

ISSN 0044-4669

Том 63, Номер 9

Сентябрь 2023



ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ

www.sciencejournals.ru



СОДЕРЖАНИЕ

Том 63, номер 9, 2023 год

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

Конструктивный алгоритм векторизации произведения $P \otimes P$
для симметричной матрицы P

А. И. Глуценко, К. А. Ласточкин 1415

Вычисление с минимальной погрешностью n -й производной
по данным измерения функции

А. С. Демидов, А. С. Кочуров 1428

Численный метод оценки скорости роста ошибок округления
в равномерной метрике

М. И. Зуев, С. И. Сердюкова 1438

О единственности решения систем линейных алгебраических уравнений,
к которым редуцируются обратные задачи гравиметрии и магнитометрии:
региональный вариант

И. И. Колотов, Д. В. Лукьяненко, И. Э. Степанова, А. В. Щепетилов, А. Г. Ягола 1446

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

Безградиентные методы федеративного обучения с l_1 и l_2 -рандомизацией
для задач негладкой выпуклой стохастической оптимизации

Б. А. Альашкар, А. В. Гасников, Д. М. Двинских, А. В. Лобанов 1458

Численный алгоритм определения источника диффузионно-логистической
модели по данным интегрального типа, основанный на тензорной оптимизации

Т. А. Звонарева, С. И. Кабанихин, О. И. Криворотько 1513

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

О взаимодействии граничных особых точек в одной эллиптической краевой задаче

А. М. Боговский 1524

Constructing Solutions of Cauchy Type Integral Equations by Using Four Kinds of Basis

M. Yaghobifar, F. Hosseini Shekarabi 1531

МАТЕМАТИЧЕСКАЯ ФИЗИКА

Добавление к классическому результату А.Н. Тихонова об электромагнитных
зондированиях для среды с тонкими слоями

А. С. Барашков 1532

Обратные задачи для диффузионно-дрейфовой модели зарядки
неоднородного полярного диэлектрика

Р. В. Бризицкий, Н. Н. Максимова, А. Г. Масловская 1537

Three-Dimensional Numerical Transfer Model Using a Monotonized Z-Scheme

N. M. Kashchenko, S. A. Ishanov, E. V. Zubkov, L. V. Zinin 1553

Универсальный подход к анализу диссипативных свойств
численного метода решения уравнений газодинамики

П. В. Чувахов, И. О. Погорелов, К. В. Шубин 1554

The Effect of Nonlocal Scale Value and Phase Lags on Thermoelastic Waves
in a Multilayered LEMV/CFRP Composite Cylinder

S. Mahesh, R. Selvamani, Farzad Ebrahimi

1564

Влияние демпфирования податливой поверхности на невязкую
неустойчивость несжимаемого пограничного слоя над ней

И. В. Савенков

1565

ИНФОРМАТИКА

Агрегирование и декомпозиция систем дифференциальных уравнений
с частными производными и систем управления с распределенными параметрами

В. И. Елкин

1575

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.61

КОНСТРУКТИВНЫЙ АЛГОРИТМ ВЕКТОРИЗАЦИИ ПРОИЗВЕДЕНИЯ $P \otimes P$ ДЛЯ СИММЕТРИЧНОЙ МАТРИЦЫ P^1

© 2023 г. А. И. Глушенко^{1,*}, К. А. Ласточкин^{1,**}¹ 117997 Москва, ул. Профсоюзная, 65, Институт проблем управления им. В.А. Трапезникова РАН, Россия

*e-mail: aiglush@ipu.ru

**e-mail: lastconst@ipu.ru

Поступила в редакцию 20.02.2022 г.
Переработанный вариант 20.02.2023 г.
Принята к публикации 29.05.2023 г.

В работе предложен конструктивный алгоритм вычисления матриц исключения $\bar{L}$ и дублирования $\bar{D}$ для операции векторизации произведения $P \otimes P$ при $P = P^T$. Матрица $\bar{L}$, получаемая в соответствии с алгоритмом, позволяет формировать из упомянутого произведения вектор, содержащий только уникальные элементы. Матрица $\bar{D}$, в свою очередь, позволяет выполнять обратное преобразование. Предложена программная реализация процедуры вычисления матриц $\bar{L}$ и $\bar{D}$. На основе отмеченных результатов предложена новая операция $\text{vesu}(\cdot)$, определенная над произведением $P \otimes P$ при $P = P^T$ и описаны ее свойства. Показаны отличия и преимущества разработанной операции от известных операций $\text{vec}(\cdot)$ и $\text{vech}(\cdot)$ ($\text{vecd}(\cdot)$) в случае их применения для векторизации произведения $P \otimes P$ при $P = P^T$. На примере параметризации алгебраического уравнения Риккати продемонстрирована эффективность операции $\text{vesu}(\cdot)$ для снижения перепараметризации задачи идентификации неизвестных параметров. Библ. 14. Фиг. 3.

Ключевые слова: векторизация, матрица исключения, матрица дублирования, произведение Кронекера, уникальные элементы матрицы, снижение размерности, перепараметризация, уравнение Риккати.

DOI: 10.31857/S0044466923090090, EDN: ROMVVI

1. ВВЕДЕНИЕ

Матричный оператор $\text{vec}(\cdot)$, векторизующий квадратную матрицу, располагая ее столбцы друг под другом, широко используется для облегчения математических вычислений с матрицами [1, 2]. В частности, линейное матричное уравнение с матрицей неизвестных P с помощью $\text{vec}(\cdot)$ преобразуется в систему линейных алгебраических уравнений, к которой применим весь известный арсенал аналитических и численных методов поиска ее решения. В [1, 2] также вводятся два варианта данной операции для случая, когда такая матрица является симметричной $P = P^T$: $\text{vech}(\cdot)$ и $\text{vecd}(\cdot)$. Обе операции возвращают вектор-столбец, который содержит в себе только уникальные элементы матрицы P , т.е. ее ниже-треугольную часть. Для $\text{vech}(\cdot)$ и $\text{vecd}(\cdot)$ различается лишь порядок расположения таких элементов, при этом операции обладают сходными свойствами, ввиду чего в дальнейшем тексте работы без потери общности будет упоминаться только операция $\text{vech}(\cdot)$. Применение таких операций позволяет существенно сократить размерность получаемой после векторизации системы уравнений, а также в принципе гарантировать возможность получения решения системы уравнений (см. лемму 7.1 в [1]). Однако существуют задачи, в которых в матричных уравнениях матрица неизвестных параметров обладает более сложной симметрией по сравнению с уже упомянутым случаем $P = P^T$. Например, если такая матрица является симметричной и состоит из блоков равной размерности, каждый из ко-

¹ Работа выполнена при частичной финансовой поддержке фонда Президента РФ (код проекта МД-1787.2022.4).

торых также является симметричной квадратной матрицей — $P \otimes P$, где $P = P^T$. В этом случае при применении операции векторизации размерность решаемой задачи можно было бы сократить еще сильнее, поскольку повторяющихся элементов в такой матрице больше, чем в обычной симметричной. Однако ни одна из имеющихся операций $\text{vec}(\cdot)$, $\text{vech}(\cdot)$ не позволяет этого сделать. Ниже рассмотрим один из классов таких задач, в которых в качестве матрицы неизвестных возникает $P \otimes P$ при $P = P^T$.

В настоящее время теория адаптивного управления предоставляет множество инструментов для решения проблем управления в условиях параметрической неопределенности, немоделируемой динамики, возмущений и пр. [3–5]. Известные подходы обеспечивают асимптотическую и экспоненциальную сходимость по параметрам и ошибке слежения, сходимость за конечное время.

При применении многих подобных методов первичной задачей является приведение исходной системы к линейному регрессионному уравнению (LRE) относительно неизвестных параметров [6]:

$$y = \omega\theta, \quad (1)$$

где $y \in \mathbb{R}^{N \times 1}$ — вектор измеримых выходных величин, $\omega \in \mathbb{R}^{N \times M}$ — измеримый регрессор, $\theta \in \mathbb{R}^{M \times 1}$ — вектор неизвестных параметров. Разумеется, желательно, чтобы размерность θ была минимально возможной для решения поставленной задачи. Для улучшения динамических свойств регрессора ω возможно применение различных схем предобработки, таких как DRE [7], MRE [8] для получения матричного квадратного регрессора $\Omega \in \mathbb{R}^{M \times M}$, а также DREM [6] для получения скалярного регрессора $\Delta \in \mathbb{R}$. Независимо от размерности регрессора, для идентификации неизвестных параметров θ обычно используются классические алгоритмы идентификации, такие как, gradient descent или least squares.

В целом методы теории адаптивного управления обладают рядом недостатков, одним из основных среди которых является отсутствие гарантий на качество получаемых в системе переходных процессов [9]. Например, асимптотическая или экспоненциальная сходимость не гарантирует отсутствие перерегулирования и/или колебательности. Поэтому становится актуальным применение методов теории адаптивного управления для задач, традиционно решаемых в рамках теории оптимального управления.

Теория оптимального управления позволяет гарантировать наилучшее возможное качество управления с точки зрения целевой функции, выбранной пользователем [10]. Для линейных систем классическим решением задачи оптимального управления является синтез регулятора, минимизирующего линейно-квадратичный (LQ) функционал от состояний и управлений на бесконечном времени [11, 12]. Параметры такого регулятора вычисляются на основе матрицы P , являющейся решением алгебраического уравнения Риккати (ARE).

Для синтеза LQ регулятора в адаптивной постановке одним из подходов является приведение ARE к виду (1). Для этого осуществляют [13] подстановку в него вместо неизвестных матриц объекта A и B их динамических оценок $\hat{A}$, $\hat{B}$ или некоторых измеримых сигналов. На основе идентификации неизвестных параметров полученной таким образом регрессии (1) и определяют искомую матрицу P . Однако задача идентификации матрицы P осложняется тем обстоятельством, что после приведения ARE к (1) вектор неизвестных параметров θ , кроме искомой матрицы P , включает в себя также перепараметризованную часть, определяемую выражением $\text{vec}(P \otimes P)$. Как хорошо известно, решение задачи идентификации неизвестных параметров при наличии перепараметризации обладает рядом существенных недостатков (неудовлетворительное качество оценок, потеря возбуждения регрессором и пр.), поэтому актуальным является уменьшение числа перепараметризованных параметров или их полное исключение. Как уже упоминалось выше, единственным существующим на сегодняшний день подходом к уменьшению числа параметров в перепараметризованной части уравнения ARE, записанного в виде (1), является использование вместо $\text{vec}(P \otimes P)$ процедуры частичной векторизации $\text{vech}(P \otimes P)$. Однако в силу симметрии матрицы P и симметрии произведения $P \otimes P$ перепараметризованная часть в данной задаче содержит повторы и в нижне-треугольной части, а поэтому использование процедуры $\text{vech}(P \otimes P)$ не является оптимальным и возвращает вектор с повторяющимися элементами. В литературе авторам не удалось обнаружить алгоритма операции векторизации $P \otimes P$, результатом которого является вектор, содержащий только уникальные элементы. Также не были обна-

ружены оценки на число уникальных элементов в произведении $P \otimes P$ для матрицы $P = P^T$ произвольно выбранного размера. Необходимо упомянуть, что в работах [13, 14] декларативно упоминается необходимость наличия такой операции, но конструктивный алгоритм ее выполнения не приводится, а полученная в [13] формула для оценки числа уникальных параметров при некоторых значениях размера матрицы $P = P^T$ выдает нецелочисленный ответ. Некоторые дополнительные пояснения по результатам работ [13, 14] приведены в разд. 2 настоящей статьи.

Таким образом, в развитие результатов работ [1, 2] для случая более сложной симметрии матрицы неизвестных параметров, целью настоящей работы является разработка алгоритма векторизации произведения $(P \otimes P)$ с целью получения вектора, включающего в себя только уникальные элементы такого произведения. Дополнительно будет получена зависимость числа уникальных элементов рассматриваемого произведения от размерности исходной матрицы P . Основным результатом работы является операция векторизации произведения $P \otimes P$ с исключением повторений, которая может быть использована как в задаче адаптивного LQ синтеза, так и в других задачах, в которых приходится оперировать выражением вида $P \otimes P$, $P = P^T$.

1.1. Определения

Выпишем основные определения из [1], которые будут использоваться в настоящей работе.

Определение 1. Операция векторизации $\text{vec}(\cdot)$ заключается в преобразовании произвольной матрицы $M \in R^{q \times l}$ в вектор $\text{vec}(M) \in R^{ql}$ путем записи столбцов один под другим.

Основные свойства операции $\text{vec}(\cdot)$:

- 1) $\text{vec}(MNC) = (C^T \otimes M) \text{vec}(N)$;
- 2) $\text{vec}(MN) = (I_m \otimes M) \text{vec}(N) = (N^T \otimes I_q) \text{vec}(M)$,

где $I_q \in R^{q \times q}$ — единичная матрица, $N \in R^{l \times m}$, $C \in R^{m \times n}$.

