,p}^{\delta,1} - 3b_{3,p}^{\delta,1/2}$, $b_{3,p}^{\delta,5/2} = 1 + 4a_{3,p}^{\delta,0} + 2a_{3,p}^{\delta,1} + 3b_{3,p}^{\delta,1/2}$, $b_{3,p}^{\delta,7/2} = -2a_{3,p}^{\delta,0} - a_{3,p}^{\delta,1} - b_{3,p}^{\delta,1/2}$.

Интегрирование разностных уравнений по времени может осуществляться двумя разными способами. Если размеры узлов сетки достаточно крупные и не слишком отличаются между собой, то целесообразно применять явный метод Рунге–Кутты 4-го порядка по времени. Соотношение временного и пространственного шагов при этом составляет 0.1–0.2. При наличии же тонких сдвиговых и пограничных слоев, требующих большого количества узлов для их подробного описания и значительных размерах области расчета, данный подход становится слишком затратным. В этом случае используется неявный метод Гаусса релаксации в линиях 2-го порядка. Отношение временного и пространственного шагов здесь близко к единице, а затраты на один шаг по времени меньше, чем для метода Рунге–Кутты.

Анализ свойств мультиоператорных разностей (3) проведем применительно к линейному скалярному уравнению $\partial f/\partial t + \partial f/\partial x = 0$. Рассмотрим поведение таких характеристик, как дисперсия и диссипация, взяв за пример работы [3], [4]. Полагая, что решение уравнения в каждой точке j на сетке $\omega_h = \{x_j = jh,\ h = \text{const}\}$ имеет вид $U(t)\exp(ikhj)$, где i – мнимая единица, а k – волновое число, получаем обыкновенное дифференциальное уравнение $U_t' + \omega(kh)U = 0$. Дискретизация временной производной не рассматривается. Для каждой разностной схемы посредством применения преобразования Фурье к исходным операторам T_n, Δ_1 и Δ_2, имеет место соответствующая комплексная функция $\omega(kh)$ и свои выражения для фазовой скорости (дисперсии) $D_\alpha = \mathrm{Re}\left[\omega(kh)/ikh\right]$ и диссипации $D_s = \mathrm{Re}\left[\omega(kh)\right]$. Дисперсия D_α, а точнее, выражение $1 - D_\alpha$, характеризует собой фазовую ошибку, вносимую заменой дифференциального оператора разностным. Диссипация D_s отражает способность разностной схемы противостоять возникающим паразитным осцилляциям решения. Поскольку для составных компактных схем разрешающая способность и стабилизирующие свойства не зависят одно от другого, целесообразно проследить за их изменением с ростом порядка аппроксимации используемых конечных разностей.

На фиг. 1 представлены зависимости схемного волнового числа α^* и диссипации D_s от волнового числа $\alpha = kh$ для центральных разностей 2-го, 6-го (кривые *1*, *2*) и мультиоператорных 14-го и

22-го порядков (кривые *3* и *4*). Параметр α^* представляет собой произведение фазовой скорости (дисперсии) схемы и физического волнового числа α. Повышение порядка аппроксимации разностных схем ведет к монотонному улучшению их разрешающей способности. Свойства конечных разностей при этом асимптотически приближаются к свойствам разностей дифференциальных. Вместе с тем отметим, что приращение чувствительности схем с повышением их порядка аппроксимации постоянно уменьшается, и говорить о разностных операторах, свойства которых полностью совпадали бы со свойствами дифференциальных, не приходится. В данном случае схема 22-го порядка демонстрирует отличие от идеальной теоретической кривой $\alpha^* = \alpha$, отмеченной цифрой *5*, лишь в области самых коротких волн, что характеризует ее значительно лучшую чувствительность по сравнению со схемами 2-го и 6-го порядков.

Операторы центральных разностей сами по себе не обладают какими-либо диссипативными свойствами, необходимыми для проведения устойчивых расчетов при численном моделировании различных задач вычислительной физики. Преодолевается этот недостаток за счет введения в дифференциальные схемы стабилизирующего механизма в виде отнормированных и ориентированных против потока операторов вида Δ_2^m, представляющих собой операторы четных разностей. При этом общий порядок схемы может понижаться до $(2m-1)$-го. В отличие от аппроксимирующих компонентов составных схем, для диссипативных операторов не используется мультиоператорный подход. В силу естественного устройства разностей диффузного типа повышение их порядка связано с укрупнением используемого сеточного шаблона.

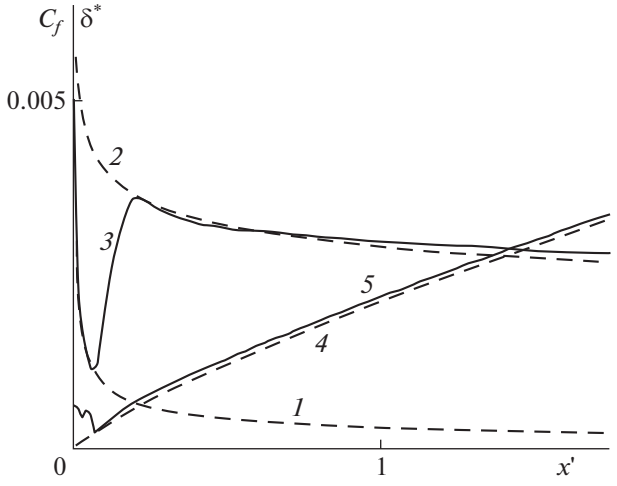
Зависимости диссипации D_s от волнового числа α для схем, использующих для стабилизации различные четные разности, представлены на фиг. 1б: кривая *1* соответствует 2-й разности, *2* – 8-й, *3* – 16-й, *4* – 24-й. С ростом значения m наблюдается сильное снижение диссипативных свойств разностных схем в областях длинных и средних волн. С одной стороны, это обстоятельство накладывает существенные ограничения на их применение для задач, характеризующихся наличием сильных градиентов параметров потока в поле течения. С другой стороны, низкий уровень схемной вязкости не только допустим, но и желателен на участках течения с достаточно гладкими распределениями рассчитываемых величин. Повысить диссипативные свойства составных схем возможно путем использования линейных комбинаций четных разностей заданного и более высокого порядков на основе подхода, описанного в [23]. На фиг. 1б – это кривая *5*, полученная для комбинации 22-й и 24-й разностей. Ее диссипативные свойства практически не отличаются от таковых 14-й разности.

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

4.1. Турбулентный пограничный слой на плоской пластине

Рассмотрим формирование турбулентного пограничного слоя на плоской пластине потоком вязкого газа с числом Маха 0.003 и числом Рейнольдса, посчитанном по расстоянию от передней кромки пластины до расчетного сечения: 5×10^6. При температуре воздуха 300 K скорость течения соответствует физическому значению, примерно, 1 м/с, что больше напоминает течение жидкости, чем газа. Кинетическая энергия турбулентности в набегающем потоке $k_\infty = 10^{-6}$, а вихревая вязкость $\mu_{t,\infty} = 0.5\mu_{l,\infty}$. В данном случае использование уравнения состояния дает вычислительную величину давления в среде, почти на пять порядков превышающую квадрат скорости. Расчет проводится на сетке с 300 узлами вдоль пластины и 200 узлами поперек. За характерный линейный размер принимается расстояние от передней кромки пластины до расчетного сечения $x' = x/L = 1$. Входная граница области расчета отстоит от передней кромки пластины на расстояние, равное 50 характерным размерам, а верхняя и выходная – на 10. Расстояние от поверхности пластины до ближайшего к ней узла внутренней области расчета составляет 10^{-5} от характерного размера.

Для расчета параметров потока в переходной области течения используется коэффициент C_{lt}, при этом его значения не вычисляются, а задаются в зависимости от значения числа Re и расстояния от передней кромки пластины x'. В том случае, когда течение считается ламинарным, коэффициент C_{lt}, стоящий перед источниковыми членами уравнения турбулентной вязкости, равен нулю. Переход задается за счет постепенного подключения в расчете вектора источниковых членов. При изменении величины $\mathrm{Re}\, x'$ от 5×10^5 до 10^6 значение коэффициента C_{lt} меняется линейно от 0 до 1. Далее по потоку течение турбулентное, $C_{lt} = 1$.



Фиг. 2. Распределения коэффициента трения и толщины вытеснения пограничного слоя.

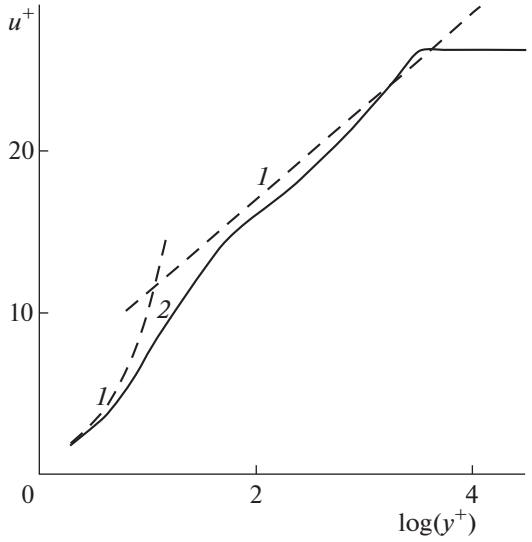
На фиг. 2 представлено распределение местного коэффициента трения $C_f = 2\mu\dfrac{\partial u}{\partial y}(\rho_\infty u_\infty^2)^{-1}$. Цифрами *1* и *2* отмечены кривые ламинарного и турбулентного распределения коэффициента трения из [20]. Полученное в расчете распределение поверхностного трения с учетом обозначено цифрой *3*. Оно согласуется с ламинарным распределением на начальном участке пластины и турбулентным на ее основном участке. Используемая модель турбулентной вязкости позволяет воспроизводить поверхностное трение на участке переходного течения. На фиг. 2 также представлены графики толщины вытеснения пограничного слоя δ^* – теоретическая кривая *4* и расчетная *5*.

Полученные в расчетах профили коэффициента турбулентной вязкости μ_t и горизонтальной компоненты вектора скорости $u^+ = u/u_\tau$ представлены на фиг. 3. По оси абсцисс отложен параметр $\log(y^+)$, где $y^+ = yu_\tau\rho_w(\mathrm{Re}\,\mu_w)^{-1}$ – универсальное расстояние, а по оси ординат $u_\tau = [\mu_w(\mathrm{Re}\,\rho_w)^{-1}\partial u/\partial y]^{1/2}$ – динамическая скорость. Цифрой *1* отмечены теоретические кривые ламинарного подслоя и логарифмического участка пограничного слоя из [20]. Полученный в расчете профиль (цифра *2*) достаточно хорошо с ними согласуется.

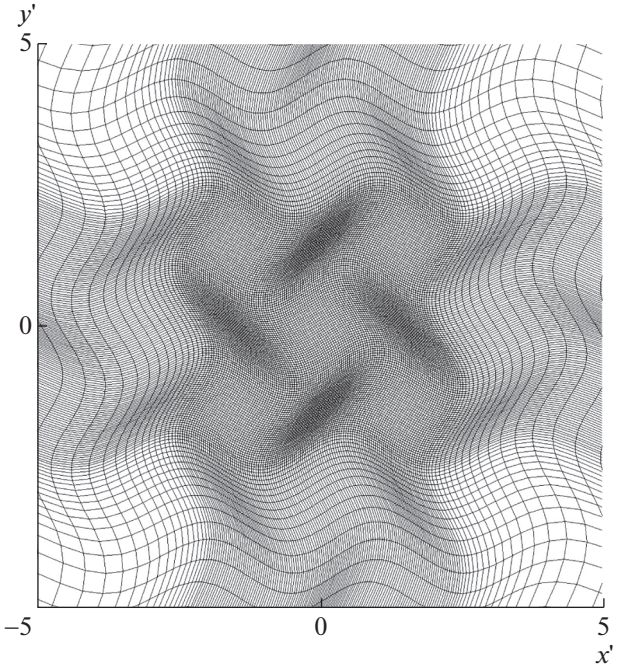
4.2. Естественная деформация плоского изолированного вихря

Звуковое излучение вихрей эллиптической формы на основе решения уравнений Эйлера исследовалось в [24], [25]. Расчеты М.В. Липавского и А.И. Толстых с использованием мультиоператорных схем (см. [16]) впервые продемонстрировали превращение круглого вихря в эллиптический на относительно длительных временах расчета без какого-либо внешнего воздействия. В данном случае рассматривается одиночный круглый вихрь на искривленной сетке, азимутальная скорость v_c которого на расстоянии от центра r_c (характерные скорость и длина) соответствует числу Маха 0.2. Радиальная составляющая скорости вихря, вращающегося по часовой стрелке, в начальный момент времени полагается равной нулю, азимутальная при $r' = r/r_c \le 1$ пропорциональна радиусу $v = v_c r'$, а при $r' > 1$ задается формулой $v = v_c \exp[0.5(1 - r'^2)]$ из [26]. Остальные параметры рассчитываются из адиабатических соотношений в предположении наличия на бесконечности условий торможения потока, что дает в качестве начального приближения решение, близкое к точному. Для стабилизации используется 24-я разность со значением нормирующей константы перед ней 0.2 вместо обычно используемой единицы. Интегрирование по времени проводится на основе метода Рунге–Кутты четвертого порядка. Соотношение шага по времени к минимальному расстоянию между узлами разностной сетки равняется 0.025.

Расчеты выполняются на сетке с 200 узлами по каждому пространственному направлению. Фрагмент ее центральной части приведен на фиг. 4. В отличие от предыдущих расчетов подоб-



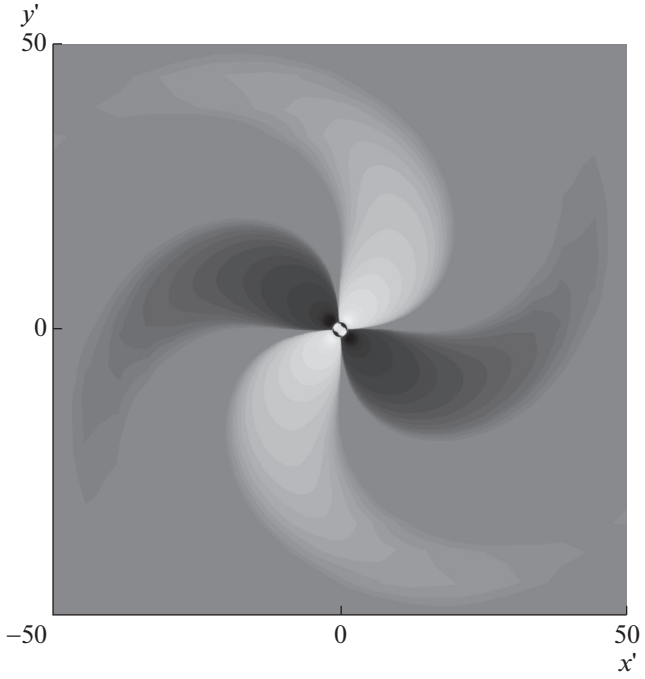
Фиг. 3. Профиль скорости в пограничном слое.



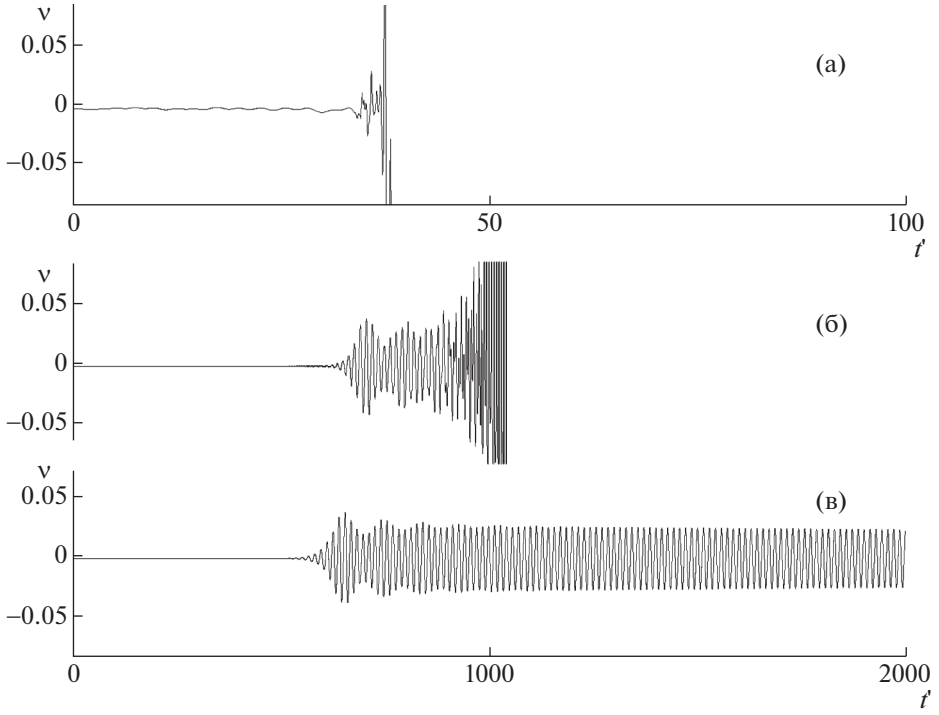
Фиг. 4. Фрагмент разностной сетки.

ной задачи, здесь используется криволинейная сетка. Присутствие в уравнениях ненулевых метрических коэффициентов служит дополнительной проверкой работоспособности мультиоператорных схем.

Полученное на момент времени $t' = tr_c/v_c = 2000$ поле относительного давления $\Delta p' = (p - p_i)/p_0$, где p_i – давление на начальный момент времени, а p_0– давление торможения, представлено на фиг. 5. Поле содержит 35 уровней в диапазоне изменения относительного давления от –0.0005 до 0.0005. Результаты получены на основе схемы 22-го порядка. При вращении вихря эллиптической формы перед его большей полуосью возникает зона повышенного давления, а за ней – пониженного. Данные области давления распространяются в окружающем про-



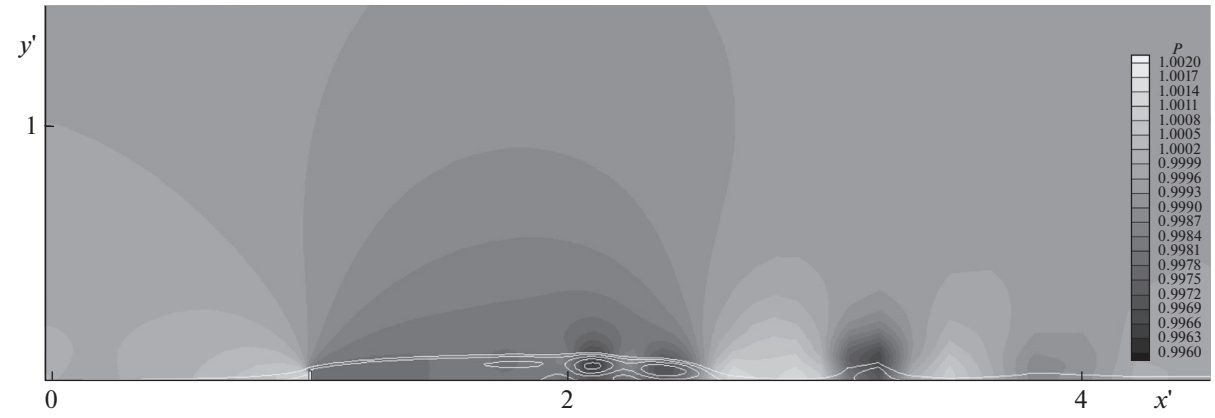
Фиг. 5. Нестационарное относительное давление.



Фиг. 6. Изменение по времени вертикальной компоненты вектора скорости.

странстве. Сам вихрь при этом продолжает вращение вокруг своей оси. В результате имеет место картина распространения звукового давления, напоминающая четырехлопастной вентилятор.

Изменение по времени t' вертикальной компоненты вектора скорости v в точке $x' = 0$, $y' = 2$ представлено на фиг. 6: (а) – центрально-разностная схема 2-го порядка, (б) – компактная схема



Фиг. 7. Поле давления и линии тока отрывного течения.

6-го и (в) – мультиоператорная схема 22-го порядков. Фрагменты (б) и (в) выполнены в одном временном масштабе, фрагмент (а) – в большем. Это вызвано тем, что уже при $t' = 40$ происходит развал решения, выполняемого по схеме 2-го порядка. Схема 6-го порядка разваливается при t' немногим более 1000, а использование разности 22-го порядка позволяет проводить устойчивый расчет задачи по крайней мере вдвое дольше. Можно сделать вывод, что используемый стабилизатор в виде 24-й разности не справляется с ошибками аппроксимации, имеющими место для схем 2-го и 6-го порядков. В то же время схема 22-го за счет своей точности в более сильной коррекции не нуждается.

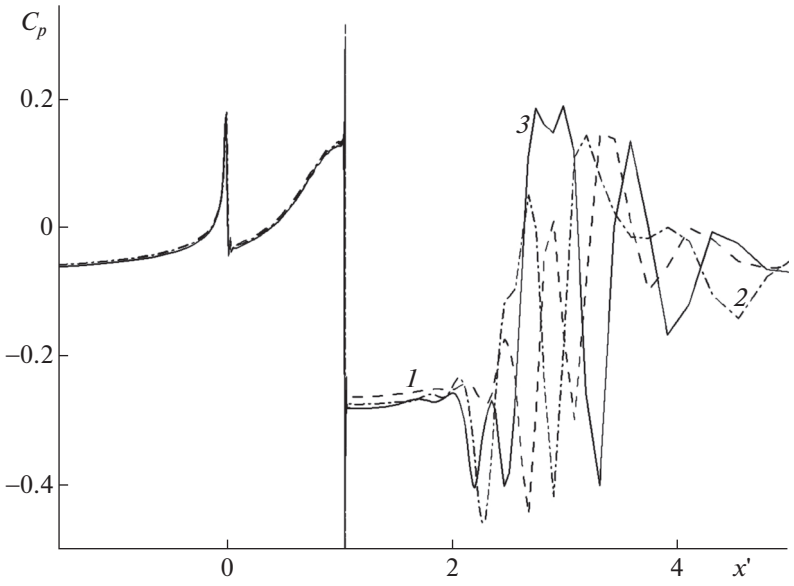
Причину данной деформации круглого вихря следует искать во внутренних волновых процессах и сжимаемости газа. В процессе вращения вихря постепенно нащупывается резонансная частота, примерно на порядок меньшая частоты кругового вращения, под действием которой круглый вихрь превращается в эллиптический. Следует отметить, что начало трансформации вихря при расчетах схемами 6-го и 22-го порядков наблюдается примерно в одно время. На волновую природу явления указывают также результаты других расчетов, согласно которым снижение числа Маха на границе ядра M_c приводит к увеличению времени начала трансформации, а его рост – к снижению.

4.3. Обтекание поперечного ребра жесткости на пластине

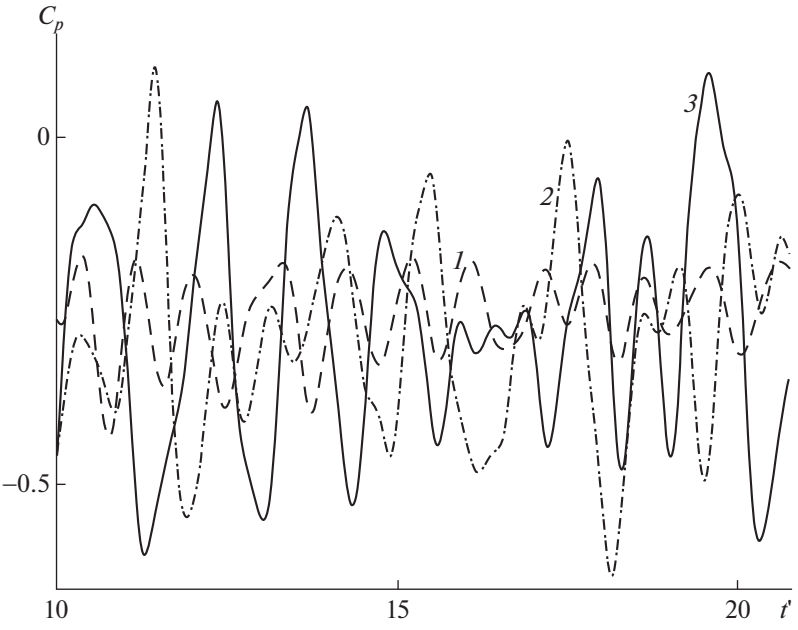
Дозвуковые течения являются нестационарными по своей природе. Важную роль в них играют силы вязкости, ответственные за вихреобразование и отрыв пограничного слоя. За счет того, что каждая точка поля чувствует изменение давления в любой другой точке пространства, течение отличается высокой вариативностью, т.е. отсутствует полная повторяемость в периодических процессах, по крайней мере, при достаточно высоких числах Re. Возникает вопрос, как могут отличаться результаты расчетов с использованием схем разного порядка аппроксимации и одинаковыми стабилизирующими операторами. При этом влияние самих стабилизирующих добавок должно быть минимальным за счет применения в них четных разностей высокого порядка.

Рассмотрим дозвуковое обтекание вязким газом невысокого ребра жесткости, расположенного на плоской пластине поперек набегающего потока. Число Маха набегающего потока $M_\infty = 0.1$. Число Рейнольдса, посчитанное по длине L – расстоянию от передней кромки пластины до ребра жесткости, 10^4. В данном двумерном случае ребро жесткости представляет собой прямоугольник высотой $h = 0.05L$ и толщиной $0.01L$. Толщина пограничного слоя перед выступом примерно равняется его высоте h, а число Рейнольдса $\mathrm{Re}_h \approx 500$. На первом этапе расчета на основе схемы 6-го порядка формируется поле течения. Далее используются схемы 6-го, 14-го и 22-го порядков и расчеты ведутся до времени $t' = tL/U_\infty = 20$, после чего полученные результаты сравниваются. Во всех трех случаях в качестве стабилизатора используется комбинация 22-й и 24-й разностей.

Мгновенное поле относительного давления $p' = p/p_\infty$, полученное с применением схемы 22-го порядка, представлено на фиг. 7. Уровни нанесены в диапазоне от 0.996 до 1.002 с шагом 0.003. Там же приведены линии тока в виде линий постоянного значения функции тока $\psi = \int \rho u dy - \int \rho v dx$. Всего нанесено пять линий в диапазоне от –0.015 до 0.005 с шагом между ни-



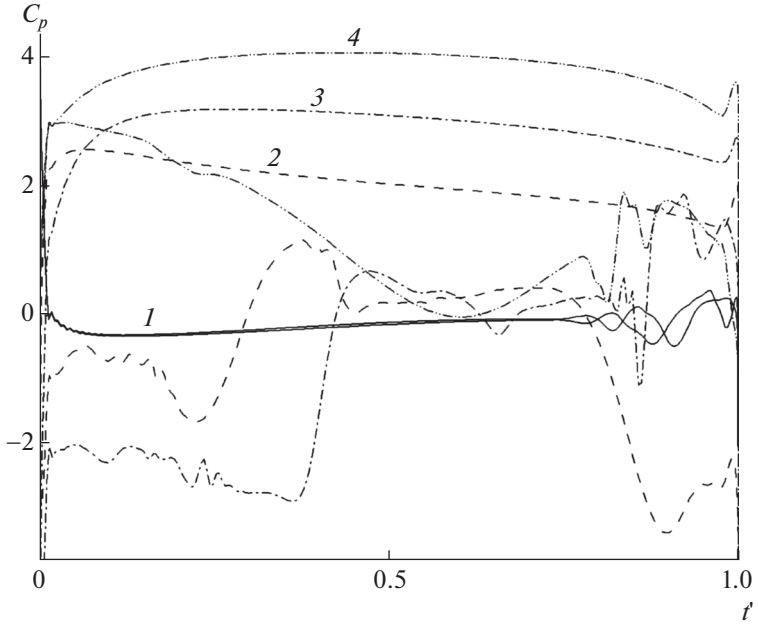
Фиг. 8. Распределение коэффициента давления.



Фиг. 9. Изменение по времени коэффициента давления.

ми 0.005. Повышенные уровни давления наблюдаются у передней кромки пластины, перед ребром жесткости и в зоне присоединения потока за ним, а области пониженного давления – в вихревых зонах ниже препятствия. Пограничный слой отрывается перед ребром жесткости и присоединяется к его передней поверхности ниже верхней кромки. Далее он снова отрывается с верхней поверхности ребра и образует нестационарную отрывную зону, протяженность которой более чем в 30 раз больше высоты обтекаемого препятствия. Удаленная от препятствия область отрыва совершает колебания, отделившиеся вихревые структуры сносятся вниз по потоку.

На фиг. 8 представлены распределения вдоль координаты $x' = x/L$ коэффициента давления $C_p = 2(p - p_\infty)/\rho_\infty U_\infty^2$. Кривая *1* получена с использованием схемы 6-го порядка, *2* – 14-го и *3* – 22-го. В областях перед препятствием, где течение стационарно, результаты расчетов разными схемами хорошо совпадают между собой. Там же, где имеют место пульсации потока в следе за ребром жесткости, наблюдаются отличия. Это видно и на фиг. 9, где представлены зависимости по вре-



Фиг. 10. Распределения коэффициента давления на поверхности.

мени коэффициента давления на поверхности пластины в точке $x' = 2.25$. Минимальные амплитуды колебания коэффициента давления имеют место в расчете схемой 6-го порядка, максимальные – в расчете схемой 22-го порядка. При примерно одинаковом числе Струхаля колебания в данной точке $St = fh/U_\infty = 0.0168$ (f – частота колебания) амплитуда звукового давления при использовании схемы 6-го порядка составляет 123 dB, 14-го и 22-го соответственно 131 и 134.6 dB.

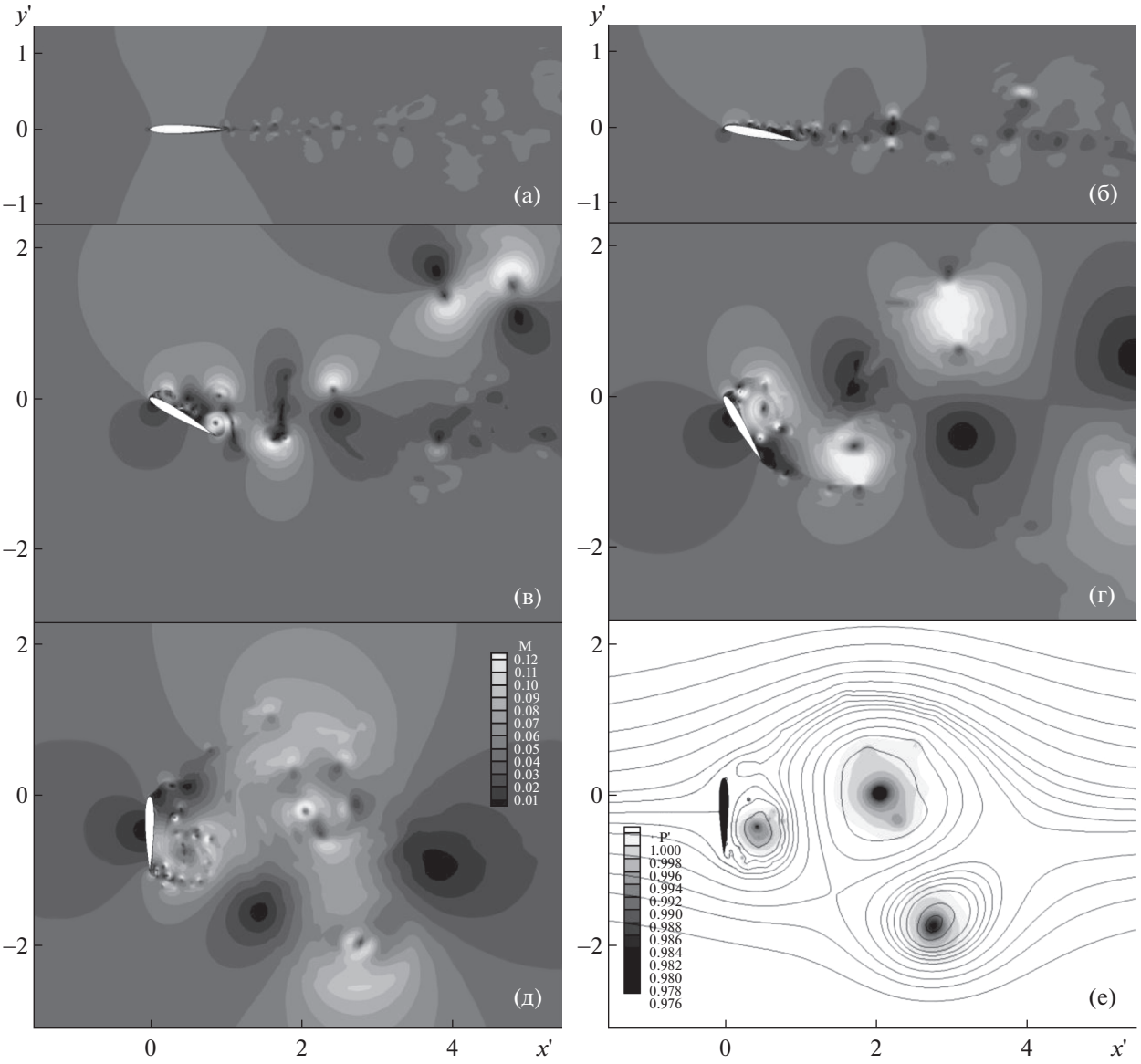
4.4. Обтекание аэродинамического профиля

На аэродинамической поверхности, движущейся под углом к встречному потоку воздуха, может возникать отрыв пограничного слоя, способный сильно влиять на ее подъемную силу. При этом общая картина течения существенно зависит от уровня вязкости в потоке. Шум аэродинамического профиля, возникающий при его ламинарном обтекании, исследуется в [5], [7] и многих других работах, механизм его возникновения рассматривается, в частности, в [27], [28]. Источником нестационарных эффектов является отрыв пограничного слоя, возникающий на его поверхности под влиянием положительного градиента давления. Эмпирические критерии отрыва ламинарного и турбулентного пограничных слоев сформулированы в [29]. В отличие от турбулентного режима течения, отрыв ламинарного пограничного слоя требует меньшего перепада давления, к тому же снижающегося с увеличением числа Рейнольдса. Подобные условия реализуются, например, в области применения малоразмерных беспилотных летательных аппаратов.

Рассмотрим течение, формирующееся около симметричного профиля NACA0012, при числе Маха набегающего потока 0.05 и числе Рейнольдса, посчитанном по хорде c, 10^5. В расчетах используется сетка типа O с 500 узлами в каждом пространственном направлении. Расстояние до внешней границы составляет $20c$, а от поверхности тела до ближайшего узла – $2.5 \times 10^{-5}c$. На фиг. 10 представлены полученные графики распределения коэффициента давления по координате $x' = x/c$ для следующих значений угла атаки набегающего потока: *1* – 0°, *2* – 30°, *3* – 60° и *4* – 90°. Значение $x' = 0$ соответствует передней кромке профиля, $x' = 1$ – задней. Отрыв пограничного слоя возникает уже при нулевом значении угла атаки на сужающихся частях как верхней, так и нижней сторонах профиля. Значение сформулированного в [29] критерия отрыва ламинарного пограничного слоя

$$C_{lam} = \frac{\mathrm{Re}\,\delta^{*2}}{2}\frac{dC_p}{dx}$$

составляет 1.1. В данном расчете получено значение 1.3. Возникающие отрывы пограничного слоя формируют колеблющийся ближний след за профилем и волновые структуры, распростра-

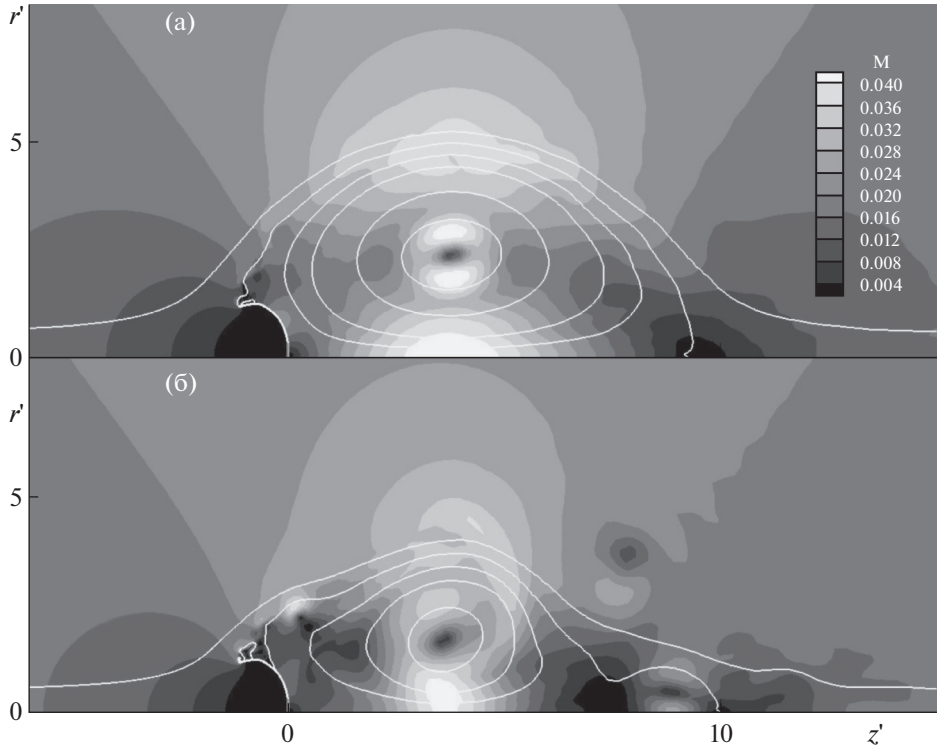


Фиг. 11. Поля равных уровней числа Маха (а)–(д), давления и линии тока (е).

няющиеся вверх по потоку. Давление повышается и понижается попеременно на верхней и нижней поверхностях.

При ненулевых углах атаки на наветренной стороне профиля давление повышенное и имеет форму плато. В то же время на подветренной стороне распределение давления неравномерно, что вызвано формирующимися здесь нестационарными отрывными течениями. Возникающий характер обтекания профиля неблагоприятно сказывается на его устойчивости.

Поля числа Маха представлены на фиг. 11: (а) – отмечен случай $\alpha = 0°$, (б) – 10°, (в) – 30°, (г) – 60° и (д) – 90°. Диапазон изменения изолиний от 0.01 до 0.12 с шагом 0.01. Об интенсивности течения в следе за профилем можно судить по насыщенности его крупными и мелкими вихревыми структурами. Если при нулевом угле атаки значения чисел Маха не превышают 0.07–0.08, то при 60° они достигают максимальных зафиксированных значений 0.135. При угле атаки 10° отрыв потока происходит у передней кромки, далее вихревые зоны смещаются вниз вдоль верхней поверхности профиля, по границе пограничного слоя наблюдаются локальные ускорения потока. С увеличением угла атаки растут и размеры вихревых структур. Отрывы пограничного слоя происходят на передней и задней кромках профиля, взаимодействуют между собой и сносятся вниз по потоку, формируя вихревой след. Поток у наветренной поверхности тормозится, а за профилем образуется течение, напоминающее дорожку Кармана. На фиг. 11д, по времени



Фиг. 12. Поля равных уровней числа Маха и линии тока.

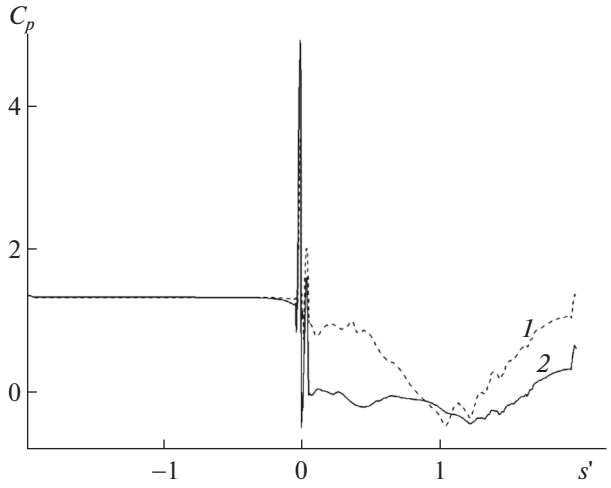
относящейся к началу формирования подобного следа при его поперечном обтекании профиля, видны мелкие вихри, цепочкой выстраивающиеся вокруг зоны отрыва в нижней его части. Остатки таких же структур просматриваются и вокруг отрывной зоны, расположенной выше и правее предыдущей.

На фиг. 11е представлено поле относительного давления $p' = p/p_\infty$ и линии тока в виде изолиний функции тока ψ, полученные при угле атаки 90°. Диапазон изменения относительного давления от 0.976 до 1 с шагом 0.002. Предельные значения изолиний функции тока -1.1 и 1.4. Минимальный шаг изолиний вблизи нулевого значения 0.01, далее он увеличивается. Области пониженного давления совпадают с центрами вихрей. Искривление изолиний у подветренной поверхности профиля подтверждает сложность течения и насыщенность донной области взаимодействующими вихрями.

4.5. Обтекание парашютного купола

Парашютные системы широко применяются на практике и имеют богатую историю разработок. Их проектирование сегодня не обходится без этапа математического моделирования, которое включает в себя совместное решение ряда нелинейных задач, одной из которых является нестационарное обтекание парашютного купола встречным потоком воздуха. В свою очередь, течение около парашютного купола моделируется на основе метода дискретных вихрей (см. [30]) или решения уравнений Эйлера невязкого газа методами конечных разностей (см. [31], [32]). Есть также примеры расчетов с использованием уравнений Навье–Стокса вязкой несжимаемой жидкости (см. [33], [34]).

Рассматривается осесимметричное обтекание жесткого непроницаемого парашютного купола потоком вязкого газа. Число Маха набегающего потока 0.02, что соответствует скорости 6–7 м/с. Поверхность купола представляет собой дугообразный участок границы расчетной области с нулевой толщиной, поскольку везде, за исключением ближайшей окрестности кромки, координаты точек наветренной и подветренной поверхностей купола попарно совпадают друг с другом. За характерный линейный размер принимается длина половины дуги купола R. В расчете используется сетка с 400 узлами вдоль поверхности и 460 – в радиальном направлении. Минимальный размер между узлами составляет $10^{-3}R$, а внешняя граница отодвигается от поверхности на расстояние 10^2R.