Определение 2. Операция частичной векторизации $\text{vech}(\cdot)$ заключается в преобразовании некоторой матрицы $M = M^T \in R^{n \times n}$ в вектор $\text{vech}(M) \in R^k$ путем записи столбцов ниже-треугольной части матрицы M один под другим:

$$\text{vech}(M) = [m_{1,1}, m_{2,1}, \dots, m_{n,1}, m_{2,2}, \dots, m_{n,2}, \dots, m_{n-1,n-1}, m_{n-1,n}, m_{n,n}]^T,$$

где $k = \frac{n^2 + n}{2}$ — число уникальных элементов матрицы M .

Операция $\text{vech}(\cdot)$ выполняется над матрицей M путем умножения $\text{vec}(M)$ на матрицу исключения (elimination matrix), а связь между операциями $\text{vec}(M)$ и $\text{vech}(M)$ задается с помощью матрицы возвращения (дублирования, duplication matrix):

$$\text{vech}(M) = L \text{vec}(M), \quad \text{vec}(M) = D \text{vech}(M),$$

где матрицы $D \in R^{n^2 \times k}$ и $L \in R^{k \times n^2}$ обладают следующими свойствами:

- 1) L имеет полный строчный ранг k ;
- 2) $LL^T = LD = I_k$;
- 3) $L^\dagger = L^T$, где $L^\dagger$ — псевдоинверсия Мура-Пенроуза матрицы L ;
- 4) $\text{vec}(M) = DL \text{vec}(M) = D \text{vech}(M)$.

Определение 3. Матрица исключения $L \in R^{k \times n^2}$ для $M = M^T \in R^{n \times n}$ определяется следующим образом:

$$L = \sum_{i \geq j} (u_{ij} \otimes e_j^T \otimes e_i^T), \quad i, j = \overline{1, n},$$

где $e_i \in R^n$ — вектор, на i -й позиции в котором стоит единица, а на остальных — нули, $u_{ij} \in R^k$ — вектор, на $(j-1)n + i - \frac{1}{2}j(j-1)$ позиции которого стоит единица, а на остальных позициях — нули.

Определение 4. Матрица возвращения $D \in R^{n^2 \times k}$ для $M = M^T \in R^{n \times n}$ определяется следующим образом:

$$D^T = \sum_{i \geq j} (u_{ij} \text{vec}^T(T_{ij})), \quad i, j = \overline{1, n}, \quad (2)$$

где $T_{ij} \in R^{n \times n}$ — матрица, заполненная нулями и содержащая единицы в позициях (i, j) и (j, i) .

Пример 1. Выполним над $M = \begin{bmatrix} m_1 & m_2 \\ m_2 & m_4 \end{bmatrix}$ операции $\text{vec}(\cdot)$ и $\text{vech}(\cdot)$:

$$\begin{aligned} \text{vec}(M) &= D \text{vech}(M) = [m_1 \ m_2 \ m_2 \ m_4]^T; \\ \text{vech}(M) &= L \text{vec}(M) = [m_1 \ m_2 \ m_4]^T, \end{aligned}$$

$$\text{где } L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \text{ а } D = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}^T.$$

2. ПОСТАНОВКА ЗАДАЧИ

Следуя работам [13, 14], рассмотрим классическое ARE, к решению которого сводится задача конструирования оптимального LQ регулятора:

$$PA + A^T P - PBR^{-1}B^T P = -Q. \quad (3)$$

Здесь $A \in R^{n \times n}$ — матрица состояний, $B \in R^{n \times m}$ — матрица коэффициентов усиления, $R = R^T > 0$, $Q = Q^T > 0$ — матрицы соответствующей размерности, являющиеся параметрами целевого функционала LQ оптимизации, $P = P^T > 0$ — матрица соответствующей размерности, являющаяся решением ARE.

Применив $\text{vec}(\cdot)$ к выражению (3), получим

$$(A^T \otimes I_n + I_n \otimes A^T) \text{vec}(P) - (P \otimes P) \text{vec}(BR^{-1}B^T) = -\text{vec}(Q).$$

Использував второе свойство операции векторизации, имеем

$$(A^T \otimes I_n + I_n \otimes A^T) \text{vec}(P) - [\text{vec}^T(BR^{-1}B^T) \otimes I_{n^2}] \text{vec}(P \otimes P) = -\text{vec}(Q). \quad (4)$$

Так как $P = P^T$, $Q = Q^T$, $BR^{-1}B^T = (BR^{-1}B^T)^T$, то $\text{vec}(P \otimes P)$, $\text{vec}^T(BR^{-1}B^T)$, $\text{vec}(Q)$ содержат повторы, а поэтому для исключения неуникальных элементов рационально перейти с учетом свойств, описанных в определении 2, от $\text{vec}(\cdot)$ к $\text{vech}(\cdot)$.

Утверждение 1. Уравнение Риккати (4) относительно матрицы $P = P^T > 0$ может быть записано в виде линейного уравнения с перепараметризацией:

$$\left[L(I_n \otimes A^T + A^T \otimes I_n) D L \left((D \text{vech}(BB^T))^T \otimes I_{n^2} \right) D_1 \right] \begin{bmatrix} L \text{vec}(P) \\ L_1 \text{vec}(P \otimes P) \end{bmatrix} = -\text{vech}(Q), \quad (5)$$

в котором матрицы $L \in R^{k \times n^2}$, $D \in R^{n^2 \times k}$, $L_1 \in R^{r \times n^4}$, $D_1 \in R^{n^4 \times r}$ задают операции исключения и возвращения над Q и $P \otimes P$ и по определению 2 удовлетворяют условиям $L = (D^T D)^{-1} D^T$, $L_1 = (D_1^T D_1)^{-1} D_1^T$, $r = 0.5n^2(n^2 + 1)$, $k = 0.5n(n + 1)$.

Доказательство утверждения непосредственно следует из применения операции частичной векторизации к $\text{vec}(P \otimes P)$, $\text{vec}(P)$, $\text{vec}^T(BR^{-1}B^T)$, $\text{vec}(Q)$ в уравнении (4) с учетом свойств матриц возвращения и исключения. Рассмотрим иллюстративный пример.

Пример 2. Пусть $n = 2$, $m = 1$, а матрицы ARE определены следующим образом:

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad P = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix}, \quad Q = \begin{bmatrix} q_1 & 0 \\ 0 & q_4 \end{bmatrix}, \quad R = 1.$$

Тогда ARE в форме (4) примет вид:

$$\begin{bmatrix} 2a_1 & a_3 & a_3 & 0 \\ a_2 & a_1 + a_4 & 0 & a_3 \\ a_2 & 0 & a_1 + a_4 & a_3 \\ 0 & a_2 & a_2 & 2a_4 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_2 \\ p_4 \end{bmatrix} - VP_{\text{kron}} = - \begin{bmatrix} q_1 \\ 0 \\ 0 \\ q_4 \end{bmatrix}, \quad (6)$$

где

$$V = \begin{bmatrix} b_1^2 I_{n^2} & b_1 b_2 I_{n^2} & b_1 b_2 I_{n^2} & b_2^2 I_{n^2} \end{bmatrix},$$

$$P_{\text{kron}} = \begin{bmatrix} p_1^2 & p_1 p_2 & p_1 p_2 & p_2^2 & p_1 p_2 & p_1 p_4 & p_2^2 & p_2 p_4 & p_1 p_2 & p_2^2 & p_1 p_4 & p_2 p_4 & p_2^2 & p_2 p_4 & p_2 p_4 & p_4^2 \end{bmatrix}^T.$$

Очевидно, что второе и третье уравнения (6) совпадают, как и то, что $P_{\text{kron}} = \text{vec}(P \otimes P)$ содержит множество повторов и всего шесть уникальных элементов.

Теперь рассмотрим запись уравнения (3) с использованием частичной векторизации (5):

$$\begin{bmatrix} 2a_1 & 2a_3 & 0 \\ a_2 & a_1 + a_4 & a_3 \\ 0 & 2a_2 & 2a_4 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_4 \end{bmatrix} - VP_{\text{kron}} = \begin{bmatrix} q_1 \\ 0 \\ q_4 \end{bmatrix}. \quad (7)$$

В данном случае

$$V = \begin{bmatrix} b_1^2 & b_1 b_2 & b_1 b_2 & b_2^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b_1^2 & 0 & 0 & b_1 b_2 & b_1 b_2 & b_2^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1^2 & 0 & 0 & b_1 b_2 & 0 & b_1 b_2 & b_2^2 \end{bmatrix},$$

а вектор $P_{\text{kron}} = \begin{bmatrix} p_1^2 & p_1 p_2 & p_1 p_2 & p_2^2 & p_1 p_4 & p_2^2 & p_2 p_4 & p_1 p_4 & p_2 p_4 & p_4^2 \end{bmatrix}^T$.

Из сравнения (7) и (6) следует, что число уравнений (по определению k из утверждения 1) сократилось до трех (повторяющееся уравнение было отброшено), а длина вектора P_{kron} (по определению r из утверждения 1) — до 10. Тем не менее в P_{kron} все также присутствует четыре повтора, число которых, очевидно, будет существенно возрастать с ростом размерности n .

Таким образом, пример 2 демонстрирует, что операция $\text{vch}(\cdot)$ не позволяет сформировать из P_{kron} вектор уникальных параметров минимальной длины, поскольку не уникальные элементы в $P \otimes P$ находятся не только в верхнетреугольной части, но и в нижнетреугольной.

В работах [13, 14] выражения, сходные с (5), рассматриваются как линейная регрессия, вектор параметров которой идентифицируется on-line. В такой задаче вектор параметров должен иметь минимально возможную перепараметризацию — фактически, состоять только из уникальных составляющих. Однако работы [13, 14] ограничены наличием только процедуры $\text{vch}(\cdot)$, что приводит к проблемам, продемонстрированным в примере 2. Поэтому в [13] сделано допущение о том, что желаемая процедура векторизации существует и уже применена к (5), не поясняя, как именно ее выполнить. Кроме того, они дают оценку на количество уникальных элементов в

$P \otimes P$, $P = P^T$: $\frac{n^4 + n^3 + 7n^2 + 6n}{8}$, которая, как будет продемонстрировано ниже, является неверной.

Поэтому актуальной является задача разработки операции, выполняющей преобразование произведения $P \otimes P$ в вектор, наполненный только уникальными элементами.

Цель. В настоящей работе требуется, во-первых, построить алгоритм формирования матрицы исключения $\bar{L} \in R^{r \times n^4}$, выполняющей преобразование $\text{vec}(P \otimes P)$ в вектор, состоящий из уникальных элементов (оптимизируя тем самым размерность r), а во-вторых, получить аналитическую формулу зависимости числа уникальных элементов в произведении $P \otimes P$ от размерности матрицы P .

3. ОСНОВНОЙ РЕЗУЛЬТАТ

Чтобы сформулировать алгоритм исключения из $P \otimes P$ повторяющихся элементов, рассмотрим подробнее представленный на фиг. 1 результат умножения по Кронекеру P саму на себя в случае $n = 2$.

Таким образом, поскольку результат произведения $P \otimes P$ при $P = P^T$ является симметричной матрицей, то элементы верхнетреугольной формы (1,2), (1,3), (1,4), (2,3), (2,4), (3,4) считаются неуникальными. Кроме того, если результат произведения $P \otimes P$ разбить на блоки размером $n \times n$, то в каждом из полных блоков нижнетреугольной части в силу $P = P^T$ наблюдается симметрия относительно главной диагонали блока, а тогда элемент (3,2) считается неуникальным. Дополнительно, чем правее и ниже расположен блок в нижнетреугольной части, тем больше в нем повторов элементов с блоками, стоящими левее и/или выше — межблочная симметрия (элементы (4,3) и (3,1) считаются неуникальными). На фиг. 1 в красные рамки взяты элементы, находящиеся ниже главной диагонали $P \otimes P$, но при этом являющиеся неуникальными при обходе нижнетреугольной части по блокам сверху вниз и слева направо.

На основе описанной и отмеченной на фиг. 1 естественной симметрии в матрице $P \otimes P$ формализуем в виде следующего утверждения алгоритм формирования матрицы исключения $\bar{L} \in R^{r \times n^4}$ и возвращения $\bar{D} \in R^{n^4 \times r}$.

Утверждение 2. Матрица исключения $\bar{L} \in R^{r \times n^4}$, обеспечивающая $\bar{L} \text{vec}(P \otimes P) \in R^r$, $r = \left(\frac{n^4}{12} + \frac{n^3}{3} + \frac{5n^2}{12} + \frac{n}{6}\right)$ может быть сформирована следующим конструктивным алгоритмом:

$$\bar{D}^T = \sum_{l=1}^r u_l \text{vec}^T(E_{\wp_l}), \quad \bar{D} \in R^{n^4 \times r}, \quad \bar{L} = (\bar{D}^T \bar{D})^{-1} \bar{D}^T,$$

$$E_{\wp} = \{E_{\wp_l} \in R^{n^2 \times n^2} : l = \overline{1, r}\},$$

$$\wp := \{(i, j, p, q) : i \geq j > 0, p \geq q > 0, p \leq i \leq n, q \leq j \leq n\},$$

где p — номер вектор-строки с числом строк n в блочной матрице $E_{\wp_l}$, q — номер вектор-столбца с числом столбцов n в блочной матрице $E_{\wp_l}$, i — номер строки в блоке (p, q) , j — номер столбца в блоке (p, q) , $u_l \in R^r$ — вектор, заполненный нулями и содержащий единицу на l -й позиции, $E_{\wp}$ — множество мощностью r всех возможных матриц с учетом $(i, j, p, q) \in \wp$, заполненных нулями и содержащих единицы в позициях (i, j, p, q) , (j, i, p, q) , (i, j, q, p) , (j, i, q, p) , (p, q, i, j) , (q, p, i, j) , (p, q, j, i) , (q, p, j, i) .

Доказательство. Поскольку алгоритм (8) является конструктивным, то справедливость утверждения 2 следует из отсутствия логических противоречий в рассуждениях, использованных для его формирования. Далее кратко остановимся на основных рассуждениях, на основе которых и был предложен алгоритм (8).

$$P \otimes P = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix} \otimes \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix} = \begin{bmatrix} p_1 \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix} & p_2 \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix} \\ p_2 \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix} & p_4 \begin{bmatrix} p_1 & p_2 \\ p_2 & p_4 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} p_1^2 & p_1 p_2 & p_1 p_2 & p_2^2 \\ p_1 p_2 & p_1 p_4 & p_2^2 & p_2 p_4 \\ p_1 p_2 & p_2^2 & p_1 p_4 & p_2 p_4 \\ p_2^2 & p_2 p_4 & p_2 p_4 & p_4^2 \end{bmatrix}$$

Фиг. 1. Симметрия в $P \otimes P$.

В первую очередь, матрицу $P \otimes P \in R^{n^2 \times n^2}$ разобьем на n^2 блоков размерностью $n \times n$. Пусть каждый блок имеет уникальный составной индекс (p, q) , где p — номер вектор-строки с числом строк n в блочной матрице $P \otimes P$, q — номер вектор-столбца с числом столбцов n в блочной матрице $P \otimes P$. Для обращения к каждому элементу блока также введем уникальный составной индекс (i, j) , где i — номер строки в блоке (p, q) , j — номер столбца в блоке (p, q) . Таким образом, каждый элемент матрицы $P \otimes P$ может быть однозначно определен уникальным составным индексом (i, j, p, q) .

Теперь, пользуясь симметрией, отмеченной на фиг. 1, определим позиции уникальных элементов в матрице $P \otimes P$, которые находятся на или ниже главной диагонали.

Ввиду общей симметричности матрицы $P \otimes P$, такие элементы должны находиться в блоках, находящихся ниже главной диагонали $P \otimes P$, или включающих ее. Таким образом, при учете $p = 1, \dots, n, q = 1, \dots, n$, необходимым условием уникальности является $p \geq q$.

Учитывая, что каждый блок матрицы $P \otimes P$ сам представляет собой симметричную матрицу, необходимо считать уникальными только элементы блока, которые находятся на или ниже его главной диагонали. Таким образом, при учете $i = 1, \dots, n, j = 1, \dots, n$, необходимым условием является $i \geq j$.

Принимая во внимание блочную структуру матрицы $P \otimes P$ и тот факт, что $P = P^T$, для элемента с индексом (i, j, p, q) обязательно будет симметричным элемент (p, q, i, j) . Ведь элемент (i, j, p, q) — это фактически, следуя определению операции произведения Кронекера, произведение (i, j) элемента матрицы P на ее элемент (p, q) , а (p, q, i, j) — это произведение (p, q) элемента матрицы P на элемент (i, j) . Поэтому необходимым условием оптимальности является $p \leq i, q \leq j$.

Таким образом, объединив сформулированные условия, может быть сформировано множество индексов уникальных элементов матрицы $P \otimes P$: $\wp := \{(i, j, p, q) : i \geq j > 0, p \geq q > 0, p \leq i \leq n, q \leq j \leq n\}$. Примем пока, что таких уникальных элементов r , а обоснование формулы для подсчета их числа приведем ниже.

Для формирования матрицы $\bar{D}$ путем, аналогичным (2) [1], необходимо для каждого l -го ($l = 1, \dots, r$) уникального элемента матрицы $P \otimes P$ (l -го члена множества $\wp$, который обозначим $\wp_l$) сформировать матрицу $E_{\wp_l} \in R^{n^2 \times n^2}$, которая содержит единицы на всех позициях, соответствующих позициям, в которых в исходной матрице $P \otimes P$ встречается элемент $\wp_l = (i^{(l)}, j^{(l)}, p^{(l)}, q^{(l)})$.

Таких позиций может быть максимум восемь: сама позиция элемента $(i^{(l)}, j^{(l)}, p^{(l)}, q^{(l)})$; симметричный элемент относительно главной диагонали блока $(p, q) - (j^{(l)}, i^{(l)}, p^{(l)}, q^{(l)})$; как уже упоминалось выше, элемент $(p^{(l)}, q^{(l)}, i^{(l)}, j^{(l)})$; также элемент, симметричный $(p^{(l)}, q^{(l)}, i^{(l)}, j^{(l)})$ относительно главной диагонали блока $(i, j) - (q^{(l)}, p^{(l)}, i^{(l)}, j^{(l)})$, а также четыре элемента, симметричных уже упомянутым четырем относительно главной диагонали матрицы $P \otimes P - (i^{(l)}, j^{(l)}, q^{(l)}, p^{(l)}), (j^{(l)}, i^{(l)}, q^{(l)}, p^{(l)}), (p^{(l)}, q^{(l)}, j^{(l)}, i^{(l)}), (q^{(l)}, p^{(l)}, j^{(l)}, i^{(l)})$. Для некоторых элементов в матрице $P \otimes P$ часть из этих восьми индексов могут указывать на один и тот же элемент (например, для элемента $(1, 1, 1, 1)$). В этом нет противоречия, поскольку это будет лишь означать, что установить единицу в некоторый элемент матрицы $E_{\wp_l}$ необходимо несколько раз.

Транспонированная векторизация матрицы $E_{\wp_l}$ представляет собой вектор размерностью $1 \times n^4$, который должен занять l -й столбец в матрице $\bar{D}$, что сходно с (2). Заявленная в утверждении процедура формирования вектора $u_l \in R^r$ и формула

$$\bar{D}^T = \sum_{l=1}^r u_l \text{vec}^T(E_{\wp_l})$$

позволяют сформировать именно такую матрицу.

От полученной матрицы дублирования с помощью псевдообращения Мура-Пенроуза можно получить матрицу исключения $\bar{L}$. При этом существование $\bar{L}$ гарантируется взаимной ортогональностью $\text{vec}(E_{\varphi_l}) \forall l = \overline{1, r}$, в свою очередь обеспечиваемой $(i, j, p, q) \in \varphi$, т.е. тем фактом, что элементы сформированного множества φ уникальны.

Для подсчета числа уникальных элементов в матрице $P \otimes P$ воспользуемся следующим подходом. В первую очередь, число уникальных элементов матрицы $P \otimes P$ не больше числа уникальных элементов в ее ниже-треугольной части, определяемого по формуле $0.5(n^4 + n^2)$. Из этой величины необходимо вычесть число элементов, расположенных выше главной диагонали своего блока для всех блоков, расположенных полностью ниже главной диагонали матрицы $P \otimes P$. Таких блоков ровно $n^2 - 0.5(n^2 + n)$, поскольку всего их n^2 , а расположенных на или выше главной диагонали — $0.5(n^2 + n)$. В каждом из таких блоков по $n^2 - 0.5(n^2 + n)$ элементов выше главной диагонали. Поэтому из $0.5(n^4 + n^2)$ необходимо вычесть $(n^2 - 0.5(n^2 + n))^2$.

Наконец, необходимо подсчитать число неуникальных элементов в ниже-треугольных частях всех блоков матрицы $P \otimes P$, расположенных на или ниже главной диагонали $P \otimes P$. Рассмотрение матрицы $P \otimes P$ необходимо вести по блочным вектор-столбцам $q = 1, \dots, n$. В первом таком столбце по мере увеличения номера вектор-строки $p = 1, \dots, n$, количество неуникальных элементов будет возрастать в виде ряда, члены которого представляют собой сумму элементов арифметической прогрессии $0, 1, 2, 3, 4, \dots$, т.е. $0, 1, 3, 6, 10, \dots$ ввиду упомянутого выше совпадения элементов (i, j, p, q) и (p, q, i, j) для всех допустимых i, j, p, q . Сумма такой последовательности вычисляется как $(n-1)n(n+1)/6$. По мере роста q число неуникальных элементов в каждом последующем столбце будет возрастать, поскольку при формировании множества φ были выставлены требования $p \leq i, q \leq j$. То есть при возрастании q на единицу в каждом из $(n-q)$ блоков соответствующего вектор-столбца рассматривается на один столбец элементов меньше, что, при учете, что $p \geq q$ означает отбрасывание еще $(nq - (q-1)q/2)$ неуникальных элементов. При этом для оставшихся в рассмотрении элементов каждого блока действует все то же правило суммы элементов арифметической прогрессии. Таким образом, общее количество неуникальных элементов в ниже-треугольных частях всех блоков матрицы $P \otimes P$, расположенных на или ниже главной диагонали $P \otimes P$, рассчитывается по формуле:

$$N_n = \frac{n(n-1)(n+1)}{6} + \sum_{q=1}^{n-1} \left[(n-q) \left(nq - \frac{q(q-1)}{2} \right) + \frac{(n-q-1)(n-q)(n-q+1)}{6} \right].$$

А общее число уникальных элементов в матрице $P \otimes P$ вычисляется по формуле

$$r = \frac{n^4 + n^2}{2} - \frac{n(n-1)(n+1)}{6} - \sum_{q=1}^{n-1} \left[(n-q) \left(nq - \frac{q(q-1)}{2} \right) + \frac{(n-q-1)(n-q)(n-q+1)}{6} \right] - \left(n^2 - \frac{n^2 + n}{2} \right)^2 = \left(\frac{n^4}{12} + \frac{n^3}{3} + \frac{5n^2}{12} + \frac{n}{6} \right), \quad (9)$$

что завершает доказательство утверждения.

На основе конструктивного алгоритма формирования матрицы исключения $\bar{L}$ введем определение операции исключения повторений из $\text{vec}(P \otimes P)$.

Определение 5. Операция исключения повторений $\text{vecu}(\cdot)$ из вектора $\text{vec}(M \otimes M) \in R^{n^4}$, $M = M^T \in R^{n \times n}$, задается следующим образом:

$$\text{vecu}(M \otimes M) = \bar{L} \text{vec}(M \otimes M),$$

где $\bar{L}$ и r определены в утверждении 2, а $\bar{L}$ обладает следующими свойствами:

- 1) $\bar{L}$ имеет полный строчный ранг r ;
- 2) $\bar{L}^\dagger = \bar{L}^T$, где $\bar{L}^\dagger$ — псевдоинверсия Мура-Пенроуза матрицы $\bar{L}$.

Существует матрица $\bar{D}$ такая, что:

- 3) $\bar{L}\bar{L}^T = \bar{L}\bar{D} = I_r$;

$$4) \text{vec}(M) = \bar{D}\bar{L} \text{vec}(M) = \bar{D} \text{vec}_u(M).$$

Справедливость определения 5 непосредственно следует из доказательства утверждения 2.

Замечание 1. Результатом выполнения операции $\text{vec}_u(\cdot)$ является вектор уникальных элементов $P \otimes P$, а также матрицы исключения и дублирования, поэтому введение отдельных дополнительных массивов для длительного хранения индексов уникальных элементов в $P \otimes P$ (которые упоминались в доказательстве) не требуется.