Фиг. 13. Распределения коэффициента давления.

Полученные поля числа Маха при ламинарном (число Рейнольдса 10^6) и турбулентном ($\mathrm{Re}_{R,\infty} = 3\times10^6$, $k_\infty = 10^{-6}$, $\mu_{t,\infty}/\mu_{l,\infty} = 0.3$) обтекании представлены на фиг. 12 в цилиндрических координатах $z' = zR^{-1}$, $r' = rR^{-1}$. Диапазон изменения равных значений числа Маха 0.004 до 0.04 с шагом 0.004. Там же приведены изолинии функции тока ψ в диапазоне от -3 до 0.5 с постоянным шагом 0.5. В обоих случаях за непроницаемым куполом формируются зоны возвратно-циркуляционного течения. Протяженность отрывной области значительно превышают размеры самого купола. В ламинарном случае, представленном на фиг. 12а, зона рециркуляции больше, чем в турбулентном (фиг. 12б). Это находит объяснение в более высоких значениях эффективной вязкости при использовании модели турбулентности. В зоне торможения перед куполом течение стационарно, у кромки наблюдаются колебания. В момент времени, зафиксированный на фиг. 12б, происходит отрыв вихря малого размера от основной возвратно-циркуляционной зоны.

Распределение коэффициента давления C_f вдоль поверхности купола представлено на фиг. 13. Нулевое значение координаты $s' = sR^{-1}$ соответствует кромке купола, отрицательные относятся к наветренной поверхности, положительные – подветренной. Результаты расчета ламинарного течения отмечены цифрой *1*, турбулентного – цифрой *2*. Различия между ними наблюдаются в районе парашютной кромки и на подветренной поверхности. Уровень донного давления для случая турбулентного течения ниже, чем ламинарного. Это означает, что в турбулентном течении коэффициент сопротивления выше. Используемые схемы высокого порядка и метод построения расчетной области могут служить основой для более совершенных расчетов на подвижных сетках с учетом проницаемости купола и напряжений в его оболочке.

ЗАКЛЮЧЕНИЕ

Рассмотренные мультиоператорные разностные схемы обладают высокими разрешающими способностями, что является их преимуществом при решении задач динамики вязкого сжимаемого газа. Максимальный достигнутый порядок аппроксимации схем составляет 22 при пяти используемых операторах. Диссипативные добавки на основе 22-й и 24-й разностей не имеют сильных стабилизирующих свойств. Они не дают возможности расчета задач с сильными градиентами параметров потока в поле течения, но позволяют моделировать дозвуковые течения вязкого сжимаемого газа при достаточно низких числах Маха. Данный класс задач имеет свою специфику и представляет интерес для практики. В качестве примера рассмотрены задачи обтекания вязким газом аэродинамического профиля в широком диапазоне изменения угла атаки и отрывного течения около непроницаемого парашютного купола.

СПИСОК ЛИТЕРАТУРЫ

1. *Русанов В.В.* Разностные схемы третьего порядка точности для сквозного счета разрывных решений // Докл. АН СССР. 1968. Т. 180. № 6. С. 1303–1305.

2. *Толстых А.И.* О методе численного решения уравнений Навье–Стокса сжимаемого газа в широком диапазоне чисел Рейнольдса // Докл. АН СССР. 1973. Т. 210. № 1. С. 48–51.
3. *Tolstykh A.I.* High accuracy non-centered compact schemes for fluid dynamics applications. Singapore: World Scient., 1994.
4. *Lele S.K.* Compact finite difference schemes with spectral-like resolution // J. Comput. Phys. 1992.V. 102. P. 16–42.
5. *Visbal M.R., Gaitonde D.V.* On the use of high-order finite-difference schemes on curvilinear and deforming meshes // J. Comput. Phys. 2002. V. 181. P. 155–185.
6. *Mardsen O., Bogey C., Bailly C.* High-order curvilinear simulations of flows around non-cartesian bodies // J. Comput. Acoustics. 2005. V. 13. № 4. P. 731–748.
7. *Tam C.K.W., Ju H.* Numerical simulation of the generation of airfoil tones at a moderate Reynolds number // AIAA Paper № 2006–2502. 2006. 23 p.
8. *Sandberg R.D., Jones L.E., Sandham N.D., Joseph P.F.* Direct numerical simulations of noise generated by airfoil trailing edges // AIAA Paper № 2007–3469. 2007. 15 p.
9. *Harten A.* ENO schemes with subcell resolution // J. Comput. Phys. 1989. V. 83. P. 148–184.
10. *Harten A., Engquist B., Osher S., Chakravarthy S.R.* Uniformly high order essentially non-oscillatory schemes III // J. Comput. Phys. 1987. V. 71. P. 231–303.
11. *Hu C., Shu C.W.* Weighted essentially non-oscillatory schemes on triangular meshes // J. Comput. Phys. 1999. V. 150. P. 97–127.
12. *Dumbser M., Kaser M., Titarev V.A., Toro E.F.* Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems // J. Comput. Phys. 2007. V. 226. P. 204–243.
13. *Abalakin A., Bachwalow P., Kozubskaya T.* Edge-based reconstruction schemes for prediction of near field flow region in complex aeroacoustic problems // Int. J. Aeroacoustics. 2014. V. 13. № 3–4. P. 207–234.
14. *Савельев А.Д.* Составные компактные схемы высокого порядка для моделирования течений вязкого газа // Ж. вычисл. матем. и матем. физ. 2007. Т. 47. № 8. С. 1389–1403.
15. *Толстых А.И.* Мультиоператорные схемы произвольного порядка, использующие нецентрированные компактные аппроксимации // Докл. АН. 1999. Т. 366. № 3. С. 319–322.
16. *Толстых А.И.* О мультиоператорном методе построения аппроксимаций и схем произвольно высокого порядка // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. № 1. С. 56–73.
17. *Савельев А.Д.* О мультиоператорном представлении составных компактных схем // Ж. вычисл. матем. и матем. физ. 2014. Т. 54. № 10. С. 1580–1593.
18. *Савельев А.Д.* О разностных схемах 18-го и 22-го порядков для уравнений с конвективными и диффузными членами // Матем. моделирование. 2017. Т. 29. № 6. С. 35–47.
19. *Menter F.R.* Zonal two equation $k - \omega$ turbulence models for aerodynamic flows // AIAA Paper 93–2906. 1993. 21 p.
20. *Лойцянский Л.Г.* Механика жидкости и газа. М.: Наука, 1987. 840 с.
21. *Steger J.L.* Implicit finite-difference simulation of flow about arbitrary two-dimensional geometries // AIAA J. 1978. V. 16. P. 678–685.
22. *Steger J.L., Warming R.F.* Flux vector splitting of the inviscid gasdynamic equations with applications to finite-difference methods // J. Comput. Phys. 1981. V. 40. P. 263–293.
23. *Савельев А.Д.* О структуре внутренней диссипации составных компактных схем для решения задач вычислительной газовой динамики // Ж. вычисл. матем. и матем. физ. 2009. Т. 49. № 12. С. 2232–2246.
24. *Chan W.M., Sheriff K., Pulliam T.H.* Instabilities of two-dimensional inviscid compressible vortices // J. fluid mech. 1993. V. 253. P. 173–209.
25. *Яковлев П.Г.* Излучение звука плоским локализованным вихрем // Акустический журнал. 2012. Т. 58. № 4. С. 563–568.
26. *Yee H.C., Sandham N.D., Djomehri M.J.* Low dissipation high order shock-capturing methods using characteristic-based filters // J. Comput. Phys. 1999. V. 150. P. 199–238.
27. *Desquenes G., Terracol M., Sagaut P.* Numerical investigation of the tone noise mechanism over laminar airfoils // J. Fluid Mech. 2007. V. 591. P. 155–182.
28. *Савельев А.Д.* Мультиоператорные компактные схемы высокого порядка в численном моделировании нестационарного дозвукового обтекания аэродинамического профиля // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 2. С. 291–303.
29. *Бам-Зеликович Г.М.* Расчет отрыва пограничного слоя // Изв. АН СССР. ОТН. 1954. № 12. С. 68–85.
30. *Белоцерковский С.М., Ништ М.И., Пономарев А.Т., Рысев О.В.* Исследование парашютов и дельтапланов на ЭВМ. М.: Машиностр., 1987. 240 с.
31. *Днепров И.В., Пономарев А.Т., Рысев О.В., Семушин С.А.* Исследование процессов нагружения и деформирования парашютов // Матем. моделирование. 1993. Т. 5. № 3. С. 97–109.
32. *Морозов В.И., Пономарев А.Т.* Моделирование нагружения парашютов с учетом изменения формы и проницаемости купола // Вестн. Харьковского нац. университета. 2008. Сер. Матем. моделирование. № 809. С. 148–161.
33. *Tutt B., Charles R., Roland S., Noetscher G.* Development of parachute simulation techniques in LS-DYNA // 11th Internat. LS-DYNA Users Conf. 2010. Detroit. P. 19–25.
34. *Yunpeng M., Jinge Z.* The simulation of canopy fabric air permeability's influence on the round parachute during the landing process // Int. Industrial Informat. Computer Engineer. Conf. 2015. Xi'an. Shaanxi. China. January 10–11. P. 2156–2159.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 303–311

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 517.5

ОБРАТНАЯ ЗАДАЧА ДЛЯ УРАВНЕНИЙ СЛОЖНОГО ТЕПЛООБМЕНА С ФРЕНЕЛЕВСКИМИ УСЛОВИЯМИ СОПРЯЖЕНИЯ[1)]

© 2021 г. А. Ю. Чеботарев

690041 Владивосток, ул. Радио, 7, Институт прикладной математики ДВО РАН, Россия

e-mail: cheb@iam.dvo.ru

Поступила в редакцию 12.02.2019 г.
Переработанный вариант 20.08.2020 г.
Принята к публикации 16.09.2020 г.

Рассматривается обратная задача для системы полулинейных эллиптических уравнений, моделирующих радиационный теплообмен с френелевскими условиями сопряжения на поверхностях разрыва коэффициента преломления. Задача состоит в отыскании правой части уравнения теплопроводности, являющейся линейной комбинацией данных функционалов, по заданным значениям этих функционалов на решении. Разрешимость обратной задачи доказана без ограничений малости. Представлено достаточное условие единственности решения. Библ. 41.

Ключевые слова: стационарные уравнения радиационного теплообмена, френелевские условия сопряжения, обратная задача, нелокальная разрешимость.

DOI: 10.31857/S0044466921020058

1. ВВЕДЕНИЕ

Задачи радиационно-кондуктивного (сложного) теплообмена представляют интерес в связи с инженерными и медицинскими приложениями (см. [1]–[5]). В [6]–[22] выполнен анализ краевых задач и задач оптимального управления для уравнений сложного теплообмена с диффузионным P_1 приближением уравнения переноса излучения. Анализ различных краевых задач, связанных с радиационным теплообменом, представлен в [23]–[28].

В [21] представлены построение и анализ стационарной модели сложного теплообмена в рамках P_1 приближения для многокомпонентной трехмерной области с учетом эффектов отражения и преломления на поверхностях разрыва коэффициента преломления. Настоящая работа посвящена анализу обратной задачи для указанной нелинейной модели сложного теплообмена. Задача заключается в отыскании неизвестных интенсивностей тепловых источников (объемных или поверхностных), а также соответствующих полей температуры и теплового излучения, по заданным значениям некоторых функционалов на решении краевой задачи. Близкие обратные задачи для стационарных уравнений сложного теплообмена рассмотрены в [29] и для квазистационарных уравнений в [31], [41].

Задачи восстановления неизвестных функций источников в эллиптических и параболических уравнениях и системах с интегральными и точечными условиями переопределения рассматривались в работах С.Г. Пяткова и др. (см. [32]–[35]). Обратные задачи с конечномерным переопределением для уравнений Навье–Стокса, уравнений тепловой конвекции и других моделей сплошных сред представлены в [36]–[41].

Статья организована следующим образом. В разд. 2 ставится прямая краевая задача с условиями сопряжения, определяются пространства и операторы, обратная задача формулируется в виде системы уравнений с операторными коэффициентами. Разрешимость обратной задачи доказана в разд. 3. Достаточное условие единственности решения получено в разд. 4, а в разд. 5 представлено доказательство вспомогательных результатов.

[1)]Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (проект 075-15-2019-1878).

2. ПОСТАНОВКА И ФОРМАЛИЗАЦИЯ КРАЕВОЙ ЗАДАЧИ С ФРЕНЕЛЕВСКИМИ УСЛОВИЯМИ СОПРЯЖЕНИЯ

Процесс сложного теплообмена рассматривается в ограниченной липшицевой области $\Omega \subset \mathbb{R}^3$. В области Ω содержится конечное число липшицевых подобластей Ω_j, $j = 1, 2, \ldots, p$, замыкания которых не пересекаются.

Подобласть

$$\Omega_0 = \Omega \backslash \left(\bigcup_{j=1}^{p} \bar{\Omega}_j \right)$$

является внешней, при этом $\Gamma = \partial\Omega \subset \Gamma_0 = \partial\Omega_0$, $\Gamma_j = \partial\Omega_j \subset \Gamma_0$, $j = 1, 2, \ldots, p$.

Процесс описывается следующими функциями: θ – нормализованная температура и φ – нормализованная интенсивность теплового излучения, усредненная по всем направлениям. Указанные функции в каждой из областей Ω_j, $j = 0, 1, \ldots, p$, удовлетворяют уравнениям

$$-a\Delta\theta + b(\theta^3|\theta| - \varphi) = f_s, \quad -\alpha\Delta\varphi + \beta(\varphi - \theta^3|\theta|) = 0. \tag{1}$$

Положительные физические параметры a, b, α и β, описывающие свойства среды, определяются стандартным образом (см. [21]). Отметим, что указанные параметры, так же как и коэффициент преломления $n > 0$, принимают постоянные значения в областях Ω_j, $j = 0, 1, \ldots, p$, и при этом, что важно, $b = \sigma\beta n^2$, $\sigma = \text{const} > 0$. Функция f_s моделирует тепловые источники.

На внешней границе $\Gamma = \partial\Omega$ заданы краевые условия

$$\{a\partial_\nu\theta + c(\theta - \theta_b)\}\big|_\Gamma = 0, \quad \{\alpha\partial_\nu\varphi + \gamma(\varphi - \theta_b^4)\}\big|_\Gamma = 0, \tag{2}$$

где θ_b – заданная граничная температура, c – коэффициент теплопередачи, $0 < \gamma \le 1/2$ – параметр, зависящий от коэффициента излучения. Через ∂_ν обозначаем производную в направлении внешней нормали ν к границе.

Условия сопряжения для температуры $\theta_j = \theta|_{\Omega_j}$ и интенсивности излучения $\varphi_j = \varphi|_{\Omega_j}$ на внутренних границах $\Gamma_j = \partial\Omega_j$, $j = 1, 2, \ldots, p$, выведенные в [21], имеют следующий вид:

$$\theta_0 = \theta_j, \quad a_0\partial_\nu\theta_0 = a_j\partial_\nu\theta_j, \tag{3}$$

$$n_0^2\alpha_0\partial_\nu\varphi_0 = n_j^2\alpha_j\partial_\nu\varphi_j, \quad h_j(\varphi_j - \varphi_0) = \alpha_0\partial_\nu\varphi_0. \tag{4}$$

Здесь $\{a_j, \alpha_j, n_j\} = \{a, \alpha, n\}|_{\Omega_j}$, $h_j > 0$ – параметры, зависящие от коэффициентов отражения на внутренних границах (см. [21]).

Для формализации краевой задачи будем использовать пространства Лебега L^s, $1 \le s \le \infty$, и пространства Соболева $H^s = W_2^s$. Обозначим $H = L^2(\Omega)$, $V = H^1(\Omega)$ и

$$W = \{w \in H : w_j = w|_{\Omega_j} \in H^1(\Omega_j),\ j = 0, 1, \ldots, p\} \subset L^6(\Omega).$$

Отождествляя пространство H с сопряженным пространством H', получаем включения $V \subset W \subset H = H' \subset W' \subset V'$. Здесь W', V' – пространства, сопряженные с W и V соответственно. Через (f, v) будем обозначать значение функционала $f \in V'$ на элементе $v \in V$ и скалярное произведение в H, если $f, v \in H$. Кроме того,

$$\|v\|^2 = (v, v), (v, w)_j = (v, w)_{L^2(\Omega_j)}, \quad \|v\|_j^2 = (v, v)_j, (v, w)_W = \sum_{j=0}^{p} (v, w)_{H^1(\Omega_j)}.$$

Далее будем предполагать, что исходные данные удовлетворяют следующим условиям, естественным с физической точки зрения:

(i) $c, \gamma \in L^\infty(\Gamma)$, $c \ge c_0 > 0$, $\gamma \ge \gamma_0 > 0$, $c_0, \gamma_0 = \text{const}$;

(ii) $\{a, b, \alpha, \beta, n\}|_{\Omega_j} = \{a_j, b_j, \alpha_j, \beta_j, n_j\}$, $b = \sigma\beta n^2$, $\sigma = \text{const} > 0$;

(iii) $0 \le \theta_b \in L^\infty(\Gamma)$; $f_s \in V'$.

Определим следующие операторы и функционалы $A_1 : V \to V'$, $A_2 : W \to W'$, $f \in V'$, $g \in W'$, используя равенства

$$(A_1\theta, \eta) = (a\nabla\theta, \nabla\eta) + \int_\Gamma c\theta\eta\, d\Gamma,$$

$$(A_2\varphi, w) = \sigma\sum_{j=0}^{p} \alpha_j n_j^2 (\nabla\varphi, \nabla w)_j + \sigma n_0^2 \int_\Gamma \gamma\varphi w\, d\Gamma + \sigma n_0^2 \sum_{j=1}^{p} h_j \int_{\Gamma_j} (\varphi_0 - \varphi_j)(w_0 - w_j)\, d\Gamma,$$

$$(f, \eta) = \int_\Gamma c\theta_b\eta\, d\Gamma, \quad (g, w) = \sigma n_0^2 \int_\Gamma \gamma\theta_b^4 w\, d\Gamma,$$

которые справедливы для всех $\theta, \eta \in V$ и $\varphi, w \in W$. Здесь $\{\varphi_j, w_j\} = \{\varphi, w\}\big|_{\Omega_j}$.

Отметим сразу, что билинейная форма $(A_1 u, v)$ определяет норму, эквивалентную стандартной норме пространства V, и поэтому в дальнейшем полагаем $\|v\|_V^2 = (A_1 v, v)$. В дальнейшем будем использовать следующие неравенства непрерывности вложений $V \subset L^s(\Omega)$, $W \subset L^s(\Omega)$, $1 \le s \le 6$:

$$\|v\|_{L^s(\Omega)} \le K_1 \|v\|_V, \quad v \in V, \quad \|w\|_{L^s(\Omega)} \le K_2 \|w\|_W, \quad w \in W, \quad 1 \le s \le 6.$$

Пусть $[t]^q = |t|^q \operatorname{sign} t$, $q > 0$, $t \in \mathbb{R}$. Указанная функция является монотонной и при этом $d[t]^q/dt = q|t|^{q-1}$.

Определение. Пара $\{\theta, \varphi\} \in V \times W$ называется слабым решением задачи (1)–(4), если

$$A_1\theta + b([\theta]^4 - \varphi) = f + f_s, \quad A_2\varphi + b(\varphi - [\theta]^4) = g. \tag{5}$$

Указанная слабая формулировка выводится обычным образом. Достаточно умножить уравнения (1) на тестовые функции $\eta \in V$ и $\sigma n^2\psi \in W$ соответственно и проинтегрировать по частям по областям Ω_j, применяя краевые условия (2) и условия сопряжения (3), (4).

Для постановки обратной задачи рассмотрим линейно независимую систему функционалов $\{f_1, \ldots, f_m\}$ из V' и предположим, что $f_s \in V'$ в первом уравнении (5) имеет вид $f_s = \sum_1^m q_j f_j$. Интенсивности источников $q_j \in \mathbb{R}$ считаются неизвестными, но задаются значения указанных функционалов на решении $\theta \in V$. Таким образом, приходим к следующей постановке.

Задача (IP). Найти $q = (q_1, \ldots, q_m) \in \mathbb{R}^m$, $\theta \in V$, $\varphi \in W$ такие, что

$$A_1\theta + b([\theta]^4 - \varphi) = f + \sum_1^m q_j f_j, \quad A_2\varphi + b(\varphi - [\theta]^4) = g, \quad (f_j, \theta) = r_j, \quad j = 1, 2, \ldots, m. \tag{6}$$

Здесь вектор $r = (r_1, \ldots, r_m) \in \mathbb{R}^m$ является заданным.

Типичным примером обратной задачи, которая возникает при моделировании процессов лазерной абляции (см. [5]), является задача нахождения интенсивностей тепловых источников, локализованных в Ω_j, $j = 1, 2, \ldots, p$, по средним значениям температуры в этих подобластях. В этом случае $f_j(x) = 1$, если $x \in \Omega_j$, и $f_j(x) = 0$, если $x \in \Omega\backslash\Omega_j$, а условия переопределения имеют вид $\int_{\Omega_j} \theta dx = r_j$, $j = 1, 2, \ldots, p$.

3. РАЗРЕШИМОСТЬ ОБРАТНОЙ ЗАДАЧИ

Предварительно сформулируем следующие вспомогательные результаты, доказательство которых приводится в конце статьи.

Лемма 1. *Для каждого $\eta \in W'$ существует единственное решение $\varphi \in W$ уравнения*

$$A_2\varphi + b\varphi = \eta. \tag{7}$$

Лемма 2. *Матрица с элементами $\sigma_{kj} = (f_j, A_1^{-1} f_k)$, $j, k = 1, 2, \ldots, m$, является невырожденной.*

Лемма 3. *Пусть для* $\zeta \in W$

$$E(\zeta) = \sigma\sum_{j=0}^{p} \alpha_j n_j^2 \|\nabla\zeta\|_j^2 + \sigma n_0^2 \int_\Gamma \gamma\zeta^2 d\Gamma + \sigma n_0^2 \sum_{j=1}^{p} h_j \int_{\Gamma_j} ([\zeta_0]^{8/5} - [\zeta_j]^{8/5})([\zeta_0]^{2/5} - [\zeta_j]^{2/5}) d\Gamma.$$

Тогда величина $K = \inf\{E(\zeta) : \|\zeta\| = 1\}$ *строго положительна и справедливо неравенство*

$$K\|w\|^2 \le E(w) \quad \forall w \in W.$$

Сведем обратную задачу (IP) к операторному уравнению относительно неизвестной функции $\theta \in V$. Пусть $\{q, \theta, \varphi\} \in \mathbb{R}^m \times V \times W$ – решение задачи (IP). В силу леммы 1 заключаем, что $\varphi = (A_2 + bI)^{-1}(g + b[\theta]^4)$. Далее, умножим скалярно уравнение для θ на $A_1^{-1} f_k$ и учтем, что $(A_1\theta, A_1^{-1} f_k) = (f_k, \theta) = r_k$. Поэтому интенсивности $q = (q_1, \ldots, q_m) \in \mathbb{R}^m$ являются решением системы

$$\sum_1^m \sigma_{kj} q_j = r_k + (b([\theta]^4 - \varphi) - f, A_1^{-1} f_k), \quad k = 1, 2, \ldots, m, \quad \sigma_{kj} = (f_j, A_1^{-1} f_k), \tag{8}$$

которая, в силу леммы 2, однозначно разрешима для заданных $\theta \in V$, $\varphi \in W$. Таким образом, функция $\theta \in V$ является решением операторного уравнения

$$A_1\theta + b([\theta]^4 - \varphi) = f + \sum_1^m q_j f_j, \quad \text{где} \quad \varphi = (A_2 + bI)^{-1}(g + b[\theta]^4), \tag{9}$$

а числа $q_1, \ldots, q_m$ являются решением системы (8). Очевидно, что задача (8), (9) эквивалентна задаче (IP).

Определим нелинейный оператор $F : V \to V$, используя равенство

$$(F(\theta), z)_V = \left(f + \sum_1^m q_j f_j - b([\theta]^4 - \varphi), z \right) \quad \forall z \in V, \tag{10}$$

где $\varphi = (A_2 + bI)^{-1}(g + b[\theta]^4)$, $q_1, \ldots, q_m$, – решение системы (8). Учитывая определение скалярного произведения в пространстве V, $(u, v)_V = (A_1 u, v)$, заключаем, что задача (8), (9) сводится к уравнению $\theta = F(\theta)$.

Лемма 4. *Оператор* $F : V \to V$ *вполне непрерывен.*

Доказательство. Пусть $\theta_{1,2} \in V$, $\|\theta_{1,2}\|_V \le \rho$, $\hat\theta = \theta_1 - \theta_2$. Из определения оператора F следует равенство

$$(F(\theta_1) - F(\theta_2), z)_V = \left(\sum_1^m \hat q_j f_j - b([\theta_1]^4 - [\theta_2]^4) + b\hat\varphi, z \right) \quad \forall z \in V, \tag{11}$$

где $\hat\varphi$ и $\hat q_j$ таковы, что

$$A_2\hat\varphi + b\hat\varphi = b([\theta_1]^4 - [\theta_2]^4), \quad \sum_1^m \sigma_{kj}\hat q_j = (b([\theta_1]^4 - [\theta_2]^4) - b\hat\varphi, A_1^{-1} f_k), \quad k = 1, 2, \ldots, m. \tag{12}$$

Для оценки правой части в (11) используем неравенства

$$\left|(b([\theta_1]^4 - [\theta_2]^4), z)\right| \le 2\max b \left(\|\theta_1\|_{L^6(\Omega)}^3 + \|\theta_2\|_{L^6(\Omega)}^3 \right) \|\hat\theta\|_{L^4(\Omega)} \|z\|_{L^4(\Omega)} \le C_1 \|\hat\theta\|_{L^4(\Omega)} \|z\|_{L^4(\Omega)}.$$

Здесь $C_1 = 4\max b K_1^3 \rho^3$.

Первое уравнение в (12) умножим скалярно на $\hat\varphi$ и отбросим неотрицательные граничные интегралы в левой части. Тогда

$$\min\{\sigma\alpha n^2, b\} \|\hat\varphi\|_W^2 \le C_1 \|\hat\theta\|_{L^4(\Omega)} \|\hat\varphi\|_{L^4(\Omega)} \le C_1 K_2 \|\hat\theta\|_{L^4(\Omega)} \|\hat\varphi\|_W.$$

Следовательно, $\|\hat\varphi\|_W \le C_2 \|\hat\theta\|_{L^4(\Omega)}$, где $C_2 = C_1 K_2 / \min\{\sigma\alpha n^2, b\}$.

Далее,

$$\left(\sum_1^m \hat{q}_j f_j, z\right) \le mK_3C_1 \max \left\|A_1^{-1} f_k\right\|_{L^4(\Omega)} \max \|f_j\| \left\|\hat{\theta}\right\|_{L^4(\Omega)} \|z\| \le C_3 \left\|\hat{\theta}\right\|_{L^4(\Omega)} \|z\|_V .$$

Здесь $K_3 > 0$ – норма матрицы, обратной к (σ_{kj}).

Полученные неравенства позволяют оценить правую часть (11):

$$(F(\theta_1) - F(\theta_2), z)_V \le (C_3 + C_1K_2 + \max bK_1K_2C_2) \left\|\hat{\theta}\right\|_{L^4(\Omega)} \|z\|_V .$$

Полагая в последнем неравенстве $z = F(\theta_1) - F(\theta_2)$, получаем оценку

$$\|F(\theta_1) - F(\theta_2)\|_V \le (C_3 + C_1K_2 + \max bK_1K_2C_2) \|\theta_1 - \theta_2\|_{L^4(\Omega)} .$$

Поскольку вложение $V \subset L^4(\Omega)$ непрерывно и компактно, из полученной оценки следует, что оператор F вполне непрерывен.

Теорема 1. *Пусть выполняются условия* (i)–(iii). *Тогда существует решение задачи* (IP).

Доказательство. Для доказательства существования неподвижной точки вполне непрерывного оператора F достаточно, на основании принципа Лере–Шаудера, показать равномерную по $\lambda \in (0, 1]$ ограниченность в пространстве V множества решений операторного уравнения $\theta = \lambda F(\theta)$. Из определения оператора F следуют равенства

$$\frac{1}{\lambda} A_1\theta + b([\theta]^4 - \varphi) = f + \sum_1^m q_j f_j, A_2\varphi + b(\varphi - [\theta]^4) = g, \tag{13}$$

где q_j, $j = 1, 2, \ldots, m$, – решение системы (8). Заметим сразу, что из (13) следует равенство $(f_k, \theta) = \lambda r_k$.

Выберем $r \in V$ так, что $(f_k, r) = r_k$, $k = 1, 2, \ldots, m$. Поскольку $C^\infty(\bar{\Omega})$ плотно в пространстве $H^1(\Omega) = V$, в дальнейшем считаем, что $r \in C^\infty(\bar{\Omega})$. Умножим скалярно первое уравнение в (13) на $(\theta - \lambda r)$. Тогда

$$\frac{1}{\lambda}(A_1\theta, \theta - \lambda r) + (b([\theta]^4 - \varphi), \theta - \lambda r) = (f, \theta - \lambda r).$$

Поскольку $\frac{1}{\lambda}(A_1\theta, \theta - \lambda r) \ge \frac{1}{\lambda}(A_1\theta, \theta) - \frac{1}{2}(A_1\theta, \theta) - \frac{1}{2}(A_1 r, r) \ge \frac{1}{2}(A_1\theta, \theta) - \frac{1}{2}(A_1 r, r)$, получаем неравенство

$$\frac{1}{2}(A_1\theta, \theta) + (b([\theta]^4 - \varphi), \theta - \lambda r) \le (f, \theta - \lambda r) + \frac{1}{2}(A_1 r, r). \tag{14}$$

Далее, пусть $\varepsilon > 0$,

$$\mu_\varepsilon(t) = \begin{cases} t - \varepsilon, & t > \varepsilon, \\ 0, & |t| \le \varepsilon, \\ t + \varepsilon, & t < -\varepsilon; \end{cases} \qquad \psi_\varepsilon = \mu_\varepsilon([\varphi]^{1/4}) \in W.$$

Умножая скалярно второе уравнение в (13) на $(\psi_\varepsilon - \lambda r)$, получаем равенство

$$(A_2\varphi, \psi_\varepsilon - \lambda r) + (b(\varphi - [\theta]^4), \psi_\varepsilon - \lambda r) = (g, \psi_\varepsilon - \lambda r).$$

В полученном равенстве перейдем к пределу при $\varepsilon \to +0$. Тогда аналогично [29, теорема 1] заключаем, что $\zeta = [\varphi]^{5/8} \in W$ и справедливо равенство

$$\begin{aligned} &\frac{16\sigma}{25} \sum_{j=0}^p \alpha_j n_j^2 \|\nabla\zeta\|_j^2 + \sigma n_0^2 \int_\Gamma \gamma\zeta^2 d\Gamma + \sigma n_0^2 \sum_{j=1}^p h_j \int_{\Gamma_j} ([\zeta_0]^{8/5} - [\zeta_j]^{8/5})([\zeta_0]^{2/5} - [\zeta_j]^{2/5}) d\Gamma + \\ &+ (b(\varphi - [\theta]^4), [\varphi]^{1/4} - \lambda r) = \lambda\sigma \sum_{j=0}^p \alpha_j n_j^2 \int_{\Omega_j} \nabla[\zeta]^{8/5} \nabla r dx + \lambda\sigma n_0^2 \int_\Gamma \gamma[\zeta]^{8/5} r d\Gamma + \sigma n_0^2 \int_\Gamma \gamma\theta_b^4([\zeta]^{2/5} - \lambda r) d\Gamma. \end{aligned} \tag{15}$$

Из равенства (15), учитывая определение функционала E (лемма 3), выводим неравенство

$$\begin{aligned}&\frac{3}{5}E(\zeta)+\frac{\sigma}{25}\sum_{j=0}^{p}\alpha_j n_j^2\|\nabla\zeta\|_j+\frac{2}{5}\sigma n_0^2\int_\Gamma\gamma\zeta^2 d\Gamma+(b(\varphi-[\theta]^4),[\varphi]^{1/4}-\lambda r)\le\\ &\le\frac{8}{5}\sigma\sum_{j=0}^{p}\alpha_j n_j^2\int_{\Omega_j}|\zeta|^{3/5}|\nabla\zeta||\nabla r|\,dx+\sigma n_0^2\int_\Gamma\gamma|\zeta|^{8/5}|r|\,d\Gamma+\sigma n_0^2\int_\Gamma\gamma\theta_b^4(|\zeta|^{2/5}+|r|)d\Gamma.\end{aligned}\tag{16}$$

Учтем, что в силу леммы 3, справедливо неравенство $K\|\zeta\|^2\le E(\zeta)$, а подынтегральные выражения в правой части (16) оценим, используя следующие неравенства Юнга с параметром $\delta>0$:

$$|\zeta|^{3/5}|\nabla\zeta||\nabla r|\le\frac{\delta^2}{2}|\nabla\zeta|^2+\frac{3}{10}\delta^{10/3}\zeta^2+\frac{1}{5}\delta^{-10}|\nabla r|^5,$$

$$|\zeta|^{8/5}|r|\le\frac{4}{5}\delta^{5/4}\zeta^2+\frac{1}{5}\delta^{-5}|r|^5,\quad \theta_b^4|\zeta|^{2/5}\le\frac{1}{5}\delta^5\zeta^2+\frac{4}{5}\delta^{-5/4}\theta_b^5.$$

Выбрав параметр $\delta=\delta(K,\alpha,n)$ достаточно малым, выводим из (16) неравенство

$$(b(\varphi-[\theta]^4),[\varphi]^{1/4}-\lambda r)\le C_0.$$

Здесь постоянная C_0 зависит только от величин K, α, n и функций θ_b, r и, что важно, не зависит от $\lambda\in(0,1]$. Сложив полученное неравенство с (14), получаем

$$\frac{1}{2}(A_1\theta,\theta)+(b([\theta]^4-\varphi),\theta-[\varphi]^{1/4})\le C_0+(f,\theta-\lambda r)+\frac{1}{2}(A_1 r,r).$$

В последней оценке можно опустить второе слагаемое в левой части, поскольку оно неотрицательно. Далее, используя неравенство $(f,\theta)\le\|f\|_{V'}^2+(1/4)\|\theta\|_V^2$, получаем оценку равномерной по λ ограниченности множества решений операторного уравнения $\theta=\lambda F(\theta)$:

$$\|\theta\|_V^2=(A_1\theta,\theta)\le 4C_0+4\|f\|_{V'}^2+4|(f,r)|+2\|r\|_V^2.\tag{17}$$

Из оценки (17) следует разрешимость обратной задачи (IP).

4. УСЛОВИЯ ЕДИНСТВЕННОСТИ РЕШЕНИЯ ОБРАТНОЙ ЗАДАЧИ

Обозначим через $\mathcal{R}\subset\mathcal{V}$ множество решений задачи (9), т.е. множество θ-компонент решений задачи (IP). Из доказательства теоремы 1 следует, что при выполнении условий (i)–(iii), указанное множество ограничено в V и в пространстве $L^6(\Omega)$. Оценим разность двух функций из множества $\mathcal{R}$.

Пусть $\theta_{1,2}\in\mathcal{R}$, $\theta=\theta_1-\theta_2$, $\xi=([\theta_1]^4-[\theta_2]^4)/(\theta_1-\theta_2)$. Заметим, что, в силу неравенства

$$0\le\xi\le 2(|\theta_1|^3+|\theta_2|^3),$$

функция $\xi\in L^2(\Omega)$ и $M=\sup\{\|\xi\|,\theta_{1,2}\in\mathcal{R}\}<+\infty$.

Из уравнения (9) для θ_1 вычтем аналогичное уравнение для θ_2 и умножим первое уравнение скалярно на θ, учитывая, что $(f_k,\theta)=0$, $k=1,2,\ldots,m$. Тогда

$$(A_1\theta+b(\xi\theta-\varphi),\theta)=0,\quad\text{где}\quad\varphi=(A_2+bI)^{-1}(b\xi\theta).\tag{18}$$

Заметим, что

$$(b\varphi,\theta)=((A_2+bI)^{-1}(b\xi\theta),b\theta)=(b\xi\theta,\psi)\le(b\xi\theta,\theta)+\frac{1}{4}(b\xi\psi,\psi),$$

где

$$A_2\psi+b\psi=b\theta.\tag{19}$$

Тогда из (18) следует неравенство

$$\|\theta\|_V^2 = (A_1\theta,\theta) \le \frac{1}{4}(b\xi\psi,\psi) \le \frac{1}{4}\left\|b^{1/2}\xi\right\|\left\|b^{1/4}\psi\right\|_{L^4(\Omega)}^2 \le \frac{M\max b^{1/2}}{4}\left\|b^{1/4}\psi\right\|_{L^4(\Omega)}^2. \tag{20}$$

Умножим уравнение (19) скалярно на ψ^3 и отбросим неотрицательное слагаемое $(A_2\psi,\psi^3)$. Используя неравенство Гёльдера, получаем оценку

$$\left\|b^{1/4}\psi\right\|_{L^4(\Omega)}^4 \le (b\theta,\psi^3) \le \left\|b^{1/4}\theta\right\|_{L^4(\Omega)}\left\|b^{1/4}\psi\right\|_{L^4(\Omega)}^3, \quad \left\|b^{1/4}\psi\right\|_{L^4(\Omega)} \le \left\|b^{1/4}\theta\right\|_{L^4(\Omega)}.$$

Справедливы также мультипликативное неравенство для нормы в $L^4(\Omega)$ и неравенство вложения V в $L^4(\Omega)$:

$$\|\theta\|_{L^4(\Omega)}^2 \le \|\theta\|^{1/2}\|\theta\|_{L^6(\Omega)}^{3/2}, \quad \|\theta\|_{L^6(\Omega)} \le K_1\|\theta\|_V.$$

Таким образом, из (20) следует неравенство

$$\|\theta\|_V^2 \le K_1^6\left(\frac{M\max b}{4}\right)^4\|\theta\|^2. \tag{21}$$

В силу теоремы Гильберта–Шмидта, собственные функции $\{w_j\}$ оператора A_1, определяемые из условий $A_1w_j = \lambda_j w_j$, $j = 1, 2, \ldots$, $(w_i, w_j) = \delta_{ij}$, $0 < \lambda_1 \le \lambda_2 \le \ldots$, образуют базис пространств H и V, причем $\lambda_j \to +\infty$ при $j \to +\infty$. Поскольку $\lambda_1\|\theta\|^2 \le (A_1\theta,\theta)$, то из оценки (21) следует единственность решения задачи (IP), если выполняется условие

$$\lambda_1 > K_1^6\left(\frac{M\max b}{4}\right)^4. \tag{22}$$

В общем случае из оценки (21), аналогично [29], в силу компактности вложения пространства V в H, следует конечномерная структура множества $\mathcal{R}$.

Теорема 2. *Пусть выполняются условия* (i). *Тогда множество решений задачи* (IP) *непусто и гомеоморфно компакту, лежащему в конечномерном пространстве, а если выполняется условие* (22), *то решение единственно*.

5. ДОКАЗАТЕЛЬСТВО ЛЕММ 1–3

Лемма 1. *Для каждого* $\eta \in W'$ *существует единственное решение* $\varphi \in W$ *уравнения*

$$A_2\varphi + b\varphi = \eta. \tag{23}$$

Доказательство. Утверждение следует из леммы Лакса–Мильграма, поскольку билинейная форма $a(\varphi,\psi) = (A_2\varphi + b\varphi,\psi)$ является непрерывной, симметричной и положительно-определенной в пространстве W.

Лемма 2. *Матрица с элементами* $\sigma_{kj} = (f_j, A_1^{-1}f_k)$, $j,k = 1, 2, \ldots, m$, *является невырожденной.*

Доказательство. Достаточно проверить, что следующая однородная линейная система

$$\sum_{j=1}^m (f_j, A_1^{-1}f_k)c_j = 0, \quad k = 1, 2, \ldots, m,$$

имеет только тривиальное решение. Умножим k-е уравнение системы на c_k и просуммируем. Тогда для $z = \sum_1^m c_j f_j$ получим $(z, A_1^{-1}z) = 0$. Следовательно, $z = 0$ и в силу линейной независимости функционалов $f_1, \ldots, f_m$ получаем $c_1 = \ldots = c_m = 0$.

Лемма 3. *Пусть для* $\zeta \in W$

$$E(\zeta) = \sigma\sum_{j=0}^p \alpha_j n_j^2\|\nabla\zeta\|_j^2 + \sigma n_0^2\int_\Gamma \gamma\zeta^2 d\Gamma + \sigma n_0^2\sum_{j=1}^p h_j\int_{\Gamma_j}([\zeta_0]^{8/5} - [\zeta_j]^{8/5})([\zeta_0]^{2/5} - [\zeta_j]^{2/5})d\Gamma.$$

Тогда величина $K = \inf\{E(\zeta) : \|\zeta\| = 1\}$ *строго положительна и справедливо неравенство*

$$K\|w\|^2 \le E(w) \quad \forall w \in W.$$

Доказательство. Пусть $K = 0$. Тогда имеется минимизирующая последовательность $\zeta^{(k)} \in W$, $\|\zeta^{(k)}\| = 1$, $E(\zeta^{(k)}) \to 0$. Из определения функционала E вытекает, что

$$\|\nabla\zeta^{(k)}\|_{L^2(\Omega_j)} \to 0, \quad j = 0, 1, \ldots, p, \quad \int_\Gamma \gamma(\zeta^{(k)})^2 d\Gamma \to 0.$$

Следовательно, найдется $\zeta \in W$ такая, что $\zeta^{(k)} \to \zeta$ сильно в H, $\zeta^{(k)}\big|_{\Omega_j} \to \zeta|_{\Omega_j}$ сильно в $H^1(\Omega_j)$ и, кроме того,

$$\zeta|_{\Omega_j} = c_j = \text{const}_j, \quad \|\zeta\| = 1, \quad \int_\Gamma \gamma\zeta^2 d\Gamma = 0.$$

Поэтому $\zeta_0 = \zeta|_{\Omega_0} = c_0 = 0$. Далее, поскольку для $j = 1, 2, \ldots, p$

$$\int_{\Gamma_j} ([\zeta_0^{(k)}]^{8/5} - [\zeta_j^{(k)}]^{8/5})([\zeta_0^{(k)}]^{2/5} - [\zeta_j^{(k)}]^{2/5}) d\Gamma \to 0,$$

получаем в пределе, что $\int_{\Gamma_j} \zeta^2 d\Gamma = 0$. Таким образом, $c_1 = \ldots = c_p = 0$ и $\zeta = 0$, что противоречит равенству $\|\zeta\| = 1$.