Ниже приведен код на языке Matlab Script Language, реализующий в соответствии с конструктивным алгоритмом (8) формирование матриц дополнения $\bar{D}$ и исключения $\bar{L}$.

```
%порядок задачи
n=2;
%число уникальных элементов в матрице P kron P
sum=n*(n-1)*(n+1)/6;
for q=1:1:(n-1)
    sum=sum+(n-q)*(n*q-q*(q-1)/2)+(n-q)*(n-q+1)*(n-q-1)/6;
end
result_unique = (n^4+n^2)/2 - sum - (n*n-n*(n+1)/2)^2;
%код для формирования матрицы возвращения
D=zeros(n^4,result_unique);
l=1;
for q=1:1:n
    for p=q:1:n
        for j=q:1:n
            if j>p
                k=j;
            else
                k=p;
            end
            for i=k:1:n
                E=zeros(n^2,n^2);
                E((p-1)*n+i,(q-1)*n+j)=1;
                E((p-1)*n+j,(q-1)*n+i)=1;
                E((q-1)*n+i,(p-1)*n+j)=1;
                E((q-1)*n+j,(p-1)*n+i)=1;
                E((i-1)*n+p,(j-1)*n+q)=1;
                E((i-1)*n+q,(j-1)*n+p)=1;
                E((j-1)*n+p,(i-1)*n+q)=1;
                E((j-1)*n+q,(i-1)*n+p)=1;
                E=reshape(E,n^2*n^2,1);
                D(:,l)=E;
                l=l+1;
            end
        end
    end
end
end
%код для формирования матрицы исключения
L=pinv(D);
```

Оценим сложность предложенного конструктивного алгоритма. Необходимо отметить, что для подсчета числа уникальных элементов *result_unique* она является линейной $O(n)$. Для построения матрицы D циклы *for* организованы таким образом, что будет осуществлен обход только уникальных r (см. (9)) штук элементов $P \otimes P$. При этом для каждого такого элемента будет осу-

ществляться векторизация матрицы E со сложностью $O(n^4)$. То есть сложность этой части алгоритма составляет $O(n^4 \cdot r)$. Сложность вычисления матрицы L совпадает со сложностью получения псевдообратной матрицы Мура–Пенроуза, которая для данного случая будет равна $O((n^4)^2 \cdot r) = O(n^8 \cdot r)$. Наибольшую сложность имеет последняя часть алгоритма. Она же и определяет его сложность в целом.

4. ЧИСЛЕННЫЙ ПРИМЕР

Рассмотрим два иллюстративных примера ($n = 2, n = 3$), поясняющих предложенный конструктивный алгоритм формирования матрицы исключения.

Пример 3. Продолжим рассмотрение ARE из примера 2. На фиг. 2 приведено расположение уникальных элементов в произведении $P \otimes P$ в соответствии с определением множества $\wp$.

Таким образом, $p = 1, 2; q = 1, 2$. Согласно сформированным правилам, множество $\wp = \{(1, 1, 1, 1), (2, 1, 1, 1), (2, 2, 1, 1), (2, 1, 2, 1), (2, 2, 2, 1), (2, 2, 2, 2)\}$, а согласно формуле расчета (9), количество уникальных элементов $r = \left(\frac{2^4}{12} + \frac{2^3}{3} + \frac{5 \cdot 2^2}{12} + \frac{2}{6}\right) = 6$.

На основе определения множества $\wp$ составим шесть матриц $E_{\wp_i}$:

$$E_{\wp_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_{\wp_2} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_{\wp_3} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_{\wp_4} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

$$E_{\wp_5} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad E_{\wp_6} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

По формуле (8) сформируем матрицу:

$$\bar{D}^T = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{vec}^T(E_{\wp_1}) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{vec}^T(E_{\wp_2}) + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{vec}^T(E_{\wp_3}) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{vec}^T(E_{\wp_4}) +$$

$$+ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \text{vec}^T(E_{\wp_5}) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \text{vec}^T(E_{\wp_6}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Все ее столбцы уникальны, поэтому матрица $\bar{L}$ существует и имеет вид

$$\bar{L} = \bar{D}^\dagger = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P \otimes P = \begin{array}{c} \begin{array}{cc} q=1 & q=2 \end{array} \\ \begin{array}{cc} p=1 & p=2 \end{array} \end{array} \begin{bmatrix} \boxed{p_1^2} & p_1 p_2 & p_1 p_2 & p_2^2 \\ p_1 p_2 & \boxed{p_1 p_4} & p_2^2 & p_2 p_4 \\ p_1 p_2 & p_2^2 & p_1 p_4 & p_2 p_4 \\ \boxed{p_2^2} & \boxed{p_2 p_4} & p_2 p_4 & \boxed{p_4^2} \end{bmatrix}$$

Фиг. 2. Расположение уникальных элементов в $P \otimes P$ ($n = 2$).

Тогда, пользуясь четвертым свойством операции из определения 5, по аналогии с (5) получим:

$$\left[L(I_n \otimes A^T + A^T \otimes I_n) D L \left((D \text{vech}(BB^T))^T \otimes I_{n^2} \right) \bar{D} \right] \cdot \begin{bmatrix} L \text{vec}(P) \\ \bar{L} \text{vec}(P \otimes P) \end{bmatrix} = -\text{vech}(Q). \quad (10)$$

Откуда, подставив значения матриц, в рассматриваемом случае получим:

$$\begin{bmatrix} 2a_1 & 2a_3 & 0 \\ a_2 & a_1 + a_4 & a_3 \\ 0 & 2a_2 & 2a_4 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_4 \end{bmatrix} - VP_{\text{kron}} = \begin{bmatrix} q_1 \\ 0 \\ q_4 \end{bmatrix},$$

где $V = \begin{bmatrix} b_1^2 & 2b_1 b_2 & 0 & b_2^2 & 0 & 0 \\ 0 & b_1^2 & b_1 b_2 & b_1 b_2 & b_2^2 & 0 \\ 0 & 0 & 0 & b_1^2 & 2b_1 b_2 & b_2^2 \end{bmatrix}$, а вектор $P_{\text{kron}} = [p_1^2 \ p_1 p_2 \ p_1 p_4 \ p_2^2 \ p_2 p_4 \ p_4^2]^T$.

Таким образом, получена минимально возможная параметризация исходной задачи. При использовании операции $\text{vec}(P \otimes P)$ вектор параметров P_{kron} содержал 16 элементов, после применения операции $\text{vech}(P \otimes P) - 10$ (см. пример 2), а после применения разработанной в работе операции $\text{vesu}(P \otimes P) - 6$. Данный пример демонстрирует практическую ценность предлагаемой операции. В задаче идентификации неизвестных параметров регрессии типа (5) общее число неизвестных параметров было сокращено с 10 до 6, что позволяет избежать излишней перепараметризации и максимально сократить размерность решаемой задачи.

Также на примере (10) наглядно видно, что результат выполнения операции $\text{vesu}(P \otimes P)$ единожды используется для получения параметризации типа (10) для дальнейшего решения ARE. Длительное хранение матриц исключения и дублирования не требуется. Хранится лишь вектор уникальных элементов $P \otimes P$, который по количеству элементов строго меньше общего числа элементов $P \otimes P$.

Пример 4. Пусть $n = 3$. В данном примере ограничимся изображением с отметками уникальных элементов в матрице $P \otimes P$ и подсчетом их количества по формуле (9). На фиг. 3 приведено расположение уникальных элементов в матрице $P \otimes P$ в соответствии с определением множества $\mathcal{Q}$.

Количество уникальных элементов $r = \left(\frac{3^4}{12} + \frac{3^3}{3} + \frac{5 \cdot 3^2}{12} + \frac{3}{6} \right) = 20$, что подтверждается фиг. 3.

При использовании операции $\text{vec}(P \otimes P)$ вектор параметров P_{kron} содержал 81 элемент, после применения операции $\text{vech}(P \otimes P)$ содержал 45 элементов, а после применения разработанной в работе операции $\text{vesu}(P \otimes P)$ содержал 20 элементов.

Несложно убедиться, что для $n = 4$ по сравнению с операцией $\text{vec}(P \otimes P)$ при использовании $\text{vesu}(P \otimes P)$ число параметров сокращается с 256 до 50, для $n = 5$ — с 625 до 105 и т.д.

	$q = 1$	$q = 2$	$q = 3$
$p = 1$	$\boxed{p_1^2}$ p_1p_2 p_1p_3	p_1p_2 p_2^2 p_2p_3	p_1p_3 p_2p_3 p_3^2
	p_1p_2 $\boxed{p_1p_4}$ p_1p_5	p_2^2 p_2p_4 p_2p_5	p_2p_3 p_3p_4 p_3p_5
	p_1p_3 p_1p_5 $\boxed{p_1p_6}$	p_2p_3 p_2p_5 p_2p_6	p_3^2 p_1p_4 p_3p_6
$P \otimes P =$	p_1p_2 p_2^2 p_2p_3	p_1p_4 p_2p_4 p_3p_4	p_1p_5 p_2p_5 p_3p_5
	$\boxed{p_2^2}$ $\boxed{p_2p_4}$ p_2p_5	p_2p_4 $\boxed{p_4^2}$ p_4p_5	p_2p_5 p_4p_5 p_5^2
$p = 2$	p_2p_3 p_2p_5 $\boxed{p_2p_6}$	p_3p_4 $\boxed{p_4p_5}$ $\boxed{p_4p_6}$	p_3p_5 p_5^2 p_5p_6
	p_1p_3 p_2p_3 p_3^2	p_1p_5 p_2p_5 p_3p_5	p_1p_6 p_2p_6 p_3p_6
$p = 3$	p_2p_3 p_3p_4 p_3p_5	p_2p_5 p_4p_5 p_5^2	p_2p_6 p_4p_6 p_5p_6
	$\boxed{p_3^2}$ $\boxed{p_3p_5}$ $\boxed{p_3p_6}$	p_3p_5 $\boxed{p_5^2}$ $\boxed{p_5p_6}$	p_3p_6 p_5p_6 $\boxed{p_6^2}$

Фиг. 3. Расположение уникальных элементов в $P \otimes P$ ($n = 3$).

5. ЗАКЛЮЧЕНИЕ

В данном исследовании разработан конструктивный алгоритм выполнения операции произведения Кронекера симметричной матрицы на саму себя $P \otimes P$ с последующей векторизацией результата таким образом, чтобы получаемый вектор включал в себя только уникальные элементы — пары произведений элементов исходной матрицы. Данный алгоритм основан на определении операции Кронекера и симметричности исходной матрицы. Предложен алгоритм поиска индексов уникальных элементов в $P \otimes P$, а также формулы для получения матриц исключения и дублирования, которые базируются на классической работе [1] по частичной векторизации матриц. Кроме того, была получена зависимость числа таких уникальных элементов от размерности исходной матрицы.

Такой подход позволяет получать минимальную по числу параметров параметризацию уравнения Риккати в виде (1), привнося тем самым необходимое обоснование в работы [13, 14], где подобная процедура была заявлена декларативно.

Предложенная операция может применяться не только в задачах, связанных с решением ARE, но и в любых других, где возникает необходимость векторизации выражения вида $P \otimes P$, при условии $P = P^T$.

Актуальной представляется задача расширения полученного результата на случай векторизации произведений вида $A \otimes B$, где $A \in R^{n \times n}$ и $B \in R^{n \times n}$ — произвольные квадратные матрицы.

В дальнейшем авторы планируют воспользоваться предложенной операцией $\text{vecu}(\cdot)$ и полученными результатами по минимальной параметризации уравнения Риккати для построения адаптивного LQ регулятора.

СПИСОК ЛИТЕРАТУРЫ

1. Magnus J.R., Neudecker H. The elimination matrix: some lemmas and applications // SIAM Journal on Algebraic Discrete Methods. 1980. V. 1 (4). P. 422–449.
2. Nagakura D. On the relationship between the matrix operators, vech and vecd // Communications in Statistics-Theory and Methods. 2018. V. 47(13). P. 3252–3268.
3. Sastry S., Bodson M. Adaptive control: stability, convergence and robustness. Courier Corporation, 2011.
4. Lavretsky E., Wise K.A. Robust adaptive control // Robust and adaptive control. London: Springer, 2013.
5. Ioannou P.A., Sun J. Robust adaptive control. Courier Corporation, 2012.
6. Ortega R., Nikiforov V., Gerasimov D. On modified parameter estimators for identification and adaptive control. A unified framework and some new schemes // Annual Reviews in Control. 2020. V. 50. P. 278–293.
7. Lion P.M. Rapid identification of linear and nonlinear systems // AIAA Journal. 1967. V. 5. P. 1835–1842.

8. *Kreisselmeier G.* Adaptive observers with exponential rate of convergence // IEEE Transactions on Automatic Control. 1977. V. 22 (1). P. 2–8.
9. *Aranovskiy S.* Parameter Estimation with Enhanced Performance. Habilitation. Rennes: Université de Rennes 1, 2021.
10. *Lewis F., Syrmos V.* Optimal control. John Wiley & sons, INC., 2nd edition, 1995.
11. *Kalman R.E. et al.* Contributions to the theory of optimal control // Bol. Soc. Mat. Mexicana. 1960. V. 5 (2). P. 102–119.
12. *Polyak B.T., Shcherbakov P.S.* Hard Problems in Linear Control Theory: Possible Approaches to Solution // Automation and Remote Control. 2005. V. 66(5). P. 681–718.
13. *Jha S.K., Roy S.B., Bhasin S.* Data-driven adaptive LQR for completely unknown LTI systems // IFAC-PapersOnLine. 2017. V. 50 (1). P. 4156–4161.
14. *Jiang Y., Jiang Z.P.* Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics // Automatica. 2012. V. 48 (10). P. 2699–2704.

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.653

ВЫЧИСЛЕНИЕ С МИНИМАЛЬНОЙ ПОГРЕШНОСТЬЮ n -Й ПРОИЗВОДНОЙ ПО ДАННЫМ ИЗМЕРЕНИЯ ФУНКЦИИ

© 2023 г. А. С. Демидов^{1,2,3,*}, А. С. Кочуров^{1,4,**}¹ 119234 Москва, ул. Колмогорова, 1, МГУ, Россия² 123098 Москва, ул. Максимова, 4, МФТИ, Россия³ 115419 Москва, ул. Орджоникидзе, 3, РУДН, Россия⁴ 119991 Москва, МЦФПМ-МГУ, Россия

*e-mail: demidov.alexandre@gmail.com

**e-mail: kchryas@yandex.ru

Поступила в редакцию 01.02.2023 г.

Переработанный вариант 09.03.2023 г.

Принята к публикации 29.05.2023 г.

Предложено решение вопроса, возникающего во всех задачах, где по экспериментальным дискретным данным априори гладкой функции требуется приближенно вычислить ее производные. Вся проблема сводится к поиску “оптимального” шага разностной аппроксимации. Эту проблему исследовали многие математики. Оказалось, что для выбора “оптимального” шага аппроксимации производной k -го порядка надо знать как можно более точную оценку модуля производной порядка $k + 1$. Предложенный в статье алгоритм, дающий такую оценку, применен к задаче о концентрации тромбина, который определяет динамику свертываемости крови. Эта динамика представлена графиками и дает интересующий биофизиков ответ о концентрации тромбина. Библ. 8. Фиг. 7.

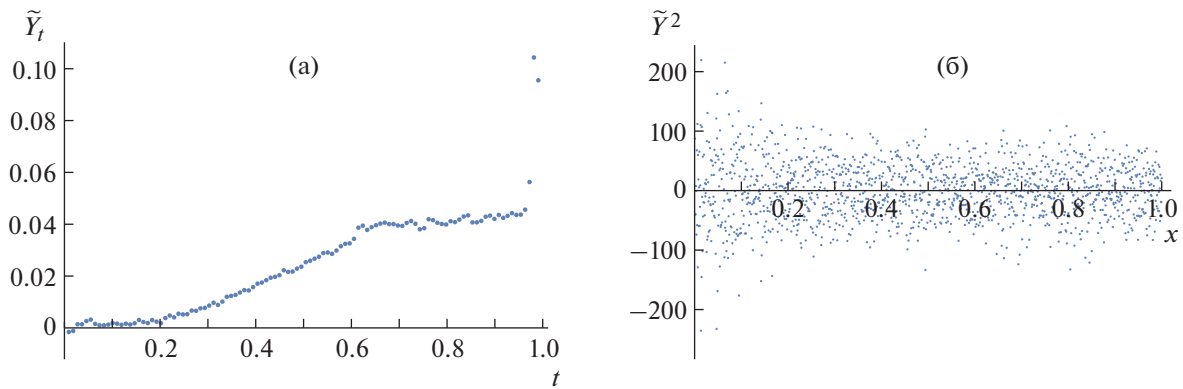
Ключевые слова: восстановление n -й производной, концентрация тромбина, оптимальный шаг аппроксимации, оценка модуля старших производных.

DOI: 10.31857/S0044466923090077, EDN: XMHVDC

1. ВВЕДЕНИЕ

Задача, которой посвящена данная работа, инициирована проблемой, возникшей в ходе исследований, проводимых в лаборатории физической биохимии Национального медицинского исследовательского центра гематологии под руководством академика РАН Ф.И. Атауллаханова. Имеются в виду работы по нахождению концентрации S тромбина (фактора свертывания II) по данным интенсивности I , введенного в кровь фермента, активирующего ее свертываемость. Измерения интенсивности I размерности $[I]$, как функции времени $\theta \in [0, T]$ и глубины $\xi \in [0, X]$ проникновения в кровь, проводились 110 раз с шагом 31.2 сек за время $T = 109 \times 31.2[T] \approx 56.7 \text{ min}$, где $[T] = \text{сек}$ — размерность T . А именно, через каждые 31.2 сек фиксировалась фотография интенсивности проникновения фермента в кровь в 1485 точках, расположенных с шагом $4.3[X]$ вплоть до глубины $X = 1484 \times 4.3[X] \approx 6.4 \text{ mm}$, где $[X] = \text{mkm}$ — размерность X .

Следует отметить, что многие трудные и пока не решенные био-физиологические проблемы свертывания крови были освещены в недавнем обзоре [1] по современным методам компьютерного моделирования тромбоза. Но математическая проблема, с которой столкнулись в лаборатории Ф.И. Атауллаханова, в этом обзоре не обозначена. Как отмечается в статье [2], “в современной концепции свертывания крови является растущее согласие с тем, что для понимания регуляции свертывания крови *in vivo* и нарушений ее функции необходимо принимать во внимание ее пространственную неоднородность, диффузию и поток.” Это утверждение базируется, в частности, на одном из наиболее значимых, и часто цитируемом исследовании [3] о свертывании крови, где проведен детальный анализ соответствующей системы дифференциальных уравнений с частными производными. В простейшей модельной ситуации измеряемая интенсив-



Фиг. 1. (а) график $\tilde{Y}_t(t, x_0)$; (б) график $\tilde{Y}_{xx}(t_0, x)$.

ность I введенного в кровь фермента и искомая концентрация C тромбина связаны уравнением диффузии

$$\frac{\partial I}{\partial \theta} - D \frac{\partial^2 I}{\partial \xi^2} = C(\theta, \xi), \quad \theta \in [0, T], \quad \xi \in [0, X], \quad D = 350 \frac{(\text{mkm})^2}{\text{sec}}.$$

При переходе к безразмерным параметрам

$$t = \frac{\theta}{T}, \quad x = \frac{\xi}{X}, \quad a = D \frac{T}{X^2}$$

получаем для нормированной интенсивности $Y = \frac{I}{T}$ размерности $\frac{[I]}{[T]}$ следующее уравнение:

$$Y_t(t, x) - a Y_{xx}(t, x) = f(t, x), \quad \text{где} \quad f(t, x) = C(tT, xX), \quad \text{а} \quad (t, x) \in [0, 1] \times [0, 1]. \quad (1)$$

Чтобы найти искомую концентрацию C тромбина, представленную функцией f , остается найти с приемлемой точностью производные $Y_t(t, x)$ и $Y_{xx}(t, x)$ функции $Y = \frac{I}{T}$. При этом

$$Y(t_i, x_k) = \tilde{Y}(t_i, x_k) + \delta(t_i, x_k), \quad |\delta(t_i, x_k)| \leq \sigma |\tilde{Y}(t_i, x_k)|, \quad (2)$$

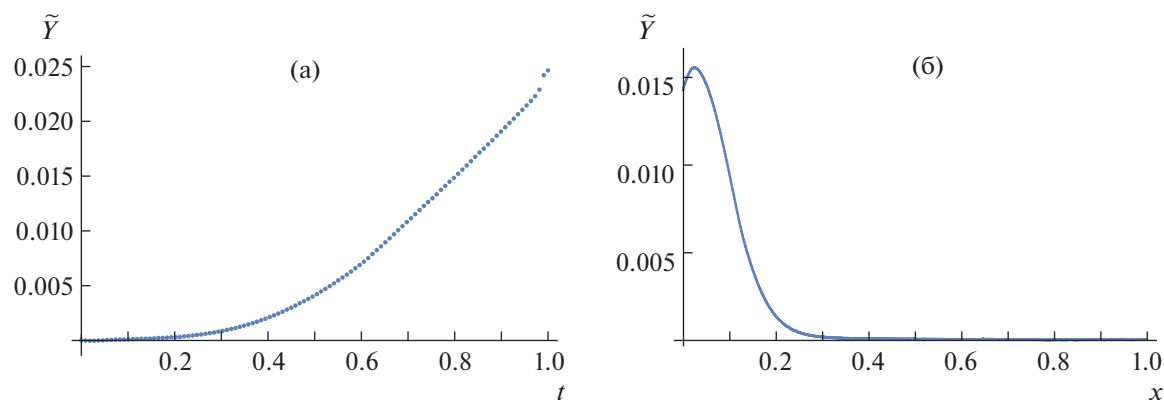
где $\tilde{Y}(t_i, x_k)$ – экспериментальные значения в точках $(t_i, x_k) = (\theta_i/T, \xi_k/T) \in [0, 1] \times [0, 1]$ нормированной интенсивности введенного в кровь фермента, а $\sigma = 0.05$ – относительная точность измерения прибора.

Казалось бы, все просто: искомая концентрация тромбина задана простой формулой $f = \tilde{Y}_t - a \tilde{Y}_{xx}$, а дискретные данные функции $\tilde{Y}$ известны.

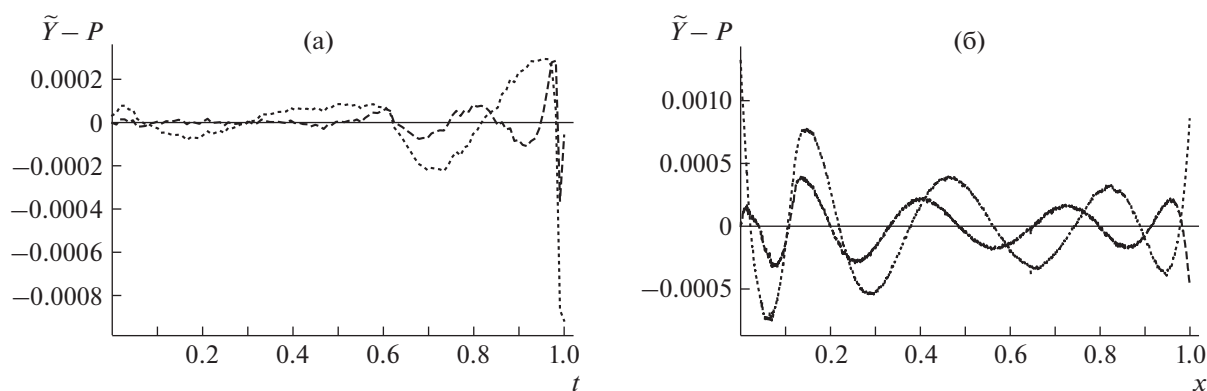
Но как найти с приемлемой точностью производные $\tilde{Y}_t(t_i, x_k)$ и $\tilde{Y}_{xx}(t_i, x_k)$?

С этим вопросом к авторам статьи обратился Ф.И. Атауллаханов. И вопрос этот вовсе не праздный. Его трудность иллюстрируется на фиг. 1а и особенно на фиг. 1б “невразумительными” графиками разностных аппроксимаций производных $\tilde{Y}_t(t, x)|_{x=0.202}$ и $\tilde{Y}_{xx}(t, x)|_{t=0.367}$, полученных непосредственно по “вполне вразумительным” данным измерений $\tilde{Y}(t, 0.202)$ и $\tilde{Y}(0.367, x)$, представленным на фиг. 2.

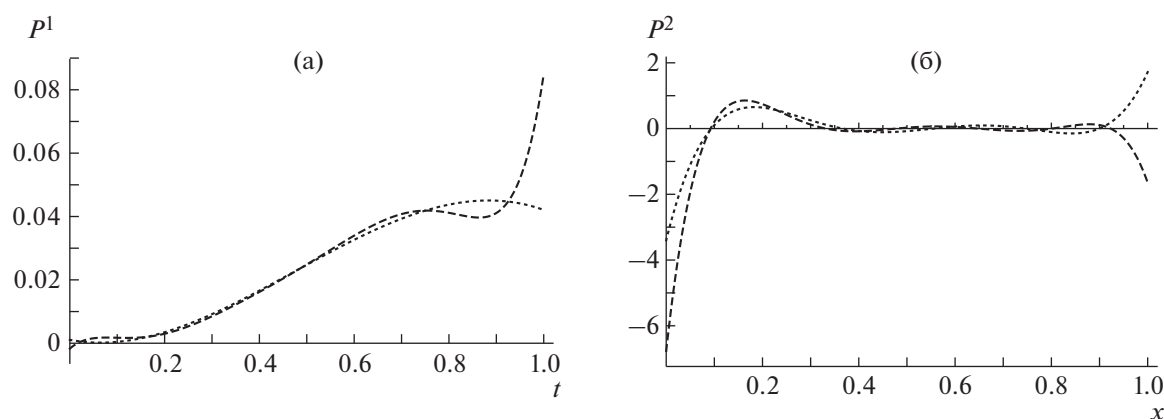
Данные измерений $\tilde{Y}$ гладкой функции Y необходимо дискретны, а приближение дискретнозначной функции $\tilde{Y}$ той или иной гладкой функцией P не влечет близости $P_t(t, x)$ и $P_{xx}(t, x)$ к искомому производным $Y_t(t, x)$ и $Y_{xx}(t, x)$. Этот общеизвестный факт проиллюстрирован на фиг. 3, 4. А именно, на фиг. 3а даны графики разности между $\tilde{Y}(t, 0.202)$ и аппроксимирующими полиномами степени $n = 4$ и $n = 8$ (с относительной точностью приближения порядка 0.037 и 0.014 соответственно), а на фиг. 3б графики разности между $\tilde{Y}(0.367, x)$ и соответствующими ап-



Фиг. 2. (а) график $\tilde{Y}(t, x_0)$; (б) график $\tilde{Y}(t_0, x)$.