СПИСОК ЛИТЕРАТУРЫ

1. *Larsen E.W., Thömmes G., Klar A., Seaïd M., Götz M.* Simplified P_N approximations to the equations of radiative heat transfer and applications // J. Comput. Phys. 2002. V. 183. № 2. P. 652–675.
2. *Modest M.F.* Radiative Heat Transfer. New York: Academic Press, 2003.
3. *Thömmes G., Pinnau R., Seaïd M., Götz M., Klar A.* Numerical methods and optimal control for glass cooling processes // Transport Theory Statistic. Phys. 2002. V. 31. № 4–6. P. 513–529.
4. *Tse O., Pinnau R.* Optimal control of a simplified natural convection-radiation model // Commun. Math. Sci. 2013. V. 11. № 3. P. 679–707.
5. *Ковтанюк А.Е., Гренкин Г.В., Чеботарев А.Ю.* Использование диффузионного приближения для моделирования радиационных и тепловых процессов в кожном покрове // Оптика и спектроскопия. 2017. Т. 123. 2. С. 194–199.
6. *Pinnau R.* Analysis of optimal boundary control for radiative heat transfer modeled by SP_1-system // Commun. Math. Sci. 2007. V. 5. № 4. P. 951–969.
7. *Гренкин Г.В., Чеботарев А.Ю.* Нестационарная задача сложного теплообмена // Ж. вычисл. матем. и матем. физ. 2014. Т. 54. № 11. С. 1806–1816.
8. *Гренкин Г.В., Чеботарев А.Ю.* Неоднородная нестационарная задача сложного теплообмена // Сиб. электрон. матем. изв. 2015. Т. 12. С. 562–576.
9. *Гренкин Г.В., Чеботарев А.Ю.* Нестационарная задача свободной конвекции с радиационным теплообменом // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 2. С. 275–282.
10. *Grenkin G.V., Chebotarev A.Yu., Kovtanyuk A.E., Botkin N.D., Hoffmann K.-H.* Boundary optimal control problem of complex heat transfer model // J. Math. Anal. Appl. 2016. V. 433. № 2. P. 1243–1260.
11. *Kovtanyuk A.E., Chebotarev A.Yu.* An iterative method for solving a complex heat transfer problem // Appl. Math. Comput. 2013. V. 219. № 17. P. 9356–9362.
12. *Kovtanyuk A.E., Chebotarev A.Yu., Botkin N.D., Hoffmann K.-H.* The unique solvability of a complex 3D heat transfer problem // J. Math. Anal. Appl. 2014. V. 409. № 2. P. 808–815.
13. *Ковтанюк А.Е., Чеботарев А.Ю.* Стационарная задача сложного теплообмена // Ж. вычисл. матем. и матем. физ. 2014. Т. 54. № 4. С. 711–719.
14. *Ковтанюк А.Е., Чеботарев А.Ю.* Стационарная задача свободной конвекции с радиационным теплообменом // Дифференц. ур-ния. 2014. Т. 50. № 12. С. 1590–1597.
15. *Kovtanyuk A.E., Chebotarev A.Yu., Botkin N.D., Hoffmann K.-H.* Theoretical analysis of an optimal control problem of conductive-convective-radiative heat transfer // J. Math. Anal. Appl. 2014. V. 412. № 1. P. 520–528.
16. *Kovtanyuk A.E., Chebotarev A.Yu., Botkin N.D., Hoffmann K.-H.* Unique solvability of a steady-state complex heat transfer model // Commun. Nonlinear Sci. Numer. Simul. 2015. V. 20. № 3. P. 776–784.

17. *Kovtanyuk A.E., Chebotarev A.Yu., Botkin N.D., Hoffmann K.-H.* Optimal boundary control of a steady-state heat transfer model accounting for radiative effects // J. Math. Anal. Appl. 2016. V. 439. № 2. P. 678–689.

18. *Chebotarev A.Yu., Kovtanyuk A.E., Grenkin G.V., Botkin N.D., Hoffmann K.-H.* Nondegeneracy of optimality conditions in control problems for a radiative-conductive heat transfer model // Appl. Math. Comput. 2016. V. 289. P. 371–380.

19. *Ковтанюк А.Е., Чеботарев А.Ю.* Нелокальная однозначная разрешимость стационарной задачи сложного теплообмена // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 5. С. 816–823.

20. *Chebotarev A.Yu., Grenkin G.V., Kovtanyuk A.E.* Inhomogeneous steady-state problem of complex heat transfer // ESAIM Math. Model. Numer. Anal. 2017. V. 51. № 6. P. 2511–2519.

21. *Chebotarev A.Y., Grenkin G.V., Kovtanyuk A.E., Botkin N.D., Hoffmann K.-H.* Diffusion approximation of the radiative-conductive heat transfer model with Fresnel matching conditions // Communicat. Nonlin. Sci. Numeric. Simulat. 2018. V. 57. P. 290–298.

22. *Chebotarev A.Y., Kovtanyuk A.E., Botkin N.D.* Problem of radiation heat exchange with boundary conditions of the Cauchy type // Communicat. Nonlin. Sci. Numeric. Simulat. 2019. V. 75. P. 262–269.

23. *Амосов А.А.* Глобальная разрешимость одной нелинейной нестационарной задачи с нелокальным краевым условием типа теплообмена излучением // Дифференц. ур-ния. Т. 41. № 1. 2005. С. 93–104.

24. *Amosov A.A.* Stationary nonlinear nonlocal problem of radiative-conductive heat transfer in a system of opaque bodies with properties depending on the radiation frequency // J. Math. Sci. 2010. V. 164. № 3. P. 309–344.

25. *Amosov A.* Unique Solvability of a Nonstationary Problem of Radiative-Conductive Heat Exchange in a System of Semitransparent Bodies // Rus. J. Math. Phys. 2016. V. 23. № 3. P. 309–334.

26. *Amosov A.A.* Unique Solvability of Stationary Radiative-Conductive Heat Transfer Problem in a System of Semitransparent Bodies // J. Math. Sci. (United States). 2017. V. 224. № 5. P. 618–646.

27. *Amosov A.A.* Asymptotic Behavior of a Solution to the Radiative Transfer Equation in a Multilayered Medium with Diffuse Reflection and Refraction Conditions // J. Math. Sci. 2020. V. 244. P. 541–575.

28. *Amosov A.A., Krymov N.E.* On a Nonstandard Boundary Value Problem Arising in Homogenization of Complex Heat Transfer Problems // J. Math. Sci. (United States). 2020. V. 244. P. 357–377.

29. *Chebotarev A.Yu., Grenkin G.V., Kovtanyuk A.E., Botkin N.D., Hoffmann K.-H.* Inverse problem with finite overdetermination for steady-state equations of radiative heat exchange // J. Math. Anal. Appl. 2018. V. 460. № 2. P. 737–744.

30. *Chebotarev A.Yu., Pinnau R.* An inverse problem for a quasi-static approximate model of radiative heat transfer // J. Math. Anal. Appl. 2019. V. 472. № 1. P. 314–327.

31. *Гренкин Г.В., Чеботарев А.Ю.* Обратная задача для уравнений сложного теплообмена // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 8. С. 1420–1430.

32. *Пятков С.Г., Сафонов Е.И.* О некоторых классах линейных обратных задач для параболических систем уравнений // Сиб. электрон. матем. изв. 2014. Т. 11. С. 777–799.

33. *Пятков С.Г., Уварова М.В.* Об определении функции источника в задачах тепломассопереноса по интегральным условиям переопределения // Сиб. журн. индустр. матем. 2016. Т. 19. № 4. С. 93–100.

34. *Пятков С.Г.* О некоторых классах обратных задач об определении функции источника в системах конвекции–диффузии // Дифференц. ур-ния. 2017. Т. 53. № 10. С. 1385.

35. *Пятков С.Г., Ротко В.В.* Обратные задачи для некоторых квазилинейных параболических систем с точечными условиями переопределения // Матем. тр. 2019. Т. 22. № 1. С. 178–204.

36. *Chebotarev A.Yu.* Subdifferential inverse problems for stationary systems of Navier–Stokes type // J. Inverse Ill-Posed Probl. 1995. V. 3. № 4. P. 268–277.

37. *Чеботарев А.Ю.* Определение правой части системы Навье–Стокса и обратные задачи для уравнений тепловой конвекции // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. № 12. С. 2279–2287.

38. *Чеботарев А.Ю.* Стабилизация сторонними токами равновесных МГД конфигураций // Ж. вычисл. матем. и матем. физ. 2012. Т. 52. № 12. С. 2238–2246.

39. *Чеботарев А.Ю.* Обратная задача для систем Навье–Стокса с конечномерным переопределением // Дифференц. ур-ния. 2012. Т. 48. № 8. С. 1166.

40. *Чеботарев А.Ю.* Обратные задачи для стационарных систем Навье–Стокса // Ж. вычисл. матем. и матем. физ. 2014. Т. 54. № 3. С. 519–528.

41. *Kovtanyuk A.E., Chebotarev A.Y., Botkin N.D., Turova V.L., Sidorenko I.N., Lampe R.* Continuum model of oxygen transport in brain // J. Math. Anal. Appl. 2019. V. 474. P. 1352–1363.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 312–344

ИНФОРМАТИКА

УДК 519

ЧИСЛЕННЫЕ МЕТОДЫ ДЛЯ ЗАДАЧИ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ В КОМПЬЮТЕРНОЙ СЕТИ[1)]

© 2021 г. Е. А. Воронцова[1], А. В. Гасников[1,2,3], П. Е. Двуреченский[3,2], А. С. Иванова[4,*], Д. А. Пасечнюк[1]

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, Московский физико-технический институт (национальный исследовательский университет), Россия

[2] 127051 Москва, Большой Каретный пер., 19, стр. 1, Институт проблем передачи информации им. А.А. Харкевича РАН, Россия

[3] Институт прикладного анализа и стохастики им. К. Вейерштрасса, Берлин, Германия

[4] 109028 Москва, Покровский бульвар, 11, Национальный исследовательский университет "Высшая школа экономики", Россия

**e-mail: asivanova@hse.ru*

Поступила в редакцию 29.11.2019 г.
Переработанный вариант 10.09.2020 г.
Принята к публикации 16.09.2020 г.

Рассматривается задача распределения ресурсов в компьютерных сетях с большим числом соединений. Соединения используют для своих целей потребители (пользователи), число которых также может быть очень большим. Для решения двойственной задачи предлагаются следующие численные методы оптимизации: быстрый градиентный метод, стохастический метод проекции субградиента, метод эллипсоидов и метод экстраполяции случайного градиента. Для каждого метода получена оценка скорости сходимости. Также приведены алгоритмы распределенного вычисления шагов рассматриваемых методов при условии приложения их к компьютерным сетям. Отдельное внимание уделено прямо двойственности предложенных алгоритмов. Библ. 38. Фиг. 1. Табл. 2.

Ключевые слова: распределение ресурсов, сети связи, максимизация полезности сети, прямо двойственность, быстрый градиентный метод, стохастический метод проекции субградиента, метод эллипсоидов, метод экстраполяции случайного градиента.

DOI: 10.31857/S0044466921020149

1. ВВЕДЕНИЕ

1.1. Мотивация

Рассматривается задача управления современными сетями связи с точки зрения оптимизации и стохастического моделирования. Для решения задач такого рода необходимо представить и проанализировать математическую модель, возникающую при симуляции процесса работы крупномасштабных широкополосных сетей. Ожидается, что в будущих сетях связи появятся приложения, которые смогут изменять свои скорости передачи данных в соответствии с доступной пропускной способностью в сети. В качестве примера такой сети можно привести TCP-трафик через сеть Интернет.

Ключевой вопрос, который мы рассматриваем в этой статье, касается того, как доступная пропускная способность в сети должна быть распределена между конкурирующими потоками. При этом контроль над использованием потребителями доступных пропускных способностей осуществляется посредством корректировки цены на соединение.

Таким образом, в работе рассматривается задача оптимизации распределения ресурсов в компьютерных сетях с большим числом соединений. Соединения используют для своих целей по-

[1)]Работа выполнена при финансовой поддержке РФФИ авторов: А.В. Гасникова поддержана грантами РФФИ 18-31-20005 мол_а_вед и 19-31-51001 Научное наставничество, работа П.Е. Двуреченского поддержана грантом РФФИ 18-29-03071 мк. Работа Е.А. Воронцовой была выполнена при поддержке Минобрнауки РФ (госзадание) № 075-00337-20-03, номер проекта 0714-2020-0005.

требители (пользователи), число которых также может быть очень большим. Цель работы состоит в определении механизма управления ресурсами, которые в контексте данной задачи являются доступными пропускными способностями соединений. При этом необходимо обеспечить стабильную работу системы и предотвратить перегрузки. В качестве критерия оптимальности используется сумма полезностей всех пользователей компьютерной сети.

Первоначально стандартные задачи распределения ресурсов, сводящиеся к максимизации совокупной полезности производителей при совместном использовании имеющихся ресурсов, были рассмотрены в [1], также задача распределения ресурсов в компьютерных сетях исследовалась в недавней работе [2]. Предложенные в монографии [3] механизмы децентрализованного распределения ресурсов с тех пор привлекают большое внимание в экономических исследованиях, см., например, [4]–[6] и ссылки в них. В данной работе, следуя [7], [8], мы рассматриваем различные механизмы корректировки цен. Практическую ценность предложенные подходы имеют за счет своей децентрализованности, что означает, что для установления и корректировки цены в отдельном соединении необходима только реакция пользователей, которые используют это соединение, а не реакция всех пользователей сети. При таком механизме корректировки все соединения действуют независимо.

Кроме того, один из предложенных в статье подходов, основанный на использовании стохастического метода проекции субградиента, обходит следующую проблему, возникающую в реальных сетях. Дело в том, что в реальных сетях пакеты с данными, которые поступают от пользователей на соединение, приходят не одновременно, поэтому на практике суммарный трафик на соединении неизвестен. Для решения этой проблемы предлагается использовать стохастические методы, для работы которых требуется не точное значение суммарного трафика, а только его оценка, которую можно получить на основе трафика одного из пользователей. Идея использовать для решения данной проблемы стохастический метод проекции субградиента предложена в работе [2].

1.2. Содержание

Статья организована следующим образом. В разд. 2 описаны постановка задачи и построение двойственной к ней. Также приводятся все необходимые предположения для прямой задачи. В разд. 3 рассматривается решение предложенной задачи быстрым градиентным методом Нестерова [9], получена оценка сложности данного метода, которая имеет порядок $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$. В разд. 4 описано решение данной задачи с использованием стохастического метода проекции субградиента и приведены оценки сложности, которые имеют порядок $O\left(\frac{1}{\varepsilon^2}\right)$.

В разд. 5 описано решение задачи с использованием метода эллипсоидов, который хорошо подходит для задач небольшой размерности, также приведен алгоритм построения сертификата точности для данного метода. Приведены оценки сложности, которые имеют порядок $O\left(m^2 \ln\frac{1}{\varepsilon}\right)$, где m – число соединений в сети. В разд. 6 описана техника регуляризации, которая позволяет восстанавливать решение прямой задачи по решению двойственной для не прямо двойственного метода. В разд. 7 описано решение регуляризованной задачи с использованием метода экстраполяции случайного градиента. Приведены оценки сложности, которые имеют порядок $O\left(\frac{1}{\sqrt{\varepsilon}}\ln\frac{1}{\varepsilon}\right)$, где логарифмический множитель возникает за счет регуляризации двойственной задачи.

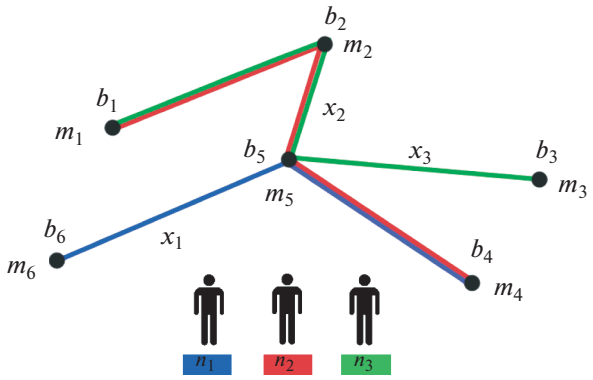
В разд. 8 приведено описание вычислительных экспериментов, подтверждающих на практике теоретические результаты, полученные в предыдущих разделах.

Также для каждого алгоритма описано его распределенное вычисление в контексте рассматриваемой задачи.

2. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим компьютерную сеть с m соединениями и n пользователями (или узлами), см. фиг. 1.

Пользователи обмениваются пакетами через фиксированный набор соединений. Структура сети задана матрицей маршрутизации $C = (C_i^j) \in \mathbf{R}^{m\times n}$. Столбцы матрицы $\mathbf{C}_i \neq 0$, $i = 1,\ldots,n$ – бу-



Фиг. 1. Пример компьютерной сети с $m = 6$ и $n = 3$.

левы m-мерные векторы такие, что $C_i^j = 1$ в случае использования узлом i соединения j, в противном случае $C_i^j = 0$. Ограничения на пропускную способность соединений задаются вектором $\mathbf{b} \in \mathbf{R}^m$ со строго положительными компонентами.

Пользователи оценивают качество работы сети с помощью функций полезности $u_k(x_k)$, $k = 1,\ldots,n$, где $x_k \in \mathbf{R}_+$ – скорость передачи данных k-го пользователя. За критерий оптимальности системы принята сумма функций полезностей для всех пользователей [1].

Задача максимизации суммарной полезности сети при заданных ограничениях на пропускную способность соединений формулируется следующим образом:

$$\max_{\left\{C\mathbf{x}=\sum_{k=1}^{n}\mathbf{C}_k x_k\right\}\le\mathbf{b}} \left\{U(\mathbf{x}) = \sum_{k=1}^{n} u_k(x_k)\right\}, \tag{1}$$

где $\mathbf{x} = (x_1,\ldots,x_n) \in \mathbb{R}_+^n$. Решением данной задачи будет оптимальное распределение ресурсов $\mathbf{x}^*$.

Рассмотрим стандартный переход к двойственной задаче для (1). Пусть задан вектор двойственных множителей $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m) \in \mathbb{R}_+^m$, который можно интерпретировать как вектор цен соединений. Определим двойственную целевую функцию

$$\varphi(\boldsymbol{\lambda}) = \max_{\mathbf{x}\in\mathbf{R}_+^n}\left\{\sum_{k=1}^{n} u_k(x_k) + \left\langle \boldsymbol{\lambda}, \mathbf{b} - \sum_{k=1}^{n}\mathbf{C}_k x_k \right\rangle\right\} = \langle \boldsymbol{\lambda}, \mathbf{b}\rangle + \sum_{k=1}^{n}\left(u_k(x_k(\boldsymbol{\lambda})) - \langle \boldsymbol{\lambda}, \mathbf{C}_k x_k(\boldsymbol{\lambda})\rangle\right), \tag{2}$$

причем пользователи выбирают оптимальные скорости передачи информации x_k, решая следующую задачу оптимизации:

$$x_k(\boldsymbol{\lambda}) = \arg\max_{x_k\in\mathbf{R}_+}\{u_k(x_k) - x_k\langle \boldsymbol{\lambda}, \mathbf{C}_k\rangle\}. \tag{3}$$

Обозначим также за $\mathbf{x}(\boldsymbol{\lambda})$ вектор с компонентами $x_k(\boldsymbol{\lambda})$. Тогда для нахождения оптимальных цен $\boldsymbol{\lambda}^*$ требуется решить задачу

$$\min_{\boldsymbol{\lambda}\in\mathbf{R}_+^m}\varphi(\boldsymbol{\lambda}). \tag{4}$$

Предположим, что для прямой задачи выполняется условие Слейтера, тогда в силу сильной двойственности и прямая, и двойственная задачи будут иметь решение. Используя условие Слейтера, можно компактифицировать решение двойственной задачи. Будем предполагать, что для решения двойственной задачи верна следующая оценка:

$$\|\boldsymbol{\lambda}^*\|_2 \le R.$$

При этом значение R никак не влияет на работу рассматриваемых алгоритмов, а только присутствует в оценках скорости их сходимости.

Основная идея данной работы заключается в применении различных оптимизационных методов для решения двойственной задачи (4) с добавлением прямо двойственного анализа этих методов, позволяющего также восстановить решение прямой задачи (1). В этом смысле мы развиваем подход наших предыдущих работ [10]–[21]. Основное отличие состоит в рассмотрении ограничений-неравенств, а также в анализе стохастических алгоритмов в смысле оценок с большой вероятностью, а не в среднем.

2.1. Сильно вогнутые функции полезности

В некоторых разделах мы будем предполагать, что функции полезности $u_k(x_k)$, $k = 1,\ldots,n$, являются *сильно вогнутыми* с константой μ. В данном подразделе опишем, какими свойствами будет обладать двойственная задача при таком предположении.

Утверждение 1 (Теорема Демьянова–Данскина–Рубинова, см. [22], [23]). *Пусть для любого* $\boldsymbol{\lambda} \in \mathbb{R}_+$ *выполняется* $\varphi(\boldsymbol{\lambda}) = \max\limits_{\mathbf{x}\in X} F(\mathbf{x},\boldsymbol{\lambda})$, *где* $F(\mathbf{x},\boldsymbol{\lambda})$ *– выпуклая и гладкая по* $\boldsymbol{\lambda}$ *функция и максимум достигается в единственной точке* $x(\boldsymbol{\lambda})$. *Тогда* $\nabla\varphi(\boldsymbol{\lambda}) = \nabla_\lambda F(x(\boldsymbol{\lambda}),\boldsymbol{\lambda})$.

Утверждение 2 (см. [24]). *Пусть для любого* $k = 1,\ldots,n$ *функции* $u_k(x_k)$ *являются* μ*-сильно вогнутыми. Тогда функция* (2), *где* $x_k(\boldsymbol{\lambda})$, $k = 1,\ldots,n$, *являются решением задачи* (3), *будет* nm^2/μ*-гладкой, т.е. градиент функции* $\varphi(\boldsymbol{\lambda})$ *будет удовлетворять условию Липшица с константой* $L = nm^2/\mu$:

$$\left\|\nabla\varphi(\boldsymbol{\lambda}^2) - \nabla\varphi(\boldsymbol{\lambda}^1)\right\|_2 \le L\left\|\boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^1\right\|_2.$$

Доказательство утверждения можно найти ниже в Приложении.

2.2. Вогнутые функции полезности

Теперь предположим, что функции полезности $u_k(x_k)$, $k = 1,\ldots,n$, являются *вогнутыми, но не сильно вогнутыми*. Тогда двойственная задача не является гладкой. В данном подразделе описаны некоторые свойства субградиентов двойственной задачи при таких предположениях.

Субградиент двойственной задачи (4) определяется следующим образом:

$$\nabla\varphi(\boldsymbol{\lambda}) = \mathbf{b} - C\mathbf{x}.$$

Учитывая, что $\mathbf{x}$ – ограниченная скорость передачи информации и вектор $\mathbf{b}$ также ограничен, получаем, что субградиенты двойственной задачи в данном случае ограничены. Таким образом, существует такая положительная константа M, что верна следующая оценка:

$$\left\|\nabla\varphi(\boldsymbol{\lambda})\right\|_2 \le M. \tag{5}$$

В качестве грубой оценки сверху для константы M из (5) можно использовать $O(n\sqrt{m})$. Множитель n возникает из-за того, что есть n слагаемых, а $\sqrt{m}$ используется как оценка зависимости 2-нормы от размерности вектора m.

3. БЫСТРЫЙ ГРАДИЕНТНЫЙ МЕТОД

В данном разделе мы предполагаем, что функции полезности $u_k(x_k)$, $k = 1,\ldots,n$, являются *сильно вогнутыми* с константой μ, следовательно, двойственная задача будет гладкой.

Для решения двойственной задачи (4) применим быстрый градиентный метод (БГМ) Нестерова в следующем варианте (метод PDFGM, см. алгоритм 1).

Алгоритм 1. Прямо двойственный быстрый градиентный метод (PDFGM)

Вход: $u_k(\mathbf{x})$, $k = 1,\ldots,n$ – сильно вогнутые функции полезности для каждого пользователя; $\boldsymbol{\lambda}^0$ – вектор начальных цен, $\alpha_t := \dfrac{t+1}{2}$, $A_{-1} := 0$, $A_t := A_{t-1} + \alpha_t = \dfrac{(t+1)(t+2)}{4}$, $\tau_t := \dfrac{\alpha_{t+1}}{A_{t+1}} = \dfrac{2}{t+3}$, $t = 0,1,\ldots,N-1$.

1: **for** $t = 0,1,\ldots,N-1$

2: Вычислить $\varphi(\boldsymbol{\lambda}^t)$, $\nabla\varphi(\boldsymbol{\lambda}^t)$

3: $\mathbf{y}^t := \left[\boldsymbol{\lambda}^t - \frac{1}{L}\left(\mathbf{b} - \sum_{k=1}^n \mathbf{C}_k x_k(\boldsymbol{\lambda}^t)\right)\right]_+$

4: $\mathbf{z}^t := \left[\boldsymbol{\lambda}^0 - \frac{1}{L}\sum_{j=0}^t \alpha_j \left(\mathbf{b} - \sum_{k=1}^n \mathbf{C}_k x_k(\boldsymbol{\lambda}^j)\right)\right]_+.$

5: $\boldsymbol{\lambda}^{t+1} := \tau_t \mathbf{z}^t + (1-\tau_t)\mathbf{y}^t$

6: $\hat{\mathbf{x}}^{t+1} := \frac{1}{A_{t+1}}\sum_{j=0}^{t+1} \alpha_j \mathbf{x}(\boldsymbol{\lambda}^j)$

7: **end for**

8: **return** $\boldsymbol{\lambda}^N$, $\hat{\mathbf{x}}^N$

3.1. Распределенный метод

Для решения рассматриваемой задачи можно также применить БГМ в распределенном варианте, что означает, что каждое соединение может вычислять оптимальную скорость передачи данных только исходя из реакции пользователей, которые используют данное соединение, и никак не взаимодействовать с другими соединениями.

Опишем процесс, происходящий на t-й итерации для соединения j.

1. На основе информации, полученной от пользователей после предыдущей итерации с номером $t-1$ (вектор $\mathbf{x}^t = \mathbf{x}(\boldsymbol{\lambda}^t)$), соединение j вычисляет

$$y_j^t = \left[\lambda_j^t - \frac{1}{L}\left(b_j - \sum_{k=1}^n C_k^j x_k^t\right)\right]_+.$$

При этом $C_k^j \neq 0$ только для тех пользователей, которые используют соединение j. Поэтому для вычисления этого шага соединению нужна только информация от пользователей, которые используют это соединение.

2. Далее соединение j аналогично вычисляет

$$z_j^t = \left[\lambda_j^0 - \frac{\alpha_j}{L}\left(b_j - \sum_{k=1}^n C_k^j x_k^t\right)\right]_+.$$

3. Получив значения на предыдущих двух шагах, соединение j вычисляет цену для следующей итерации $t+1$:

$$\lambda_j^{t+1} = \tau_t z_j^t + (1-\tau_t) y_j^t$$

и посылает эту информацию всем пользователям, которые с ним соединены.

4. Далее пользователи вычисляют оптимальные скорости передачи информации $\hat{\mathbf{x}}^{t+1}$, в частности, для пользователя k получаем

$$x_k(\boldsymbol{\lambda}^{t+1}) = \arg\max_{x_k \in \mathbf{R}_+}\left(u_k(x_k) - x_k \sum_{j=1}^m \lambda_j^{t+1} C_k^j\right),$$

где в силу определения матрицы C пользователю необходима информация только от соединений, которые он использует. Далее пользователь вычисляет оптимальную скорость

$$\hat{x}_k^{t+1} = \frac{A_t \hat{x}_k^t + x_k^{t+1}}{A_{t+1}}.$$

Замечание 1. Минусом данного алгоритма является то, что каждому соединению необходимо знать реакции всех пользователей, которые его используют на каждой итерации. К сожалению, в реальных сетях пользователи отправляют данные неодновременно, поэтому достаточно сложно собрать данную информацию для соединения. Однако наличие полной информации о пользователях позволяет соединению быстрее устанавливать равновесную цену.

3.2. Оценка скорости сходимости БГМ

Прежде чем привести доказательство сходимости БГМ для рассматриваемой задачи, сформулируем ключевую лемму, необходимую для получения оценок невязки в ограничениях и зазора двойственности после работы прямо двойственного метода PDFGM.

Лемма 1. *Пусть алгоритм* 1 *начинает свою работу с начальной точки* $\boldsymbol{\lambda}^0$, *которая лежит в евклидовом шаре радиуса* R *с центром в начале координат. Тогда после* N *итераций алгоритма* 1 *будет выполняться следующее неравенство*:

$$A_N\varphi(\mathbf{y}^N) - A_N U(\hat{\mathbf{x}}^N) + 2\hat{R}A_N\left\|(C\hat{\mathbf{x}}^N - \mathbf{b})_+\right\|_2 \le \frac{37L\hat{R}^2}{9}, \tag{6}$$

где $\hat{\mathbf{x}}^N = \frac{1}{A_N}\sum_{t=0}^{N-1}\alpha_t\mathbf{x}(\boldsymbol{\lambda}^t)$ *и* $\hat{R} = 3R$.

Доказательство леммы можно найти в Приложении.

Теперь сформулируем теорему об оценке скорости сходимости алгоритма 1.

Теорема 1. *Пусть алгоритм* 1 *начинает свою работу с начальной точки* $\boldsymbol{\lambda}^0$, *которая лежит в евклидовом шаре радиуса* R *с центром в начале координат. Тогда после*

$$N = \left\lceil \frac{2\hat{R}}{3}\sqrt{\frac{37L}{\varepsilon}} \right\rceil$$

итераций алгоритма 1 *будут выполняться следующие неравенства*:

$$U(\mathbf{x}^*) - U(\hat{\mathbf{x}}^N) \le \varepsilon, \quad \left\|(C\hat{\mathbf{x}}^N - \mathbf{b})_+\right\|_2 \le \frac{\varepsilon}{\hat{R}},$$

где $\hat{\mathbf{x}}^N = \frac{1}{A_N}\sum_{t=0}^{N-1}\alpha_t\mathbf{x}(\boldsymbol{\lambda}^t)$, $\mathbf{x}^*$ *— оптимальное решение задачи* (1), $\hat{R} = 3R$.

Доказательство. Обозначим оптимальное значение исходной прямой задачи (1) $\mathrm{Opt}[P]$, а оптимальное значение двойственной задачи (4) — $\mathrm{Opt}[D]$. В силу слабой двойственности имеем

$$\mathrm{Opt}[D] \ge \mathrm{Opt}[P].$$

Кроме того, для всех $\mathbf{x} \in \mathbb{R}^n_+$ и оптимального решения двойственной задачи (4) $\boldsymbol{\lambda}^*$ получаем

$$\mathrm{Opt}[P] \ge U(\mathbf{x}) - \left\langle \boldsymbol{\lambda}^*, \left(\sum_{k=1}^n \mathbf{C}_k x_k - \mathbf{b}\right)_+ \right\rangle \ge U(\mathbf{x}) - \hat{R}\left\|(C\mathbf{x} - \mathbf{b})_+\right\|_2. \tag{7}$$

Тогда

$$\varphi(\mathbf{y}^N) - U(\hat{\mathbf{x}}^N) = \varphi(\mathbf{y}^N) - U(\hat{\mathbf{x}}^N) + \mathrm{Opt}[P] - \mathrm{Opt}[P] + \mathrm{Opt}[D] - \mathrm{Opt}[D] = \underbrace{(\mathrm{Opt}[D] - \mathrm{Opt}[P])}_{\ge 0} +$$

$$+ (\mathrm{Opt}[P] - U(\hat{\mathbf{x}}^N)) + \underbrace{(\varphi(\mathbf{y}^N) - \mathrm{Opt}[D])}_{\ge 0} \overset{(7)}{\ge} -\left\langle \boldsymbol{\lambda}^*, (\mathbf{b} - C\hat{\mathbf{x}}^N)_+ \right\rangle \ge -\hat{R}\left\|(C\hat{\mathbf{x}}^N - \mathbf{b})_+\right\|_2.$$

После подстановки последнего неравенства в (6) получим следующую оценку:

$$\hat{R}\left\|(C\hat{\mathbf{x}}^N - \mathbf{b})_+\right\|_2 \le \frac{37L\hat{R}^2}{9A_N}.$$

Следовательно, $\varphi(\mathbf{y}^N) - U(\hat{\mathbf{x}}^N) \ge -\frac{37L\hat{R}^2}{9A_N}$. С другой стороны, из (6) следует, что $\varphi(\mathbf{y}^N) - U(\hat{\mathbf{x}}^N) \le \frac{37L\hat{R}^2}{9A_N}$. Поэтому

$$\left|\varphi(\mathbf{y}^N) - U(\hat{\mathbf{x}}^N)\right| \le \frac{37L\hat{R}^2}{9A_N}.$$

Так как $\varphi(\mathbf{y}^N) \geq \mathrm{Opt}[D] = \varphi(\mathbf{y}^*) \geq \mathrm{Opt}[P] = U(\mathbf{x}^*)$, выполняется следующее неравенство:

$$U(\mathbf{x}^*) - U(\hat{\mathbf{x}}^N) \leq \frac{37L\hat{R}^2}{9A_N} = \frac{148L\hat{R}^2}{9(N+1)(N+2)} \leq \frac{148L\hat{R}^2}{9N^2} \leq \varepsilon.$$

Выражая из последнего неравенства N, получаем оценку из условия теоремы.

4. СТОХАСТИЧЕСКИЙ МЕТОД ПРОЕКЦИИ СУБГРАДИЕНТА

Рассмотрим исходную задачу (1), но теперь предположим, что функции полезности $u_k(x_k)$, $k = 1,\ldots,n$, вогнуты, но не сильно вогнуты. В этом случае двойственная задача (4) становится негладкой. Поэтому для ее решения предлагается применить стохастический метод проекции субградиента. Впервые идея использовать для решения данной проблемы стохастический метод проекции субградиента предложена в работе [2].

Рассмотрим вероятностное пространство $(\Omega, \mathcal{F}, \mathbb{P})$. Пусть на нем определена последовательность независимых случайных величин $\{\xi^t\}_{t=0}^{\infty}$, равномерно распределенных на $\{1,\ldots,n\}$, т.е.

$$\mathbb{P}(\xi^t = i) = \frac{1}{n}, \quad i \in \{1,\ldots,n\}.$$

Если доступен оракул, выдающий стохастический субградиент двойственной функции $\nabla\varphi(\boldsymbol{\lambda}, \xi)$:

$$\nabla\varphi(\boldsymbol{\lambda}, \xi) = \mathbf{b} - nC_\xi x_\xi(\boldsymbol{\lambda}),$$

то имеем

$$\mathbb{E}[\mathbf{b} - nC_{\xi_t} x_{\xi_t}(\boldsymbol{\lambda}^t) | \xi^t] = \mathbf{b} - \sum_{k=1}^{n} \mathbf{C}_k x_k(\boldsymbol{\lambda}^t) = \nabla\varphi(\boldsymbol{\lambda}^t)$$

Алгоритм 2. Прямо двойственный стохастический метод проекции субградиента (PDSSGM), версия 1

Вход: $u_k(\mathbf{x})$, $k = 1,\ldots,n$ – вогнутые функции полезности для каждого пользователя; β – шаг метода.

1: $\boldsymbol{\lambda}^0 := \mathbf{0}$
2: **for** $t = 1,\ldots,N-1$
3: Вычислить $\nabla\varphi(\boldsymbol{\lambda}^{t-1}, \xi)$
4: $\boldsymbol{\lambda}^t := [\boldsymbol{\lambda}^{t-1} - \beta(\mathbf{b} - nC_{\xi^{t-1}} x_{\xi^{t-1}}(\boldsymbol{\lambda}^{t-1}))]_+$
5: $\hat{\mathbf{x}}^{t+1} := \frac{1}{t+1}\sum_{j=0}^{t} \mathbf{x}(\boldsymbol{\lambda}^j)$
6: $\hat{\boldsymbol{\lambda}}^{t+1} := \frac{1}{t+1}\sum_{j=0}^{t} \boldsymbol{\lambda}^j$
7: **end for**
8: **return** $\hat{\boldsymbol{\lambda}}^N$, $\hat{\mathbf{x}}^N$

Следовательно, стохастический субградиент является несмещенной оценкой субградиента.

Оптимальное решение задачи (2) будем искать с помощью прямо двойственного стохастического метода проекции субградиента PDSSGM. Приведем описание двух версий данного метода, см. алгоритм 2 и алгоритм 3. В алгоритме 2 используется полная модель восстановления вектора $\mathbf{x}(\boldsymbol{\lambda})$ на каждой итерации. Однако вычисление вектора $\mathbf{x}(\boldsymbol{\lambda})$ по сложности практически эквивалентно вычислению полного субградиента $\varphi(\boldsymbol{\lambda})$. Поэтому основным алгоритмом является алгоритм 3, в котором для восстановления вектора $\mathbf{x}(\boldsymbol{\lambda})$ используется неполная, стохастическая модель, что означает, что на каждой итерации пересчитывается только одна компонента вектора $\mathbf{x}(\boldsymbol{\lambda})$, остальные остаются без изменений. При доказательстве теоремы сходимости мы сначала

показываем оценку скорости сходимости для алгоритма 2, а затем показываем, что приближенное решение прямой задачи, полученное в алгоритме 3, близко по точности к решению, полученному в алгоритме 2.

Алгоритм 3. Прямо двойственный стохастический метод проекции субградиента (PDSSGM), версия 2

Вход: $u_k(\mathbf{x})$, $k = 1,\ldots,n$ – вогнутые функции полезности для каждого пользователя; β – шаг метода.

1: $\boldsymbol{\lambda}^0 := \mathbf{0}$

2: **for** $t = 1,\ldots,N-1$

3: Вычислить $\nabla\varphi(\boldsymbol{\lambda}^{t-1}, \xi)$

4: $\boldsymbol{\lambda}^t := [\boldsymbol{\lambda}^{t-1} - \beta(\mathbf{b} - nC_{\xi^{t-1}}x_{\xi^{t-1}}(\boldsymbol{\lambda}^{t-1}))]_+$

5: $\tilde{\mathbf{x}}^{t+1}_{\xi^t} := \frac{t}{t+1}\tilde{\mathbf{x}}^t_{\xi^t} + \frac{1}{t+1}nx_{\xi^t}(\lambda^t)$, $\tilde{\mathbf{x}}^{t+1}_j := \tilde{\mathbf{x}}^t_j$ при $j \neq \xi^t$

6: $\hat{\boldsymbol{\lambda}}^{t+1} := \frac{1}{t+1}\sum_{j=0}^{t}\boldsymbol{\lambda}^j$

7: **end for**

8: **return** $\hat{\boldsymbol{\lambda}}^N$, $\tilde{\mathbf{x}}^N$

4.1. Распределенный метод

Рассмотрим, как можно применить для решения рассматриваемой задачи стохастический метод проекции субградиента в распределенном варианте.

Опишем процесс, происходящий на t-й итерации для соединения j:

1. На основе информации, полученной от соединений после предыдущей итерации с номером $t-1$, случайный пользователь ξ^t передает данные с оптимальной скоростью

$$x_{\xi^t}(\boldsymbol{\lambda}^{t+1}) = \arg\max_{x_{\xi^t}\in\mathbf{R}_+}\left(u_{\xi^t}(x_{\xi^t}) - x_{\xi^t}\sum_{j=1}^{m}\lambda_j^{t+1}C^j_{\xi^t}\right),$$

где в силу определения матрицы C пользователю необходима информация только от соединений, которые он использует.

2. Далее соединение j вычисляет цену на следующую итерацию на основе реакции этого пользователя:

$$\lambda_j^{t+1} = [\lambda_j^t - \beta(b_j - nC^j_{\xi^t}x^t_{\xi^t})]_+.$$

При этом $C^j_{\xi^t} \neq 0$ только для тех пользователей, которые используют соединение j. Поэтому цена поменяется только для актуальных соединений пользователя, который передал данные.

Замечание 2. Главное преимущество данного метода заключается в том, что соединение меняет цену только на основе реакции одного пользователя, что гораздо больше приближает постановку задачи к ситуации в реальных сетях, в которых пользователи отправляют данные не одновременно.

4.2. Оценка скорости сходимости для стохастического метода проекции субградиента

Прежде чем перейти к доказательству основной теоремы об оценках скорости сходимости предложенных методов, приведем необходимые предположения для решаемой задачи. Предположим, что существует положительная константа $M = O(n\sqrt{m})$ такая, что верна следующая оценка:

$$\|\nabla\varphi(\lambda,\xi)\|_2 \leq M. \tag{8}$$

Данное предположение корректно в силу того, что скорость передачи информации $\mathbf{x}$ ограничена и вектор пропускных способностей $\mathbf{b}$ также ограничен в силу физических соображений. Поэтому, исходя из определения стохастического субградиента, он ограничен.

Также предположим, что

$$\mathbb{E}\left[\exp\left(\frac{\|\nabla\varphi(\lambda,\xi)-\nabla\varphi(\lambda)\|_2^2}{\sigma^2}\right)\right]\le\exp(1),$$

где σ – положительная числовая константа, порядок зависимости от n и m такой же, как у M.