Фиг. 3.



Фиг. 4. (а) $P^1 = P_t$ — для $n = 4$ и $n = 8$; (б) $P^2 = P_{xx}$ — для $n = 7$ и $n = 8$.

проксимирующими полиномами степени $n = 7$ и $n = 8$ (с относительной точностью приближения порядка 0.082 и 0.029 соответственно).

И несмотря на такую относительно приемлемую аппроксимацию функций $\tilde{Y}(t, 0.202)$ и $\tilde{Y}(0.367, x)$ аппроксимирующими полиномами, различие производных этих полиномов, графич-

ки которых представлены на фиг. 4а и фиг. 4б, достигает 100%, что не позволяет сказать что-либо определенное об искомым производных $Y_t(t, x)$ и $Y_{xx}(t, x)$.

Обозначенная проблема построения производных функции по данным ее экспериментальных измерений относится к задаче о наилучшем приближении неограниченного линейного оператора линейными операторами на классе элементов банахова пространства. Эта задача разрабатывается со второй половины 20-го века многими видными математиками. Она и родственные ей экстремальные задачи подробно освещены в недавнем обзоре [4]. Но собственно обозначенная проблема, конкретизация которой представлена в разд. 1, там не исследовалась. Частично ответ на нее дан в разд. 2, где используется приведенный в разд. 3 алгоритм.

2. ВОПРОС АТАУЛЛАХАНОВА КАК ПРОБЛЕМА ВЫБОРА ШАГА АППРОКСИМАЦИИ

Экспериментальные измерения функции I проводились в лаборатории физической биохимии Гематологического Научного Центра РАМН посредством регистрации фото- или кино-камерой интенсивности люминесценции, введенного в кровь фермента, активирующего свертываемость крови. Эти, как и любые экспериментальные измерения, необходимо дискретны по времени и пространственной переменной. В данном случае по экспериментальным измерениям $\tilde{I}(\theta_i/T, \xi_k/X)$ были получены значения $\tilde{Y}(t_i, x_k)$ в дискретных точках $t_i = i\tau$ ($i = 0, \dots, 109$) и $x_k = kh$ ($k = 0, \dots, 1484$), где

$$\tau = 31.2/3400.8 = 1/109 \quad \text{и} \quad h = 4.3/6381.2 = 1/1484,$$

соответствующие шагу измерения 31.2 sec по времени и шагу 4.3 mkm по глубине проникновения фермента I .

По значениям $\tilde{Y}$ в точках $(t_i, x_k) \in [0, 1] \times [0, 1]$

$$(t_i, x_k) = (i\tau, kh), \quad (t_i \pm j\tau, 0), x_k \pm l(0, h) (t_i \pm j\tau, x_k \pm lh),$$

где j и l — некоторые натуральные числа, вычислялись приближения $\tilde{Y}_{j(\tau, 0)}^1(t, x)$ и $\tilde{Y}_{l(0, h)}^2(t, x)$ частных производных

$$Y_t(t, x) \approx \tilde{Y}_{j(\tau, 0)}^1(t, x) \stackrel{\text{def}}{=} \frac{\tilde{Y}(t + j\tau, x) - \tilde{Y}(t - j\tau, x)}{2j\tau},$$

$$Y_{xx}(t, x) \approx \tilde{Y}_{l(0, h)}^2(t, x) \stackrel{\text{def}}{=} \frac{\tilde{Y}(t, x + lh) - 2\tilde{Y}(t, x) + \tilde{Y}(t, x - lh)}{(lh)^2}.$$

В том случае, когда $j = l = 1$, т.е. когда разностные аппроксимации производных вычисляются с шагом проведенных экспериментальных измерений, результаты вычислений $\tilde{Y}_{j(\tau, 0)}^1(t, x_0)$ и $\tilde{Y}_{l(0, h)}^2(t_0, x)$ были выше представлены на фиг. 1 (для $k = 300$, $i = 40$, т.е. для $x_0 = kh = \frac{300}{1484} \approx 0.202$, $t_0 = i\tau = \frac{40}{109} \approx 0.367$). Бросается в глаза особенная по сравнению с графиком $\tilde{Y}^1 = \tilde{Y}_{(\tau, 0)}^1(t, x_0)$

“невразумительность” графика $\tilde{Y}^2 = \tilde{Y}_{(0, h)}^2(t_0, x)$. Это связано не столько с тем, что это графики производным различных порядков (а именно, 1-го по t и 2-го по x), сколько с существенным различием величины шагов $\tau = 1/109$ и $h = 1/1484$. Действительно, чем меньше шаг, тем больше отклонение от истинного значения производной за счет слагаемого в виде разностного отношения для погрешности измерения, в знаменателе которого стоит шаг аппроксимации. С другой стороны, с ростом шага аппроксимации растет остаточный член в аппроксимационной формуле Тейлора, что увеличивает погрешность аппроксимации рассматриваемой производной.

Проблема выбора шага аппроксимации обычно решается последовательным его увеличением (например, в два раза) и этот процесс завершается тогда, когда графическое изображение разностной аппроксимации визуально будет мало отличаться от графика гладкой функции. Но при этом есть немалый риск сгладить существенные особенности искомой производной. А там, где производная мала, на первый план выступают погрешности измерений в виде “шерстки”, с которыми можно мириться, поскольку это не затрагивает принципиальные особенности производной. Кроме того, конечно, желательно иметь более обоснованный выбор шага аппроксимации и избавиться от необходимости визуального наблюдения. Следует при этом сказать, что во-

прос выбора оптимального шага аппроксимации s_{opt} при реконструкции производной n -го порядка функции одной переменной $y : s \mapsto y(s)$, равной, аналогично (2), сумме измеренной и погрешности $\delta(s)$, был рассмотрен в работе [5] для $n = 1$ и $n = 2$, в [6] для $n = 3$ и $n = 4$, а для произвольного n – в [7]. Так, С.Б. Стечкин доказал, что шаг аппроксимации для функции s_{opt} в следующих оценках:

$$\left| \dot{y}(s) - \frac{\tilde{y}(s + s_{\text{opt}}) - \tilde{y}(s - s_{\text{opt}})}{2s_{\text{opt}}} \right| \leq \frac{\|\delta\|}{s_{\text{opt}}} + \frac{s_{\text{opt}}}{2} \|\ddot{y}\| = \sqrt{2\|\delta\| \cdot \|\ddot{y}\|}, \quad \text{где} \quad s_{\text{opt}} = \sqrt{\frac{2\|\delta\|}{\|\ddot{y}\|}}, \quad (3)$$

$$\left| \ddot{y}(s) - \frac{\tilde{y}(s + s_{\text{opt}}) - 2\tilde{y}(s) + \tilde{y}(s - s_{\text{opt}})}{s_{\text{opt}}^2} \right| \leq \frac{4\|\delta\|}{s_{\text{opt}}^2} + \frac{s_{\text{opt}}}{3} \|y^{(3)}\|, \quad \text{где} \quad s_{\text{opt}} = \left(\frac{24\|\delta\|}{\|y^{(3)}\|} \right)^{1/3} \quad (4)$$

для функции $y \in C^{n+1}$ (соответственно, $n = 1, 2$), будет оптимальным, в том смысле, что неравенства (3), (4), где $\|\delta\| = \sup_s |\delta(s)| \approx \sigma \|\ddot{y}\|$, выполняются для всех функций $y \in C^{n+1}$ и обращаются в равенства для некоторых из них.

Поскольку функция $\tilde{y}$ измеряется в дискретном наборе точек s_i , параметр s_{opt} необходимо заменяется на некоторое ближайшее к нему. В том случае, когда $s_i = i\eta$, параметр s_{opt} заменяется на

$$\Delta_{\text{opt}} = \mu\eta \in [(\mu - 1/2)\eta, (\mu + 1/2)\eta], \quad (5)$$

где натуральное число μ соответствует оптимальному шагу $\mu\eta$ аппроксимации рассматриваемой производной.

В реальных задачах, кроме принадлежности функции $y : \mathbb{R} \ni s \mapsto y(s)$ к классу C^n , практически всегда имеется дополнительная информация, например, бывает известно, где функция сильно (мало) меняется, где она принимает большие (малые) значения. Таким образом, возникает чисто математическая проблема: улучшить оценки (3) и (4) при наличии той или иной дополнительной информации. Впрочем и формулы (3), (4) дают ориентир для выбора подходящего шага приближения производной $y^{(n)}$. Но для этого требуется как можно более точно оценить норму производной $y^{(n+1)}$. Алгоритм, позволяющий получить такую оценку для большинства реальных задач, дан в разд. 3.

3. ОТВЕТ НА ВОПРОС АТАУЛЛАХАНОВА

В случае упомянутой во введении задачи о концентрации тромбина, приближенные значения $\tilde{Y}'_t(t, x)$ и $\tilde{Y}''_x(t, x)$ производных $Y'_t(t, x)$ и $Y''_x(t, x)$ в левой части уравнения (1) задавались, соответственно, в виде

$$\tilde{Y}'_t(t, x_*) = \dot{y}(t) \quad \text{для} \quad x_* = kh, \quad h = 1/1484, \quad k = 0, \dots, 1484,$$

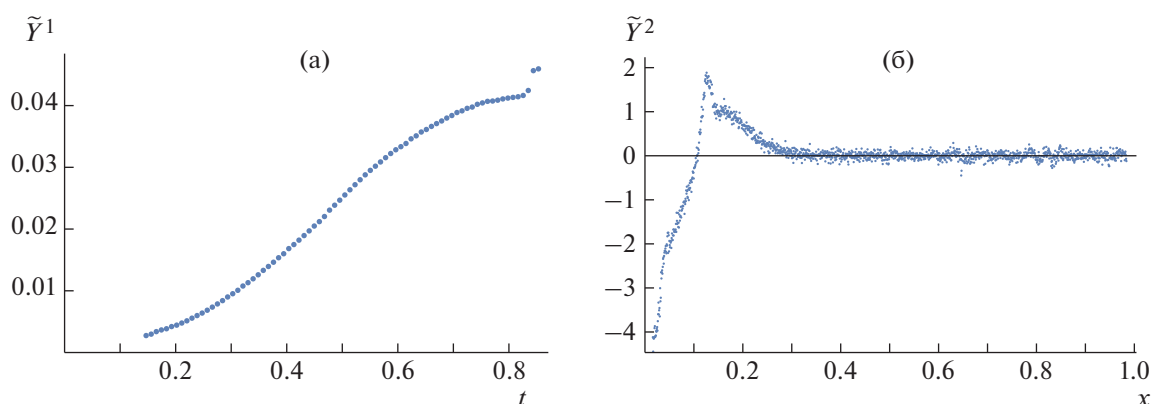
$$\tilde{Y}''_x(t_*, x) = \ddot{y}(t) \quad \text{для} \quad t_* = i\tau, \quad \tau = 1/109, \quad i = 0, \dots, 109,$$

где производные $\dot{y}(s)$ и $\ddot{y}(s)$ функции $y : [0, 1] \ni s \mapsto y(s)$ вычислялись по формулам (3) и (4) с заменой s на $s_i = i\eta$ и s_{opt} на $\Delta_{\text{opt}} \stackrel{(5)}{=} \mu\eta$. При этом нормы $\|\ddot{y}\|$ и $\|y^{(3)}\|$ оценивались, следуя алгоритму разд. 4.

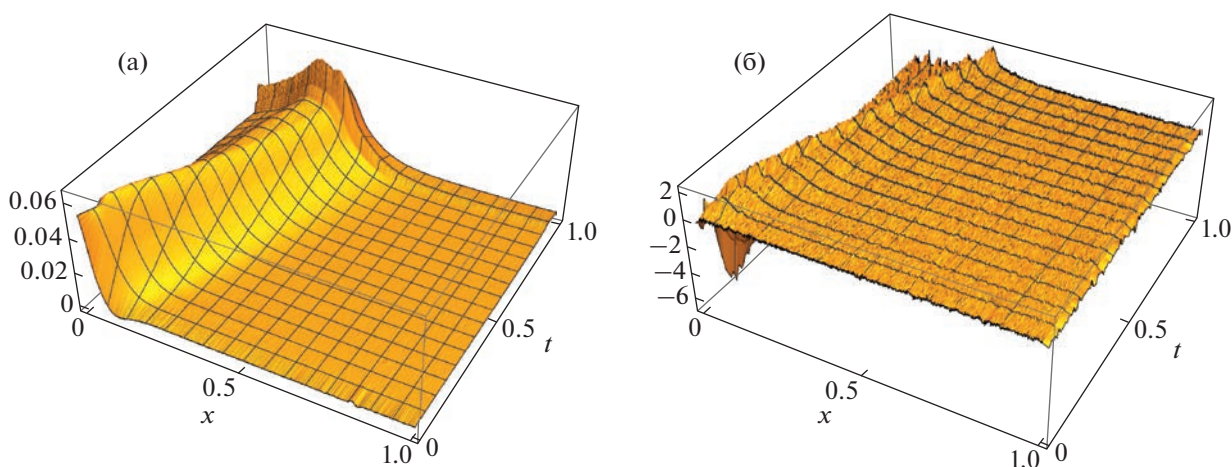
Графики $\tilde{Y}'_t(t, x_*)$ и $\tilde{Y}''_x(t_*, x)$ изображены на фиг. 5. Сравнение этих графиков с представленными на фиг. 1 демонстрируют важность выбора оптимального шага аппроксимаций производных.