Для получения оценки скорости сходимости алгоритма 9 необходимо предположение о том, что функции $u_k(x_k), k=1,\ldots,n$, являются липшицевыми с константой M_{u_k}, тогда функция $U(\mathbf{x})$ будет липшицевой с некоторой константой M_U:

$$\forall\mathbf{x},\mathbf{y}\quad |U(\mathbf{x})-U(\mathbf{y})|\le M_U\|\mathbf{x}-\mathbf{y}\|_2,$$

при этом $M_U=O(\sqrt{n})$. Может оказаться, что функция $u_k(x_k)$ липшицева везде, кроме, например, точки 0. Примером такой функции является одна из наиболее распространенных функций полезности пользователей $u_k(x_k)=\ln x_k$. Но в силу специфики рассматриваемой задачи всегда существуют $\overline{\varepsilon}>0$, $\underline{\varepsilon}>0$ такие, что $x_k^*\ge\underline{\varepsilon}$, $x_k^*\le\overline{\varepsilon}$. Тогда задачу можно решать на компакте $Q=\{\mathbf{x}:\underline{\varepsilon}\le x_k\le\overline{\varepsilon},\ k=1,\ldots,n\}$, и рассматриваемая функция $u_k(x_k)=\ln x_k$ становится липшицевой на Q. В общем случае вогнутая функция полезности $u(x)$ будет липшицевой на компакте, лежащем в относительной внутренности области определения $u(x)$.

Пусть

$$\mathbb{E}\left[\exp\left(\frac{\|\mathbf{x}(\lambda\lambda,\xi)-\mathbf{x}(\lambda)\|_2^2}{\sigma_x^2}\right)\right]\le\exp(1),$$

где $\sigma_x=O(\sqrt{n})$ – положительная числовая константа и

$$\mathbf{x}(\lambda,\xi)=(0,\ldots,nx_\xi(\lambda),\ldots,0)^{\mathrm{T}}.$$

Сформулируем ключевую лемму, необходимую для получения оценок скорости сходимости невязки в ограничениях и зазора двойственности после работы прямо двойственного метода PDSSGM.

Лемма 2. *Пусть алгоритм* 3 *начинает свою работу с начальной точки* $\lambda^0=0$ *и с шагом* β. *Тогда после* N *итераций алгоритма* 3 *с вероятностью* $1-4\delta$ *выполняется неравенство*:

$$\varphi(\hat{\lambda}^N)-U(\tilde{\mathbf{x}}^N)+2R\left\|[C\tilde{\mathbf{x}}^N-\mathbf{b}]_+\right\|_2\le C_1\frac{R^2\sigma\sqrt{g(N)J}}{\sqrt{N}}+\frac{2R^2}{\beta N}+\frac{\beta M^2}{2}+$$
$$+\frac{\sqrt{2}\left(1+\sqrt{3\ln\frac{1}{\delta}}\right)}{\sqrt{N}}\left(M_U\sigma_x+2R\left(\sigma+\sigma_x\sqrt{\lambda_{\max}(C^{\mathrm{T}}C)}\right)\right),$$

где

$$\hat{\lambda}^N=\frac{1}{N}\sum_{t=0}^{N-1}\lambda^t$$

и

$$\tilde{\mathbf{x}}^N=\frac{1}{N}\sum_{t=0}^{N-1}\mathbf{x}(\lambda^t,\xi^t),$$

C_1 – *положительная числовая константа*, $g(N)=\ln\left(\frac{N}{\delta}\right)+\ln\ln\left(\frac{F}{f}\right)$,

$$F=2\sigma^2N(2\beta)^N\left(2R^2+2\beta^2M^2+\beta R^2+24\ln\frac{N}{\delta}\beta\sigma^2N\right),$$

$f = \sigma^2 R^2$ *и*

$$J = \max\left\{1, \quad \frac{1}{R}\beta C_1\sqrt{\sigma^2 g(N)} + \sqrt{\frac{1}{R^2}\beta^2 C_1^2\sigma^2 g(N) + \frac{2R^2 + 2\beta^2 M^2}{R^2}}\right\},$$

а R *определяется условием* $\|\lambda^*\|_2 \le R$.

Доказательство леммы приведено в Приложении.

Теперь сформулируем теорему об оценке скорости сходимости алгоритма 3.

Теорема 2. *Пусть алгоритм* 3 *начинает свою работу с начальной точки* $\lambda^0 = 0$ *и с шагом* $\beta = \frac{R}{M\sqrt{N}}$. *Обозначим*

$$A = \sqrt{2}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right)\left(M_U\sigma_x + 2R\left(\sigma + \sigma_x\sqrt{\lambda_{\max}(C^{\mathrm{T}}C)}\right)\right) + 2.5RM.$$

Тогда после

$$N = O\left(\left\lceil \frac{A^2}{\varepsilon^2}\ln\left(\frac{MR}{\varepsilon\delta}\right)\right\rceil\right)$$

итераций алгоритма 3 *с вероятностью* $1 - 4\delta$ *будут выполняться следующие неравенства*:

$$U(x^*) - U(\tilde{x}^N) \le \varepsilon, \quad \left\|(C\tilde{x}^N - b)_+\right\|_2 \le \frac{\varepsilon}{R},$$

где $\tilde{x}^N = \frac{1}{N}\sum_{t=0}^{N-1} x(\lambda^t, \xi^t)$, x^* *— оптимальное решение задачи* (1).

Доказательство. Начало доказательства такое же, как в теореме 1, но с использованием оценки из леммы 2. В результате для шага $\beta = \frac{R}{M\sqrt{N}}$ получим, что

$$\frac{\sqrt{2}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right)}{\sqrt{N}}\left(M_U\sigma_x + 2R\left(\sigma + \sigma_x\sqrt{\lambda_{\max}(C^{\mathrm{T}}C)}\right)\right) + \frac{5RM}{2\sqrt{N}} + C_1\frac{R^2\sigma\sqrt{g(N)J}}{\sqrt{N}},$$

при этом с точностью до констант $g(N) \approx \ln\left(\frac{N}{\delta}\right)$, $J \approx \max\left\{1, \beta\sqrt{g(N)}\right\}$. Далее найдем N, при котором оценка становится меньше ε.

Введем следующие обозначения:

$$A = \sqrt{2}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right)\left(M_U\sigma_x + 2R\left(\sigma + \sigma_x\sqrt{\lambda_{\max}(C^{\mathrm{T}}C)}\right)\right) + 2.5RM,$$

$$B = C_1R^2\sigma.$$

Необходимо получить минимальную оценку на число итераций N для достижения заданной точности ε. Для $J = 1$ получаем, что

$$\sqrt{N} = \left\lceil \frac{A + B\sqrt{\ln\left(\frac{N}{\delta}\right)}}{\varepsilon} \right\rceil. \tag{9}$$

Подставляя рекурсивно значение N, из (9) получим следующую оценку сложности:

$$N = O\left(\left\lceil \frac{A^2}{\varepsilon^2}\ln\left(\frac{MR}{\varepsilon\delta}\right)\right\rceil\right).$$

Для $J = \beta\sqrt{g(N)} = \dfrac{R\sqrt{g(N)}}{M\sqrt{N}}$ потребуем, чтобы

$$\frac{A}{\sqrt{N}} + \frac{Bg(N)R}{MN} = \frac{A}{\sqrt{N}} + \frac{\bar{B}g(N)}{N} \le \varepsilon.$$

Так как нам нужно найти минимальное N, то, заменив последнее неравенство на равенство и решая полученное уравнение, получаем, что

$$\sqrt{N} = \left\lceil \frac{A + \sqrt{A^2 + 4\varepsilon\bar{B}\ln\left(\dfrac{N}{\delta}\right)}}{2\varepsilon} \right\rceil.$$

Аналогично случаю $J = 1$ из последнего равенства получается следующая оценка:

$$N = O\left(\left\lceil \frac{A^2}{\varepsilon}\ln\left(\frac{MR}{\varepsilon\delta}\right)\right\rceil\right).$$

Худшая из оценок сложности для $J = 1$ и для $J = \beta\sqrt{g(N)}$ и будет оценкой из условия теоремы.

5. МЕТОД ЭЛЛИПСОИДОВ

В этом разделе для решения исходной задачи (1) предлагается применить метод эллипсоидов [25]. Данный метод можно использовать при небольшой размерности (m) двойственной задачи или в случае, когда требуется высокая точность решения. Данный метод является прямо двойственным, т.е. по решению двойственной задачи можно восстановить решение прямой задачи.

Рассмотрим исходную задачу (1) и двойственную к ней (2). Как и в предыдущем разделе, считаем, что функции $u_k(x_k)$, $k = 1,\ldots,n$, являются вогнутыми, но не сильно вогнутыми. Также предположим, что решение двойственной задачи лежит в евклидовом шаре радиуса R с центром в начале координат, т.е. $\|\boldsymbol{\lambda}^*\|_2 \le R$. В качестве начальной точки метода возьмем нулевой вектор: $\boldsymbol{\lambda}^0 = 0$. Задачу будем решать на множестве:

$$\Lambda_{2R} = \left\{\boldsymbol{\lambda} \in \mathbb{R}^m_+ : \|\boldsymbol{\lambda}\|_2 \le 2R\right\}.$$

Приведем описание метода эллипсоидов (алгоритм 4), который будем применять для решения двойственной задачи.

Алгоритм 4. Метод эллипсоидов

Вход: $u_k(x_k), k = 1,\ldots,n$ – вогнутые функции полезностей

1: $B_0 := 2R \cdot I_n$, I_n – единичная матрица

2: **for** $t = 0,\ldots,N-1$

3: Вычислить $\nabla\varphi(\boldsymbol{\lambda}^t)$

4: $\mathbf{q}_t := B_t^{\mathrm{T}}\nabla\varphi(\boldsymbol{\lambda}^t)$

5: $\mathbf{p}_t := \dfrac{B_t^{\mathrm{T}}\mathbf{q}_t}{\sqrt{\mathbf{q}_t^{\mathrm{T}}B_tB_t^{\mathrm{T}}\mathbf{q}_t}}$

6: $B_{t+1} := \dfrac{m}{\sqrt{m^2-1}}B_t + \left(\dfrac{m}{m+1} - \dfrac{m}{\sqrt{m^2-1}}\right)B_t\mathbf{p}_t\mathbf{p}_t^{\mathrm{T}}$

7: $\boldsymbol{\lambda}^{t+1} := \boldsymbol{\lambda}^t - \dfrac{1}{m+1}B_t\mathbf{p}_t$

8: **end for**

9: **return** $\boldsymbol{\lambda}^N$

Для восстановления решения прямой задачи по решению двойственной необходимо определить сертификат точности ξ для метода эллипсоидов. Напомним, что сертификатом точности называется последовательность $\xi = \{\xi_t\}_{t=0}^{N-1}$ весов таких, что

$$\xi_t \geq 0, \quad \sum_{t=0}^{N-1} \xi_t = 1.$$

Построение сертификата точности в нашем случае будет осуществляться в процессе работы метода эллипсоидов, см. алгоритм 12, общая схема которого такова [26]:

1. Находим "наиболее узкую полоску", содержащую эллипсоид Q_N, остающийся после итерации N, т.е. такой вектор $\mathbf{h}$, что на Q_N выполняется неравенство:

$$\max_{\boldsymbol{\lambda} \in Q_N} \langle \mathbf{h}, \boldsymbol{\lambda} \rangle - \min_{\boldsymbol{\lambda} \in Q_N} \langle \mathbf{h}, \boldsymbol{\lambda} \rangle \leq 1. \tag{10}$$

Для метода эллипсоидов все Q_N представлены в виде

$$Q_N = \{B_N \mathbf{z} + \boldsymbol{\lambda}^N : \mathbf{z}^{\mathrm{T}} \mathbf{z} \leq 1\}.$$

Тогда для решения (10) необходимо сделать сингулярное разложение матрицы $B_N = UDV$, где U и V – ортогональные матрицы и D – диагональная матрица с положительными элементами на диагонали. Далее искомый вектор $\mathbf{h}$ определяется следующим образом: $\mathbf{h} = 1/(2\sigma^{i}*) \cdot U\mathbf{e}^{i}*$, где i_* – индекс наименьшего диагонального элемента матрицы D, $\sigma^{i}*$ – соответствующее этому индексу значение элемента матрицы D, $\mathbf{e}^i$ – векторы стандартного базиса.

2. Для векторов $\mathbf{h}^+ = \left[\mathbf{h}, -\langle \mathbf{h}, \boldsymbol{\lambda}^N \rangle\right]$ и $\mathbf{h}^- = -\mathbf{h}^+$ находим разложения следующего вида:

$$\mathbf{h}^+ = \sum_{t=0}^{N-1} \nu_t \left[\nabla\varphi(\boldsymbol{\lambda}^t), -\langle \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^t \rangle\right] + \mathbf{y}^+,$$

$$\mathbf{h}^- = \sum_{t=0}^{N-1} \mu_t \left[\nabla\varphi(\boldsymbol{\lambda}^t), -\langle \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^t \rangle\right] + \mathbf{z}^+,$$

существование которых следует из утверждения 4.1 [26]. Этот шаг описывают п. 6–13 рассмотренного ниже алгоритма 5.

3. Из коэффициентов разложения ν_t и μ_t векторов $\mathbf{h}^+$ и $\mathbf{h}^-$, соответственно, получаем выражения для ξ_t, $t \in I_N$, где

$$I_N = \{t \leq N-1 : \boldsymbol{\lambda}^t \in \operatorname{int} \Lambda_{2R}\}.$$

Коэффициенты разложения определяются только для тех точек, получаемых в процессе работы алгоритма 5, которые принадлежат допустимому множеству.

Алгоритм 5. Построение сертификата точности для метода эллипсоидов

Вход: $N-1$ – номер итерации, на которой производится вычисление сертификата точности, $\left\{B_t, \boldsymbol{\lambda}^t, \nabla\varphi(\boldsymbol{\lambda}^t)\right\}_{t=0}^{N-1}$ – протокол работы метода эллипсоидов после N итераций

1: **если** $\nabla\varphi(\boldsymbol{\lambda}^{N-1}) = 0$, **то**

2: $\quad \xi_t := 0$ для всех $t = 0, \ldots, N-2$

3: $\quad \xi_{N-1} := 1$

4: **иначе**

5: $\quad \mathbf{h} := 1/(2\sigma^{i}*) \cdot U\mathbf{e}^{i}*$

6: $\quad \mathbf{g}_\nu := \mathbf{h}$, $\mathbf{g}_\mu := -\mathbf{h}$

7: $\quad$ **for** $t = 0, \ldots, N-1$

8: $\mathbf{q} := B_t^{\mathrm{T}} \nabla \varphi(\lambda^t)$

9: $\nu_t := [\mathbf{g}_\nu^{\mathrm{T}} B_t \mathbf{q}]_+ / \|\mathbf{q}\|_2^2$

10: $\mathbf{g}_\nu := \mathbf{g}_\nu - \nu_t \nabla \varphi(\boldsymbol{\lambda}^t)$

11: $\mu_t := [\mathbf{g}_\mu^{\mathrm{T}} B_t \mathbf{q}]_+ / \|\mathbf{q}\|_2^2$

12: $\mathbf{g}_\mu := \mathbf{g}_\mu - \mu_t \nabla \varphi(\boldsymbol{\lambda}^t)$

13: **end for**

14: $\xi_t := (\nu_t + \mu_t) / \sum_{i \in I_N} (\nu_i + \mu_i)$, $t \in I_N$

15: **end если**

16: **return** $\{\xi_t\}_{t=0}^{N-1}$

Замечание 3. Отметим, что, в отличие от БГМ и стохастического метода проекции субградиента, в методе эллипсоидов для вычисления шагов 4–6 алгоритма 4 необходима информация о всех компонентах градиента, т.е. нужна информация от всех пользователей. Поэтому необходим общий центр для всех соединений, который будет собирать информацию со всех соединений и выполнять эти вычисления.

Сформулируем теорему об оценке скорости сходимости метода эллипсоидов для решаемой задачи.

Теорема 3 (см. [26]). *Пусть алгоритм* 4 *начинает свою работу с начальным шаром* $B_0 = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \|\boldsymbol{\lambda}\| \le 2R\}$ *и сертификат точности* ξ *формируется в соответствии с алгоритмом* 5. *Тогда после*

$$N = 2m(m+1)\left\lceil \ln\left(\frac{32 \cdot 4MR}{\varepsilon}\right)\right\rceil \tag{11}$$

итераций будут выполняться следующие неравенства:

$$U(\mathbf{x}^*) - U(\hat{\mathbf{x}}^N) \le \varepsilon, \quad \left\|[C\hat{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 \le \frac{\varepsilon}{R},$$

где

$$\hat{\mathbf{x}}^N = \sum_{t \in I_N} \xi_t \mathbf{x}(\lambda^t), \quad I_N = \left\{t \le N-1 : \lambda^t \in \operatorname{int} \Lambda_{2R}\right\}.$$

Доказательство теоремы можно найти ниже в Приложении.

6. РЕГУЛЯРИЗАЦИЯ ДВОЙСТВЕННОЙ ЗАДАЧИ

В предыдущих разделах мы рассматривали прямо двойственные методы для решения двойственной задачи. Однако существует стандартный подход, который позволяет по решению двойственной задачи не прямо двойственным методом восстанавливать решение прямой задачи. Ключевой идеей данного подхода является регуляризация двойственной задачи, которая гарантирует сильную выпуклость регуляризованной задачи. Далее опишем подробно предлагаемый подход и сформулируем леммы, связывающие решения прямой и двойственной задачи.

Регуляризуем функционал (2) по Тихонову

$$\varphi_\delta(\boldsymbol{\lambda}) = \varphi(\boldsymbol{\lambda}) + \frac{\delta}{2}\|\boldsymbol{\lambda}\|_2^2$$

и вместо задачи (4) будем решать регуляризованную задачу

$$\min_{\boldsymbol{\lambda} \in \mathbf{R}_+^m} \varphi_\delta(\boldsymbol{\lambda}).$$

Параметр δ будет оптимально подобран позже. Как и в предыдущем разделе, предполагается, что задача решается на множестве

$$\Lambda_{2R} = \{\lambda \in \mathbb{R}_+^m : \|\boldsymbol{\lambda}\|_2 \le 2R\}.$$

Для полученной регуляризованной функции сформулируем следующую лемму о гладкости регуляризованной задачи.

Лемма 3. *Пусть функция* $\varphi(\lambda)$ *—* L*-гладкая, тогда регуляризованная функция* $\varphi_\delta(\lambda)$ *является* $(L+\delta)$*-гладкой, т.е. для любых* $\lambda^1, \lambda^2 \in \mathbf{R}^m_+$

$$\left\|\nabla\varphi_\delta(\lambda^1) - \nabla\varphi_\delta(\lambda^2)\right\|_2 \le (L+\delta)\left\|\lambda^1 - \lambda^2\right\|_2. \tag{12}$$

Доказательство. Градиент регуляризованной функции

$$\nabla\varphi_\delta(\lambda) = \nabla\varphi(\lambda) + \delta\lambda.$$

Следовательно, имеем оценку

$$\left\|\nabla\varphi_\delta(\lambda^1) - \nabla\varphi_\delta(\lambda^2)\right\|_2 = \left\|\nabla\varphi(\lambda^1) - \nabla\varphi(\lambda^2) + \delta(\lambda^1 - \lambda^2)\right\|_2 \le \left\|\nabla\varphi(\lambda^1) - \nabla\varphi(\lambda^2)\right\|_2 + \delta\left\|\lambda^1 - \lambda^2\right\|_2.$$

Откуда в силу утверждения 2 следует (12).

Также для оценки сходимости алгоритма для прямой задачи нам понадобится следующая вспомогательная лемма о связи между оценкой на градиент для двойственной задачи и оценками на сходимость по функции и невязку в ограничении для прямой задачи.

Лемма 4 (см. [10]). *Пусть* $\mathbf{x}^*$ *— решение прямой задачи* (1). *Тогда выполняются следующие неравенства*:

$$\left\|C\mathbf{x}(\lambda) - \mathbf{b}\right\|_2 \le \left\|\nabla\varphi_\delta(\lambda)\right\|_2 + \delta\left\|\lambda\right\|_2, \tag{13}$$

$$U(\mathbf{x}^*) - U(\mathbf{x}(\lambda)) \le \left\|\nabla\varphi_\delta(\lambda)\right\|_2 \cdot \left\|\lambda\right\|_2 + \delta\left\|\lambda\right\|_2^2, \tag{14}$$

где $\mathbf{x}(\lambda)$ *определяется в* (3).

Доказательство. В силу (3) имеем

$$U(\mathbf{x}(\lambda)) + \left\langle\lambda, \mathbf{b} - C\mathbf{x}(\lambda)\right\rangle \ge U(\mathbf{x}^*) + \left\langle\lambda, \mathbf{b} - C\mathbf{x}^*\right\rangle \ge U(\mathbf{x}^*).$$

Откуда

$$U(\mathbf{x}(\lambda)) \ge U(\mathbf{x}^*) - \left\langle\lambda, \mathbf{b} - C\mathbf{x}(\lambda)\right\rangle = U(\mathbf{x}^*) - \left\langle\lambda, \nabla\varphi(\lambda)\right\rangle.$$

Так как $\varphi(\lambda) = \varphi_\delta(\lambda) - \dfrac{\delta}{2}\|\lambda\|_2^2$, имеем

$$\left\|\nabla\varphi(\lambda)\right\|_2 = \left\|\nabla\varphi_\delta(\lambda) - \delta\lambda\right\|_2 \le \left\|\nabla\varphi_\delta(\lambda)\right\| + \delta\left\|\lambda\right\|_2.$$

Из последнего неравенства, учитывая соотношение $\nabla\varphi(\lambda) = \mathbf{b} - C\mathbf{x}(\lambda)$, получаем (13).

Далее, оценка (14) следует из

$$U(\mathbf{x}_*) - U(\mathbf{x}(\lambda)) \le \left\langle\lambda, \nabla\varphi(\lambda)\right\rangle \le \left\|\nabla\varphi(\lambda)\right\|_2 \cdot \left\|\lambda\right\|_2 \le$$
$$\le \left\|\lambda\right\|_2 \cdot \left(\left\|\nabla\varphi_\delta(\lambda)\right\|_2 + \delta\left\|\lambda\right\|_2\right) \le \left\|\nabla\varphi_\delta(\lambda)\right\|_2 \cdot \left\|\lambda\right\|_2 + \delta\left\|\lambda\right\|_2^2.$$

Кроме этого, понадобится лемма о сходимости по градиенту для регуляризованной функции.

Лемма 5. *Пусть* λ^*_δ *— решение регуляризованной двойственной задачи. Имеет место следующее неравенство*:

$$\left\|\nabla\varphi_\delta(\lambda^N)\right\|_2 \le (L+\delta)\left\|\lambda^N - \lambda^*_\delta\right\|_2.$$

Доказательство немедленно следует из леммы 3 и равенства

$$\nabla\varphi_\delta(\lambda^*_\delta) = 0.$$

Мы сформулировали необходимые леммы для регуляризованной задачи и в следующем разделе рассмотрим на примере применение этого подхода.

7. МЕТОД ЭКСТРАПОЛЯЦИИ СЛУЧАЙНОГО ГРАДИЕНТА

Рассмотрим метод экстраполяции случайного градиента [27]. Отметим, что данный метод не требует пересчета градиента на каждой итерации, необходимо пересчитывать только одну его компоненту на каждой итерации, что значительно сокращает вычисления, особенно для задач

большой размерности. Так как данный метод не является прямо двойственным, то необходимо применять алгоритм 6 к регуляризованной задаче.

Параметры α, η, τ и θ_t выбираются следующим образом:

$$\bar{\alpha} = 1 - \frac{1}{n + \sqrt{n^2 + 16nL/\delta}}, \tag{15}$$

$$\alpha = n\bar{\alpha}, \quad \eta = \frac{\delta\bar{\alpha}}{1-\bar{\alpha}}, \quad \tau = \frac{1}{n(1-\bar{\alpha})} - 1, \quad \theta_t = \bar{\alpha}^{-t}. \tag{16}$$

7.1. Распределенный метод

В данной секции опишем распределенную версию рассмотренного метода. Для начала отметим, что векторы $\underline{\boldsymbol{\lambda}}_1^0, \ldots, \underline{\boldsymbol{\lambda}}_n^0$ хранятся у соответствующих пользователей и влияют на формирование оптимального значения трафика для соответствующего пользователя. Как было отмечено при описании распределенного БГМ, на определение оптимального трафика пользователя влияют только цены соединений, через которые данный пользователь обменивается пакетами. Поэтому можно считать, что в векторе $\underline{\boldsymbol{\lambda}}_k^t$ ненулевые только те компоненты, номера которых совпадают с номерами используемых соединений.

Алгоритм 6. Метод экстраполяции случайного градиента (RGEM)

Вход: Параметры α, η, τ, $\{\theta_t\}_{t=1}^N$

$\boldsymbol{\lambda}^0 := \mathbf{0}$

$\underline{\boldsymbol{\lambda}}_i^0 := \boldsymbol{\lambda}^0$, $i = 1, \ldots, n$

$y_{-1} = y_0 = \mathbf{0}$

for $t = 1, \ldots, N$

  Выбрать случайным образом k_t из множества $\{1, \ldots, n\}$ равномерно по всем значениям

  $\tilde{\mathbf{y}}_k^t := \mathbf{y}_k^{t-1} + \alpha(\mathbf{y}_k^{t-1} - \mathbf{y}_k^{t-2})$, $k = 1, \ldots, n$

  $\boldsymbol{\lambda}^t := \left[\eta\boldsymbol{\lambda}^{t-1} - \frac{1}{n}\sum_{k=1}^n \tilde{\mathbf{y}}_k^t\right]_+ / (\delta + \eta)$

  $\underline{\boldsymbol{\lambda}}_{k_t}^t := (\boldsymbol{\lambda}^t + \tau\underline{\boldsymbol{\lambda}}_{k_t}^{t-1})/(1+\tau)$

  $\underline{\boldsymbol{\lambda}}_k^t := \underline{\boldsymbol{\lambda}}_k^{t-1}$, $k \in \{1, \ldots n\} \backslash \{k_t\}$

  $\mathbf{y}_{k_t}^t := \mathbf{b} - n\mathbf{C}_{k_t} x_{k_t}(\underline{\boldsymbol{\lambda}}_{k_t}^t)$

  $\mathbf{y}_k^t := \mathbf{y}_k^{t-1}$, $k \in \{1, \ldots, n\} \backslash \{k_t\}$

end for

$\bar{\boldsymbol{\lambda}}^N := \left(\sum_{t=0}^{N-1} \theta_t \boldsymbol{\lambda}^t\right) / \sum_{t=1}^N \theta_t$

return $\bar{\boldsymbol{\lambda}}^N$

Опишем распределенный алгоритм на t-й итерации.

1. Используя информацию, собранную с пользователей на предыдущей итерации, соединение j вычисляет

$$\tilde{y}_{k,j}^t := y_{k,j}^{t-1} + \alpha(y_{k,j}^{t-1} - y_{k,j}^{t-2}) = b_j - nC_k^j x_k(\underline{\boldsymbol{\lambda}}_k^{t-1}) + \alpha(nC_k^j x_k(\underline{\boldsymbol{\lambda}}_k^{t-2}) - nC_k^j x_k(\underline{\boldsymbol{\lambda}}_k^{t-1})).$$

Заметим, что в силу определения матрицы C соединению j нужна информация только от пользователей, которые обмениваются пакетами через это соединение.

2. Далее соединение j меняет цену по следующему правилу:

$$\lambda_j^t = \left[\eta\lambda_j^{t-1} - \frac{1}{n}\sum_{k=1}^{n}\tilde{y}_{k,j}^t\right]_+ /(\delta+\eta).$$

3. Один из пользователей k_t реагирует на изменение цен и в качестве локального вектора цен сохраняет

$$\underline{\boldsymbol{\lambda}}_{k_t}^t = (\boldsymbol{\lambda}^t + \tau\underline{\boldsymbol{\lambda}}_{k_t}^{t-1})/(1+\tau),$$

остальные пользователи никак не изменяют локальные цены, т.е. $\underline{\boldsymbol{\lambda}}_{k_t}^t = \underline{\boldsymbol{\lambda}}_{k_t}^{t-1}$.

4. Пользователь k_t вычисляет

$$x_{k_t}^t(\underline{\lambda}_{k_t,j}^t) = \arg\max_{x_k\in\mathbf{R}_+}\left(u_k(x_k) - x_k\sum_{j=1}^{m}\underline{\lambda}_{k_t,j}^t C_k^j\right)$$

и отправляет эту информацию соединениям, которые использует.

5. Соединение j обновляет у себя информацию для k_t пользователя

$$y_{k_t,j}^t = b_j - nC_{k_t}^j x_{k_t}(\underline{\boldsymbol{\lambda}}_{k_t}^t).$$

Данную информацию соединение будет обновлять только в случае, когда через него обменивается пакетами пользователь k_t.

7.2. Оценка скорости сходимости метода экстраполяции случайного градиента

В данном разделе мы, так же, как в разд. 3, рассматриваем задачу (2) с μ-сильно вогнутыми функциями затрат $u_k(x_k), k = 1,\ldots,n$. Напомним, что в силу сильной вогнутости функций затрат двойственная задача (4) будет гладкой с константой Липшица $L = \frac{nm^2}{\mu}$.

Для оценки скорости сходимости метода необходима следующая оценка для невязки по аргументу из теоремы 2.1 [27]:

$$\mathbb{E}\left[\left\|\boldsymbol{\lambda}_\delta^* - \boldsymbol{\lambda}^N\right\|_2^2\right] \le \frac{4\Delta(\overline{\alpha})^N}{\delta}, \tag{17}$$

где $\Delta = \delta\left\|\boldsymbol{\lambda}_\delta^* - \boldsymbol{\lambda}^0\right\|_2^2 + \frac{B}{n\delta} + \varphi_\delta(\lambda^0) - \varphi_\delta(\lambda_\delta^*)$, $B = \|\mathbf{b}\|_2^2$.

Используя (17), можно доказать теорему об оценке скорости сходимости метода непосредственно для задачи (11).

Теорема 4. *Пусть для решения регуляризованной двойственной задачи* (11) *применяется метод RGEM с параметрами* (15), (16) *и* $\delta = \frac{\varepsilon}{8R^2}$ *для*

$$N = \left\lceil 2\left(n + \sqrt{n^2 + \frac{128nLR^2}{\varepsilon}}\right)\ln\left(\frac{4RA}{\varepsilon}\right)\right\rceil$$

итераций, где $A = 2\left(LR + \frac{\varepsilon}{8R}\right)\sqrt{6 + \frac{16LR^2n + 8B}{n\varepsilon}}$. *Тогда*

$$\mathbb{E}[U(\mathbf{x}^*) - U(\mathbf{x}(\lambda^N))] \le \varepsilon, \quad \mathbb{E}\left[\left\|C\mathbf{x}(\lambda^N) - \mathbf{b}\right\|_2\right] \le \frac{\varepsilon}{2R}.$$

Доказательство. Из леммы 4 имеем оценки для невязки по ограничениям (13) и по целевой функции (14). Ввиду предположения $\boldsymbol{\lambda} \in \Lambda_{2R}$, выполняются следующие неравенства:

$$\left\|C\mathbf{x}^N - \mathbf{b}\right\|_2 \le \left\|\nabla\varphi_\delta(\boldsymbol{\lambda}^N)\right\|_2 + 2\delta R, \tag{18}$$

$$U(\mathbf{x}^*) - U(\mathbf{x}^N) \le 2R\left\|\nabla\varphi_\delta(\boldsymbol{\lambda}^N)\right\|_2 + 4\delta R^2, \tag{19}$$

где $\mathbf{x}^N = \mathbf{x}(\boldsymbol{\lambda}^N)$. Из леммы 5 и неравенства (17) получаем следующую оценку для $\left\|\nabla\varphi_\delta(\boldsymbol{\lambda}^N)\right\|_2$:

$$\mathbb{E}\left[\left\|\nabla\varphi_\delta(\boldsymbol{\lambda}^N)\right\|_2\right] \le 2(L+\delta)\sqrt{\frac{\Delta}{\delta}}(\bar{\alpha})^{N/2}.$$

Оценим Δ. Для функции φ_δ с липшицевым градиентом выполняется неравенство

$$\varphi_\delta(\boldsymbol{\lambda}^0) - \varphi_\delta(\boldsymbol{\lambda}_\delta^*) \le \left\langle\nabla\varphi_\delta(\boldsymbol{\lambda}_\delta^*), \boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\right\rangle + \frac{L+\delta}{2}\left\|\boldsymbol{\lambda}_\delta^* - \boldsymbol{\lambda}^0\right\|_2^2.$$

Учитывая, что $\nabla\varphi_\delta(\boldsymbol{\lambda}_\delta^*) = 0$, получаем

$$\Delta \le \delta\left\|\boldsymbol{\lambda}_\delta^* - \boldsymbol{\lambda}^0\right\|_2^2 + \frac{B}{n\delta} + \frac{L+\delta}{2}\left\|\boldsymbol{\lambda}_\delta^* - \boldsymbol{\lambda}^0\right\|_2^2 \le (6\delta + 2L)R^2 + \frac{B}{n\delta}.$$

Пусть δ подбирается так, чтобы выполнялось $4\delta R^2 = \frac{\varepsilon}{2}$, тогда $\delta = \frac{\varepsilon}{8R^2}$. Отсюда имеем

$$4\delta R^2 = \frac{\varepsilon}{2}, \quad 2\delta R = \frac{\varepsilon}{4R}.$$

Потребуем выполнения неравенства $U(\mathbf{x}^*) - U(\mathbf{x}(\boldsymbol{\lambda}^N)) \le \varepsilon$. Тогда, в силу (18), (19) должно выполняться неравенство:

$$\left\|\nabla\varphi_\delta(\boldsymbol{\lambda}^N)\right\|_2 \le \frac{\varepsilon}{4R}.$$

Откуда имеем

$$2(L+\delta)\sqrt{\frac{\Delta}{\delta}}(\bar{\alpha})^{N/2} \le \frac{\varepsilon}{4R}.$$

Учитывая, что

$$2(L+\delta)\sqrt{\frac{\Delta}{\delta}} \le 2\left(LR + \frac{\varepsilon}{8R}\right)\sqrt{6 + \frac{16LR^2n + 8B}{n\varepsilon}},$$

получаем следующую оценку на число итераций:

$$N = \left\lceil 2\left(n + \sqrt{n^2 + \frac{128nLR^2}{\varepsilon}}\right)\ln\left(\frac{4RA}{\varepsilon}\right)\right\rceil,$$

где $A = 2\left(LR + \frac{\varepsilon}{8R}\right)\sqrt{6 + \frac{16LR^2n + 8B}{n\varepsilon}}$.

Замечание 3. Отметим, что оценку сложности алгоритма 6 можно также представить в виде $O\left(\max\left\{n, \sqrt{nLR^2/\varepsilon}\right\}\ln\left(\frac{1}{\varepsilon}\right)\right)$, где логарифмический множитель появляется за счет необходимости регуляризации двойственной задачи. При этом на каждой итерации вычисляется только одна компонента вектора реакции пользователя на изменение цены, соответственно, арифметическая сложность операции лучше, чем при вычислении всех компонент этих векторов. Для БГМ предположения для целевой функции аналогичны, но за счет того, что на каждой итерации необходимо вычислять полный градиент, стоимость работы алгоритма, соответственно, $O\left(n\sqrt{LR^2/\varepsilon}\right)$. Таким образом, несмотря на то, что теоретическая оценка скорости сходимости для RGEM имеет тот же порядок, что и для БГМ, на практике выигрыш получается за счет более дешевых вычислений в рамках одной итерации.

8. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Программный код для численных экспериментов был написан на языках программирования Python версии 3.6 и C++ стандарта C++14. Исходный код для экспериментов и рассмотренных в статье методов опубликован на GitHub и доступен по ссылке: https://github.com/dmivilensky/network-resource-allocation. Измерение времени работы производилось на компьютере с

Таблица 1. Сравнение количества итераций и времени работы БГМ и метода экстраполяции случайного градиента для сильно выпуклых (квадратичных) функций полезности

Сеть	FGM		RGEM	
	Итерации	Время	Итерации	Время
$m = 2$, $n = 1500$, $\varepsilon = 10^{-2}$	350	24.5 с	3000	21.1 с
$m = 5$, $n = 1500$, $\varepsilon = 10^{-2}$	380	42.7 с	6700	36.9 с
$m = 70$, $n = 5000$, $\varepsilon = 10^{-2}$	400	150.0 с	7800	132.6 с
$m = 70$, $n = 5000$, $\varepsilon = 10^{-3}$	1070	374.5 с	9180	283.7 с
$m = 100$, $n = 5000$, $\varepsilon = 10^{-2}$	417	175.1 с	8200	164.0 с
$m = 70$, $n = 7000$, $\varepsilon = 10^{-2}$	421	218.9 с	8600	206.4 с
$m = 100$, $n = 7000$, $\varepsilon = 10^{-2}$	427	290.3 с	9200	276.0 с
$m = 100$, $n = 7000$, $\varepsilon = 10^{-3}$	1120	761.6 с	10130	638.2 с

2-ядерным процессором Intel Core i5-5250U с тактовой частотой 1.6 ГГц на каждое ядро, ОЗУ компьютера составляла 8 Гб.

8.1. Сильно выпуклые (квадратичные) функции полезности

Рассмотрим задачу (1) для функций полезности следующего вида:

$$u_k(x_k) = a_k x_k - \frac{\sigma n}{2} x_k^2, \quad a_k \sim \mathcal{U}(0,100), \quad \sigma = 0.1,$$

где a_k – независимые и одинаково распределенные случайные величины. Тогда задачу (3) можно решить явно:

$$\mathbf{x}(\lambda) = \frac{[\mathbf{a} - C\boldsymbol{\lambda}]_+}{n\sigma}.$$

Для малого числа пользователей ($n = 1500$) пропускные способности соединений выбираются одинаковыми (в данном случае $\mathbf{b} = (5,\ldots,5)^{\mathrm{T}}$), а спрос на передачу информации равномерен ($c_{ij} = 1$ для любых i, j). Для большего числа пользователей вектор пропускных способностей генерируется случайно, так что $b_i \sim \mathcal{U}(1,6)$. Также случайно и независимо выбираются элементы матрицы спроса, так что $c_{ij} = 1$ с вероятностью $p = 0.5$ и $c_{ij} = 0$ с вероятностью $q = 0.5$.

В табл. 1 представлены число итераций и время работы быстрого градиентного метода (FGM) и метода экстраполяции случайного градиента (RGEM) для различных конфигураций сети (с числом соединений m), числа пользователей n и требуемой точности ε. В таблице выделены те случаи, в которых метод экстраполяции случайного градиента сходится к решению за меньшее время, чем быстрый градиентный метод, несмотря на большее число итераций по сравнению с БГМ. Действительно, для $n \gg 0$ число требующихся запросов оптимального решения $x_k(\boldsymbol{\lambda})$ от пользователей, в случае метода экстраполяции случайного градиента, будет меньшим, чем при использовании других алгоритмов, так как за одну итерацию метода экстраполяции случайного градиента запрос отправляется только к одному случайному пользователю.

8.2. Выпуклые (логарифмические) функции полезности

Рассмотрим работу стохастического субградиентного метода (алгоритм 9) и метода эллипсоидов (алгоритм 4) для функции полезности следующего вида:

$$u_k(x_k) = \ln x_k.$$

В таком случае явное решение задачи (3) выглядит следующим образом (операция $1/\cdot$ применительно к вектору понимается поэлементно):

$$\mathbf{x}(\boldsymbol{\lambda}) = \frac{1}{C\boldsymbol{\lambda}}.$$

Таблица 2. Сравнение количества итераций и времени работы стохастического субградиентного метода и метода эллипсоидов для выпуклых (логарифмических) функций полезности

Сеть	Метод эллипсоидов		SGM	
	Итерации	Время	Итерации	Время
$m = 2, n = 1500, \varepsilon = 10^{-2}$	40	0.02 с	2000	0.2 с
$m = 5, n = 1500, \varepsilon = 10^{-2}$	85	0.06 с	2500	0.3 с
$m = 70, n = 5000, \varepsilon = 10^{-2}$	120	1.9 с	4000	1.3 с
$m = 70, n = 5000, \varepsilon = 10^{-3}$	800	5.4 с	9020	2.4 с
$m = 100, n = 5000, \varepsilon = 10^{-2}$	300	9.0 с	5000	3.1 с
$m = 70, n = 7000, \varepsilon = 10^{-2}$	250	8.7 с	5590	5.5 с
$m = 100, n = 7000, \varepsilon = 10^{-2}$	380	19.0 с	6480	10.8 с
$m = 100, n = 7000, \varepsilon = 10^{-3}$	1830	91.5 с	17970	30.6 с

Для малого числа пользователей ($n = 1500$) пропускные способности соединений выбираются одинаковыми (в данном случае $\mathbf{b} = (5, \ldots, 5)^{\mathrm{T}}$), а спрос на передачу информации равномерен ($c_{ij} = 1$ для любых i, j). Для большего числа пользователей вектор пропускных способностей генерируется случайно, так что $b_i \sim \mathcal{U}(1, 6)$. Также случайно и независимо выбираются элементы матрицы спроса, так что $c_{ij} = 1$ с вероятностью $p = 0.5$ и $c_{ij} = 0$ с вероятностью $q = 0.5$.

В табл. 2 представлены число итераций и время работы стохастического субградиентного метода (SGM) и метода эллипсоидов для различных конфигураций сети, числа пользователей и требуемой точности. В таблице выделены те случаи, в которых стохастический субградиентный метод сходится к решению за меньшее время, чем метод эллипсоидов.

Отметим, что аналогично методу экстраполяции случайного градиента, стохастический субградиентный метод требует вычисления лишь одной компоненты вектора $\mathbf{x}(\boldsymbol{\lambda})$ реакций пользователей на установившиеся цены на каждой итерации. Таким образом, при большем числе итераций метода число вычисляемых компонент $x_k(\boldsymbol{\lambda}^t)$ будет меньшим по сравнению с другими алгоритмами, например, с методом эллипсоидов, также как и коммуникационная сложность в случае распределенной реализации.

9. ЗАКЛЮЧЕНИЕ

В заключение отметим некоторые возможные развития данной работы и кратко опишем методы без подробного анализа оценки скорости сходимости.

В разд. 5 мы рассматривали метод эллипсоидов для задач небольшой размерности, который является прямо двойственным. Существуют и другие методы, которые дают высокую точность и хорошо подходят для задач небольшой размерности. Одним из таких методов является метод Вайды [28]. Однако для восстановления решения прямой задачи при решении двойственной задачи методом Вайды необходимо, чтобы была сходимость по норме градиента для двойственной задачи. Для этого требуется гладкость двойственной задачи, что порождается сильной выпуклостью целевой функции прямой задачи (утверждение 2). Если сильной выпуклости в прямой задаче нет, то ее можно регуляризовать, как было описано в разд. 6, но при этом оценка сходимости логарифмически ухудшится.

Также для решения двойственной задачи можно применить методы высокого порядка [29], [30], если двойственная задача достаточно гладкая. При этом шаги данных методов можно будет считать распределенно, так как в данной задаче рассматривается централизованная архитектура с точки зрения взаимодействия соединения и использующих его пользователей. Однако отметим, что оптимальные методы высокого порядка, требующие одномерного поиска и не обладающие прямо двойственностью, необходимо применять только после регуляризации двойственной задачи.

Еще одно направление – методы типа редукции дисперсии, например, [31], [32], которые являются промежуточными между стохастическим градиентным методом и БГМ. Однако данные

методы тоже не прямо двойственные, поэтому их необходимо применять к предварительно регуляризованной двойственной задаче.

Особый интерес для авторов представляют метод Hogwild! [33] и методы с использованием мини-батчинга. Заметим, что при этом одновременно данные передают не все пользователи, но и не только один пользователь, как это получается при использовании стохастических методов. Выбирая в качестве размера батча количество пользователей, которые отправляют данные в один момент времени, можно учесть специфику работы реальных сетей.

ПРИЛОЖЕНИЕ

Вспомогательные результаты

Приведем некоторые леммы из других работ, на которые мы будем ссылаться в доказательствах. Также приведем доказательства утверждений о свойствах двойственной функции, которые используются при доказательстве основных теорем.

Лемма 6 (см. [34, лемма 2]). *Для случайного вектора $\xi \in \mathbf{R}^n$ следующие утверждения эквивалентны с точностью до константного множителя у σ.*

1. *Хвосты*: $\mathbb{P}\{\|\xi\|_2 \geq \gamma\} \leq 2\exp\left(-\dfrac{\gamma^2}{2\sigma^2}\right)\ \forall \gamma \geq 0$.

2. *Моменты*: $(\mathbb{E}[\xi^p])^{1/p} \leq \sigma\sqrt{p}$ *для любого положительного целого* p.

3. *Предположение легких хвостов*: $\mathbb{E}\left[\exp\left(\dfrac{\|\xi\|_2^2}{\sigma^2}\right)\right] \leq \exp(1)$.

Лемма 7 (см. [34, следствие 8]). *Пусть $\{\xi^k\}_{k=1}^N$ – последовательность случайных векторов из $\mathbf{R}^n$ таких, что для $k = 1,\ldots,N$ и для любых $\gamma \geq 0$*

$$\mathbb{E}[\xi^k \,|\, \xi^1,\ldots,\xi^{k-1}] = 0, \quad \mathbb{E}\left[\|\xi^k\|_2 \geq \gamma \,|\, \xi^1,\ldots,\xi^{k-1}\right] \leq \exp\left(-\frac{\gamma^2}{2\sigma_k^2}\right) \quad \textit{почти наверное,}$$

где σ_k^2 принадлежит $\sigma(\xi^1,\ldots,\xi^{k-1})$ для всех $k = 1,\ldots,N$. Пусть $S_N = \sum_{k=1}^N \xi^k$. Тогда существует константа C_1 такая, что для любых фиксированных $\delta > 0$ и $B > b > 0$ с вероятностью $1 - \delta$ выполняется:

$$\textit{либо } \sum_{k=1}^N \sigma_k^2 \geq B,$$

$$\textit{либо } \quad \|S_N\|_2 \leq C_1\sqrt{\max\left\{\sum_{k=1}^N \sigma_k^2, b\right\}\left(\ln\frac{2n}{\delta} + \ln\ln\frac{B}{b}\right)}.$$

Лемма 8 (см. [35, следствие из теор. 2.1, случай (ii)]). *Пусть $\{\xi^k\}_{k=1}^N$ – последовательность случайных векторов из $\mathbf{R}^n$ удовлетворяет условию*

$$\mathbb{E}[\xi^k \,|\, \xi^1,\ldots,\xi^{k-1}] = 0 \quad \textit{почти наверное,} \quad k = 1,\ldots,N$$

и пусть $S_N = \sum_{k=1}^N \xi^k$. Пусть последовательность $\{\xi^k\}_{k=1}^N$ удовлетворяет "light-tail"-условию:

$$\mathbb{E}\left[\exp\left(\frac{\|\xi^k\|_2^2}{\sigma_k^2}\right) \,|\, \xi^1,\ldots,\xi^{k-1}\right] \leq \exp(1) \quad \textit{почти наверное,} \quad k = 1,\ldots,N,$$

где $\sigma_1,\ldots,\sigma_N$ – положительные числа. Тогда для всех $\gamma \geq 0$ выполняется

$$\mathbb{P}\left\{\|S_N\| \geq (\sqrt{2} + \sqrt{2\gamma})\sqrt{\sum_{k=1}^N \sigma_k^2}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Доказательство утверждения 2. Представим двойственную функцию в следующем виде:

$$\varphi(\boldsymbol{\lambda}) = \sum_{k=1}^{n}\left\{u_k(x_k(\boldsymbol{\lambda})) - \langle \boldsymbol{\lambda}, \mathbf{C}_k\rangle x_k(\boldsymbol{\lambda}) + \frac{1}{n}\langle \boldsymbol{\lambda}, \mathbf{b}\rangle\right\} = \sum_{k=1}^{n}\varphi_k(\boldsymbol{\lambda}),$$

при этом из утверждения 1 получаем, что

$$\nabla\varphi(\boldsymbol{\lambda}) = \sum_{k=1}^{n}\nabla\varphi_k(\boldsymbol{\lambda}) = \sum_{k=1}^{n}\left(\frac{1}{n}\mathbf{b} - \mathbf{C}_k x_k(\boldsymbol{\lambda})\right).$$

Определим

$$x_k(\boldsymbol{\lambda}^1) = \arg\max_{x_k\in\mathbf{R}_+}\left\{u_k(x_k) - x_k\left\langle \boldsymbol{\lambda}^1, \mathbf{C}_k\right\rangle\right\},$$
$$x_k(\boldsymbol{\lambda}^2) = \arg\max_{x_k\in\mathbf{R}_+}\left\{u_k(x_k) - x_k\left\langle \boldsymbol{\lambda}^2, \mathbf{C}_k\right\rangle\right\}.$$

Запишем необходимые условия максимума первого порядка

$$\left\langle \nabla u_k(x_k(\boldsymbol{\lambda}^1)) - \left\langle \boldsymbol{\lambda}^1, \mathbf{C}_k\right\rangle, x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\rangle \geq 0,$$
$$\left\langle \nabla u_k(x_k(\boldsymbol{\lambda}^2)) - \left\langle \boldsymbol{\lambda}^2, \mathbf{C}_k\right\rangle, x_k(\boldsymbol{\lambda}^2) - x_k(\boldsymbol{\lambda}^1)\right\rangle \geq 0.$$

Складывая эти неравенства, получаем

$$\left\langle \nabla u_k(x_k(\boldsymbol{\lambda}^2)) - \nabla u_k(x_k(\boldsymbol{\lambda}^1)), x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\rangle \leq \left\langle \left\langle \boldsymbol{\lambda}^2, \mathbf{C}_k\right\rangle - \left\langle \boldsymbol{\lambda}^1, \mathbf{C}_k\right\rangle, x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\rangle.$$

В силу сильной вогнутости $u_k(x_k)$ для любых x_k^1 и x_k^2, $k = 1,\ldots,n$, выполняется

$$\left\langle \nabla u_k(x_k^2) - \nabla u_k(x_k^1), x_k^1 - x_k^2\right\rangle \geq \mu\left\|x_k^1 - x_k^2\right\|_2^2.$$

Отсюда получаем, что

$$\mu\left\|x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\|_2^2 \leq \left\langle \left\langle \boldsymbol{\lambda}^2, \mathbf{C}_k\right\rangle - \left\langle \boldsymbol{\lambda}^1, \mathbf{C}_k\right\rangle, x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\rangle \leq \left\|\mathbf{C}_k\right\|_2\cdot\left\|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\right\|_2\cdot\left\|x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\|_2.$$

Тогда можно получить следующую оценку для всех $\nabla\varphi_k$ градиента

$$\left\|\nabla\varphi_k(\boldsymbol{\lambda}^1) - \nabla\varphi_k(\boldsymbol{\lambda}^2)\right\|_2 \leq \left\|\mathbf{C}_k\right\|_2\cdot\left\|x_k(\boldsymbol{\lambda}^1) - x_k(\boldsymbol{\lambda}^2)\right\|_2 \leq \frac{1}{\mu}\left\|\mathbf{C}_k\right\|_2^2\cdot\left\|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\right\|_2.$$

Для матрицы C с учетом ее структуры верна оценка $\left\|\mathbf{C}_k\right\|_2 \leq m$. Тогда для градиента двойственной функции

$$\left\|\nabla\varphi(\boldsymbol{\lambda}^1) - \nabla\varphi(\boldsymbol{\lambda}^2)\right\|_2 \leq \sum_{k=1}^{n}\left\|\nabla\varphi_k(\boldsymbol{\lambda}^1) - \nabla\varphi_k(\boldsymbol{\lambda}^2)\right\|_2 \leq \frac{m^2 n}{\mu}\left\|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\right\|_2.$$

Доказательство леммы 1. Для доказательства леммы 1 сначала сформулируем и докажем одну техническую лемму.

Обозначим $d_L(\boldsymbol{\lambda}) = \frac{L}{2}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_2^2$ и рассмотрим последовательности

$$l_t(\boldsymbol{\lambda}) = \sum_{j=0}^{t}\alpha_j\left[\varphi(\boldsymbol{\lambda}^j) + \left\langle \nabla\varphi(\boldsymbol{\lambda}^j), \boldsymbol{\lambda} - \boldsymbol{\lambda}^j\right\rangle\right]$$

и

$$\psi_t(\boldsymbol{\lambda}) = l_t(\boldsymbol{\lambda}) + d_L(\boldsymbol{\lambda}), \quad t = 0,1,\ldots,$$

где $\{\boldsymbol{\lambda}^j\}_{j\geq 0}$ – последовательность точек, генерируемых алгоритмом 1.

Лемма 9. *После N шагов алгоритма* 1 *выполняется следующее неравенство*:

$$A_N\varphi(\mathbf{y}^N) \leq \min_{\boldsymbol{\lambda}\in\mathbf{R}_+^m}\psi_N(\boldsymbol{\lambda}) = \psi_N(\mathbf{z}^N). \tag{П.20}$$

Доказательство леммы 9. Докажем по индукции, что (П.20) верно. При $t = 0$ неравенство (П.20) выполняется. Действительно,

$$\psi_0 = \min_{\lambda\in\mathbf{R}_+^m}\left\{\alpha_0\left[\varphi(\lambda^0) + \left\langle\nabla\varphi(\lambda^0), \lambda - \lambda^0\right\rangle\right] + \frac{L}{2}\left\|\lambda - \lambda^0\right\|_2^2\right\} \overset{①}{\geq}$$
$$\overset{①}{\geq} \alpha_0 \min_{\lambda\in\mathbf{R}_+^m}\left\{\varphi(\lambda^0) + \left\langle\nabla\varphi(\lambda^0), \lambda - \lambda^0\right\rangle + \frac{L}{2}\left\|\lambda - \lambda^0\right\|_2^2\right\} \overset{②}{\geq} \alpha_0\varphi(y_0),$$

где ① выполняется, так как $\alpha_0 = 1/2 \leq 1$, а ② – в силу того, что функция $\varphi(\lambda)$ имеет липшицев градиент (см. утверждение 2 и [36, лемма 1.2.3]). Итак, $A_0\varphi(\mathbf{y}^0) = \frac{1}{2}\varphi(\mathbf{y}^0) \leq \psi_0$.

Пусть (П.20) верно при t:

$$A_t\varphi(\mathbf{y}^t) \leq \psi_t(\mathbf{z}^t). \tag{П.21}$$

Докажем, что (П.20) верно при $t+1$. Действительно, имеем,

$$\psi_{t+1}(\mathbf{z}^{t+1}) = \min_{\lambda\in\mathbf{R}_+^m}\left\{\psi_t(\lambda) + \alpha_{t+1}\left[\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle\right]\right\} \overset{①}{\geq}$$
$$\overset{①}{\geq} \min_{\lambda\in\mathbf{R}_+^m}\left\{\psi_t(\mathbf{z}^t) + \frac{L}{2}\left\|\lambda - \mathbf{z}^t\right\|_2^2 + \alpha_{t+1}\left[\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle\right]\right\} \overset{②}{\geq}$$
$$\overset{②}{\geq} \min_{\lambda\in\mathbf{R}_+^m}\left\{A_t\varphi(\mathbf{y}^t) + \frac{L}{2}\left\|\lambda - \mathbf{z}^t\right\|_2^2 + \alpha_{t+1}\left[\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle\right]\right\} \overset{③}{\geq}$$
$$\overset{③}{\geq} \min_{\lambda\in\mathbf{R}_+^m}\left\{A_t\left(\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle\right) + \frac{L}{2}\left\|\lambda - \mathbf{z}^t\right\|_2^2 + \alpha_{t+1}\left[\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle\right]\right\},$$

где ① выполняется в силу сильной выпуклости прокс-функции $\frac{1}{2}\left\|\lambda - \lambda^0\right\|_2^2$ и свойств экстремума в точке $\mathbf{z}^t$, ② следует из (П.21), ③ – в силу выпуклости функции $\varphi(\lambda)$.

Так как между коэффициентами A_t и α_t БГМ есть следующая зависимость: $A_{t+1} = \sum_{j=0}^{t+1}\alpha_j = A_t + \alpha_{t+1}$ и $\tau_t = \alpha_{t+1}/A_{t+1}$, соотношение $\lambda^{t+1} = \tau_t\mathbf{z}^t + (1-\tau_t)\mathbf{y}^t$ из алгоритма 1 можно переписать в виде:

$$A_{t+1}\lambda^{t+1} = \alpha_{t+1}\mathbf{z}^t + A_t\mathbf{y}^t.$$

Используя последние соотношения, можно сделать следующие преобразования:

$$A_t\left\langle\nabla\varphi(\lambda^{t+1}), \mathbf{y}^t - \lambda^{t+1}\right\rangle + \alpha_{t+1}\left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle = -A_{t+1}\left\langle\nabla\varphi(\lambda^{t+1}), \lambda^{t+1}\right\rangle +$$
$$+ \alpha_{t+1}\left\langle\nabla\varphi(\lambda^{t+1}), \lambda\right\rangle + A_t\left\langle\nabla\varphi(\lambda^{t+1}), \mathbf{y}^t\right\rangle = \alpha_{t+1}\left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \mathbf{z}^t\right\rangle.$$

Тогда имеем

$$A_t\left(\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \mathbf{y}^t - \lambda^{t+1}\right\rangle\right) + \frac{L}{2}\left\|\lambda - \mathbf{z}^t\right\|_2^2 + \alpha_{t+1}\left[\varphi(\lambda^{t+1}) + \left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \lambda^{t+1}\right\rangle\right] =$$
$$= A_{t+1}\varphi(\lambda^{t+1}) + \frac{L}{2}\left\|\lambda - \mathbf{z}^t\right\|_2^2 + \alpha_{t+1}\left\langle\nabla\varphi(\lambda^{t+1}), \lambda - \mathbf{z}^t\right\rangle. \tag{П.23}$$

После замены последнего выражения в (П.22) на (П.23) можно воспользоваться расширенным вариантом неравенства Фенхеля для сопряженных функций [37]:

$$\langle\mathbf{g}, \mathbf{s}\rangle + \frac{\xi}{2}\|\mathbf{s}\|^2 \geq -\frac{1}{2\xi}\|\mathbf{g}\|_*^2, \quad \mathbf{g}\in\mathbb{E}^*, \quad \mathbf{s}\in\mathbb{E},$$

где $\mathbb{E}$ – конечномерное вещественное векторное пространство, $\mathbb{E}^*$ – пространство линейных функций на $\mathbb{E}$ (двойственное пространство), норма в двойственном пространстве $\|\mathbf{g}\|_* = \max\limits_{\mathbf{x}}\{\langle \mathbf{g}, \mathbf{x}\rangle \,|\, \|\mathbf{x}\|_E = 1\}$. В нашем случае $\mathbf{g} = \nabla\varphi(\lambda^{t+1})$, $\mathbf{s} = \lambda - \mathbf{z}^t$, $\xi = \dfrac{L}{\alpha_{t+1}}$. Следовательно,

$$\psi_{t+1}(\mathbf{z}^{t+1}) \geq A_{t+1}\varphi(\boldsymbol{\lambda}^{t+1}) - \frac{\alpha_{t+1}^2}{2L}\left\|\nabla\varphi(\boldsymbol{\lambda}^{t+1})\right\|_2^2. \tag{П.24}$$

Для завершения доказательства леммы требуется показать, что $A_{t+1}\varphi(\mathbf{y}^{t+1})$ меньше, чем правая часть неравенства в (П.24).

В силу L-гладкости функции $\varphi(\boldsymbol{\lambda})$ (см. утверждение 2)

$$\varphi(\mathbf{y}^{t+1}) \leq \varphi(\boldsymbol{\lambda}^{t+1}) + \left\langle \nabla\varphi(\boldsymbol{\lambda}^{t+1}), \mathbf{y}^{t+1} - \boldsymbol{\lambda}^{t+1}\right\rangle + \frac{L}{2}\left\|\mathbf{y}^{t+1} - \boldsymbol{\lambda}^{t+1}\right\|_2^2 =$$
$$= \min_{\boldsymbol{\lambda}}\left\{\varphi(\boldsymbol{\lambda}^{t+1}) + \left\langle \nabla\varphi(\boldsymbol{\lambda}^{t+1}), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1}\right\rangle + \frac{L}{2}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1}\right\|_2^2\right\} = \varphi(\boldsymbol{\lambda}^{t+1}) - \frac{1}{2L}\left\|\nabla\varphi(\boldsymbol{\lambda}^{t+1})\right\|_2^2.$$

После умножения обеих частей полученного неравенства на A_{t+1}:

$$A_{t+1}\varphi(\mathbf{y}^{t+1}) \leq A_{t+1}\varphi(\boldsymbol{\lambda}^{t+1}) - \frac{A_{t+1}}{2L}\left\|\nabla\varphi(\boldsymbol{\lambda}^{t+1})\right\|_2^2.$$

В силу того, что для коэффициентов БГМ $\alpha_{t+1}^2 \leq A_{t+1}$, получаем

$$A_{t+1}\varphi(\mathbf{y}^{t+1}) \leq A_{t+1}\varphi(\boldsymbol{\lambda}^{t+1}) - \frac{\alpha_{t+1}^2}{2L}\left\|\nabla\varphi(\boldsymbol{\lambda}^{t+1})\right\|_2^2. \tag{П.25}$$

Следовательно, в силу (П.24) и (П.25) $A_{t+1}\varphi(\mathbf{y}^{t+1}) \leq \psi_{t+1}(\mathbf{z}^{t+1})$, что и требовалось доказать.

Доказательство леммы 1. Определим следующее множество:

$$\Lambda_{2\hat{R}} = \left\{\boldsymbol{\lambda} \in \mathbb{R}_+^m : \|\boldsymbol{\lambda}\|_2 \leq 2\hat{R}\right\}.$$

$\hat{R}$ определяется в силу следующих неравенств:

$$\left\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\right\|_2 + \left\|\boldsymbol{\lambda}^0\right\|_2 \leq \left\|\boldsymbol{\lambda}^*\right\|_2 + 2\left\|\boldsymbol{\lambda}^0\right\|_2 \leq 3R = \hat{R}.$$

При этом все $\boldsymbol{\lambda}^t$ будут принадлежать $\Lambda_{2\hat{R}}$, так как

$$\left\|\boldsymbol{\lambda}^t\right\|_2 \leq \left\|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^*\right\|_2 + \left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2 + \left\|\boldsymbol{\lambda}^0\right\|_2 \leq 2\left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2 + \left\|\boldsymbol{\lambda}^0\right\|_2 \leq 2\left\|\boldsymbol{\lambda}^*\right\|_2 + 3\left\|\boldsymbol{\lambda}^0\right\|_2 \leq 5R \leq 2\hat{R},$$

где для второго неравенства учитывалось, что $\left\|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^*\right\|_2 \leq \left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2$, $t = 0, 1, \ldots$.

Последнее неравенство можно доказать следующим образом. Для любого $\boldsymbol{\lambda} \in \mathbf{R}_+^m$ в силу леммы 9 и сильной выпуклости функции $\psi_t(\boldsymbol{\lambda})$ с константой L верно

$$A_t\varphi(\mathbf{y}^t) + \frac{L}{2}\left\|\boldsymbol{\lambda} - \mathbf{z}^t\right\|_2^2 \leq \psi_t(\mathbf{z}^t) + \frac{L}{2}\left\|\boldsymbol{\lambda} - \mathbf{z}^t\right\|_2^2 \leq \psi_t(\boldsymbol{\lambda}) =$$
$$= \sum_{j=0}^{t}\alpha_j\left[\varphi(\boldsymbol{\lambda}^j) + \left\langle \nabla\varphi(\boldsymbol{\lambda}^j), \boldsymbol{\lambda} - \boldsymbol{\lambda}^j\right\rangle\right] + \frac{L}{2}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_2^2. \tag{П.26}$$

Последнее выражение в (П.26) в силу выпуклости функции $\varphi(\boldsymbol{\lambda})$ можно оценить сверху как $A_t\varphi(\boldsymbol{\lambda}) + \dfrac{L}{2}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_2^2$. Тогда при $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$

$$\frac{L}{2}\left\|\boldsymbol{\lambda}^* - \mathbf{z}^t\right\|_2^2 \leq A_t(\varphi(\mathbf{y}^t) - \varphi(\boldsymbol{\lambda}^*)) + \frac{L}{2}\left\|\boldsymbol{\lambda}^* - \mathbf{z}^t\right\|_2^2 \leq \frac{L}{2}\left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2^2.$$

Следовательно,

$$\left\|\boldsymbol{\lambda}^* - \mathbf{z}^t\right\|_2 \leq \left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2. \tag{П.27}$$

Поскольку $\mathbf{y}^t$ в алгоритме 1 определяется с помощью шага метода проекции градиента для выпуклой функции $\varphi(\boldsymbol{\lambda})$, последовательность генерируемых алгоритмом точек $\mathbf{y}^t$, $t = 0,1,\ldots$ также будет ограничена (доказательство этого факта см., например, в [38, лемма 9.17, с. 183] или в [28, с. 265]):

$$\left\|\boldsymbol{\lambda}^* - \mathbf{y}^t\right\|_2 \le \left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2. \tag{П.28}$$

Далее

$$\left\|\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^*\right\|_2 = \left\|\tau_t(\mathbf{z}^t - \boldsymbol{\lambda}^*) + (1-\tau_t)(\mathbf{y}^t - \boldsymbol{\lambda}^*)\right\|_2 \le \tau_t\left\|\mathbf{z}^t - \boldsymbol{\lambda}^*\right\|_2 + (1-\tau_t)\left\|\mathbf{y}^t - \boldsymbol{\lambda}^*\right\|_2.$$

Из последнего неравенства с помощью (П.27), (П.28) получаем нужный результат:

$$\left\|\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^*\right\|_2 \le \left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\right\|_2, \quad t = -1, 0, 1, \ldots$$

В силу леммы 9

$$\begin{aligned} A_N\varphi(\mathbf{y}^N) &\le \min_{\boldsymbol{\lambda}\in\mathbf{R}^m_+}\left\{\frac{L}{2}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_2^2 + \sum_{t=0}^N \alpha_t\left[\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right]\right\} \le \\ &\le \min_{\boldsymbol{\lambda}\in\Lambda_{2\hat{R}}}\left\{\frac{L}{2}\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|_2^2 + \sum_{t=0}^N \alpha_t\left[\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right]\right\} \le \\ &\overset{①}{\le} \min_{\boldsymbol{\lambda}\in\Lambda_{2\hat{R}}}\left\{\sum_{t=0}^N \alpha_t\left[\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right]\right\} + \frac{37L\hat{R}^2}{9}, \end{aligned}$$

где ① выполняется, так как

$$\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\right\|^2 \le 2\|\boldsymbol{\lambda}\|^2 + 2\left\|\boldsymbol{\lambda}^0\right\|^2 \le 8\hat{R}^2 + \frac{2}{9}\hat{R}^2 = \frac{74}{9}\hat{R}^2. \tag{П.29}$$

После подстановки определений двойственной целевой функции $\varphi(\boldsymbol{\lambda}^t)$ (2) и ее градиента $\nabla\varphi(\boldsymbol{\lambda}^t)$ (см. утверждение 1):

$$\begin{aligned} &\sum_{t=0}^N \alpha_t\left[\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right] = \\ &= \sum_{t=0}^N \alpha_t\left(\left\langle\boldsymbol{\lambda}^t, \mathbf{b}\right\rangle + \sum_{k=1}^n\left(u_k(x_k^t(\boldsymbol{\lambda}^t)) - \left\langle\boldsymbol{\lambda}^t, \mathbf{C}_k x_k^t(\boldsymbol{\lambda}^t)\right\rangle\right) + \left\langle\mathbf{b} - \sum_{k=1}^n \mathbf{C}_k x_k^t(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right) = \\ &= \sum_{t=0}^N \alpha_t\left(\sum_{k=1}^n u_k(x_k^t(\boldsymbol{\lambda}^t)) + \left\langle\boldsymbol{\lambda}, \mathbf{b} - \sum_{k=1}^n \mathbf{C}_k x_k^t(\boldsymbol{\lambda}^t)\right\rangle\right) \le A_N\left(U(\hat{\mathbf{x}}^N) + \left\langle\boldsymbol{\lambda}, \mathbf{b} - C\hat{\mathbf{x}}^N\right\rangle\right), \end{aligned}$$

где последнее неравенство выполняется в силу вогнутости функций полезности.

Итак,

$$\begin{aligned} A_N\varphi(\mathbf{y}^N) &\le A_N U(\hat{\mathbf{x}}^N) + \frac{37L\hat{R}^2}{9} + A_N\min_{\boldsymbol{\lambda}\in\Lambda_{2\hat{R}}}\left\{\left\langle\boldsymbol{\lambda}, \mathbf{b} - C\hat{\mathbf{x}}^N\right\rangle\right\} = A_N U(\hat{\mathbf{x}}^N) + \frac{37L\hat{R}^2}{9} - \\ &- A_N\max_{\boldsymbol{\lambda}\in\Lambda_{2\hat{R}}}\left\{\left\langle\boldsymbol{\lambda}, C\hat{\mathbf{x}}^N - \mathbf{b}\right\rangle\right\} = A_N U(\hat{\mathbf{x}}^N) + \frac{37L\hat{R}^2}{9} - 2\hat{R}A_N\left\|\left(C\hat{\mathbf{x}}^N - \mathbf{b}\right)_+\right\|_2. \end{aligned}$$

Из этого получаем оценку (6).

Доказательство леммы 2. Для доказательства леммы 2 сначала приведем доказательства нескольких вспомогательных технических лемм.

Лемма 10. *Пусть A, B и $\{r_t\}_{t=0}^N$ – неотрицательные числа такие, что для любого $l = 1,\ldots,N$ выполняется неравенство*

$$\frac{1}{2}r_l^2 \le Ar_0^2 + Br_0\sqrt{\sum_{t=0}^{l-1} r_t^2}. \tag{П.30}$$

Тогда верно следующее неравенство:

$$r_l \le Cr_0, \tag{П.31}$$

где C – положительная константа, для которой выполняется $C^2 \ge \max\{1, 2A + 2BC\sqrt{N}\}$, т.е., в частности, можно выбрать

$$C = \max\left\{1, B\sqrt{N} + \sqrt{B^2N + 2A}\right\}.$$

Доказательство. Докажем (П.31) по индукции. Для $l = 0$ неравенство выполнено, так как $C \ge 1$. Пусть (П.31) выполнено для всех $l < N$. Докажем, что оно выполнено и для $l + 1$. Действительно,

$$r_{l+1} \overset{(\text{П.30})}{\le} \sqrt{2}\sqrt{Ar_0^2 + Br_0\sqrt{\sum_{t=0}^{l} r_t^2}} \overset{(\text{П.31})}{\le} r_0\sqrt{2}\sqrt{A + BC\sqrt{N}} = r_0\underbrace{\sqrt{2A + 2BC\sqrt{N}}}_{\le C} \le Cr_0.$$

Лемма 11. *Пусть для последовательностей неотрицательных коэффициентов $\{R_t\}_{t\ge 0}$ и случайных векторов $\{\boldsymbol{\eta}^t\}_{t\ge 0}$, $\{\mathbf{a}^t\}_{t\ge 0}$ для всех $l = 1, \ldots, N$ выполняется неравенство*

$$\frac{1}{2}R_l^2 \le A + u\sum_{t=0}^{l-1}\left\langle \boldsymbol{\eta}^t, \mathbf{a}^t\right\rangle, \tag{П.32}$$

где A – неотрицательная константа, $d \ge 1$ – положительная константа, $\left\|\mathbf{a}^t\right\|_2 \le \widetilde{R}_t d$ и $\widetilde{R}_t = \max\{\widetilde{R}_{t-1}, R_t\}$ для всех $t \ge 1$, $\widetilde{R}_0 = R_0$, $\widetilde{R}_t$ зависит только от $\boldsymbol{\eta}^0, \ldots, \boldsymbol{\eta}^t$. Пусть также вектор $\mathbf{a}^t$ – это функция от $\boldsymbol{\eta}^0, \ldots, \boldsymbol{\eta}^{t-1}$ $\forall t \ge 1$, a^0 – постоянный вектор и для любого $t \ge 0$

$$\mathbb{E}[\eta^t \,|\, \{\boldsymbol{\eta}^k\}_{k=0}^{t-1}] = 0, \quad \mathbb{E}\left[\exp\left(\left\|\boldsymbol{\eta}^t\right\|_2^2\sigma^{-2}\right)|\,\{\boldsymbol{\eta}^k\}_{k=0}^{t-1}\right] \le \exp(1).$$

Тогда с вероятностью $1 - 2\delta$ выполняются следующие неравенства:

$$\widetilde{R}_l \le JR_0 \quad \text{и} \quad A + u\sum_{t=0}^{l-1}\left\langle \boldsymbol{\eta}^t, \mathbf{a}^t\right\rangle \le A + udD\sqrt{\sigma^2 g(N)NJ}\widetilde{R}_0^2$$

$\forall l = 1, \ldots, N$ одновременно, где D – положительная константа,

$$F = 2\sigma^2 d^2 N(2ud)^N\left(2A + ud\widetilde{R}_0^2 + 12ud\ln\frac{N}{\delta}\sigma^2 N\right),$$

$f = d^2\sigma^2\widetilde{R}_0^2$, $g(N) = \ln\left(\frac{N}{\delta}\right) + \ln\ln\left(\frac{F}{f}\right)$ u

$$J = \max\left\{1, \frac{1}{\widetilde{R}_0}udD\sqrt{\sigma^2 g(N)} + \sqrt{\frac{1}{\widetilde{R}_0^2}u^2d^2C_1^2\sigma^2 g(N) + \frac{2A}{R_0^2}}\right\}.$$

Доказательство. Применим ко второму слагаемому из правой части (П.32) неравенство Коши–Буняковского:

$$\frac{1}{2}R_l^2 \le A + ud\sum_{t=0}^{l-1}\left\|\boldsymbol{\eta}^t\right\|_2\widetilde{R}_t \le A + \frac{ud}{2}\sum_{t=0}^{l-1}\widetilde{R}_t^2 + \frac{ud}{2}\sum_{t=0}^{l-1}\left\|\boldsymbol{\eta}^t\right\|_2^2. \tag{П.33}$$

По теореме 2.1 из [35]

$$(\forall N \ge 1,\ \forall\gamma \ge 0): \quad \mathbb{P}\left\{\left\|\sum_{t=0}^{N-1}\boldsymbol{\eta}^t\right\|_2 \ge (\sqrt{2} + \sqrt{2}\gamma)\sqrt{\sum_{t=0}^{N-1}\sigma_t^2}\right\} \le \exp\left(-\frac{\gamma^2}{3}\right). \tag{П.34}$$

Тогда с вероятностью не меньшей, чем

$$1 - \frac{\delta}{N} = 1 - \exp\left(-\frac{\gamma^2}{3}\right) \tag{П.35}$$

выполняется следующее неравенство:

$$\left\|\boldsymbol{\eta}^t\right\|_2 \le \sqrt{2}\left(1+\sqrt{3\ln\frac{N}{\delta}}\right)\sigma \le 2\sqrt{6\ln\frac{N}{\delta}}\sigma. \tag{П.36}$$

Действительно, выражая из (П.35) γ, получаем, что $\gamma = \sqrt{3\ln\frac{N}{\delta}}$. После подстановки данного выражения в (П.34) и выбора единого $\sigma \in \mathbf{R}_+$ вместо последовательности σ_t, $t = 0,\ldots,N-1$, получаем оценку (П.36).

Объединяя полученные неравенства, получаем, что с вероятностью, большей или равной $1-\delta$, неравенство

$$\frac{1}{2}R_l^2 \le A + \frac{ud}{2}\sum_{t=0}^{l-1}\widetilde{R}_t^2 + 12ud\ln\frac{N}{\delta}\sigma^2 l$$

выполняется для всех $l = 1,\ldots,N$ одновременно. Заметим, что последнее слагаемое в полученной оценке – неубывающая функция от l. Определим $\hat{l}$ как наибольшее целое число, для которого выполнено $\hat{l} \le l$ и $\widetilde{R}_{tl} = R_{\hat{l}}$. Тогда получаем, что $R_{\hat{l}} = \widetilde{R}_{\hat{l}} = \widetilde{R}_{\hat{l}+1} = \ldots = \widetilde{R}_l$, и, следовательно, с вероятностью $\ge 1-\delta$ имеем

$$\frac{1}{2}\widetilde{R}_l^2 \le A + \frac{ud}{2}\sum_{t=0}^{\hat{l}-1}\widetilde{R}_t^2 + 12ud\ln\frac{N}{\delta}\sigma^2\hat{l} \le A + \frac{ud}{2}\sum_{t=0}^{l-1}\widetilde{R}_t^2 + 12ud\ln\frac{N}{\delta}\sigma^2 l \quad \forall l = 1,\ldots,N.$$

Получаем, что с вероятностью $\ge 1-\delta$ верна следующая оценка:

$$\widetilde{R}_l^2 \le 2A + ud\sum_{t=0}^{l-1}\widetilde{R}_k^2 + 24ud\ln\frac{N}{\delta}\sigma^2 l \le 2A\underbrace{(1+ud)}_{\le 2ud} + \underbrace{(ud+u^2d^2)}_{\le 2u^2d^2}\sum_{t=0}^{l-2}\widetilde{R}_t^2 +$$

$$+ 24ud\ln\frac{N}{\delta}\sigma^2\underbrace{(l+ud(l-1))}_{\le 2udl} \le 2ud\left(2A + ud\sum_{t=0}^{l-2}\widetilde{R}_t^2 + 24ud\ln\frac{N}{\delta}\sigma^2 l\right) \quad \forall l = 1,\ldots,N.$$

Применяя данную оценку рекурсивно, получаем, что с вероятностью $\ge 1-\delta$ верно

$$\widetilde{R}_\partial^2 \le (2ud)^l\left(2A + ud\widetilde{R}_0^2 + 24ud\ln\frac{N}{\delta}\sigma^2 l\right).$$

Далее рассмотрим последовательность случайных величин $\xi^t = \left\langle \boldsymbol{\eta}^t, \mathbf{a}^t\right\rangle$. Заметим, что $\mathbb{E}[\xi^t\,|\,\xi^0,\ldots,\xi^{t-1}] = \left\langle \mathbb{E}[\boldsymbol{\eta}^t\,|\,\boldsymbol{\eta}^0,\ldots,\boldsymbol{\eta}^{k-1}], \mathbf{a}^t\right\rangle = 0$, тогда, используя неравенство Коши–Буняковского, получаем,

$$\mathbb{E}\left[\exp\left(\frac{(\xi^t)^2}{\sigma^2 d^2\widetilde{R}_t^2}\right)|\,\xi^0,\ldots,\xi^{t-1}\right] \le \mathbb{E}\left[\exp\left(\frac{\left\|\boldsymbol{\eta}^t\right\|_2^2 d^2\widetilde{R}_t^2}{\sigma^2 d^2\widetilde{R}_t^2}\right)|\,\boldsymbol{\eta}^0,\ldots,\boldsymbol{\eta}^{t-1}\right] =$$

$$= \mathbb{E}\left[\exp\left(\frac{\left\|\boldsymbol{\eta}^t\right\|_2^2}{\sigma^2}\right)|\,\boldsymbol{\eta}^0,\ldots,\boldsymbol{\eta}^{t-1}\right] \le \exp(1).$$

Определим $\hat{\sigma}_t^2 = \sigma^2 d^2\widetilde{R}_t^2$, тогда с вероятностью $\ge 1-\delta$ выполняется

$$\sum_{t=0}^{l-1}\hat{\sigma}_t^2 \le \sigma^2 d^2 l(2ud)^l\left(2A + ud\widetilde{R}_0^2 + 24ud\ln\frac{N}{\delta}\sigma^2 l\right) \le$$

$$\le \sigma^2 d^2 N(2ud)^N\left(2A + ud\widetilde{R}_0^2 + 24ud\ln\frac{N}{\delta}\sigma^2 N\right) := \frac{F}{2}$$

для всех $l = 1,\ldots,N$ одновременно, где

$$F = 2\sigma^2 d^2 N(2ud)^N \left(2A + ud\widetilde{R}_0^2 + 24ud\ln\frac{N}{\delta}\sigma^2 N\right).$$

Используя следствие 8 из [34] для $b = \hat{\sigma}_0^2$, получаем для любого $l = 1,\ldots,N$ с вероятностью $\geq 1 - \frac{\delta}{N}$ следующую оценку:

$$\text{либо } \sum_{t=0}^{l-1}\hat{\sigma}_t^2 \geq F, \quad \text{либо } \left|\sum_{t=0}^{l-1}\xi^t\right| \leq C_1\sqrt{\sum_{t=0}^{l-1}\hat{\sigma}_t^2\left(\ln\left(\frac{N}{\delta}\right) + \ln\ln\left(\frac{F}{f}\right)\right)}, \tag{П.37}$$

где $C_1 > 0$ – константа, которая не зависит от F и f.

Далее, объединяя полученные оценки, получаем, что оценка (П.37) с вероятностью $\geq 1 - \delta$ верна для всех $l = 1,\ldots,N$ одновременно.

Учитывая выбор F, получаем, что с вероятностью $\geq 1 - 2\delta$

$$\left|\sum_{t=0}^{l-1}\xi^t\right| \leq C_1\sqrt{\sum_{t=0}^{l-1}\hat{\sigma}_t^2\left(\ln\left(\frac{N}{\delta}\right) + \ln\ln\left(\frac{F}{f}\right)\right)}$$

для всех $l = 1,\ldots,N$ одновременно.

Для удобства дальнейших рассуждений обозначим $g(N) := \ln\left(\frac{N}{\delta}\right) + \ln\ln\left(\frac{F}{f}\right) \approx \ln\left(\frac{N}{\delta}\right)$, пренебрегая константой. Используя $\hat{\sigma}_t^2 = \sigma^2 d^2 \widetilde{R}_t^2$, получаем, что с вероятностью $\geq 1 - 2\delta$ справедлива следующая оценка:

$$\frac{1}{2}\widetilde{R}_l^2 \leq A + u\sum_{t=0}^{l-1}\underbrace{\left\langle \boldsymbol{\eta}^t, \mathbf{a}^t\right\rangle}_{\xi^t} \leq A + udD\sqrt{\sigma^2 g(N)}\sqrt{\sum_{t=0}^{l-1}\widetilde{R}_t^2} \tag{П.38}$$

для всех $l = 1,\ldots,N$ одновременно. Выбирая в качестве $A = \frac{A}{\widetilde{R}_0^2}$, $B = \frac{1}{\widetilde{R}_0}udC_1\sqrt{\sigma^2 g(N)}$, $r_t = \widetilde{R}_t$, из леммы 10 получаем, что с вероятностью $1 - 2\delta$ выполняется

$$\widetilde{R}_l \leq JR_0$$

для всех $l = 1,\ldots,N$ одновременно, где

$$J = \max\left\{1, \frac{1}{\widetilde{R}_0}udC_1\sqrt{\sigma^2 g(N)} + \sqrt{\frac{1}{\widetilde{R}_0^2}u^2 d^2 C_1^2\sigma^2 g(N) + \frac{2A}{R_0^2}}\right\}.$$

Отсюда получаем, что с вероятностью $1 - 2\delta$ оценка

$$A + u\sum_{t=0}^{l-1}\left\langle \boldsymbol{\eta}^t, \mathbf{a}^t\right\rangle \leq A + udC_1\sqrt{\sigma^2 g(N)lJ}\widetilde{R}_0^2 \leq A + udC_1\sqrt{\sigma^2 g(N)NJ}\widetilde{R}_0^2$$

верна для всех $l = 1,\ldots,N$ одновременно.