На фиг. 6 представлены графики

$$\tilde{Y}_t : [0, 1] \times [0, 1] \ni (t, x) \mapsto \tilde{Y}_t(t, x) \quad \text{и} \quad \tilde{Y}_{xx}(t, x) : [0, 1] \times [0, 1] \ni (t, x) \mapsto \tilde{Y}_{xx}(t, x).$$



Фиг. 5.



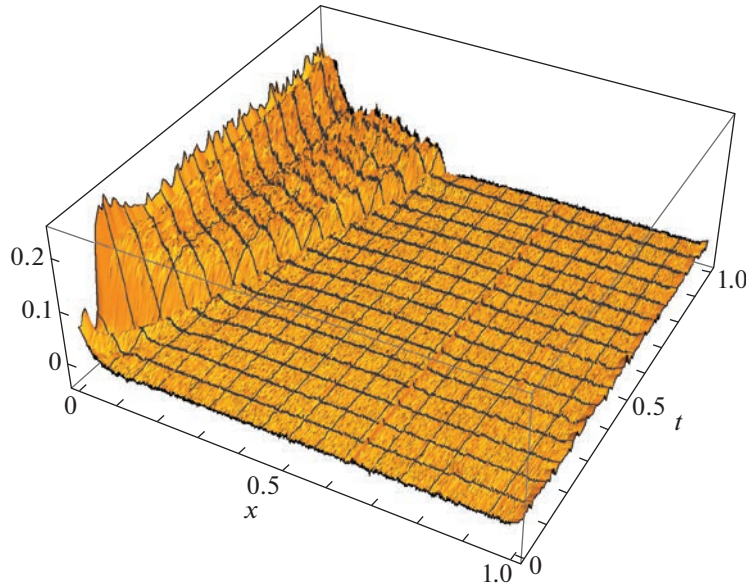
Фиг. 6.

А на фиг. 7 представлен график искомого приближения правой части $f(t, x) = C(tT, xX)$ формулы (1). При этом шаг для вычисления производной по каждой переменной определялся в два этапа. Сначала при каждом фиксированном значении x находился шаг по t в соответствии с представленным в п. 3 алгоритмом. А затем в качестве итогового шага по t был взят шаг, равный среднему арифметическому шагов, вычисленных при каждом фиксированном значении x . Он оказался равным 10τ , в то время как минимальный (соответственно, максимальный) составлял 7τ (соответственно, 18τ). Шаг по переменной x выбирался переменной местами t и x . Этот средний шаг оказался равным $24h$, в то время как минимальный (соответственно, максимальный) составил $7h$ (соответственно, $35h$).

Отметим также, что на прямоугольнике $[\Delta_{\text{орт}_t}, 1 - \Delta_{\text{орт}_t}] \times [\Delta_{\text{орт}_x}, 1 - \Delta_{\text{орт}_x}]$ производные были подсчитаны по симметричным разностям, а вне этого прямоугольника по соответствующим левым или правым разностям.

4. ОЦЕНКА НОРМЫ ПРОИЗВОДНЫХ ФУНКЦИИ ОДНОЙ ПЕРЕМЕННОЙ ПО ЕЕ ДИСКРЕТНЫМ ЗНАЧЕНИЯМ

Обозначим через N число точек $s_i = i\eta \in [0, 1]$, $i = 0, \dots, N-1$, в которых известны приближенные значения $\tilde{y} : s \mapsto y(s) - \delta(s)$ гладкой функции y . Для выбора оптимального шага аппрок-



Фиг. 7.

симации производной $y^{(n)}$ порядка n требуется, как было отмечено выше, знание весьма точной оценки нормы

$$\|y^{(n+1)}\| = \max_{s \in [0,1]} |y^{(n+1)}(s)|.$$

С этой целью введем двумерный массив отрезков $[s_i, s_j]$ по $i = 0, \dots, N-1$ и $j = i + (n+1)\ell \leq N-1$, где ℓ – натуральное число, такое, что $s_j = \eta j \leq 1$. Заметим, что

$$\|y^{(n+1)}\| = \max_{i,j} \max_{s \in [s_i, s_j]} |y^{(n+1)}(s)|.$$

Для функции $\tilde{y} = s \mapsto y(s) - \delta(s)$ рассмотрим так называемую (см., например, [8] стр. 231) разделенную разность $(n+1)$ -го порядка в точке s_i с шагом $\ell\eta$, относящуюся к отрезку $[s_i, s_{i+(n+1)\ell}]$. Она задается формулой

$$\tilde{y}_{i,j}^{(n+1)} \stackrel{\text{def}}{=} \frac{\sum_{k=0}^{n+1} (-1)^{n+1-k} C_{n+1}^k \tilde{y}(s_{i+k\ell})}{(\ell\eta)^{n+1}} = \frac{\sum_{k=0}^{n+1} (-1)^{n+1-k} C_{n+1}^k y(s_{i+k\ell})}{(\ell\eta)^{n+1}} - \frac{\sum_{k=0}^{n+1} (-1)^{n+1-k} C_{n+1}^k \delta(s_{i+k\ell})}{(\ell\eta)^{n+1}} \quad (6)$$

и удовлетворяет условию

$$y^{(n+1)}(\zeta) = \tilde{y}_{i,j}^{(n+1)} + O(\|\delta\|/(\ell\eta)^{n+1}), \quad (7)$$

где ζ – некоторая точка на $[s_i, s_j] = [s_i, s_{i+(n+1)\ell}]$, а $\|\delta\| = \sup_s |\delta(s)|$ есть норма погрешности измерения. Учитывая (7), будем принимать за искомое значение нормы $\|y^{(n+1)}\|$ оценку такой величины

$$\|\tilde{y}^{(n+1)}\| \stackrel{\text{def}}{=} \max_{i,j} |\tilde{y}_{i,j}^{(n+1)}|, \quad j = i + (n+1)\ell.$$

Как следует из формулы (6), величины $|y_{i,j}^{(n+1)}|$, а тем более, $|\tilde{y}_{i,j}^{(n+1)}|$ очень сильно зависят от положения и длины отрезка $[s_i, s_j]$. Это затрудняет поиск устойчивого максимума этих величин, который можно было бы принять за приемлемую оценку $\|\tilde{y}^{(n+1)}\|$. Ниже излагается алгоритм из трех последовательных действий (шагов), позволяющий преодолеть эту трудность. На первых двух

шагах элементы $|y_{i,j}^{(n+1)}|$ выстраиваются в неубывающую последовательность, в “хвосте” которой концентрируются разностные отношения, вычисленные с большой погрешностью. Это построение осуществляется с помощью функций $\mathcal{Y}$ и $\mathcal{M}$, формальное определение которых иллюстрируется простым примером. “Хвост” последовательности отделяется на 3-м шаге алгоритма параметром $\hat{\kappa}$, а за приемлемую оценку $\|\tilde{y}^{(n+1)}\|$ принимается $\mathcal{M}(\hat{\kappa})$.

Шаг 1. Из двумерного массива элементов $|\tilde{y}_{i,j}^{(n+1)}|$ формируется набор

$$\mathcal{Y}^{(n+1)} = \{\mathcal{Y}_0^{(n+1)}, \mathcal{Y}_1^{(n+1)}, \dots, \mathcal{Y}_{N-n-2}^{(n+1)}\}$$

из семейств

$$\mathcal{Y}_i^{(n+1)} = \{|\tilde{y}_{i,i+1}^{(n+1)}|, |\tilde{y}_{i,i+2}^{(n+1)}|, |\tilde{y}_{i,i+3}^{(n+1)}|, \dots, |\tilde{y}_{i,i+k_i}^{(n+1)}|\}, \quad i = 0, \dots, N-n-2.$$

В семействе $\mathcal{Y}_i^{(n+1)}$ есть ровно $k_i = \lfloor (N-1-i)/(n+1) \rfloor$ элементов. В частности, в последнем семействе лишь один элемент $|\tilde{y}_{N-n-2,N-1}^{(n+1)}|$, а общее число элементов в наборе $\mathcal{Y}^{(n+1)}$ равно

$$\mathcal{K} = k_0 + k_1 + \dots + k_{N-n-2}.$$

Пусть $\kappa \in [1, \mathcal{K}]$ — *порядковый* номер элемента $|\tilde{y}_{i,j}^{(n+1)}|$ в последовательном одномерном (линейном) наборе $\mathcal{Y}^{(n+1)}$. Будем ассоциировать набор $\mathcal{Y}^{(n+1)}$ с функцией

$$\mathcal{Y}^{(n+1)} : [1, \mathcal{K}] \ni \kappa \mapsto \mathcal{Y}^{(n+1)}(\kappa),$$

где $\mathcal{Y}^{(n+1)}(\kappa)$ есть элемент $|\tilde{y}_{i,j}^{(n+1)}|$ набора $\mathcal{Y}^{(n+1)}$ с номером κ .

Шаг 2. Положим $\bar{\omega} = (\omega_1, \omega_2, \dots, \omega_{\mathcal{K}})$, так, что $\mathcal{Y}(\omega_{\kappa}) \leq \mathcal{Y}(\omega_m)$, если $\kappa < m$. Введем функцию

$$\mathcal{M} = \mathcal{M}_{\tilde{y}}^{(n+1)} : [1, \mathcal{K}] \ni \kappa \mapsto \mathcal{M}_{\tilde{y}}^{(n+1)}(\kappa) = \mathcal{Y}^{(n+1)}(\omega_{\kappa}).$$

Пример с $\|\delta\| = 0$ при $n = 1$, $N = 10$ (тем самым, $\mathcal{K}$), $y : [0, 1] \ni s \mapsto y(s) = 3s^3/2$. Тогда

$$|\tilde{y}_{i,i+2k}^{(n+1)}| = |y_{i,i+2k}^{(n+1)}| = 9 \frac{i+k}{N-1} = i+k \text{ и на шаге 1 получаем}$$

$$\mathcal{Y}^{(2)} = \{(1, 2, 3, 4)_0, (2, 3, 4, 5)_1, (3, 4, 5, 6)_2, (4, 5, 6, 7)_3, (5, 6, 7, 8)_4, (6, 7, 8, 9)_5, (7, 8, 9, 10)_6, (8, 9, 10, 11)_7\}.$$

Здесь нижний индекс i у скобок соответствует семейству $\mathcal{Y}_i^{(2)}$. Шаг 2 в этом примере дает

$$\bar{\omega} = (1, 2, 5, 3, 6, 9, 4, 7, 10, 12, 8, 11, 13, 15, 14, 16, 17, 18, 19, 20) \quad \text{и}$$

$$\mathcal{M}_{\tilde{y}}^{(2)}(\kappa) \in \{1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8\}.$$

Шаг 3. Приближим график функции $\mathcal{M}_{\tilde{y}}^{(n+1)}$ графиком кусочно-линейной функции

$$k \mapsto V_{a_1, b_1, a_2, b_2, \kappa}(k) = \begin{cases} a_1 k + b_1 & \text{при } 1 \leq k \leq \kappa, \\ a_2 k + b_2 & \text{при } \kappa < k \leq \mathcal{K}, \end{cases}$$

параметризованной числами $a_1, b_1, a_2, b_2, \kappa$. Приближение достигается минимизацией

$$S(a_1, b_1, a_2, b_2, \kappa) = \sum_{k=1}^{\kappa} (\tilde{y}(k) - a_1 k - b_1)^2 + \sum_{k=\kappa+1}^{\mathcal{K}} \left(k - \frac{\tilde{y}(k) - b_2}{a_2} \right)^2 (\tilde{y}(k) - \tilde{y}(k-1)) \quad (8)$$

по параметрам $a_1, b_1, a_2, b_2, \kappa$.

Предложение 1. Для нахождения минимума функции (8) требуется линейно зависящее от $\mathcal{K}$ число операций сложения и умножения.

В самом деле, достаточно решить для каждого κ две задачи минимизации:

$$\min_{a_1, b_1} \sum_{k=1}^{\kappa} (\tilde{y}(k) - a_1 k - b_1)^2, \quad \min_{a_2, b_2} \sum_{k=\kappa+1}^{\mathcal{K}} \left(k - \frac{\tilde{y}(k) - b_2}{a_2} \right)^2 (\tilde{y}(k) - \tilde{y}(k-1)).$$

Если $\kappa = \mathcal{H}$, то остается лишь первая из этих задач минимизации и вторая, если $\kappa = 0$.