Доказательство леммы 2. Для $\boldsymbol{\lambda} \in \mathbf{R}_+^m$

$$\left\|\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}\right\|_2^2 \left\|[\boldsymbol{\lambda}^t - \beta\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t)]_+ - \boldsymbol{\lambda}\right\|_2^2 \leq \left\|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}\right\|_2^2 - 2\beta\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t), \boldsymbol{\lambda}^t - \boldsymbol{\lambda}\right\rangle + \beta^2\left\|\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t)\right\|_2^2,$$

т.е.

$$0 \leq \frac{1}{2\beta}\left(\left\|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}\right\|_2^2 - \left\|\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}\right\|_2^2\right) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle + \frac{\beta}{2}\left\|\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t)\right\|_2^2. \tag{П.39}$$

После прибавления к обеим сторонам неравенства (П.39) $\varphi(\lambda^t)$, умножения на N и суммирования от 0 до $N-1$:

$$\frac{1}{N}\sum_{t=0}^{N-1}\varphi\lambda^t) \le \frac{1}{N}\sum_{t=0}^{N-1}\left\{\varphi(\lambda^t) + \left\langle\nabla\varphi(\lambda^t,\xi^t),\lambda-\lambda^t\right\rangle + \frac{\beta}{2}\left\|\nabla\varphi(\lambda^t,\xi^t)\right\|_2^2 + \right.$$
$$\left. + \frac{1}{2\beta}\left(\left\|\lambda^t-\lambda\right\|_2^2 - \left\|\lambda^{t+1}-\lambda\right\|_2^2\right)\right\}. \tag{П.40}$$

В силу выпуклости $\varphi(\lambda)$ для $\hat{\lambda}^N = \frac{1}{N}\sum_{t=0}^{N-1}\lambda^t$ получаем, что

$$N\varphi(\hat{\lambda}^N) \le \sum_{t=0}^{N-1}\left\{\varphi(\lambda^t) + \left\langle\nabla\varphi(\lambda^t,\xi^t),\lambda-\lambda^t\right\rangle\right\} + \frac{\beta}{2}\left\|\nabla\varphi(\lambda^t,\xi^t)\right\|_2^2 + \frac{1}{2\beta}\left(\left\|\lambda^0-\lambda\right\|_2^2 - \left\|\lambda^N-\lambda\right\|_2^2\right). \tag{П.41}$$

Выбираем $\lambda = \lambda^*$ и прибавляем и вычитаем справа $\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\lambda^t),\lambda^*-\lambda^t\right\rangle$. Получаем

$$N\varphi(\hat{\lambda}^N) \le \sum_{t=0}^{N-1}\left\{\varphi(\lambda^t) + \left\langle\nabla\varphi(\lambda^t),\lambda^*-\lambda^t\right\rangle\right\} + \frac{\beta}{2}\left\|\nabla\varphi(\lambda^t,\xi^t)\right\|_2^2 +$$
$$+ \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\lambda^t,\xi^t) - \nabla\varphi(\lambda^t),\lambda^*-\lambda^t\right\rangle + \frac{1}{2\beta}\left(\left\|\lambda^0-\lambda^*\right\|_2^2 - \left\|\lambda^N-\lambda^*\right\|_2^2\right). \tag{П.42}$$

Из выпуклости $\varphi(\lambda)$ имеем

$$\sum_{t=0}^{N-1}\left\{\varphi(\lambda^t) + \left\langle\nabla\varphi(\lambda^t),\lambda^*-\lambda^t\right\rangle\right\} \le \sum_{t=0}^{N-1}\left\{\varphi(\lambda^t) + \varphi(\lambda^*) - \varphi(\lambda^t)\right\} \le \sum_{t=0}^{N-1}\varphi(\lambda^*) \le N\varphi(\lambda^*).$$

Подставляя полученную оценку в (П.42), получаем

$$\frac{1}{2\beta}\left\|\lambda^N-\lambda^*\right\|_2^2 \le \frac{1}{2\beta}\left\|\lambda^0-\lambda^*\right\|_2^2 + \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\lambda^t,\xi^t) - \nabla\varphi(\lambda^t),\lambda^*-\lambda^t\right\rangle + \frac{\beta}{2}\left\|\nabla\varphi(\lambda^t,\xi^t)\right\|_2^2. \tag{П.43}$$

Определим $R_t = \left\|\lambda^t-\lambda^*\right\|_2$ и $\widetilde{R}_t = \max\{\widetilde{R}_{t-1}, R_t\}$, причем $R_0 = \widetilde{R}_0$ и, так как $\lambda^0 = \mathbf{0}$ и $\left\|\lambda^*\right\|_2 \le R$, то $R_0 = R$. При этом по построению получаем, что $\lambda^t \in B_{\widetilde{R}_t}(\lambda^*)$. Так же определим $\left\|\mathbf{a}^t\right\|_2 = \left\|\lambda^t-\lambda^*\right\|_2 \le \widetilde{R}_t$. Тогда (П.43) можно переписать в следующем виде:

$$\frac{1}{2\beta}\widetilde{R}_N^2 \le \frac{1}{2\beta}\widetilde{R}_0^2 + \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\lambda^t,\xi^t) - \nabla\varphi(\lambda^t),\mathbf{a}^t\right\rangle + \frac{\beta}{2}\left\|\nabla\varphi(\lambda^t,\xi^t)\right\|_2^2.$$

Обозначим $\eta^t = \nabla\varphi(\lambda^t,\xi^t) - \nabla\varphi(\lambda^t)$. По теореме 2.1 из [35] имеем

$$\mathbb{P}\left\{\left\|\sum_{t=0}^{N-1}\boldsymbol{\eta}^t\right\|_2 \ge (\sqrt{2}+\sqrt{2}\gamma)\sqrt{\sum_{t=0}^{N-1}\sigma_t^2}\,|\,\{\xi^t\}_{t=0}^{N-1}\right\} \le \exp\left(-\frac{\gamma^2}{3}\right). \tag{П.44}$$

Используя лемму 2 из [34], получаем, что

$$\mathbb{E}\left[\exp\left(\frac{\left\|\boldsymbol{\eta}^t\right\|_2^2}{\sigma^2}\right)|\,\{\xi^k\}_{k=0}^{t-1}\right] \le \exp(1),$$

при этом η^t зависит только от $\xi^{t-1},\ldots,\xi^0$. Используя новые обозначения и (8), имеем

$$\widetilde{R}_N^2 \le \widetilde{R}_0^2 + 2\beta\sum_{t=0}^{N-1}\left\langle\boldsymbol{\eta}^t,\mathbf{a}^t\right\rangle + \beta^2M^2.$$

Тогда из леммы 11 с константами $A = \widetilde{R}_0^2 + \beta^2 M^2$, $d = 1$ и $u = \beta$, получаем, что с вероятностью $1 - 2\delta$, где $\frac{\delta}{N} = \exp\left(-\frac{\gamma^2}{3}\right)$, верна следующая оценка:

$$\widetilde{R}_l \le JR_0 \quad \text{и} \quad \sum_{t=0}^{l-1}\left\langle \eta^t, \mathbf{a}^t\right\rangle \le D\sqrt{\sigma^2 g(N)NJ}\,\widetilde{R}_0^2 \tag{П.45}$$

$\forall l = 1,\ldots,N$ одновременно, где D – положительная константа,

$$F = 2\sigma^2 N(2\beta)^N\left(2A + \beta\widetilde{R}_0^2 + 24\ln\frac{N}{\delta}\beta\sigma^2 N\right),$$

$f = \sigma^2\widetilde{R}_0^2$, $g(N) = \ln\left(\frac{N}{\delta}\right) + \ln\ln\left(\frac{F}{f}\right)$ и

$$J = \max\left\{1, \frac{1}{\widetilde{R}_0}\beta C_1\sqrt{\sigma^2 g(N)} + \sqrt{\frac{1}{\widetilde{R}_0^2}\beta^2 C_1^2\sigma^2 g(N) + \frac{2A}{R_0^2}}\right\}.$$

Чтобы оценить зазор двойственности, используем (П.41), также отметим, что данная оценка верна для любого $\boldsymbol{\lambda} \in \mathbf{R}_+^m$. Поэтому, беря минимум по всем $\boldsymbol{\lambda}$ из множества $\Lambda_{2R} = \left\{\boldsymbol{\lambda} \in \mathbb{R}_+^m : \|\boldsymbol{\lambda}\|_2 \le 2R\right\}$, получаем

$$N\varphi(\hat{\boldsymbol{\lambda}}^N) \le \min_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right) + \frac{1}{2\beta}\left\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}\right\|_2^2\right\} + \frac{N\beta M^2}{2},$$

где для оценки последнего слагаемого использовалось предположение (8). Также учитывалось, что $\left\|\boldsymbol{\lambda}^N - \boldsymbol{\lambda}\right\|_2^2 \ge 0$. В силу (П.29) получаем следующую оценку:

$$\varphi(\hat{\boldsymbol{\lambda}}^N) \le \frac{1}{N}\min_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right)\right\} + \frac{2R^2}{\beta N} + \frac{\beta M^2}{2}.$$

Прибавим и вычтем из выражения под минимумом $\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle$. Тогда получаем

$$\min_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right)\right\} \le \min_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right)\right\} +$$
$$+ \max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}\right\rangle\right\} + \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), -\boldsymbol{\lambda}^t\right\rangle.$$

Заметим, что $-\boldsymbol{\lambda}^* \in \Lambda_{2R}$. Тогда имеем

$$\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), -\boldsymbol{\lambda}^t\right\rangle = \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^* - \boldsymbol{\lambda}^t\right\rangle +$$
$$+ \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), -\boldsymbol{\lambda}^*\right\rangle \le \max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}\right\rangle +$$
$$+ \sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^* - \boldsymbol{\lambda}^t\right\rangle.$$

Отсюда получаем следующую оценку:

$$\varphi(\hat{\boldsymbol{\lambda}}^N) \le \frac{1}{N}\min_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right)\right\} + \frac{2R^2}{\beta N} + \frac{\beta M^2}{2} \le$$
$$\le \frac{1}{N}\min_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\boldsymbol{\lambda}^t) + \left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda} - \boldsymbol{\lambda}^t\right\rangle\right)\right\} + \frac{1}{N}\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^* - \boldsymbol{\lambda}^t\right\rangle +$$
$$+ \frac{2}{N}\max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\boldsymbol{\lambda}^t, \xi^t) - \nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}\right\rangle\right\} + \frac{2R^2}{\beta N} + \frac{\beta M^2}{2}.$$

Из определения нормы получаем, что

$$\max_{\lambda\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\lambda^t,\xi^t)-\nabla\varphi(\lambda^t),\lambda\right\rangle\right\}\le 2R\left\|\sum_{t=0}^{N-1}\left(\nabla\varphi(\lambda^t,\xi^t)-\nabla\varphi(\lambda^t)\right)\right\|_2.$$

Используя (П.44), получаем, что с вероятностью $1-\delta$ выполняется

$$\left\|\sum_{t=0}^{N-1}\left(\nabla\varphi(\lambda^t,\xi^t)-\nabla\varphi(\lambda^t)\right)\right\|_2\le\sigma\sqrt{2N}\left(1+\sqrt{3\ln\frac{1}{\delta}}\right). \tag{П.47}$$

Подставим в выражение $\sum_{t=0}^{N-1}\left(\varphi(\lambda^t)+\left\langle\nabla\varphi(\lambda^t),\lambda-\lambda^t\right\rangle\right)$ из (П.46) значения $\varphi(\lambda^t)$ и $\nabla\varphi(\lambda^t)$ и получим

$$\sum_{t=0}^{N-1}\left(\left\langle\lambda^t,\mathbf{b}\right\rangle+\sum_{k=1}^{n}(u_k(x_k(\lambda^t))-\left\langle\lambda^t,\mathbf{C}_k x_k(\lambda^t)\right\rangle+\left\langle\mathbf{b}-C\mathbf{x}^t(\lambda^t),\lambda-\lambda^t\right\rangle\right)=$$
$$=\sum_{t=0}^{N-1}\left(\sum_{k=1}^{n}(u_k(x_k(\lambda^t))+\left\langle\mathbf{b}-C\mathbf{x}^t(\lambda^t),\lambda\right\rangle\right).$$

Тогда в силу вогнутости функций $u_k(x_k)$

$$\frac{1}{N}\min_{\lambda\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left(\varphi(\lambda^t)+\left\langle\nabla\varphi(\lambda^t),\lambda-\lambda^t\right\rangle\right)\right\}\le U(\hat{\mathbf{x}}^N)-\frac{1}{N}\max_{\lambda\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left\langle C\mathbf{x}^t(\lambda^t)-\mathbf{b},\lambda\right\rangle\right\}.$$

Учитывая последнее неравенство, из (П.46) получаем

$$\varphi(\hat{\lambda}^N)\le U(\hat{\mathbf{x}}^N)-\frac{1}{N}\max_{\lambda\in\Lambda_{2R}}\left\{\sum_{t=0}^{N-1}\left\langle C\mathbf{x}^t(\lambda^t)-\mathbf{b},\lambda\right\rangle\right\}+\frac{2R^2}{\beta N}+\frac{\beta M^2}{2}+$$
$$+\frac{2R}{N}\left\|\sum_{t=0}^{N-1}\left(\nabla\varphi(\lambda^t,\xi^t)-\nabla\varphi(\lambda^t)\right)\right\|_2+\frac{1}{N}\sum_{t=0}^{N-1}\left\langle\nabla\varphi(\lambda^t,\xi^t)-\nabla\varphi(\lambda^t),\lambda^*-\lambda^t\right\rangle.$$

Отсюда, учитывая оценку (П.47) и результат (П.44), получаем, что с вероятностью $1-3\delta$

$$\varphi(\hat{\lambda}^N)-U(\hat{\mathbf{x}}^N)+2R\left\|[C\hat{\mathbf{x}}^N-\mathbf{b}]_+\right\|_2\le\frac{2R\sigma\sqrt{2}\left(1+\sqrt{3\ln\frac{1}{\delta}}\right)}{\sqrt{N}}+\frac{2R^2}{\beta N}+\frac{\beta M^2}{2}+C_1\frac{\sigma\sqrt{g(N)J}R^2}{\sqrt{N}}. \tag{П.48}$$

По теореме 2.1 из [35] для всех $\gamma>0$ имеем

$$P\left\{\left\|\sum_{t=0}^{N-1}\left(\mathbf{x}(\lambda^t,\xi^t)-\mathbf{x}(\lambda^t)\right)\right\|_2\ge(\sqrt{2}+\sqrt{2}\gamma)\sqrt{\sum_{t=0}^{N-1}\sigma_x^2}\,|\,\{\xi^t\}_{t=0}^{N-1}\right\}\le\exp\left(-\frac{\gamma^2}{3}\right).$$

Выбирая $\gamma=\sqrt{3\ln\frac{1}{\delta}}$, получаем, что с вероятностью $1-\delta$

$$\left\|\tilde{\mathbf{x}}^N-\hat{\mathbf{x}}^N\right\|_2=\frac{1}{N}\left\|\sum_{t=0}^{N-1}\left(\mathbf{x}(\lambda^t,\xi^t)-\mathbf{x}(\lambda^t)\right)\right\|_2\le\sigma_x\sqrt{\frac{2}{N}}\left(1+\sqrt{3\ln\frac{1}{\delta}}\right).$$

Тогда с вероятностью $1-\delta$ выполняется следующее неравенство:

$$\left\|C\tilde{\mathbf{x}}^N-C\hat{\mathbf{x}}^N\right\|_2\le\|C\|_2\cdot\left\|\tilde{\mathbf{x}}^N-\hat{\mathbf{x}}^N\right\|_2\le\sigma_x\sqrt{\frac{2\lambda_{\max}(C^{\mathrm{T}}C)}{N}}\left(1+\sqrt{3\ln\frac{1}{\delta}}\right).$$

Заметим, что верно следующее:

$$2R\left\|[C\tilde{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 = \max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\left\langle C\tilde{\mathbf{x}}^N - \mathbf{b}, \boldsymbol{\lambda}\right\rangle + \left\langle C\hat{\mathbf{x}}^N - C\hat{\mathbf{x}}^N - \mathbf{b} + \mathbf{b}, \boldsymbol{\lambda}\right\rangle\right\} \le$$

$$\le \max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\left\langle C\hat{\mathbf{x}}^N - \mathbf{b}, \boldsymbol{\lambda}\right\rangle\right\} + \max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\left\{\left\langle C\tilde{\mathbf{x}}^N - C\hat{\mathbf{x}}^N, \boldsymbol{\lambda}\right\rangle\right\} \le 2R\left\|[C\tilde{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 + 2R\|C\tilde{\mathbf{x}}^N - C\hat{\mathbf{x}}^N\|_2 \le \quad (\text{П.49})$$

$$\le 2R\left\|[C\tilde{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 + 2R\sigma_x\sqrt{\frac{2\lambda_{\max}(C^{\mathrm{T}}C)}{N}}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right).$$

В силу липшицевости функции U получаем

$$\left|U(\tilde{\mathbf{x}}^N) - U(\hat{\mathbf{x}}^N)\right| \le M_U\left\|\tilde{\mathbf{x}}^N - \hat{\mathbf{x}}^N\right\|_2 \le M_U\sigma_x\sqrt{\frac{2}{N}}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right).$$

Тогда имеем

$$U(\hat{\mathbf{x}}^N) = U(\tilde{\mathbf{x}}^N) + (U(\hat{\mathbf{x}}^N) - U(\tilde{\mathbf{x}}^N)) \ge U(\tilde{\mathbf{x}}^N) - M_U\sigma_x\sqrt{\frac{2}{N}}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right). \quad (\text{П.50})$$

Подставляя (П.49) и (П.50) в (П.48), получаем, что с вероятностью $1 - 4\delta$ выполняется

$$\varphi(\hat{\boldsymbol{\lambda}}^N) - U(\tilde{\mathbf{x}}^N) + 2R\left\|[C\tilde{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 \le C_1\frac{\sigma\sqrt{g(N)J}R^2}{\sqrt{N}} + \frac{2R^2}{\beta N} + \frac{\beta M^2}{2} +$$

$$+ \frac{\sqrt{2}\left(1 + \sqrt{3\ln\frac{1}{\delta}}\right)}{\sqrt{N}}\left(M_U\sigma_x + 2R\left(\sigma + \sigma_x\sqrt{\lambda_{\max}(C^{\mathrm{T}}C)}\right)\right).$$

Доказательство теоремы 3. Так как $\|\nabla\varphi(\boldsymbol{\lambda})\|_2 \le M$ для любых $\boldsymbol{\lambda} \in \Lambda_{2R}$ (см. (5)), то справедлива следующая оценка:

$$\sup_{\boldsymbol{\lambda}^1,\boldsymbol{\lambda}^2\in\Lambda_{2R}}\left\langle\nabla\varphi(\boldsymbol{\lambda}^1), \boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^1\right\rangle \le M\cdot 4R.$$

Из теоремы 4.1 [26] получаем

$$\max_{\boldsymbol{\lambda}\in\Lambda_{2R}}\sum_{t=1}^{N}\xi^t\left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^t - \boldsymbol{\lambda}\right\rangle \le \varepsilon_N,$$

где $\varepsilon_N = 32\times 4MR\exp\left\{-\dfrac{N}{2m(m+1)}\right\}$. Тогда имеем

$$\forall\boldsymbol{\lambda}\in\Lambda_{2R}\sum_{t\in I_N}\xi^t\left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^t - \boldsymbol{\lambda}\right\rangle \le \sum_{t=1}^{N}\xi^t\left\langle\nabla\varphi(\boldsymbol{\lambda}^t), \boldsymbol{\lambda}^t - \boldsymbol{\lambda}\right\rangle \le \varepsilon_N.$$

Отсюда получаем, что верна следующая оценка:

$$\sum_{t\in I_N}\xi^t\left\langle\mathbf{b} - C\mathbf{x}^t, \boldsymbol{\lambda}^t\right\rangle + \max_{\boldsymbol{\lambda}\in\Lambda_R}\left\langle-\sum_{t\in I_N}\xi^t(\mathbf{b} - C\mathbf{x}^t), \boldsymbol{\lambda}\right\rangle \le \varepsilon_N,$$

которую можно переписать в следующем виде:

$$\sum_{t\in I_N}\xi^t\left\langle\mathbf{b} - C\mathbf{x}^t, \boldsymbol{\lambda}^t\right\rangle \le \varepsilon_N - 2R\left\|[C\hat{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2. \quad (\text{П.51})$$

Далее, в силу (3), для каждого $\mathbf{x} \ge 0$ и $t \in I_N$ выполнено

$$U(\mathbf{x}^t(\boldsymbol{\lambda}^t)) - \left\langle C\mathbf{x}^t(\boldsymbol{\lambda}^t) - \mathbf{b}, \boldsymbol{\lambda}^t\right\rangle \ge U(\mathbf{x}) - \left\langle C\mathbf{x} - \mathbf{b}, \boldsymbol{\lambda}^t\right\rangle.$$

Умножая t-е неравенство на ξ^t, суммируя по всем индексам из I_N и учитывая, что $\sum_{t\in I_N}\xi^t U(\mathbf{x}^t) \le U(\hat{\mathbf{x}}^N)$, в силу вогнутости функций $u_k(x_k)$, $k = 1,\ldots,N$, получаем

$$U(\mathbf{x}) - U(\hat{\mathbf{x}}^N) + \left\langle\mathbf{b} - C\mathbf{x}, \hat{\boldsymbol{\lambda}}^N\right\rangle \le \sum_{t\in I_N}\xi^t\left\langle\mathbf{b} - C\mathbf{x}^t, \boldsymbol{\lambda}^t\right\rangle,$$

где $\hat{\boldsymbol{\lambda}}^N = \sum_{t\in I_N} \xi^t \boldsymbol{\lambda}^t$. Используя оценку (П.51), получаем

$$2R\left\|[C\hat{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 + U(\mathbf{x}^*) - U(\hat{\mathbf{x}}^N) + \left\langle \mathbf{b} - C\mathbf{x}^*, \hat{\boldsymbol{\lambda}}^N \right\rangle \le \varepsilon_N. \qquad (\text{П.52})$$

Поскольку $\hat{\boldsymbol{\lambda}}^N \in \Lambda_{2R}$, и, следовательно, $\hat{\boldsymbol{\lambda}}^N \ge 0$, откуда $\left\langle \mathbf{b} - C\mathbf{x}^*\hat{\boldsymbol{\lambda}}^N \right\rangle \ge 0$, из (П.51) следует $U(\mathbf{x}^*) - U(\hat{\mathbf{x}}^N) \le \varepsilon_N$. Далее, так как для всех $\mathbf{x} \ge 0$, в силу определения $\boldsymbol{\lambda}^*$, выполняется $U(\mathbf{x}^*) \ge U(\mathbf{x}) - \left\langle \boldsymbol{\lambda}^*, C\mathbf{x} - \mathbf{b} \right\rangle$, получаем

$$U(\hat{\mathbf{x}}^N) \le U(\mathbf{x}^*) - \left\langle \boldsymbol{\lambda}^*, \mathbf{b} - C\hat{\mathbf{x}}^N \right\rangle \le U(\mathbf{x}^*) - \min_{\boldsymbol{\lambda}\in\hat{\Lambda}_R} \left\{ \left\langle \boldsymbol{\lambda}, \mathbf{b} - C\hat{\mathbf{x}}^N \right\rangle \right\} =$$
$$= U(\mathbf{x}^*) + \max_{\boldsymbol{\lambda}\in\Lambda_R} \left\{ \left\langle \boldsymbol{\lambda}, C\hat{\mathbf{x}}^N - \mathbf{b} \right\rangle \right\} \le U(\mathbf{x}^*) + R\left\|[C\hat{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2.$$

Отсюда вместе с (П.52) получаем $R\left\|[C\hat{\mathbf{x}}^N - \mathbf{b}]_+\right\|_2 \le \varepsilon_N$. Оценка (11) на количество итераций метода следует из следующей выкладки:

$$\varepsilon_N = 32\times 4MR\exp\left\{-\frac{N}{2m(m+1)}\right\} \le \varepsilon \Rightarrow -\frac{N}{2m(m+1)} \le \ln\left(\frac{\varepsilon}{32\times 4MR}\right) \Rightarrow N \ge 2m(m+1)\ln\left(\frac{32\times 4MR}{\varepsilon}\right).$$

СПИСОК ЛИТЕРАТУРЫ

1. *Kelly F.P., Maulloo A.K., Tan D.K.H.* Rate control for communication networks: shadow prices, proportional fairness and stability // J. of the Operational Research Society. 1998. V. 49. 3. P. 237–252.
2. *Рохлин Д.Б.* Распределение ресурсов в сетях связи с большим числом пользователей: стохастический метод градиентного спуска // Теория вероятностей и ее применения (в печати). 2019.
3. *Arrow K.J., Hurwicz L.* Decentralization and computation in resource allocation. Stanford University, Department of Economics, 1958.
4. *Kakhbod A.* Resource allocation in decentralized systems with strategicagents: an implementation theory approach. Springer Science & BusinessMedia, 2013.
5. *Campbell D.E.* Resource allocation mechanisms. Cambridge University Press, 1987.
6. *Friedman E.J., Oren S.S.* The complexity of resource allocation and price mechanisms under bounded rationality // Economic Theory. 1995. V. 6. 2. P. 225–250.
7. *Nesterov Yu., Shikhman V.* Dual subgradient method with averaging for optimal resource allocation // European Journal of Operational Research. 2018. V. 270. 3. P. 907–916.
8. *Ivanova A., Dvurechensky P., Gasnikov A., Kamzolov D.* Composite optimization for the resource allocation problem // arXiv preprint arXiv:1810.00595. 2018.
9. *Нестеров Ю.Е.* Метод минимизации выпуклых функций со скоростью сходимости $O(1/k^2)$ // Докл. АН СССР. 1983. Т. 269. 39. С. 543–547.
10. *Gasnikov A.V., Gasnikova E.V., Nesterov Yu.E., Chernov A.V.* Efficient numerical methods for entropy-linear programming problems // Comput. Math. and Math. Phys. 2016. V. 56. 4. P. 514–524.
11. *Chernov A., Dvurechensky P., Gasnikov A.* Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints // Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19–23, 2016 Proceedings. Springer International Publishing, 2016. P. 391–403.
12. *Dvurechensky P., Gasnikov A., Gasnikova E., Matsievsky S., Rodomanov A., Usik I.* Primal-Dual Method for Searching Equilibrium in Hierarchical Congestion Population Games // Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School (DOOR 2016) Vladivostok, Russia, September 19–23, 2016. P. 584–595. arXiv:1606.08988.
13. *Anikin A., Gasnikov A., Turin A., Chernov A.* Dual approaches to the minimization of strongly convex functionals with asimple structure under affine constraints // Comput. Math. and Math. Phys. 2017. V. 57. № 8. P. 1262–1276.
14. *Dvurechensky P., Gasnikov A., Kroshnin A.* Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm // Proceedings of the 35th International Conference on Machine Learning. 2018. V. 80. P. 1367–1376. arXiv:1802.04367.
15. *Nesterov Yu., Gasnikov A., Guminov S., Dvurechensky P.* Primal-dual accelerated gradient methods with small-dimensional relaxation oracle // arXiv:1809.05895. 2018.
16. *Guminov S., Dvurechensky P., Gasnikov A.* On Accelerated Alternating Minimization // arXiv:1906.03622. 2019.

17. *Guminov S.V., Nesterov Yu.E., Dvurechensky P.E., Gasnikov A.V.* Accelerated Primal-Dual Gradient Descent with Linesearch for Convex, Nonconvex, and Nonsmooth Optimization Problems // Doklady Mathematics. 2019. V. 99. P. 125–128.

18. *Kroshnin A., Tupitsa N., Dvinskikh D., Dvurechensky P., Gasnikov A., Uribe C.A.* On the Complexity of Approximating Wasserstein Barycenters // Proceedings of the 36th International Conference on Machine Learning. 2019. V. 97. Eds. K. Chaudhuri, R. Salakhutdinov. California, USA: PMLR, 2019. P. 3530–3540. arXiv:1901.08686.

19. *Uribe C.A., Dvinskikh D., Dvurechensky P., Gasnikov A., Nedich A.* Distributed Computation of Wasserstein Barycenters Over Networks // 2018 IEEE Conference on Decision and Control (CDC). 2018. P. 6544–6549. arXiv:1803.02933.

20. *Dvinskikh D., Gorbunov E., Gasnikov A., Dvurechensky P., Uribe C.A.* On Primal and Dual Approaches for Distributed Stochastic Convex Optimization over Networks // 2019 IEEE Conference on Decision and Control (CDC). 2019. arXiv:1903.09844.

21. *Dvurechensky P., Dvinskikh D., Gasnikov A., Uribe C.A., Nedic A.* Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters // Advances in Neural Information Processing Systems. 2018. V. 31. P. 10783–10793. arXiv:1806.03915.

22. *Данскин Д.М.* Теория максмина. М.: Советское радио, 1970.

23. *Демьянов В.Ф., Малоземов В. Н.* Введение в минимакс. М.: Наука, 1972.

24. *Nesterov Yu.* Smooth minimization of non-smooth functions // Math. Programming. 2005. V. 103. P. 127–152.

25. *Юдин Д.Б., Немировский А.С.* Информационная сложность и эффективные методы решения выпуклых экстремальных задач // Экономика и матем. методы. 1976. 2. С. 357–369.

26. *Nemirovski A., Onn S., Rothblum U.G.* Accuracy certificates for computational problems with convex structure // Math. of Operations Research. 2010. V. 35. 1. P. 52–78.

27. *Lan G., Zhou Y.* Random gradient extrapolation for distributed and stochastic optimization // SIAM Journal on Optimization. 2018. V. 28. 4. P. 2753–2782.

28. *Bubeck S.* Convex optimization: Algorithms and complexity // Foundations and Trends in Machine Learning. 2015. V. 8. № 3/4. P. 231–357.

29. *Nesterov Yu.* Implementable tensor methods in unconstrained convex optimization: tech. rep. Universite catholique de Louvain, Center for Operations Research and Econometrics (CORE). 2018.

30. *Gasnikov A., Dvurechensky P., Gorbunov E., Vorontsova E., Selikhanovych D., Uribe C.A., Jiang B., Wang H., Zhang S., Bubeck S., Jiang Q., Lee Y.T., Li Y., Sidford A.* Near Optimal Methods for Minimizing Convex Functions with Lipschitz p-th Derivatives // Proceedings of the Thirty-Second Conference on Learning Theory. 2019. V. 99. P. 1392–1393. URL: http://proceedings.mlr.press/v99/gasnikov19b.html. arXiv:1809.00382.

31. *Zhou K., Shang F., Cheng J.* A simple stochastic variance reduced algorithm with fast convergence rates // arXiv preprint arXiv:1806.11027. 2018.

32. *Zhou K.* Direct acceleration of SAGA using sampled negative momentum //arXiv preprint arXiv:1806.11048. 2018.

33. *Niu F., Recht B., Re C., Wright S.J.* Hogwild: A lock-free approach to parallelizing stochastic gradient descent // Advances in Neural Information Processing Systems. 2011. P. 693–701.

34. *Jin C., Netrapalli P., Ge R., Kakade S.M., Jordan M.I.* A short note on concentration inequalities for random vectors with subgaussian norm // arXiv preprint arXiv:1902.03736. 2019.

35. *Juditsky A., Nemirovski A.* Large deviations of vector-valued martin-gales in 2-smooth normed spaces: tech. rep. HAL: hal-00318071. 2008. URL: http://hal.archives-ouvertes.fr/hal-00318071/. arXiv:0809.0813.

36. *Nesterov Yu.* Lectures on Convex Optimization. 2nd ed. Springer, 2018.

37. *Нестеров Ю.Е.* Алгоритмическая выпуклая оптимизация: дисс. д.ф.-м.н.: 01.01.07. М.: МФТИ, 2013.

38. *Beck A.* Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB. SIAM, Philadelphia, 2014.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 345–362

ИНФОРМАТИКА

УДК 519

О ВЕРХНЕЙ ГРАНИЦЕ СЛОЖНОСТИ СОРТИРОВКИ[1)]

© 2021 г. И. С. Сергеев

125438 Москва, 4-й Лихачёвский пер., 15, ФГУП "НИИ "Квант", Россия

e-mail: isserg@gmail.com

Поступила в редакцию 18.09.2019 г.
Переработанный вариант 23.07.2020 г.
Принята к публикации 16.09.2020 г.

Показано, что для сортировки набора из n элементов линейно упорядоченного множества всегда достаточно $\log_2(n!) + o(n)$ попарных сравнений. Библ. 14. Фиг. 3.

Ключевые слова: сортировка, сложность, дерево решений, частичный порядок, симплекс.

DOI: 10.31857/S0044466921020125

1. ВВЕДЕНИЕ

Рассматривается стандартная задача сортировки набора из n элементов линейно упорядоченного множества с помощью попарных сравнений. Подробное введение в проблематику представлено в [1, гл. 3], [2, п. 5.3], [3, Ch. 2].

Алгоритм сортировки можно изобразить в виде бинарного корневого дерева с ориентацией ребер в направлении от корня. Такое дерево обычно называется деревом сравнений, а также деревом решений или бинарной решающей диаграммой. Внутренняя вершина дерева соответствует операции сравнения некоторых двух элементов. В зависимости от результата сравнения алгоритм осуществляет переход по одному из ребер к следующей вершине. Концевая вершина (лист) дерева соответствует полученному в результате сравнений упорядочению входного набора. *Сложностью* алгоритма называется глубина дерева – максимальное расстояние в ребрах между корнем и листом.

Естественно ограничить рассмотрение алгоритмами, не выполняющими избыточных сравнений (т.е. сравнений, не добавляющих новой информации о порядке элементов). В соответствующих таким алгоритмам деревьях каждая возможная перестановка связана ровно с одним листом. В частности, деревья для сортировки n-элементного набора имеют ровно $n!$ листьев.

Пусть $S(n)$ означает минимальную сложность алгоритма сортировки n-элементного набора. Поскольку глубина дерева с m листьями не меньше $\log m$, имеет место простая нижняя оценка

$$S(n) \geq \log(n!) = n\log n - \log e \cdot n + O(\log n), \tag{1}$$

$\log e \approx 1.443$. (Здесь и далее по тексту основание у двоичных логарифмов опускается.) Более того, поскольку даже средняя глубина дерева по всем его листьям не превосходит $\log m$, то оценка (1) действительна и для сложности сортировки в среднем по всем $n!$ возможным перестановкам входного набора (см. также [1]–[3]). Подобные нижние оценки называются теоретико-информационными.

С точки зрения верхней оценки, в целом лучшим среди известных остается метод Форда–Джонсона [4] (метод бинарных вставок), предложенный более 60 лет назад. Он приводит к соотношению

$$S(n) \leq \log(n!) + cn + O(\log n), \tag{2}$$

где константа c в зависимости от n варьируется от $\log(3e/8) \approx 0.028$ (в благоприятном случае $n \sim 2^k/3$) до $\log[3/(4\ln 2)] \approx 0.114$ (в неблагоприятном случае $n \sim \ln 2 \times 2^k/3$). Метод бинарных вставок доказуемо оптимален при $n \leq 15$ и при некоторых больших n (см. [5]). Усовершенствование метода, предложенное в [6] около 1980 г., позволяет уточнить константу в неблагоприят-

[1)]Работа выполнена при финансовой поддержке РФФИ (код проекта 19-01-00294а).

ном случае до $c \approx 0.105$. Достаточно большую работу по сглаживанию неравномерности оценок метода бинарных вставок проделали авторы [7]. По результатам этой работы можно указать константу для неблагоприятного случая где-то в пределах $0.06 < c < 0.07$, точнее сказать трудно[1]. Для сложности сортировки в среднем оценки сближены гораздо сильнее. Константа в верхней границе сложности недавно была понижена до $c \approx 0.032$ при любом n (см. [8], [9]).

Метод Форда–Джонсона основан на процедурах вставки элементов в линейно упорядоченное множество. Вставка каждого элемента выполняется отдельно. Эффективность метода обеспечивается путем подбора мощности целевого множества, близкой к степени двойки.

Однако известно, что совместная вставка даже двух элементов выполняется быстрее, чем раздельная, если мощность n целевого множества находится в пределах $2^k < n < \frac{17}{14} \times 2^k - 1$ (см. [10], [11], а также [2]). Еще выгоднее может быть группировка элементов по 4 или 5 с последующей их сортировкой до вставки (см. [6]). Высокая эффективность метода работы [8] (с точки зрения средней сложности) также достигается за счет того, что часть вставок выполняется совместно для пар элементов.

В развитие этой идеи мы предлагаем алгоритм групповой вставки большого числа элементов, организованный в некотором смысле как система массового обслуживания. Этот подход перекликается с концепцией массового производства (частично упорядоченных наборов), которой следуют лучшие известные алгоритмы выбора элемента заданного порядка (см. [12], [13]). Предлагаемый метод позволяет приблизить сложность вставки, приходящуюся на один элемент, к теоретико-информационной нижней границе, т.е. $\log n + o(1)$ при любом n. Как следствие, на методе сортировки с помощью групповых вставок достигается оценка

$$S(n) \le \log(n!) + o(n).$$

Разумеется, такая же оценка справедлива и для сложности сортировки в среднем.

Изложение построено следующим образом. В разд. 2 дается краткая справка о методе Форда–Джонсона. В разд. 3 приводятся некоторые элементарные соображения, лежащие в основе предлагаемого метода. Обобщенная концепция метода групповой вставки изложена в разд. 4. Центральная часть метода – стратегия выбора элементов для сравнений – описана в разд. 5. В разд. 6 представлены основные заключения о сложности групповой вставки и сортировки.

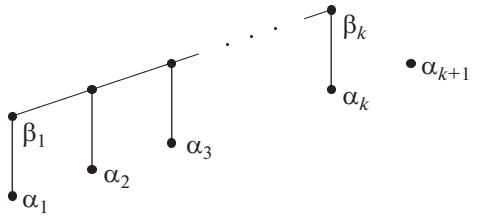
2. МЕТОД БИНАРНЫХ ВСТАВОК

Операцию сравнения элементов e_1 и e_2 будем обозначать через $e_1 ? e_2$, а отношение порядка – неравенствами $e_1 < e_2$ или $e_1 > e_2$. Если элементы e_1 и e_2 совпадают, то операция сравнения все равно возвращает либо $e_1 < e_2$, либо $e_1 > e_2$. Без ограничения общности можно считать, что все элементы входного набора различны. Для краткости линейно упорядоченный набор элементов далее будем называть *цепочкой*. Под *длиной* цепочки будем понимать число элементов в ней. Номер элемента цепочки при нумерации с 1 в порядке возрастания будем называть *рангом*.

Напомним суть метода бинарных вставок (см. [4], а также [2]): все элементы разбиваются на пары, внутри пар выполняются сравнения, большие элементы пар сортируются. В результате получается частичный порядок, диаграмма[2] которого изображена на фиг. 1: элементы в парах обозначены через α_i, β_i, где $\beta_i > \alpha_i$, нумерация в порядке возрастания β_i. В нечетном случае $n = 2k + 1$ один из элементов не имеет пары, он обозначен через α_{k+1}.

По построению элементы $\alpha_1, \beta_1, \beta_2, \ldots, \beta_k$ образуют цепочку. Далее проводится последовательная вставка оставшихся элементов α_i в эту цепочку. Первым вставляется элемент α_3 с помощью двух сравнений, затем элемент α_2 – также за два сравнения; далее вставляются все элементы, для которых достаточно трех сравнений (это α_5 и α_4 в таком порядке) и т.д. Всякий раз вставка элемента выполняется в цепочку длины $2^j - 1$, где $j = 2, 3, \ldots$, разве что в финальной серии вставок используется цепочка неполной длины.

[1] Наиболее сильная форма результата в [7] представлена графически.
[2] Диаграмма Хассе: ребра соединяют пары элементов с известным порядком (подробнее см., например, [2]).



Фиг. 1. Частичный порядок в методе бинарных вставок.

Можно проверить, что j сравнений достаточно для вставки каждого из элементов α_i с номерами $u_{j-1} < i \le u_j$, где $u_j = \dfrac{2^{j+1} + (-1)^j}{3}$. Таким образом, сложность $f(n)$ сортировки n-элементного набора методом Форда–Джонсона при $\lfloor 2^{k+1}/3 \rfloor \le n < \lfloor 2^{k+2}/3 \rfloor$ описывается выражением

$$f(n) = kn - \lfloor 2^{k+2}/3 \rfloor + \lfloor k/2 \rfloor + 1.$$

Эту формулу удобно переписать в асимптотическом виде. Пусть $n = \tau \times 2^{k+1}/3$, где $\tau \in [1, 2)$. Тогда

$$f(n) = n\log n - [2/\tau + \log(2\tau/3)]n + \log n/2 + O(1). \tag{3}$$

Коэффициент в скобках при линейном члене принимает максимальное значение $\log(8/3) \approx 1.415$ при $\tau = 1$, а минимальное значение $\log(4e\ln 2/3) \approx 1.329$ – при $\tau = 2\ln 2$. Более подробный анализ метода, см. в [4], [2].

Введем стандартное обозначение $M(m, n)$ для сложности слияния двух цепочек длины m и n. Одним из результатов работы Шульте–Мёнтинга [6] является оценка $M(5, n) \le 5k - 8$ при $n \le \dfrac{319}{448} \times 2^k - O(1)$. Она демонстрирует возможность вставки пятерки элементов в цепочку длины n за $5k - 8 + S(5) = 5k - 1$ сравнений, т.е. со средней сложностью $k - 1/5$ на один элемент. Поэтому в общем случае часть вставок финальной серии метода Форда–Джонсона, а именно, вставку элементов с номерами после u_{k-1}, выгодно произвести в цепочку меньшей длины со средней сложностью $k - 1/5$ вместо k. Это наблюдение приводит к уточнению оценки (3) для всех n, за исключением очень близко расположенных к точкам последовательности $\{2^i/3\}$. Максимальное значение константы в оценке сложности (2) при этом понижается до $c \approx 0.105$. Аналогичный результат получен в [7] весьма сложным способом: вместо вставок в финальной стадии алгоритма используются слияния длинных цепочек методом Кристена (см. [14]), но при этом оценка для неблагоприятного случая улучшена значительно.

3. ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ

Для удобства изложения введем еще несколько понятий. В задачах слияния или вставки более длинную цепочку будем называть *главной*. Множество элементов главной цепочки, расположенных между некоторыми элементами α и β, назовем *интервалом* и обозначим через (α, β): если Z – главная цепочка, то $(\alpha, \beta) = \{z \in Z \mid \alpha < z < \beta\}$. В качестве концов интервала также допускаются условные элементы $\pm\infty$: по определению, $-\infty < \alpha < +\infty$ выполнено для всех α. В отличие от длины цепочки, *длину интервала* определим как увеличенное на единицу число элементов в нем, т.е. как число возможностей для вставки нового элемента. Для длины будем использовать обозначение $|(\alpha, \beta)|$.

Обозначим через $Q(a)$ максимальное число n, при котором вставка элемента в интервал длины n выполняется со сложностью не более a или, иначе говоря, $M(1, n-1) \le a$. Тривиально,

$$Q(a) = 2^{\lfloor \log a \rfloor}. \tag{4}$$

Аналогично, через $P(a)$ обозначим максимальное n такое, что $M(2, n-1) \le 2a$: в аргументе функции P записываем число сравнений, приходящееся на один элемент вставляемой пары. Известно (см. [10], [11]), что

$$P(a) = \begin{cases} \left\lfloor \frac{17}{14} \times 2^a \right\rfloor, & a \in \mathbb{N}, \\ \left\lfloor \frac{12}{7} \times 2^{a-1/2} \right\rfloor, & (a-1/2) \in \mathbb{N}. \end{cases} \tag{5}$$

Пусть $M(m \times k, n)$ означает сложность слияния m цепочек длины k с цепочкой длины n. Введем обозначения $Q_m(a)$ и $P_m(a)$ для максимального числа n, при котором $M(m \times 1, n-1) \le am$ и соответственно $M(m \times 2, n-1) \le 2am$.

По определению, $Q_1(a) = Q(a)$ и $P_1(a) = P(a)$. Также очевидно, что $Q_m(a) \ge Q(a) - m + 1$ и $P_m(a) \ge P(a) - 2(m-1)$.

Первое следствие, оправдывающее введение функции Q_m, можно извлечь из (5). Поскольку $Q_2(a) \ge P(a-1/2)$,

$$Q_2(a) \ge \begin{cases} 2^a - 1, & a \in \mathbb{N}, \\ \left\lfloor \frac{17}{14} \times 2^{a-1/2} \right\rfloor, & (a-1/2) \in \mathbb{N}. \end{cases} \tag{6}$$

Таким образом, $Q_2(a)$ может быть существенно больше, чем $Q(a)$.

При $m = 1$ осмысленны только полуцелые аргументы у функций $P_m(a)$ и $Q_{2m}(a)$. Скажем, $P_1(r-1/4) = P_1(r-1/2) \approx \frac{6}{7} \times 2^r$, $r \in \mathbb{N}$. Однако с ростом m открываются новые возможности. Для разминки и иллюстрации основной идеи предлагаемого далее метода выведем оценку

$$P_m\left(r - \frac{1}{4} + \frac{1}{4m}\right) \ge \frac{51}{56} \times 2^r - 4m. \tag{7}$$

В главной цепочке длины $\left\lceil \frac{51}{56} \times 2^r \right\rceil - 4m$ выберем элемент α ранга $\left\lceil \frac{17}{56} \times 2^r \right\rceil - 2m + 1$. Тогда интервал $(-\infty, \alpha)$ имеет длину не более $Q_2(r-3/2) - 2(m-1)$, согласно (6), а интервал $(\alpha, +\infty)$ – не более $P(r-1) - 2(m-1)$, согласно (5).

Произвольная пара $\beta_0 < \beta_1$ обрабатывается следующим образом. Сравниваем β_0 с α. Если $\beta_0 > \alpha$, то пара вставляется в интервал $(\alpha, +\infty)$ за $2(r-1)$ сравнений методом из [10], [11]. Иначе, пара разбивается: элемент β_1 вставляется в главную цепочку бинарным методом за r сравнений, а элемент β_0 помещается во временное хранилище – *контейнер*. Если контейнер был пуст, то переходим к обработке следующей пары. Но если в контейнере оказалось 2 элемента, вставляем их в интервал $(-\infty, \alpha)$ за $2r-3$ сравнений.

По мере выполнения вставок главная цепочка и ее подынтервалы удлиняются. Выбор длин интервалов с запасом $2(m-1)$ обеспечивает возможность вставки в них элементов всех m обрабатываемых пар за планируемое число сравнений: соответственно $2(r-3/2)$ для двух элементов в интервал $(-\infty, \alpha)$ и $2(r-1)$ – для пар в интервал $(\alpha, +\infty)$.

После того, как обработка всех пар завершена, в контейнере еще может оставаться один элемент. Тривиальным способом он вставляется в цепочку за $r-1$ сравнений. Тогда общее число сравнений, выполняемых алгоритмом, в худшем случае оценивается как $m + m(r + r - 3/2) + 1/2 = 2m(r-1/4) + 1/2$. Соотношение (7) доказано.

4. ОБЩИЙ МЕТОД

В этом разделе мы опишем общий вид предлагаемого метода групповой вставки вместе со средствами его анализа, попутно вводя необходимые понятия.

Операцию сравнения в любом методе вставки или слияния можно рассматривать как шаг разбиения главной цепочки на все меньшие интервалы, в которых локализуются вставляемые элементы или группы элементов. Это дает возможность рекурсивного построения алгоритма: задача слияния с длинной цепочкой с помощью нескольких сравнений сводится к слияниям с более короткими цепочками.

Опишем основные принципы метода групповой вставки упорядоченных пар элементов в главную цепочку. В главной цепочке выделяются интервалы, которые называются *финишными*.

Вставка элемента или пары элементов в финишный интервал выполняется подходящим алгоритмом из доставляющих оценки (4)–(6). Для некоторых финишных интервалов предусмотрены контейнеры. Как в простом примере выше, любой контейнер имеет емкость 2 элемента. Когда в контейнере оказывается два элемента, то выполняется их совместная вставка в соответствующий интервал.

Пары обрабатываются по очереди. Для результата обработки пары могут быть следующие возможности: пара попадает в финишный интервал и применяется метод из [10], [11] для окончательной вставки ее в главную цепочку, либо пара в какой-то момент разбивается и элементы обрабатываются по отдельности.

Судьба одиночных элементов может быть следующей: либо элемент попадает в финишный интервал и вставляется в главную цепочку бинарным методом, либо элемент помещается в контейнер. Если в контейнере нет других элементов, то работа продолжается с новой парой. Иначе выполняется обработка двух элементов из контейнера.

После того, как все пары подверглись обработке, любой из вставляемых элементов либо включен в главную цепочку, либо находится один в некотором контейнере. Алгоритм завершается тем, что каждый из оставшихся в контейнерах элементов вставляется в свой интервал тривиальным бинарным методом.

Удельной сложностью обработки одной пары назовем число сравнений, приходящихся на элементы пары, в наихудшем случае при условии, что ни один из элементов пары не остался в контейнере (для этого можно предположить, что все контейнеры не пусты). В ходе подсчета сложности одно сравнение, выполняемое при совместной обработке пары элементов, засчитывается как полсравнения каждому элементу. Удельная сложность вставки пары в примере из предыдущего раздела равна $2(k - 1/4)$.

Итоговую сложность алгоритма групповой вставки можно оценить сверху как

$$\rho m + C\log(n + 2m), \tag{8}$$

где ρ – удельная сложность, m – число вставляемых объектов, n – длина исходной цепочки, а C – общее число контейнеров. Избыточное число сравнений при вставке последнего элемента в контейнере здесь оценивается грубо как $\log(n + 2m)$; более аккуратное рассуждение привело бы к оценке $O(1)$; в реальных алгоритмах эта величина не превосходит 1, как в примере из разд. 3.

Теперь введем комплексное понятие стратегии сравнений. Пусть задан некоторый числовой набор $y_1, \ldots, y_d \in \mathbb{R}$ с условием $0 \le y_1 < \ldots < y_d < 1$. Вводятся величины $p_{r,i}$, характеризующие длину интервала $J_{r,i}$, вставка пары в который возможна с удельной сложностью $2(r + y_i)$. Наряду с величинами $p_{r,j}$ будем также использовать удельные величины $x_{r,j} = p_{r,j} \times 2^{-r}$.

При любом $i = 1, 2, \ldots, d$ и $r \ge r_0$ задается сокращенное дерево сравнений $T_{r,i}$ для вставки пары в интервал $J_{r,i}$. Под сокращенным деревом мы понимаем поддерево обычного дерева сравнений. Концевые вершины $T_{r,i}$ соответствуют вызову процедуры вставки в интервалы меньшей длины, назовем эти интервалы *терминальными*. Разбиение пары допускается только в концевых вершинах дерева. Так определенные сокращенные деревья сравнений описывают сведение задачи вставки в длинный интервал к задачам меньшей размерности.

Концевой вершине дополнительно ставятся в соответствие ограничения на длины терминальных интервалов, обеспечивающие выполнение оценки удельной сложности вставки пары в интервал $J_{r,i}$. Если при концевой вершине на расстоянии s от корня выполняется вставка пары в терминальный интервал (α', α''), то ограничение имеет вид $|(\alpha', \alpha'')| \le p_{l,j}$, где $l + y_j + s/2 \le r + y_i$. Для одиночного элемента β, вставляемого в (α', α''), ограничение может выглядеть либо как $|(\alpha', \alpha'')| \le p_{l,j}$, либо как $|(\alpha', \alpha'')| \le 2^l$. В первом случае вклад этого элемента в удельную сложность оценивается как $l + y_j + s/2 + 1/2$, а во втором – как $l + s/2$. Сумма оценок для двух элементов пары не должна превосходить $2(r + y_i)$.

Будем считать, что ранги элементов α', α'' также выражаются линейными комбинациями величин $p_{l,j}$. Некоторые ограничения могут при этом тривиально выполняться (например, в случае тождественности правой и левой частей неравенств). Множество оставшихся определим как *систему ограничений* стратегии.

Итак, под *стратегией сравнений* будем понимать семейство деревьев (фактически алгоритмов) $T_{r,i}$ и систему ограничений, обеспечивающую удельную сложность $2(r + y_i)$ вставки пар в интервалы $J_{r,i}$. *Глубиной* стратегии назовем максимальную глубину деревьев $T_{r,i}$.

Если минимальная удельная сложность вставки пары по всем терминальным интервалам в алгоритме с деревом сравнений $T_{r,i}$ равна $2(r' + y_j)$, то положим $w_{r,i} = r - r' + 1$.

Мы ограничим рассмотрение *однородными* стратегиями – такими, в которых очередной элемент интервала $J_{r,i}$, выбираемый для сравнения, делит интервал $J_{r,i}$ в отношении, не зависящем от r (с точностью до целочисленного округления). Тогда при любом i все деревья $T_{r,i}$ подобны некоторому дереву T_i. Можно считать, что внутренней вершине T_i приписана пара (y, Δ), $y \in \{0,1\}$, $\Delta \in \mathbb{R}$, кодирующая сравнение $\beta_y \,?\, \alpha$, где (β_0, β_1) – вставляемая пара, а α – элемент ранга $\Delta|J_{r,i}|$ в интервале $J_{r,i}$. В однородной стратегии $w_{r,i} = w_i$ при всех r. Величину $\max_i w_i$ назовем *шириной* однородной стратегии. Ширина стратегии характеризует максимальную разницу сложности вставки между исходным интервалом и его терминальными подынтервалами. Ниже под стратегией всюду понимается однородная стратегия.

Далее в выкладках позволим длинам интервалов и рангам элементов цепочки принимать нецелые значения, под которыми будут пониматься округления до ближайшего целого в ту или иную сторону. В случае утверждения о сложности вставки в интервал нецелой длины x будем требовать выполнение этого утверждения для интервала длины $\lceil x \rceil$.

Наша ближайшая цель – построить переход от стратегии к конкретному алгоритму. В предположении существования пределов $z_i = \lim_{r\to\infty} x_{r,i}$ ограничения стратегии можно переписать в терминах предельных величин z_i и перейти к задаче линейного программирования. Ниже следуют описание и обоснование корректности такого перехода.

Пусть на множестве последовательностей d-мерных вещественных векторов $x_0, x_1, \ldots \in \mathbb{R}^d$, $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,d})$, определена система σ нормированных линейных неравенств ширины w вида

$$\sum_{i=1}^{d}\sum_{j=0}^{w-1} a_{i,j,k} x_{t+j,i} \le b_k, \quad \text{где} \quad \sum_{i=1}^{d}\sum_{j=0}^{w-1} |a_{i,j,k}| = 1, \quad k = 1, 2, \ldots, K, \tag{9}$$

при всех $t \ge 0$. Здесь $a_{i,j,k}$, b_k – вещественные коэффициенты. С помощью формальной подстановки $x_{j,i} = z_i$ для всех i, j в (9) получим *приведенную систему*

$$\sum_{i=1}^{d} a_{i,k} z_i \le b_k, \quad a_{i,k} = \sum_{j=0}^{w-1} a_{i,j,k}, \quad k = 1, 2, \ldots, K, \tag{10}$$

относительно переменных z_i.

Система (10) определяет симплекс $T(\sigma)$ в пространстве $\mathbb{R}^d$. Для произвольного параметра $\varepsilon \ge 0$ определим вложенный симплекс $T^\varepsilon(\sigma) \subset T(\sigma)$ системой неравенств

$$\sum_{i=1}^{d} a_{i,k} z_i \le b_k - \varepsilon, \quad k = 1, 2, \ldots, K. \tag{11}$$

Будем говорить, что подпоследовательность $x_l, x_{l+1}, \ldots, x_{l'}$ удовлетворяет (9), если ограничения (9) выполнены для всех $t = l, l+1, \ldots, l' + w - 1$ при подходящем доопределении последовательности векторами $x_{l'+1}, \ldots, x_{l'+w-1}$.

Далее через $\|\cdot\|$ обозначается l_∞-норма в векторном пространстве (максимум модуля координат вектора), а через $\langle \cdot, \cdot \rangle$ – евклидово скалярное произведение векторов. Нам понадобится простой факт о скорости приближения к заранее известному решению задачи (9).

Лемма 1. *Пусть* $u \in T^\varepsilon(\sigma)$, $v \in T(\sigma)$ *и* $0 < \varepsilon < \|v - u\|$. *Тогда на отрезке* $[u, v]$ *найдется удовлетворяющая системе неравенств* σ (9) *ширины* w *последовательность* $\{x_i\}$ *с началом* $x_0 = x_1 = \ldots = x_{w-1} = u$, *сходящаяся к точке* v *со скоростью*

$$\|v - x_i\| \le (1 - \varepsilon\Delta)^{i/w-1} \|v - u\|, \quad \Delta = \|v - u\|^{-1}. \tag{12}$$

Доказательство. Определим $\varepsilon_i = \varepsilon(1 - \varepsilon\Delta)^i$. При $i \ge 1$ и $j = 0, 1, \ldots, w-1$ положим

$$x_{iw+j} = x_{(i-1)w+j} + \varepsilon_{i-1}\Delta(v - u). \tag{13}$$

По построению

$$v - x_{iw+j} = \left[1 - (\varepsilon_0 + \ldots + \varepsilon_{i-1})\Delta\right](v - u) = (1 - \varepsilon\Delta)^i (v - u). \tag{14}$$

Таким образом, оценка (12) выполняется. Осталось проверить, что построенная последовательность удовлетворяет системе σ.

В силу $x_{lw} = x_{lw+1} = \ldots = x_{lw+w-1} \in [u, v] \subset T(\sigma)$ неравенства (9) выполнены при всех $t = lw$, где $l \in \mathbb{N}$.

Покажем, что $x_{lw} \in T^{\varepsilon_l}(\sigma)$. При $l = 0$ это гарантировано условием леммы. Обозначим $a_k = (a_{1,k}, \ldots, a_{d,k})$. Тогда при $l > 1$ и при любом $k = 1, 2, \ldots, K$, согласно (14), имеет место

$$\langle a_k, x_{lw} \rangle = \langle a_k, (1 - \varepsilon_l/\varepsilon)v + (\varepsilon_l/\varepsilon)u \rangle \le (1 - \varepsilon_l/\varepsilon)b_k + (\varepsilon_l/\varepsilon)(b_k - \varepsilon) = b_k - \varepsilon_l$$

в силу линейности скалярного произведения.

Пусть $lw < t < (l+1)w$. В этом случае в левых частях неравенств (9) используются только координаты векторов x_{lw} и $x_{(l+1)w}$. Тогда с помощью (13) и с учетом $x_{lw} \in T^{\varepsilon_l}(\sigma)$ при любом $k = 1, 2, \ldots, K$ получаем

$$\sum_{i=1}^{d}\sum_{j=0}^{w-1} a_{i,j,k} x_{t+j,i} \le \sum_{i=1}^{d}\sum_{j=0}^{w-1} a_{i,j,k} x_{lw,i} + \varepsilon_l \sum_{i=1}^{d}\sum_{j=0}^{w-1} |a_{i,j,k}| \le (b_k - \varepsilon_l) + \varepsilon_l.$$

Следовательно, построенная последовательность удовлетворяет системе неравенств σ.

Сформулируем основной результат раздела. Обозначим

$$\pi(a) = \lim_{r\to\infty} P(r + a) \times 2^{-r}. \tag{15}$$

В частности, $\pi(a) = \dfrac{17}{14}$ при $0 \le a < \dfrac{1}{2}$ и $\pi(a) = \dfrac{12}{7}$ при $\dfrac{1}{2} \le a < 1$, согласно (5).

Лемма 2. *Пусть для некоторого набора $y_1, \ldots, y_d \in \mathbb{R}$, где $0 \le y_1 < \ldots < y_d < 1$, имеется стратегия сравнений глубины h и ширины w, задающая систему ограничений σ (9) на $x_{r,i}$, где $x_{r,i} \times 2^r$ – нижняя оценка длины интервала, вставка упорядоченной пары в который выполняется с удельной сложностью $2(r + y_i)$. Пусть $v = (v_1, \ldots, v_d) \in T(\sigma)$, $u = (u_1, \ldots, u_d) \in T^{\varepsilon}(\sigma)$, где $0 < \varepsilon < \|v - u\|$, причем $u_i < v_i$ и $u_i \le \pi(y_i)$ для всех $i = 1, 2, \ldots, d$. Обозначим $u_{\min} = \min\{u_i\}$. Тогда для любого $r \in \mathbb{N}$ и любых $m \in \mathbb{N}$, $\varphi > 0$, удовлетворяющих условию*

$$m \le \min\{u_{\min}, \varepsilon/2\} \times 2^{r-\varphi-1}, \tag{16}$$

выполняется

$$P_m\left[r + y_i + \frac{(r+2)\varphi \times 2^{\varphi}}{m}\right] \ge v_i \times 2^r - \frac{R}{\varepsilon} m \times 2^{\varphi+2} - R \times 2^r e^{-\frac{\varepsilon}{2R}\left[\frac{\varphi}{w(d+1)(h+1)} - 2\right]}, \tag{17}$$

где $i = 1, 2, \ldots, d$ и $R = \|v - u\|$.

Доказательство. Зададимся некоторым значением r и допустимой парой параметров m и φ. Положим $L = \left\lfloor \dfrac{\varphi}{(h+1)(d+1)} \right\rfloor$ и $q = r - L$. Из (16) и (15) вытекает $m \le \dfrac{6}{7} \times 2^{r-\varphi}$, откуда $r > \varphi \ge L$. Следовательно, $q > 0$.

Пусть $\delta \in (0, \varepsilon)$ – параметр, который будет выбран позже. На отрезке $[u, v]$ выберем точку $v' = (1 - \delta/\varepsilon)v + (\delta/\varepsilon)u$. То, что $v' \in T^{\delta}(\sigma)$, проверяется рассуждением из леммы 1: при любом $k = 1, 2, \ldots, K$

$$\langle a_k, v' \rangle = \langle a_k, (1 - \delta/\varepsilon)v + (\delta/\varepsilon)u \rangle \le (1 - \delta/\varepsilon)b_k + (\delta/\varepsilon)(b_k - \varepsilon) = b_k - \delta.$$

По определению ширина системы ограничений не превосходит ширину соответствующей ей стратегии; можно считать, что обе величины равны w. Рассмотрим σ' – систему, получающуюся из σ вычитанием δ из правых частей неравенств (9). Тогда имеем $v' \in T(\sigma')$ и $u \in T^{\varepsilon-\delta}(\sigma')$. Приме-

няя лемму 1 к системе σ', построим последовательность $\{x_l = (x_{l,1},\ldots,x_{l,d})\}$, сходящуюся к v', с началом $x_q = \ldots = x_{q+w-1} = u$.

Обозначим $X_{l,i} = x_{q+l,i} \times 2^{q+l}$. По условию заданная стратегия сравнений обеспечивает удельную сложность $2(q + l + y_i)$ вставки пары в интервал длины $X_{l,i}$. Покажем индукцией по $dl + i$, что для вставки m пар в интервал длины

$$X'_{l,i} = X_{l,i} - 2^{(dl+i)h} \times 2m \tag{18}$$

можно построить алгоритм той же удельной сложности $2(q + l + y_i)$ с $c_{dl+i} \le (dl + i) \times 2^{(dl+i)(h+1)}$ контейнерами. (Напомним, что под интервалом длины $X'_{l,i}$ понимается интервал длины $\left\lceil X'_{l,i} \right\rceil$.)

Убедимся, что $X'_{l,i} > 0$. Действительно, согласно (16),

$$X_{l,i} \ge u_i \times 2^{q+l} \ge m \times 2^{\varphi+1-r+q+l} \ge 2m \times 2^{(h+1)(d+1)L+l-L} > 2m \times 2^{(d+1)Lh}.$$

Теперь проверим, что запас по длине и число контейнеров достаточны для вставки, затем обеспечим выполнение ограничений (9) для подпоследовательности векторов x'_{q+l}, $0 \le l \le L$, с компонентами $x'_{q+l,i} = X'_{l,i} \times 2^{-(q+l)}$.

При $0 \le l \le w - 1$ (база индукции) требуемое гарантируется условиями леммы на u_i: длины интервалов находятся в области применения оценок (5). В этом случае контейнеры не нужны. Для вставки m пар требуется запас длины $2m - 2$ по отношению к (5) с возможными дополнительными поправками: 1 на округление длины интервала до $\left\lceil X'_{0,i} \right\rceil$ и 1 на ошибку (15) при предельном переходе по отношению к (5). Поэтому взятый в (18) запас минимум в $2m$ достаточен.

При $l \ge w$, отталкиваясь от разбиения интервала длины $X_{l,i}$ на подынтервалы в соответствии со стратегией, построим разбиение интервала длины $X'_{l,i}$. По существу задача состоит в том, чтобы распределить запас длины $2^{(dl+i)h} \times 2m$ между подынтервалами. Будем передвигаться по дереву сравнений от корня к произвольному листу. Пусть в очередной вершине выполняется сравнение с элементом[3] α главной цепочки, и (α', α'') – минимальный интервал, включающий α, границы которого либо являются элементами предшествующих сравнений, либо границами исходного интервала. Тогда запас длины интервала (α', α'') поделим поровну между интервалами (α', α) и (α, α''). В итоге, поскольку дерево сравнений имеет глубину h, любой из подынтервалов, образованный точками сравнений, получит запас длины не менее $2^{(dl+i-1)h} \times 2m$ – по индуктивному предположению, достаточный запас для вставки пары с удельной сложностью $2(q + l' + y_j)$ или одиночного элемента – с удельной сложностью[4] $q + l' + y_j + 1/2$, где $l' < l$ или $j < i$. Разумеется, запас выбирается в пределах длины подынтервала, чтобы она оставалась положительной. Также заметим, что терминальные интервалы в алгоритме имеют удельную сложность вставки пары не менее $q + l - w \ge q$, поэтому рекурсивный вызов алгоритма (вставки в терминальный интервал) возможен. Формально не исключается случай, когда стратегия предполагает вставку одиночного элемента во внутренний подынтервал с удельной сложностью не менее $q + l + y_i + 1/2$. В этой ситуации просто вызывается алгоритм вставки в исходный интервал стартовой длины $X'_{l,i}$.

С использованием разбиения интервала длины $X'_{l,i}$ разбиение интервала длины $\left\lceil X'_{l,i} \right\rceil$ строится простым округлением рангов всех внутренних элементов разбиения до ближайших целых чисел сверху. При этом длины любых подынтервалов исходного разбиения, в том числе терминальных, изменяются менее чем на 1, т.е. превращаются в округления до ближайших целых сверху или снизу. Таким образом, учет округлений не приводит к изменению оценки удельной сложности алгоритма.

[3] Поскольку стратегия оперирует с нецелыми длинами интервалов, под элементом в этом абзаце понимается условная точка разбиения интервала.

[4] Под удельной сложностью вставки одиночного элемента понимается приходящаяся на него удельная сложность вставки в рамках пары.

Оценим достаточное число контейнеров. Дерево сравнений имеет не более 2^h листьев, каждому соответствует до двух терминальных интервалов, вставка в которые по предположению индукции выполняется алгоритмами, использующими не более $2c_{dl+i-1}$ контейнеров. Еще не более $2^{h+1}+1$ контейнеров может понадобиться для обслуживания непосредственно терминальных интервалов и исходного интервала. Получаем

$$c_{dl+i} \le 2^{h+1}[(dl+i-1)\times 2^{(dl+i-1)(h+1)}] + 2^{h+1} + 1 \le (dl+i)\times 2^{(dl+i)(h+1)}.$$

Теперь проверим выполнение ограничений σ для $x'_{l,i}$. Формально положим $x'_{l,i} = x_{j,i}$ при $j > r$. Определим $\delta = 2^{\varphi-r+1}m$. По условию (16) имеет место $\delta < \varepsilon/2$. Кроме того, в силу (18) и определений параметров φ и δ выполнено $x'_{l,i} - x_{j,i} \in (-\delta, 0]$ для всех $j \ge q$. Значит, при всех $t \ge q$ и любом $k = 1, 2, \ldots, K$ получаем

$$\sum_{i=1}^{d}\sum_{j=0}^{w-1} a_{i,j,k}x'_{t+l,i} \le \sum_{i=1}^{d}\sum_{j=0}^{w-1} a_{i,j,k}x_{t+j,i} + \delta\sum_{i=1}^{d}\sum_{j=0}^{w-1}|a_{i,j,k}| \le (b_k - \delta) + \delta.$$

Таким образом, последовательность $\{x'_l = (x'_{l,1}, \ldots, x'_{l,d})\}$ удовлетворяет системе ограничений σ при $t \ge q$. Тем самым доказан индуктивный переход: вставка в интервал длины $X'_{L,i}$ действительно выполняется с удельной сложностью $2(r + y_i)$ сравнений.

Остается проверить оценку (17). Выбором параметра L обеспечено

$$c_{dl+i} \le (d+1)L \times 2^{(d+1)(h+1)L} \le \varphi \times 2^{\varphi}. \tag{19}$$

Из леммы 1 следует

$$\begin{aligned} |x'_{r,i} - v_i| &\le |x'_{r,i} - x_{r,i}| + |x_{r,i} - v'_i| + |v'_i - v_i| \le \delta + (1 - (\varepsilon - \delta)/R)^{L/w-1}R + (\delta/\varepsilon)|v_i - u_i| \le \delta + \\ &+ e^{-(\varepsilon-\delta)(L/w-1)/R}R + (\delta/\varepsilon)R \le \delta + e^{-\frac{\varepsilon}{2R}\left[\frac{\varphi}{w(d+1)(h+1)}-2\right]}R + (\delta/\varepsilon)R < \frac{2R}{\varepsilon}m \times 2^{\varphi-r+1} + e^{-\frac{\varepsilon}{2R}\left[\frac{\varphi}{w(d+1)(h+1)}-2\right]}R, \end{aligned} \tag{20}$$

где также использовалось известное неравенство $(1-x)^{1/x} \le e^{-1}$ при $x \in (0,1)$. Теперь (17) получается подстановкой (19) и (20) в оценку (8), в которой можно положить $n + 2m \le X_{L,i} = x_{r,i} \times 2^r < 2^{r+2}$, поскольку из теоретико-информационных соображений $x_{r,i} \le 2^{y_i+1/2} < 4$.

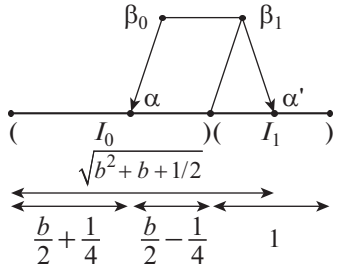
Оценки в обеих леммах весьма грубы: точность приносится в жертву простоте или общности изложения. В частности, существенно меньшее число контейнеров требуется конкретной стратегии, к которой будет применяться лемма 2.

5. УНИВЕРСАЛЬНАЯ СТРАТЕГИЯ

В этом разделе строится стратегия сравнений для доказательства асимптотической оценки высокой точности. Известно (и вполне очевидно), что для построения оптимального алгоритма сортировки некоторого частично упорядоченного множества желательно, чтобы при каждом сравнении множество возможных доупорядочиваний разбивалось на две примерно равные части в зависимости от результата сравнения. Такой ход мысли ведет к градиентному методу: сначала выполняем некоторое число сравнений, руководствуясь делением множества исходов как можно точнее пополам; затем полученный частичный порядок обрабатываем каким-то простым способом.

Теоретико-информационная нижняя оценка сложности вставки пары в интервал длины $n-1$ составляет $\log[n(n-1)/2] < 2(\log n - 1/2)$. Представим идеальный случай, предположив, что с удельной сложностью $2a$ пару можно вставить в интервал длины $Y(a) = (1-\lambda)\times 2^{a+1/2}$ при любом достаточно большом a, где $\lambda \ge 0$ – не зависящий от a параметр, который будет определен позднее.

Рассмотрим одну из градиентных стратегий для алгоритма вставки пары $\beta_0 < \beta_1$ в интервал длины $Y(a)$. Стратегия характеризуется параметром $s \in \mathbb{N}$. Сразу заметим, что при построении стратегии длины интервалов считаются действительными числами: необходимость округления до целых учитывается при переходе от стратегии к алгоритму леммой 2.



Фиг. 2. Оптимальные варианты для очередного сравнения.

Сначала отдельно рассмотрим задачу выбора очередного элемента для сравнения. Пусть в результате серии сравнений частичный порядок таков, что элемент β_1 принадлежит интервалу I_1, а β_0 принадлежит интервалу $I_0 \cup I_1$, где $I_0 \cap I_1 = \varnothing$ и $|I_0| = b|I_1|$. Тогда наилучший выбор для сравнения с β_0 – это элемент[5] α, имеющий ранг $(b/2 + 1/4)|I_1|$, считая от левого края интервала I_0. (Если $b > 1/2$, то $\alpha \in I_0$.) А наилучший выбор для сравнения с β_1 – это элемент α' ранга $\sqrt{b^2 + b + 1/2}|I_1|$ (фиг. 2). Вообще, если бы мы хотели разбить множество исходов в отношении $x : (1 - x)$, то для сравнения с β_0 и β_1 нужно было бы выбрать соответственно элементы ранга $x(b + 1/2)|I_1|$ или ранга $\sqrt{b^2 + x(2b + 1)}|I_1|$.

Исходное разбиение. Не ограничивая общности, считаем, что первое сравнение имеет вид $\beta_1 ? \alpha_1$, тогда в качестве α_1 следует взять элемент ранга $Y(a - 1/2) = (1 - \lambda) \times 2^a$. Если $\beta_1 < \alpha_1$, то вызывается алгоритм вставки пары в интервал $(-\infty, \alpha_1)$ длины $Y(a - 1/2)$. Иначе, выполняем сравнение $\beta_0 ? \alpha_2$, где α_2 – элемент оптимального ранга $(1 - \lambda)\dfrac{\sqrt{2} + 1}{4} \times 2^a$. Причина, по которой второе сравнение выполняется с элементом β_0, будет разъяснена чуть позже.

Если $\beta_0 < \alpha_2$, то элемент β_0 вставляется в интервал $(-\infty, \alpha_2)$ с удельной сложностью $a - 2 + \log(\sqrt{2} + 1)$, а элемент β_1 вставляется в интервал $(\alpha_1, +\infty)$ длины $(1 - \lambda)(\sqrt{2} - 1) \times 2^a$ с удельной сложностью $a + \log(\sqrt{2} - 1)$. Следовательно, в этом случае вставка пары выполняется за $2a$ удельных сравнений.

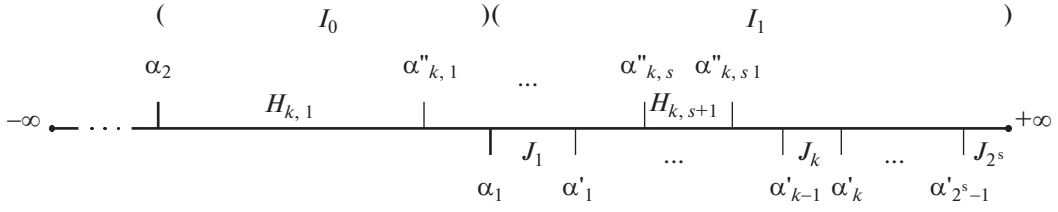
Если же $\beta_0 > \alpha_2$, то $\beta_1 \in I_1$ и $\beta_0 \in I_0 \cup I_1$, где $I_1 = (\alpha_1, +\infty)$ и $I_0 = (\alpha_2, \alpha_1)$. Пусть $b = |I_0|/|I_1| = \dfrac{2\sqrt{2} + 1}{4}$. С помощью s сравнений, разбивая каждый раз множество исходов пополам, мы локализуем β_1 в одном из 2^s подынтервалов интервала I_1. В соответствии с приведенными выше формулами границами этих подынтервалов являются элементы α'_k ранга $b_k \cdot |I_1|$, $b_k = \sqrt{b^2 + k(2b + 1) \times 2^{-s}}$, $k = 0, 1, \ldots, 2^s - 1$, и условная точка $\alpha'_{2^s} = +\infty$. Обозначим подынтервал $(\alpha'_{k-1}, \alpha'_k)$ через J_k. Общая схема разбиения показана на фиг. 3.

Пусть элемент β_1 определен в интервал J_k. Следующий этап – вставка элемента β_0. По очереди сравниваем его с элементами $\alpha''_{k,1}, \ldots, \alpha''_{k,s+1}$, определяемыми следующим образом. Положим $\alpha''_{k,0} = \alpha_2$. Обозначим через $H_{k,j}$ интервал $(\alpha''_{k,j-1}, \alpha''_{k,j})$. Длина интервала $H_{k,j}$ выбирается равной

$$|H_{k,j}| = (1 - \lambda) \times 2^{2a - s - 2 - j - \log[|J_k|/(1-\lambda)]}. \tag{21}$$

Если $\beta_0 > \alpha''_{k,1}$, то выполняем сравнение $\beta_0 ? \alpha''_{k,2}$; если $\beta_0 > \alpha''_{k,2}$, то сравниваем β_0 с $\alpha''_{k,3}$ и т.д. Как только $\beta_0 < \alpha''_{k,j}$, элементы β_0 и β_1 вставляются независимо в интервалы $H_{k,j}$ и J_k с суммарной удельной сложностью $2a - s - 2 - j$. В этом случае общая удельная сложность вставки пары составит $2a$.

[5] Здесь и ниже под элементом обычно понимается точка разбиения интервала.



Фиг. 3. Универсальное разбиение на интервалы.

Наконец, если $\beta_0 > \alpha''_{k,s+1}$, то оба элемента β_0, β_1 принадлежат интервалу $(\alpha''_{k,s+1}, \alpha'_k)$. Хотелось бы вставить пару в этот интервал ценой $2a - 2s - 3$ сравнений. Это было бы возможно в случае, когда, скажем, $\alpha''_{k,s+1} = \alpha_{k-1}$. Но равенство, вообще говоря, не выполняется, и интервал оказывается длиннее требуемого. Оценим превышение длины. Интервал J_k имеет длину

$$|J_k| = (b_k - b_{k-1})|I_1| = \frac{2b+1}{2^s(b_k + b_{k-1})}|I_1| = (1-\lambda)\frac{\sqrt{2}+1}{2^{s+1}(b_k + b_{k-1})} \times 2^a. \tag{22}$$

Тогда из (21) следует, что

$$|H_{k,j}| = (1-\lambda) \times 2^{a-1-j-\log(\sqrt{2}+1)+\log(b_k+b_{k-1})} = \frac{b_k + b_{k-1}}{2^{j+1}}|I_1|. \tag{23}$$

Значит, сумма длин интервалов $H_{k,j}$ при фиксированном k равна

$$\sum_{j=1}^{s+1}|H_{k,j}| = (1 - 2^{-s-1})\frac{b_k + b_{k-1}}{2}|I_1|. \tag{24}$$

Получаем

$$|(\alpha''_{k,s+1}, \alpha'_k)| = b_k|I_1| - \sum_{j=1}^{s+1}|H_{k,j}| = \left[b_k - (1 - 2^{-s-1})(b_k + b_{k-1})/2\right]|I_1| =$$
$$= \left(\frac{b_k - b_{k-1}}{2} + \frac{b_k + b_{k-1}}{2^{s+2}}\right)|I_1| < \left(\frac{2b+1}{2^{s+1} \times 2b} + \frac{b+1}{2^{s+1}}\right)|I_1| < \frac{7}{2^{s+2}}|I_1|.$$

Следовательно,

$$|(\alpha''_{k,s+1}, \alpha'_k)| - Y(a - s - 3/2) < \frac{7}{2^{s+2}}|I_1| - (1-\lambda) \times 2^{a-s-1} = \frac{7 - 2(\sqrt{2}+1)}{2^{s+2}}|I_1| < \frac{\ln 2}{2^s}|I_1|. \tag{25}$$

Полученная оценка понадобится чуть позже. Теперь мы модифицируем описанную выше стратегию сравнений, позволив некоторое отступление от идеала, т.е. деления множества исходов ровно пополам. А именно, мы изменим распределение длин подынтервалов J_k внутри интервала I_1, уменьшив (оценочную) сложность вставки элементов в каждый из них, тем самым увеличив допустимую сложность вставки в интервалы $H_{k,j}$ и удлинив их. Для этого у нас имеется резерв – использовать для сложности вставки элемента простую оценку $Q(a)$ вместо $Y(a - 1/2)$, которая лучше, когда значение a близко снизу к целому числу[6]. Длина интервала I_1 не будет изменена.

Заметим, что длины интервалов J_k с ростом k изменяются в пределах от $\frac{2b+1}{2b} \times 2^{-s}|I_1|$ до $\frac{2b+1}{2b+2} \times 2^{-s}|I_1|$ (см. (22)). В силу $b < 1$ выполнено $\left[\frac{2}{3}, \frac{4}{3}\right] \subset \left[1 - \frac{1}{2b+1}, 1 + \frac{1}{2b+1}\right]$, поэтому при любом (достаточно большом) a часть интервалов J_k имеет длину, близкую к степени двойки. Выполненное вторым сравнение $\beta_0 \,?\, \alpha_2$ преследовало именно цель уменьшения показателя b до величины меньше 1.

Коррекция разбиения. На этом шаге мы вносим поправки в длины интервалов J_k и $H_{k,j}$, корректируя стратегию, и одновременно дискретизируем задачу (новые интервалы будут обозна-

[6] Едва ли можно построить эффективный алгоритм вставки, не прибегая к оценкам типа $Q(a)$, что видно на примерах алгоритмов из [10], [11] и простого алгоритма из разд. 3.

чаться как J_k^* и $H_{k,j}^*$). Введем параметр дискретизации $d \in 2\mathbb{N}$ и положим $y_i = i/d$, $i = 0,1,\ldots,d-1$. По-прежнему мы ищем оценку длины интервала, достаточной для вставки пары с удельной сложностью $2(r + y_i)$, в виде $Y(r + y_i) = (1-\lambda)\times 2^{r+y_i+1/2}$.

Задавшись некоторым $a = r + y_i$, $r \in \mathbb{N}$, опишем модификацию исходного разбиения. Четность параметра d позволяет выбрать элемент α_1 для первого сравнения так же, как и в идеальной схеме. Но сложность вставки элемента в интервал $I_1 = (\alpha_1, +\infty)$ теперь приходится оценить как $a + \left\lceil d\log(\sqrt{2}-1)\right\rceil/d < a + \log(\sqrt{2}-1) + 1/d$. Как следствие, вместо α_2 приходится выбирать элемент α_2^* ранга не выше $(1-\lambda)\times 2^{a-2-\left\lceil d\log(\sqrt{2}-1)\right\rceil/d}$. Руководствуясь соображением последующего применения леммы 2, создадим дополнительный запас сложности $1/d$ на втором сравнении и выберем в качестве α_2^* элемент чуть меньшего ранга:

$$(1-\lambda)\times 2^{a-2-\left\lceil d\log(\sqrt{2}-1)\right\rceil/d-1/d} > (1-\lambda)\frac{\sqrt{2}+1}{4}\times 2^{a-2/d}. \tag{26}$$

Тогда, используя неравенство $e^x \geq 1 + x$, получаем

$$|I_0^*| = |(\alpha_2^*, \alpha_1)| \leq (1-\lambda)\times 2^a\left(1 - \frac{\sqrt{2}+1}{4}\times 2^{-2/d}\right) \leq (1-\lambda)\times 2^a\left(\frac{3-\sqrt{2}}{4} + \frac{\sqrt{2}+1}{2d}\ln 2\right).$$

Следовательно,

$$0 \leq |I_0^*| - |I_0| \leq \frac{(3+2\sqrt{2})}{2d}\ln 2\,|I_1|. \tag{27}$$

При d достаточно большом, $|I_0^*| < |I_1|$.

Выполним перераспределение интервалов J_k внутри I_1. Введем параметр $\mu > 0$. Потребуем, чтобы удельная сложность вставки элемента в J_k^* была меньше сложности вставки в J_k в идеальной стратегии хотя бы на $\left(2 - \frac{k}{2^s}\right)\mu$ сравнений, $k = 1,2,\ldots,2^s$. Напомним, что в идеальной стратегии сложность оценивается как $\log[|J_k|/(1-\lambda)]$. Прибегать к неравномерности сужения вынуждает эффект смещения правых границ интервалов J_k. При равномерном сужении в $1+\mu$ раз, ввиду $|I_0| < |I_1|$, компенсация этого смещения, вообще говоря, была бы затруднена при малых k: сумма длин интервалов $H_{k,j}$ выросла бы примерно на $\mu|I_0|$, тогда как точка α'_{k+1} (правая граница интервала J_k) могла бы сместиться вправо приблизительно на $\mu|I_1|$. Поэтому для интервалов с большими номерами выбирается меньший коэффициент сужения.

Сужение. Мы собираемся сузить почти все интервалы J_k, но это сужение будет компенсировано расширением небольшого числа интервалов, длины которых близки к степеням двойки, с опорой на простую оценку (4). Сначала оценим сверху величину θ_p общего сужения интервалов с номерами от p до 2^s. Перед этим заметим, что для величины b_k выполняются простые соотношения (проверяются возведением в квадрат)

$$b + \frac{k}{2^s}b \leq b_k \leq b + \frac{k}{2^s}\frac{2b+1}{2b^2}. \tag{28}$$

С учетом дискретизации длину суженного интервала J_k^* можно выбрать в пределах

$$2^{-\left(2-\frac{k}{2^s}\right)\mu-\frac{1}{d}}|J_k| \leq |J_k^*| \leq 2^{-\left(2-\frac{k}{2^s}\right)\mu}|J_k|. \tag{29}$$

Исходя из пессимистического предположения, что все интервалы сужаются, снова используя неравенство $e^x \geq 1 + x$, а также (22) и (28), получаем

$$\theta_p \le \sum_{k=p}^{2^s}(|J_k| - |J_k^*|) \le \sum_{k=p}^{2^s}|J_k|\left[1 - 2^{-\left(2-\frac{k}{2^s}\right)\mu-\frac{1}{d}}\right] \le \sum_{k=p}^{2^s}|J_k|[(2-k2^{-s})\mu + 1/d]\ln 2 = \left(2\mu + \frac{1}{d}\right)\times$$

$$\times (b+1-b_{p-1})\ln 2|I_1| - \mu 2^{-s}\ln 2\sum_{k=p}^{2^s}k|J_k| = \left(2\mu+\frac{1}{d}\right)(b+1-b_{p-1})\ln 2|I_1| - \mu 2^{-2s}(2b+1)\ln 2\times$$

$$\times \sum_{k=p}^{2^s}\frac{k}{b_k + b_{k-1}}|I_1| \le \left[2\mu\left(1-\frac{p-1}{2^s}b\right) - \mu 2^{-2s}\frac{2b+1}{2(b+1)}\frac{2^{2s}-p^2}{2} + \frac{1}{d}\right]\ln 2|I_1| =$$

$$= \left[\left[2 - 2b\frac{p-1}{2^s} - \frac{2b+1}{4(b+1)}\left(1-\frac{p^2}{2^{2s}}\right)\right]\mu + \frac{1}{d}\right]\ln 2|I_1|.$$

В частности, для суммарного сужения всех интервалов J_k имеем оценку

$$\theta_1 \le (2\mu + 1/d)\ln 2|I_1|, \tag{30}$$

а при $p \ge 2$ — оценку

$$\theta_p \le \left[\left[2 - \frac{2b+1}{4(b+1)}\right]\mu + \frac{1}{d}\right]\ln 2|I_1|, \tag{31}$$

поскольку $p^2/2^s \le p \le 2(p-1)$ и $\dfrac{2b+1}{4(b+1)} < b = \dfrac{2\sqrt{2}+1}{4}$.

Расширение. Теперь оценим снизу величину возможной компенсации. Если

$$(1-\lambda)\times 2^{l+2\mu} \le |J_k| < 2^l, \quad l \in \mathbb{N},$$

то сложность вставки элемента пары в интервал J_k в идеальной стратегии пока оценивалась как минимум в $l + 2\mu$. Тем не менее длину интервала можно увеличить до 2^l, при этом уменьшив оценку сложности по меньшей мере на 2μ, что и требуется.

Сначала проверим, что если λ и μ не слишком велики, то при некотором $l \in \mathbb{N}$ справедливо

$$|J_{2^s}| \le (1-\lambda)\times 2^{l+2\mu} < 2^l \le |J_1|. \tag{32}$$

При $\mu \ll \lambda$ для этого достаточно выполнения условия $|J_1|/|J_{2^s}| \ge \dfrac{2}{1-\lambda}$. Используя (22) и (28), выводим

$$\frac{|J_1|}{|J_{2^s}|} = \frac{b_{2^s} + b_{2^s-1}}{b_1 + b_0} \ge \frac{2b+2-\frac{b}{2^s}}{2b + \frac{2b+1}{2b^2\times 2^s}} \ge \frac{b+1}{b}\left[1 - \frac{b}{2(b+1)\times 2^s}\right]\left(1 - \frac{2b+1}{4b^3\times 2^s}\right) > \frac{b+1}{b}(1-2^{1-s}),$$

где также использовались простые неравенства $\dfrac{1}{1+x} \ge 1-x$ и $(1-x)(1-y) \ge 1-x-y$, справедливые при $x, y \ge 0$. Подстановкой численных значений легко проверяется, что полученная оценка превосходит $\dfrac{2}{1-\lambda}$, скажем, при любых $s \ge 10$ и $\lambda \le 1/50$.

Итак, некоторый отрезок $g = [(1-\lambda)\times 2^{l+2\mu}, 2^l]$ лежит внутри отрезка $\left[|J_{2^s}|, |J_1|\right]$. Оценим возможное увеличение длины только для тех интервалов J_k, длины которых оказались в g. Напомним, что длины интервалов монотонно убывают с ростом k (см. (22)). Длину отрезка g оценим снизу с помощью (22) и справедливого при $0 \le x \le 1$ неравенства $2^x \le 1 + x$ как

$$[1-(1-\lambda)2^{2\mu}]\times 2^l \ge [1-(1-\lambda)(1+2\mu)]|J_{2^s}| \ge L = (\lambda - 2\mu)\frac{2b+1}{2(b+1)\times 2^s}|I_1|. \tag{33}$$

Теперь оценим сверху разность длин соседних интервалов:

$$\frac{2^s}{(2b+1)|I_1|}(|J_k|-|J_{k+1}|)=\frac{1}{b_k+b_{k-1}}-\frac{1}{b_{k+1}+b_k}\le\frac{b_{k+1}-b_{k-1}}{4b^2}=\frac{2(2b+1)}{4b^2(b_{k+1}+b_{k-1})\times 2^s}\le\frac{2b+1}{4b^3\times 2^s}.$$

Следовательно, при любом k

$$|J_k|-|J_{k+1}|\le\Delta=\frac{(2b+1)^2}{4b^3\times 2^{2s}}|I_1|. \tag{34}$$

Легко проверить, что если множество точек делит отрезок длины L на подотрезки с длинами не более Δ, то сумма расстояний от этих точек до границы отрезка не меньше $\frac{L(L-\Delta)}{2\Delta}$. Подставляя в эту формулу значения L и Δ, приходим к оценке

$$\frac{L(L-\Delta)}{2\Delta}\ge\left[(\lambda-2\mu)^2\frac{b^3}{2(b+1)^2}-(\lambda-2\mu)\frac{2b+1}{4(b+1)\times 2^s}\right]|I_1|=$$

$$=(\lambda-2\mu)\left[\lambda-2\mu-\frac{(2b+1)(b+1)}{2b^3\times 2^s}\right]\frac{b^3}{2(b+1)^2}|I_1|>(\lambda-2\mu)(\lambda-2\mu-2^{2-s})\frac{b^3}{2(b+1)^2}|I_1|.$$

Так мы оценили снизу максимум суммарного возможного расширения интервалов J_k с длинами из отрезка g (до длины 2^l).

Компенсация сужения остальных интервалов J_k осуществима при $\frac{L(L-\Delta)}{2\Delta}\ge\theta_1$. Используя (30), достаточное для этого условие можно теперь записать как

$$(\lambda-2\mu)(\lambda-2\mu-2^{2-s})\frac{b^3}{2(b+1)^2}\ge\left(2+\frac{1}{d}\right)\ln 2\mu. \tag{35}$$

Пока ограничимся тривиальным замечанием: при достаточно большом s можно задать параметры в виде[7] $d\asymp 2^s$, $\lambda\asymp 2^{-s/2}$, $\mu\asymp 2^{-s}$ так, что (35) будет удовлетворено.

Сведение. Пришло время проверить, позволяют ли предпринятые меры компенсировать дефицит длины интервалов $H_{k,j}$, отраженный в соотношении (25). Обратим внимание на изменение условий, в которых выполнялась оценка (25). Для величины $Y(a-s-3/2)$ используется прежняя оценка, но при этом произошло удлинение интервала I_0 (см. (27)), а правая граница интервала J_k могла сместиться правее на величину, оцененную сверху как θ_{k+1} (см. (31)).

Таким образом, мы требуем, чтобы при любом k прирост суммарной длины интервалов $H_{k,j}$ составил не менее

$$|I_0^*|-|I_0|+\theta_{k+1}+\frac{\ln 2}{2^s}|I_1|+\delta|I_1|, \tag{36}$$

где $\delta|I_1|$ – дополнительный запас, который потребуется позже для применения леммы 2.

По построению, $|H_{k,j}^*|\ge|H_{k,j}|\times 2^{\left(2-\frac{k}{2^s}\right)\mu-\frac{1}{d}}$ с учетом поправки на дискретизацию. Поэтому, подставляя (24) и используя (28), получаем

$$\sum_{j=1}^{s+1}(|H_{k,j}^*|-|H_{k,j}|)\ge\left[2^{\left(2-\frac{k}{2^s}\right)\mu-\frac{1}{d}}-1\right]\sum_{j=1}^{s+1}|H_{k,j}|\ge\left[\left(2-\frac{k}{2^s}\right)\mu-\frac{1}{d}\right]\ln 2\left(1-\frac{1}{2^{s+1}}\right)\frac{b_k+b_{k-1}}{2}|I_1|\ge$$

$$\ge\left(2-\frac{k}{2^s}\right)\left(1-\frac{1}{2^{s+1}}\right)\left(1+\frac{k-\frac{1}{2}}{2^s}\right)\mu b\ln 2|I_1|-\frac{b+1}{d}\ln 2|I_1|\ge$$

$$\ge\left(2+\frac{k-2}{2^s}-\frac{k^2}{2^{2s}}\right)\mu b\ln 2|I_1|-\frac{b+1}{d}\ln 2|I_1|\ge\left[(2-2^{1-s})\mu b-\frac{b+1}{d}\right]\ln 2|I_1|.$$

[7] Символ $\asymp$ означает равенство порядков роста.

Используя эту оценку, а также (27) и (31), достаточное для выполнения (36) условие можно записать как

$$(2-2^{1-s})\mu b-\frac{b+1}{d}\ge\frac{(3+2\sqrt{2})}{2d}+\left[2-\frac{2b+1}{4(b+1)}\right]\mu+\frac{1}{d}+\frac{1}{2^s}+\delta\log e. \tag{37}$$

Положим $d=2^s$ и $\delta=2^{-s}$. Тогда, чтобы удовлетворить (37), можно выбрать $\mu=30\times2^{-s}$ (принимаем, что $s\ge10$). Условие (35) при этом будет выполнено, например, если $\lambda=20\times2^{-s/2}+64\times2^{-s}$. Обеспечивающее (32) частное условие $\lambda\le1/50$ выполнено при любом $s\ge20$.

Описание стратегии. Фиксируя выбранные значения параметров d, δ, μ, λ, получаем стратегию вставки с системой ограничений, которую обозначим через $\sigma[s]$. Стратегия заключается в том, что элементы главной цепочки, участвующие в сравнениях, разбивают цепочку на интервалы, вставка в которые выполняется с такой сложностью, которая получается в вышеприведенном расчете. Опишем систему ограничений стратегии.

Пусть $p(a)$ является оценкой длины интервала с удельной сложностью вставки пары $2a$ сравнений, а $p_c(J)$ – оценкой длины интервала с удельной сложностью вставки элемента на c выше, что и в некоторый интервал J из построенного ранее разбиения. Для вставки в некоторые интервалы J мы выбираем оценки типа $Q(a)$, а для остальных вставок – оценки типа $p(a)$ при подходящих a.

Первое сравнение, в силу выбора элемента α_1, не порождает ограничений. Для второго сравнения вводится неравенство

$$p(a)\le p(a-1/2)+p_{1/d}(I_1). \tag{38}$$

Здесь мы используем запас сложности, обеспеченный выбором элемента α_2^* (см. (26)), и можем позволить вставку в интервал I_1 с чуть большей удельной сложностью. Это понадобится только для формального удовлетворения условиям применения леммы 2.

Следующие s сравнений, локализующие β_1 в одном из интервалов J_k^*, относятся к внутренним вершинам дерева сравнений. Для того чтобы обеспечить в дальнейшем вставку элемента β_1 в любой из этих интервалов с заданной сложностью, вводится ограничение

$$p(a)\le p(a-1/2)+\sum_{k=1}^{2^s}p_0(J_k^*). \tag{39}$$

Ограничение всего одно, так как длины интервалов J_k^*, $k>1$, мы можем задать свободно, выбирая подходящим образом пограничные элементы $\alpha_k'^*$, тогда только длина интервала J_1^* будет определяться длиной остальных интервалов.

Длины интервалов $H_{k,j}^*$ тоже устанавливаются свободно (выбором элементов для сравнений), и лишь вставки после $2s+3$ сравнений порождают ограничения вида

$$p(a)\le p_0(I_{-1})+\sum_{j=1}^{s+1}p_0(H_{k,j}^*)+p(a-s-3/2)+\sum_{i=k+1}^{2^s}p_0(J_i^*), \tag{40}$$

где $I_{-1}=(-\infty,\alpha_2^*)$.

Таким образом, система ограничений $\sigma[s]$ складывается из (38)–(40) при всевозможных $a=r+i/d$. Для перехода к форме записи из определения стратегии следует в указанных неравенствах заменить выражения $p(x)$ и $p_c(J)$ величинами $p_{r,i}$ или числовыми значениями в случаях, когда применяется оценка $Q(a)$.

Глубина стратегии с системой ограничений $\sigma[s]$ равна $2s+3$. Проверим, что ширина (стратегии и, следовательно, системы $\sigma[s]$) не превосходит $s+4$. Действительно,

$$\frac{|I_1|}{p(a)}=\sqrt{2}-1>\frac{1}{2^2},\quad \frac{|I_{-1}|}{p(a)}\ge\frac{\sqrt{2}+1}{4\times2^{2/d}}>\frac{1}{2},$$

согласно (26) (напомним, что $p(a)=Y(a)$). Далее, в силу (29) и (22)

$$\frac{|J_k^*|}{p(a)}\ge\frac{|J_k|}{p(a)\times2^{2\mu+1/d}}\ge\frac{\sqrt{2}+1}{(b+1)\times2^{s+2+2\mu+1/d}}>\frac{1}{2^{s+2}}.$$

Наконец, из (23) следует

$$\frac{|H^*_{k,s+1}|}{p(a)} \ge \frac{|H_{k,s+1}|}{p(a)} \ge \frac{b}{2^{s+1}}\frac{|I_1|}{p(a)} > \frac{1}{2^{s+3}}.$$

Таким образом, если $p(a) = p_{r',i'}$, то при перезаписи (38)–(40) в терминах величин $p_{r,i}$, среди этих величин будут встречаться только такие, для которых $r \ge r' - s - 3$.

Лемма 3. *Пусть* $s \ge 20$, $d = 2^s$, $\mu = 30\times 2^{-s}$, $\lambda = 20\times 2^{-s/2} + 64\times 2^{-s}$ *и* $\varepsilon = \dfrac{1}{25\times 2^s}$. *Пусть также* $y_i = i/d$ *и* $v_i = (1-\lambda)\times 2^{y_i+1/2}$. *Тогда*

(i) $v = (v_0,\ldots,v_{d-1}) \in T(\sigma[s])$;

(ii) $u = v/2 \in T^{\varepsilon}(\sigma[s])$.

Доказательство. Первая часть леммы уже доказана, так как сама стратегия подогнана под заданное решение. Остается проверить (ii).

Напомним, что при переходе к канонической системе (9) мы переписываем неравенства в терминах переменных $x_{r,i} = p_{r,i}\times 2^{-r}$ и выполняем нормировку, а для перехода к приведенной системе (10) производится подстановка $x_{r,i} = z_i$.

Пусть $p(a) = p_{r',i'}$. Выполним предварительную нормировку неравенств (38)–(40) домножением на $2^{-r'}$ и перепишем их в терминах величин $x_{r,i}$. Полученные неравенства обозначим через (38')–(40'). Оценим сумму абсолютных величин коэффициентов при $x_{r,i}$ в каждом из этих неравенств для определения итоговых коэффициентов нормировки.

Вклад от слагаемого $p(a)$ в каждую сумму коэффициентов равен 1. Тогда вклад любого из слагаемых $p(a-1/2)$, $p_{1/d}(I_1)$, $p_0(I_{-1})$ и $p_0(H_{k,1})$ не выше 1, вклад $p(a-s-3/2)$ не превосходит 2^{-s-1}, при любом j вклад слагаемого $p_0(H_{k,j+1})$ вдвое меньше, чем у $p_0(H_{k,j})$. Поскольку $|J^*_k| \le |J_1| \le 2^{a-s}$, согласно (22), то вклад каждого из слагаемых $p_0(J^*_k)$ не превосходит 2^{-s}. Эти оценки означают, что суммы коэффициентов при $x_{r,i}$ в неравенствах (38'), (39') и (40') не превосходят соответственно 3, 3 и 5.

Теперь оценим запас, с которым вектор u удовлетворяет приведенным неравенствам (38')–(40'). По построению для основного решения v приведенное неравенство (38') выполнялось бы и при замене $p_{1/d}(I_1)$ на $p_0(I_1)$ в (38). Таким образом, правая часть (38) при $p_{r,i} = v_i\times 2^r$ превосходит левую на величину не менее

$$p_{1/d}(I_1) - p_0(I_1) \ge |I_1|(2^{1/d}-1) \ge \frac{\ln 2}{d}|I_1| = (1-\lambda)(\sqrt{2}-1)\ln 2\times 2^{a-s} > 2^{a-s-2}.$$

После нормировки (делением не более, чем на 3×2^a) и подстановки u вместо v величина запаса в приведенном неравенстве (38') сокращается не более, чем в 6×2^a раз, и остается не меньшей, чем $\dfrac{1}{24\times 2^s} > \varepsilon$.

Оценим свободный член, возникающий в (39') из-за того, что для сложности вставки в некоторые интервалы J^*_k используется оценка $Q(a)$. Число таких интервалов не меньше $\lfloor L/\Delta \rfloor$ (см. выше), а применяемая оценка имеет величину $2^l \ge |J_{2^s}| \ge 2^{a-s-2}$, согласно (32) и (22). Поэтому, используя значения L и Δ из (33) и (34), заключаем, что свободный член в правой части (39') не меньше

$$\frac{L-\Delta}{\Delta\times 2^{s+2}} = (\lambda - 2\mu)\frac{b^3}{2(b+1)(2b+1)} - \frac{1}{2^{s+2}} > \frac{1}{2^{s/2}}.$$

При подстановке u вместо v в приведенное неравенство (39') все члены сокращаются вдвое, тогда и свободный член можно уменьшить наполовину, т.е. на $\dfrac{1}{6\times 2^{s/2}} > \varepsilon$ с учетом нормировки.

Каждое из неравенств (40) при $p_{r,i} = v_i\times 2^r$ выполняется с заложенным в сумму длин интервалов $H^*_{k,j}$ запасом

$$\delta|I_1| = (1-\lambda)\delta(\sqrt{2}-1)\times 2^a > 2^{a-s+1}/5$$

(см. (36)). При нормировке (делением не более, чем на 5×2^a) и подстановке u вместо v запас сокращается не более чем в 10×2^a раз, до величины не менее $\varepsilon = \dfrac{1}{25\times 2^s}$.

Следовательно, при уменьшении свободных членов в приведенных неравенствах (38')–(40') на ε вектор u остается решением системы.

6. СОРТИРОВКА

Теорема 1. *При любом $a \ge 2$ и $2^{a/4} \le m \le 2^{a/2}$ выполнено*

$$P_m(a) \ge 2^{a+1/2} - O(a^{-\gamma}\times 2^a), \tag{41}$$

где γ мало, например, $\gamma = \dfrac{1}{5}$.

Доказательство. Воспользуемся результатом леммы 3 и применим лемму 2. Значения всех параметров заимствуются из леммы 3. Имеем $v \in T(\sigma[s])$ и $u \in T^{\varepsilon}(\sigma[s])$ для системы ограничений $\sigma[s]$ глубины $h = 2s+3$ и ширины $w \le s+4$. При этом справедливо $u_i < 2^{y_i - 1/2} < \pi(y_i)$ (см. (5)). Кроме того, $\varepsilon < 1-\lambda < R = \|v-u\| < \sqrt{2}$. При $a < 100$ неравенство (41) заведомо выполнено при достаточно больших значениях константы под знаком "O". Поэтому считаем, что $a \ge 100$. Положим $\varphi = \log\left(\dfrac{m\varepsilon}{4a^2}\right)$ и $s = \lceil 2\gamma \log a \rceil$. Если $\gamma \le 1/4$, то при этом $m\varepsilon \ge 2^{a/4}/(50a^{2\gamma}) \ge 4a^2$, следовательно, $\varphi > 0$. Условие (16) леммы 2 выполнено автоматически.

Пусть $r + y_i \le a - \dfrac{(a+2)\varphi\times 2^{\varphi}}{m} \le r + y_i + \dfrac{1}{d}$, где $r \in \mathbb{N}$. С помощью леммы 2 получаем оценку

$$P_m(a) \ge P_m\left[r + y_i + \frac{(r+2)\varphi\times 2^{\varphi}}{m}\right] \ge v_i\times 2^r - Rm\times 2^{\varphi+2}/\varepsilon - R\times 2^r e^{-\frac{\varepsilon}{2R}\left(\frac{\varphi}{w(d+1)(h+1)}-2\right)} \ge$$

$$\ge (1 - O(2^{-s/2}))\times 2^{a+\frac{1}{2}-\frac{(a+2)\varphi\times 2^{\varphi}}{m}-\frac{1}{d}} - O(m^2/a^2) - O\left(2^a e^{-\Theta(as^{-2}\times 2^{-2s})}\right) =$$

$$= (1 - O(2^{-s/2}))\times 2^{a+1/2-O(2^{-s})} - O(2^a/a^2) - 2^{a-O(as^{-2}\times 2^{-2s})} = 2^{a+1/2}(1 - O(a^{-\gamma})),$$

если, скажем, $\gamma \le 1/5$.

Следствие 1. При любом $a \ge 2$ и $2^{a/4} \le m \le 2^{a/2}$ выполнено

$$Q_{2m}(a) \ge 2^a - O(a^{-1/5}\times 2^a).$$

Реализованное доказательство главного технического результата (теорема 1) довольно громоздко. Поэтому для большей ясности приведем неформальное резюме схемы рассуждения.

1. Построим некоторую идеальную серию сравнений, всякий раз выполняя деление множества исходов пополам. К сожалению, она не дает сведения к подобной задаче меньшей размерности.

2. Отступление от идеала (поправочный коэффициент λ) позволяет разбалансировать длины интервалов, используя оценку (4) для сложности вставки, близкой к целому числу, и получить сходящийся рекурсивный алгоритм.

3. Дискретизация (параметр d) вводится, чтобы перейти от континуального к конечному семейству алгоритмов. В приведенном доказательстве шаги 2 и 3 совмещены. В результате мы имеем алгоритм с хорошей оценкой сложности, но который рекурсивно сводится к себе самому.

4. Обеспечим прогрессивный рост качества оценки сложности с увеличением размерности задачи, стартуя из точки, в которой оценка тривиально выполняется (лемма 2).

С технической стороны доказательство состоит в контролируемом учете поправок из трех источников: неравномерности разбиения, дискретизации и малой размерности.

Отметим, что даже упрощенный вариант стратегии из разд. 5 с выбором $s = 1$ и глубиной 6, только не универсальный, а с подбором оптимальных разбиений при каждом i, позволил бы улучшить результаты метода бинарных вставок при всех достаточно больших n. Однако для по-

лучения более точной оценки приходится выбирать параметр s растущим. Главный содержательный результат формулирует следующая

Теорема 2. *При любых n выполняется равенство*

$$S(n) = \log(n!) + O(n\log^{-1/5} n).$$

Доказательство. Начинаем как в методе бинарных вставок. Разобьем набор из n элементов на пары, упорядочим их и выполним сортировку старших элементов пар (они образуют главную цепочку). На это требуется $n/2 + S(n/2)$ сравнений. Пусть младшие элементы пар обозначаются через α_i с нумерацией в порядке возрастания ранга старших элементов в главной цепочке (фиг. 1).

Вставим первые $n_0 = n/\log n$ из младших элементов в главную цепочку методом бинарных вставок, потратив на это $n + O(n_0)$ сравнений. Оставшиеся элементы разобьем на группы по $m \approx \sqrt{n}/\log n$ штук: группа с номером k содержит элементы $\alpha_{n_0+(k-1)m+1}, \alpha_{n_0+(k-1)m+2}, \ldots, \alpha_{n_0+km}$. Группы вставляются в главную цепочку по очереди в порядке возрастания номеров методом следствия 1. Последняя группа может содержать менее m элементов – их вставляем тривиальным бинарным методом.

Группа с номером t должна быть вставлена в интервал длины $L_t = 2(n_0 + tm) - m$. Согласно следствию 1, при некотором $\rho_t = \log L_t + O(\log^{-1/5} n)$ выполнено $Q_m(\rho_t) \geq L_t$. Поэтому сложность вставки t-й группы не превосходит $\rho_t m$.

Представим идеализированный случай, когда элемент α_k мог бы быть вставлен в главную цепочку за $\log(2k-1)$ сравнений. Тогда, если вставлять элементы в порядке возрастания номеров, было бы затрачено $\log\prod_{k=1}^{\lceil n/2\rceil}(2k-1) = \log(n!) - \log((n/2)!) - n/2 + O(\log n)$ сравнений.

В групповом методе мы тратим на вставку элемента α_k не более $\log(2k+m) + O(\log^{-1/5} n) = \log(2k-1) + O(\log^{-1/5} n)$ сравнений при $n_0 < k < n/2 - m$ и не более $\log(2k-1) + O(1)$ сравнений при $k \leq n_0$ и $k \geq n/2 - m$. Общее превышение сложности вставки по отношению к идеализированной ситуации тогда можно оценить как

$$O(n_0 + m) + O((n/2 - n_0)\log^{-1/5} n).$$

Получаем рекуррентное соотношение

$$S(n) \leq n/2 + S(n/2) + \log(n!) - \log((n/2)!) - n/2 + O(n\log^{-1/5} n)$$

или, если выразить через величины $s(n) = S(n) - \log(n!)$,

$$s(n) \leq s(n/2) + O(n\log^{-1/5} n).$$

Отсюда следует $s(n) = O(n\log^{-1/5} n)$.

СПИСОК ЛИТЕРАТУРЫ

1. *Ахо А., Хопкрофт Дж., Ульман Дж.* Проектирование и анализ вычислительных алгоритмов. М.: Мир, 1979.
2. *Кнут Д.* Искусство программирования. Т. 3. Сортировка и поиск. М.: Вильямс, 2007.
3. *Mehlhorn K.* Data structures and algorithms. V. 1. Sorting and searching. Berlin, NY: Springer, 1984.
4. *Ford L.R., Johnson S.M.* A tournament problem // Amer. Math. Monthly. 1959. V. 66. № 5. P. 387–389.
5. *Peczarski M.* The Ford-Johnson algorithm still unbeaten for less than 47 elements // Inf. Process. Lett. 2007. V. 101. № 3. P. 126–128.
6. *Schulte Mönting J.* Merging of 4 or 5 elements with n elements // Theor. Comput. Sci. 1981. V. 14. P. 19–37.
7. *Manacher G.K., Bui T.D., Mai T.* Optimum combinations of sorting and merging // J. ACM. 1989. V. 36. № 2. P. 290–334.
8. *Iwama K., Teruyama J.* Improved average complexity for comparison-based sorting // Theor. Comput. Sci. 2020. V. 807. P. 201–219.
9. *Edelkamp S., Weiß A., Wild S.* QuickXsort: a fast sorting scheme in theory and practice // Algorithmica. 2020. V. 82. P. 509–588.
10. *Graham R.L.* On sorting by comparisons // in: Computers in Number Theory. London: Academic Press, 1971. P. 263–269.
11. *Hwang F.K., Lin S.* Optimal merging of 2 elements with n elements // Acta Inf. 1971. V. 1. P. 145–158.
12. *Schönhage A., Paterson M., Pippenger N.* Finding the median // J. Comp. Sys. Sci. 1976. V. 13. P. 184–199.
13. *Dor D., Zwick U.* Selecting the median // SIAM J. Comput. 1999. V. 28. № 5. P. 1722–1758.
14. *Christen C.* Improving the bounds for optimal merging // in: Proc. 19th IEEE Conf. on Found. of Comput. Sci. (Ann Arbor, USA, 16–18 October 1978). NY: IEEE, 1978. P. 259–266.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 2, с. 363–366

ПАМЯТИ ИВАНА СТАНИСЛАВОВИЧА МЕНЬШИКОВА (1952–2020)

DOI: 10.31857/S0044466921020101

26 апреля 2020 г. скоропостижно скончался Иван Станиславович Меньшиков, пионер экспериментальной экономики и экспериментальной теории игр в России.

И.С. Меньшиков – выпускник знаменитой московской математической школы № 2. Теорию игр он изучал на факультете Вычислительной математики и кибернетики МГУ под руководством Юрия Борисовича Гермейера, основателя кафедры Исследования операций на ВМК и отдела Исследования операций в Вычислительном Центре АН СССР. Окончив с отличием МГУ в 1974 г., Иван Станиславович поступил в аспирантуру к Гермейеру, а после его кончины в 1975 г. перешел под руководство будущего академика Павла Сергеевича Краснощекова, преемника Ю.Б. Гермейера на кафедре. Успешно защитив в 1978 г. диссертацию по теории игр, в работе над которой большую помощь ему оказал будущий доктор наук Николай Серафимович Кукушкин, И.С. Меньшиков до конца жизни работал в ВЦ РАН – младшим, старшим, ведущим научным сотрудником.

С начала 1980-х годов Иван Станиславович принял активное участие в организованных академиком Никитой Николаевичем Моисеевым исследованиях по теоретико-игровому моделированию экологических проблем на основе изучения модели Гермейера–Вателя. Результаты были опубликованы в ведущих журналах АН СССР [1]–[3]. От применения теории игр в экологических проблемах исследователи перешли к изучению международных отношений в целом, в частности, актуальной в это время тематике “звездных войн”. В 1991 г. была подготовлена большая коллективная работа на эти темы, завершающим аккордом которой стала модель, обосновывающая создание совместной советско-американской системы противоракетной обороны [4]. Теоретико-игровые основы проведенных исследований были рассмотрены в популярных книгах, написанных И.С. Меньшиковым вместе с коллегами [5], [6].

Большое влияние на исследования И.С. Меньшикова оказало его сотрудничество с зарубежными коллегами, которое началось в 1981 г., когда он выступил с докладом на международной конференции по теории игр, состоявшейся в Австрии в Международном институте прикладного системного анализа (International Institute for Applied Systems Analysis, IIASA). На этой конферен-

ции он познакомился с французским специалистом по теории игр Эрве Муленом. Это знакомство переросло в многолетнее сотрудничество, в результате которого на русский язык им были переведены две книги Э. Мулена. В 1990 г. Э. Мулен, который к тому времени работал в США, организовал поездку И.С. Меньшикова и его коллеги В.А. Гурвича в США, в рамках которой они приняли участие в экономической конференции в Университете Норт-вестерн, а также посетили несколько ведущих американских университетов и выступили там с докладами. В Университете Карнеги-Меллон их принимал проф. Санджей Шривастава, в Гарварде – будущий лауреат Нобелевской премии (2007 г.) проф. Эрик Маскин, в Университете Дьюк в Дареме – проф. Эрве Мулен. На конференции в университете Норт-вестерн Иван Станиславович прослушал лекции по экспериментальной экономике Вернона Смита (будущего нобелевского лауреата, 2002 г.) и выдающегося исследователя, одного из родоначальников экспериментальной экономики Чарльза Плотта, профессора Калифорнийского технологического института. Иван Станиславович установил научные контакты с Ч. Плоттом, который передал ему программное обеспечение для проведения сетевых компьютерных экспериментов как с учебной, так и с исследовательской целями.

Это поездка сыграла ключевую роль в дальнейшей работе И.С. Меньшикова, который сразу оценил актуальность лабораторных экономических экспериментов для России в условиях становления рыночных отношений. Группа сотрудников ВЦ РАН под руководством Ивана Станиславовича начала проводить лабораторные эксперименты на небольшой сети из нескольких персональных компьютеров. Поиски настоящего сетевого компьютерного класса привели Ивана Станиславовича в 1991 г. в Академию народного хозяйства при Совете Министров СССР, где при активном содействии ректора АНХ академика Абела Гезевича Аганбегяна была открыта первая в нашей стране Лаборатория экспериментальной экономики, научным руководителем которой стал И.С. Меньшиков. Коллектив лаборатории, сформировавшийся в ВЦ АН СССР, сохранял научные контакты с будущим академиком, А.А. Петровым и его отделом в ВЦ РАН. Соображения о том, что методы экспериментальной экономики, помимо учебной работы, могут активно использоваться при планировании реформ, горячо поддерживались А.А. Петровым. Но ни коллективу лаборатории, ни А.А. Петрову, не удалось привлечь внимание тех, от кого зависело проведение реформ. Так, с самого начала деятельности лаборатории Иван Станиславович предлагал проводить целенаправленные лабораторные исследования по совершенствованию экономических механизмов в России, в частности, механизма размещения государственных ценных бумаг (ГКО) в интересах стабилизации финансовой системы страны. Многократные обращения к руководству ЦБ России не привели к какому-либо прогрессу в этой области, поскольку чиновникам всегда казалось, что такие эксперименты “не ко времени”. Таким образом, хотя научные исследования на основе лабораторных компьютерных экспериментов продолжались (результаты опубликованы в статье [7]), их использование в учебных целях стало главным направлением деятельности И.С. Меньшикова в 1990-х годах.

На проходившем в Москве, в августе 1992 г., Десятом всемирном конгрессе Международной экономической ассоциации, Иван Станиславович принял предложение профессора С. Шривастава включиться в международный проект по обучению финансовым рынкам методами экспериментальной экономики FAST – Financial Analysis and Security Trading (Финансовый анализ и торговля ценными бумагами). При АНХ был создан FAST-Центр, научным руководителем которого с 1993 по 2003 г. был Иван Станиславович. В 1993, 1995, 2000 г. Иван Станиславович проходил стажировки в США, где получил сертификат Университета Карнеги Меллон, подтверждающий право обучать по программам FAST-1 и FAST-2. Иван Станиславович быстро сумел овладеть методами преподавания математики финансовых рынков и рынков ценных бумаг. Он читал лекции по программе FAST, проводил семинарские занятия, готовил учебные пособия. Всего за 10 лет по программе FAST были обучены более 3000 человек, которые высоко оценили эту программу. На основе этой деятельности И.С. Меньшиков подготовил курс лекций [8].

В 2003 г. И.С. Меньшиков стал руководителем Лаборатории экспериментальной экономики при кафедре анализа систем и решений МФТИ, созданной по инициативе академика А.А. Петрова и декана Факультета управления и прикладной математики МФТИ члена-корр. РАН А.А. Шананина. Иван Станиславович преподавал в МФТИ с 1980 г., когда академик Н.Н. Моисеев привлек его к чтению лекций по теории игр. С образованием лаборатории И.С. Меньшиков организовал в рамках экономико-математической специализации ФУПМ МФТИ курс лекций и компьютерных лабораторных работ по экспериментальной экономике, пользующийся большим успехом у студентов. Прочитанные лекции послужили основой учебных пособий [9],]10]. Возможности лаборатории позволили Ивану Станиславовичу сочетать преподавание с эксперимен-

тальным исследованием поведенческих аспектов принятия решений. Под его руководством было проведено огромное количество лабораторных экспериментов по изучению особенностей поведения в различных экономических ситуациях, изучалось влияние ограниченной рациональности, социальной общности, психологических особенностей экономических агентов на их поведение [11]–[13]. Обобщая результаты проведенных экспериментов, И.С. Меньшиков разработал новые концепции равновесия в играх с неполной информацией, которые успешно объясняли сложные аспекты поведения участников, не укладывающиеся в классические принципы рациональности [14]–[16]. Его совершенно оригинальной идеей было создание в лаборатории психофизиологического кабинета, где с помощью специального оборудования, стабилографических кресел стало возможным изучать процесс принятия решений каждым участником в реальном времени, одновременно наблюдая за его состоянием [17]–[20]. Надо подчеркнуть, что тематика экспериментальной экономики входила в госзадание отдела Математического моделирования экономических систем ВЦ РАН, в котором в 2010-е годы Иван Станиславович работал под руководством члена-корр. РАН Игоря Гермогеновича Поспелова, всемерно поддерживавшего эти исследования.

В последние годы И.С. Меньшиков разработал и внедрил принципиально новые способы проведения лекций по теории игр, которые теперь включали 15-минутные лабораторные игры с применением беспроводной локальной сети, в которых студенты могли участвовать, используя смартфон или любое другое устройство для работы с браузером. Это позволило легко перейти на удаленный режим обучения, когда возникла такая необходимость, и проводить игры дистанционно. Лекции Иван Станиславович читал буквально до последних дней.

И.С. Меньшиков руководил многочисленными проектами по экспериментальной экономике, поддержанными РФФИ. Он уделял много сил и времени работе со своими учениками, охотно работая с каждым, кто проявлял интерес к его области знаний. За время работы на Физтехе через лабораторию прошли сотни студентов ФУПМ, многие из которых выбирали Ивана Станиславовича своим научным руководителем. Его ученики работают в различных областях и разных странах. Семь его аспирантов успешно защитились в России, а три его ученика защитили диссертации за рубежом.

Иван Станиславович был всегда открыт для общения, охотно делился своими идеями с коллегами, учениками, студентами. Его уход – огромная потеря не только для Физтеха и Вычислительного центра, но и в масштабах всей страны. Память об этом светлом человеке надолго сохранится в сердцах коллег и учеников.

СПИСОК ОСНОВНЫХ ТРУДОВ И.С. МЕНЬШИКОВА

1. *Меньшиков И.С., Меньшикова О.Р.* Сильные ситуации равновесия и N-ядро в играх с иерархическим вектором интересов // Ж. вычисл. матем. и матем. физ. 1985. Т. 25. № 9. С. 1304–1313.
2. *Кукушкин Н.С., Меньшиков И.С., Меньшикова О.Р., Моисеев Н.Н.* Устойчивые компромиссы в играх со структурированными функциями выигрыша // Ж. вычисл. матем. и матем. физ. 1985. Т. 25. № 12. С. 1761–1776.
3. *Кукушкин Н.С., Меньшиков И.С., Меньшикова О.Р., Моисеев Н.Н.* Об одном классе теоретико-игровых конструкций, представляющих интерес для экологии // Докл. АН СССР. 1986. Т. 287. № 5. С. 1044–1046.
4. *Кукушкин Н.С., Меньшиков И.С., Меньшикова О.Р., Моисеев Н.Н.* Математические модели и теория “институтов согласия” // Моделирование процессов мирового развития и сотрудничества / Под ред. Д.М. Гвишиани, Е.П. Велихова, В.М. Лейбина. М.: Наука, 1991. С. 160–199.
5. *Кукушкин Н.С., Меньшиков И.С., Меньшикова О.Р.* Конфликты и компромиссы. М.: Знание, 1986. Сер. “Математика, Кибернетика”. 32 с.
6. *Гурвич В.А., Меньшиков И.С.* Институты согласия. М.: Знание, 1989. Сер. “Математика, Кибернетика”. 48 с. ISBN 5-07-000613-4
7. *Menshikov I.S., Plott C., Men’shikova O., Myagkov M.* From non-market attitude to market behavior. Experimental study // Economic Theory. USA, 1996. P. 1–50.
8. *Меньшиков И.С.* Финансовый анализ ценных бумаг: курс лекций. М.: Финансы и статистика, 1998. 352 с. ISBN 5-279-01824-4
9. *Меньшиков И.С.* Лекции по теории игр и экономическому моделированию: учеб. пос. для студ. ВУЗов по направлению “Прикладная математика и физика”. М.: МЗ Пресс, 2006. 207 с. ISBN 5-94073-099-X
10. *Меньшиков И.С.* Лекции по теории игр и экономическому моделированию. М.: МЗ Пресс, 2010. ISBN 978-5-86567-093-3

11. *Menshikov I.S., Shklover A.V., Babkina T.S., Myagkov M.G.* From rationality to cooperativeness: the totally mixed Nash equilibrium in Markov strategies in the iterated prisoner dilemma // PLoS ONE. 2017. Т. 12. № 11.
12. *Berkman E.T., Lukinova E., Myagkov M., Menshikov I.* Sociality as a natural mechanism of public goods provision // PLoS ONE. 2015. Т. 10. № 3.
13. *Lukinova E., Babkina T., Myagkov M., Sedush A., Menshikov I., Menshikova O.* Sociality is not lost with monetary transactions within social groups. In: Experimental Economics and Machine Learning. Proceedings of the Fourth Workshop on Experimental Economics and Machine Learning. 2017. С. 18–30.
14. *Селютин В.А., Меньшиков И.С.* Сравнение поведенческих концепций равновесия на примере игры "11-20" // Труды Московского физико-технического института. 2019. Т. 11. № 1 (41). С. 53–61.
15. *Меньшиков И.С., Меньшикова О.Р., Чабан А.Н., Бойко Д.К., Старков Д.М.* Лабораторный анализ социальных процессов принятия решений // Труды Московского физико-технического института. 2017. Т. 9. № 3 (35). С. 86–97.
16. *Меньшиков И.С., Платонов В.В.* Игровые модели сетевых аукционов и их лабораторные исследования // Матем. моделирование. 2009. Т. 21. № 8. С. 63–79.
17. *Бурнаев Е.В., Меньшиков И.С.* Модель функционального состояния участников лабораторных рынков // Известия Российской академии наук. Теория и системы управления. 2009. № 6. С. 159–176.
18. *Меньшиков И.С., Меньшикова О.Р.* Лабораторный анализ процесса принятия экономических решений на основе комплекса стабилографических кресел // Известия ЮФУ. Технические науки. 2008. № 6 (83). С. 162–165.
19. *Лукьянов В.И., Максакова О.А., Меньшиков И.С., Меньшикова О.Р., Сенько О.В., Чабан А.Н.* Функциональное состояние и эффективность участников лабораторных рынков // Известия Российской академии наук. Теория и системы управления. 2007. № 6. С. 150–166.
20. *Меньшиков И.С.* Анализ функционального состояния участников финансовых рынков // Психология. Журнал Высшей школы экономики. 2009. Т. 6. № 2. С. 125–152.