Рассмотрим 1-ю из этих задач для какого-либо $\kappa = 1, \dots, \mathcal{K}$. Раскрывая выражение, стоящее под знаком суммы, и производя суммирование по k получившихся слагаемых, получим

$$-2a_1 \sum_{k=1}^{\kappa} k\tilde{y}(k) - 2b_1 \sum_{k=1}^{\kappa} \tilde{y}(k) + \sum_{k=1}^{\kappa} \tilde{y}(k)^2 + \frac{1}{6}a_1^2\kappa(\kappa+1)(2\kappa+1) + a_1b_1\kappa(\kappa+1) + b_1^2\kappa.$$

Поэтому минимум по $a_1 = a_1(\kappa)$, $b_1 = b_1(\kappa)$ в 1-й задаче минимизации полностью определяется значениями

$$\alpha(\kappa) = \sum_{k=1}^{\kappa} k\tilde{y}(k), \quad \beta(\kappa) = \sum_{k=1}^{\kappa} \tilde{y}(k), \quad \gamma(\kappa) = \sum_{k=1}^{\kappa} \tilde{y}(k)^2,$$

а чтобы решить первые задачи нахождения минимума для всех κ , нужно знать все $\alpha(\kappa)$, $\beta(\kappa)$, $\gamma(\kappa)$, $\kappa = 1, \dots, \mathcal{H}$. Так как

$$\alpha(\kappa) = \alpha(\kappa-1) + \kappa\tilde{y}(\kappa), \quad \beta(\kappa) = \beta(\kappa-1) + \tilde{y}(\kappa), \quad \gamma(\kappa) = \gamma(\kappa-1) + \tilde{y}(\kappa)^2,$$

то для нахождения всех $\alpha(\kappa)$, $\beta(\kappa)$, $\gamma(\kappa)$, $\kappa = 1, \dots, \mathcal{H}$, нужно выполнить операций сложения и умножения в количестве, линейно зависящем от $\mathcal{H}$. Точно также и для вторых задач минимизации (по a_2, b_2) находятся функции $\kappa \mapsto a_2(\kappa)$ и $\kappa \mapsto b_2(\kappa)$. Для их нахождения также требуется число сложений и умножений, линейно зависящее от $\mathcal{H}$.

Положив $S^*(\kappa) \stackrel{(8)}{=} S(a_1(\kappa), b_1(\kappa), a_2(\kappa), b_2(\kappa), \kappa)$ и определив равенством

$$S^*(\kappa^*) \stackrel{\text{def}}{=} \min_{\kappa \in [1, \mathcal{H}]} S^*(\kappa),$$

число κ^* , завершаем доказательство предложения 1.

По найденным параметрам κ^* , $a_1(\kappa^*)$, $b_1(\kappa^*)$ задачи минимизации (8) определяется следующим образом искомое число $\hat{\kappa}$, отделяющее “хвост” последовательности $\mathcal{M}(\kappa)$. А именно,

$$\hat{\kappa} = \max \{ \kappa < \kappa^* : \mathcal{M}(\kappa) < a_1(\kappa^*)\kappa + b_1(\kappa^*) \}.$$

Итак, за приемлемую оценку $\|\tilde{y}^{(n+1)}\|$ принимается $\mathcal{M}(\hat{\kappa})$.

5. ЗАКЛЮЧЕНИЕ

Рассмотрена трудная и важная прикладная задача о вычислении производных априори гладкой функции по экспериментальным данным ее значений в дискретных точках. На практике этот вопрос обычно решается последовательным увеличением шага аппроксимации (например, в два раза) и этот процесс завершается тогда, когда графическое изображение разностной аппроксимации визуальное будет мало отличаться от графика гладкой функции. Однако при этом есть немалый риск сгладить существенные особенности искомой производной. Кроме того, конечно, желательно иметь более обоснованный выбор шага аппроксимации и избавиться от необходимости визуального наблюдения. Обе эти негативные особенности отсутствуют в алгоритме, предложенном в данной работе. Это позволило дать ответ на вопрос Ф.И. Атауллаханова, касающийся проблемы определения концентрации тромбина по данным интенсивности введенного в кровь фермента, активирующего ее свертываемость. Суть проблемы в том, что искомая концентрации тромбина задается правой частью уравнения диффузии, в левой части которого стоит разность производной первого порядка по времени и производной второго порядка по пространственной переменной интенсивности фермента, которая определяется экспериментально.

СПИСОК ЛИТЕРАТУРЫ

1. Dunster J.L., Gibbins J.M., Pantelev M.A., Volpert V. Modeling thrombosis in silico: Frontiers, challenges, unresolved problems and milestones // *Physics of Life Reviews*. 2018. Vol. 26–27. P. 57–95. <https://doi.org/10.1016/j.plrev.2018.02.005>

2. *Panteleev M.A., Dashkevich N.M., Ataullakhanov F.I.* Hemostasis and thrombosis beyond biochemistry: roles of geometry, flow and diffusion // *Thrombosis Research*. 2015. Vol. 136. No 4. P. 699–711. Epub 2015 Jul 29. Review. PubMed PMID: 26278966.
<https://doi.org/10.1016/j.thromres.2015.07.025>
3. *Атауллаханов Ф.И., Лобанова Е.С., Морозова О.Л., Шноль Э.Э., Ермакова Е.А., Бутылин А.А., Заикин А.Н.* Сложные режимы распространения возбуждения и самоорганизация в модели свертывания крови // *Успехи физ. наук*. 2007. Т. 177. № 1. С. 87–104.
4. *Арестов В.В., Акопян Р.Р.* Задача Стечкина о наилучшем приближении неограниченного оператора ограниченными и родственные ей задачи // *Тр. Ин-та матем. и механ. УрО РАН*. 2020. Т. 26. № 4. С. 7–31.
5. *Стечкин С.Б.* Неравенства между нормами производных произвольной функции // *Acta scient. math*. 1965. Vol. 26. № 3–4. P. 225–230.
6. *Арестов В.В.* О наилучшем приближении операторов дифференцирования // *Матем. заметки*. 1967. Т. 1. № 2. С. 149–154.
7. *Буслаев А.П.* О приближении оператора дифференцирования // *Матем. заметки*. 1981. Т. 29. № 5. С. 372–378.
8. *Бабенко К.И.* Основы численного анализа. Москва-Ижевск: НИЦ “Регулярная и хаотическая динамика”, 2002. С. 3–847.

ОБЩИЕ ЧИСЛЕННЫЕ
МЕТОДЫ

УДК 517.929

ЧИСЛЕННЫЙ МЕТОД ОЦЕНКИ СКОРОСТИ РОСТА ОШИБОК
ОКРУГЛЕНИЯ В РАВНОМЕРНОЙ МЕТРИКЕ© 2023 г. М. И. Зуев¹, С. И. Сердюкова^{1,*}¹ 141980 Дубна, М.о., ул. Жолио-Кюри, 6, ЛИТ ОИЯТ, Россия

*e-mail: sis@jinr.ru

Поступила в редакцию 16.02.2023 г.
Переработанный вариант 20.03.2023 г.
Принята к публикации 28.04.2023 г.

В настоящей работе разработан численно-аналитический алгоритм оценки ошибок округления в равномерной метрике. Установлена их ограниченность на всем интервале вычисления вольт-амперных характеристик длинных джозефсоновских переходов при использовании предлагаемой схемы второго порядка точности. На примере системы двух разностных уравнений показано, как можно исследовать численно скорость роста ошибок округления в равномерной метрике в случае степенной неустойчивости. Кроме того, получены оценки скорости роста ошибок округления в равномерной метрике для схемы Русанова третьего порядка точности. Расчеты проводились на суперкомпьютере “Говорун” с использованием системы REDUCE. Библ. 9. Фиг. 6.

Ключевые слова: конечно-разностные методы, оценка роста ошибок округления в равномерной метрике, численный метод, система REDUCE, суперкомпьютер “Говорун”.

DOI: 10.31857/S0044466923080173, EDN: WTZUZO

ВВЕДЕНИЕ

Цель этой работы — показать, как можно оценивать скорость роста ошибок округления в равномерной метрике, используя численные методы. Устойчивые в равномерной метрике разностные схемы пригодны для счета разрывных решений. Равномерная метрика является более естественной и наглядной по сравнению с l_2 метрикой. Но для получения оценок в равномерной метрике приходится использовать более сложные методы. Так, основные результаты по исследованию устойчивости разностных схем в равномерной метрике были получены с помощью метода перевала. При этом, кроме определения точек перевала и линий наискорейшего спуска, немалого труда стоило доказывать определяющую роль “асимптотических вычетов”. С появлением суперкомпьютеров возникла реальная возможность оценивать скорость роста ошибок округления путем прямого вычисления норм степеней характеристической матрицы $\|A^n\|$. Элементы A^n являются тригонометрическими многочленами $p(e^{i\phi}) = \sum_{k=-n}^n a_k e^{ik\phi}$. Норма $p(e^{i\phi})$ в метрике l_1 есть $\sum_k |a_k|$. Норма $\|A^n\|$ в метрике l_1 есть максимум норм ее элементов. В этой работе получены оценки скорости роста ошибок округления в равномерной метрике для трех конкретных примеров.

В разд. 1 получены оценки скорости роста ошибок округления в равномерной метрике для разностной схемы, которая используется при вычислении вольт-амперных характеристик (ВАХ) длинных джозефсоновских переходов. Установлена ограниченность роста ошибок округления в равномерной метрике на всем интервале вычисления ВАХ.

В разд. 2 получены оценки скорости роста ошибок округления в равномерной метрике для системы двух разностных уравнений путем прямого вычисления норм степеней характеристической матрицы в метрике l_1 . Полученные численно результаты согласуются с теоретическими: в работе [5] была установлена неустойчивость в C рассматриваемой системы порядка $O(n^{1/8})$.

В разд. 3 численно получены оценки скорости роста ошибок округления в равномерной метрике для схемы Русанова [6, 7] при различных значениях параметров схемы. Все расчеты были проведены на суперкомпьютере “Говорун” [8] с использованием системы REDUCE [9].

1. ВЫЧИСЛЕНИЕ ВАХ ДЛИННЫХ ДЖОЗЕФСОНОВСКИХ ПЕРЕХОДОВ. ЧИСЛЕННЫЙ МЕТОД ОЦЕНКИ ОШИБОК ОКРУГЛЕНИЯ В РАВНОМЕРНОЙ МЕТРИКЕ

Вычисление вольт-амперных характеристик систем n длинных джозефсоновских переходов связано (см. [1], (7)) с решением системы n существенно нелинейных дифференциальных уравнений. Был разработан алгоритм, позволяющий свести задачу к решению одного уравнения $u_{tt} = -\beta u_t + u_{xx} - \sin(u) + I$. Для решения последнего предлагается использовать разностную аппроксимацию второго порядка точности:

$$\frac{u_v^{n+1} - 2u_v^n + u_v^{n-1}}{\tau^2} = -\beta \frac{u_v^{n+1} - u_v^{n-1}}{2\tau} + \frac{u_{v+1}^n - 2u_v^n + u_{v-1}^n}{h^2} - \sin(u_v^n) + I(n),$$

τ, h — шаги сетки по t, x соответственно, ниже используются обозначения $\gamma = \tau/h, \delta = \beta\tau/2$. Линеинная часть без $I(n)$ рассматриваемого разностного уравнения эквивалентна системе разностных уравнений:

$$u_v^{n+1} - u_v^n = v_v^{n+1}, \quad \frac{v_v^{n+1} - v_v^n}{\tau} = -\beta \frac{v_v^{n+1} + v_v^n}{2\tau} + \frac{u_{v+1}^n - 2u_v^n + u_{v-1}^n}{h^2}. \quad (1)$$

Решается задача Коши, заданы начальные данные u_v^0, v_v^0 . После преобразования Фурье

$$U^n = \sum_{v=-\infty}^{\infty} e^{iv\phi} [u_v^n, v_v^n]^*$$

(знак * означает преобразование вектора-строки в вектор-столбец) и некоторых алгебраических манипуляций система (1) преобразуется в векторную систему:

$$U^{n+1} = A(e^{i\phi})U^n, \quad U^n = A^n(e^{i\phi})U^0.$$

Здесь $A(e^{i\phi})$ — характеристическая матрица:

$$A(e^{i\phi}) = \begin{bmatrix} (1 - (4\gamma^2 \sin^2(\phi/2))/(1 + \delta)) & (1 - \delta)/(1 + \delta) \\ -(4\gamma^2 \sin^2(\phi/2))/(1 + \delta) & (1 - \delta)/(1 + \delta) \end{bmatrix}.$$

Функция Грина системы (1) есть

$$\Gamma_j^n = \frac{1}{2\pi} \int_{-\pi}^{\pi} A^n(e^{i\phi}) e^{ij\phi} d\phi.$$

Порядок роста функции Грина Γ_j^n в метрике l_1 определяет скорость роста ошибок округления ϵ_j^n в равномерной метрике:

$$\sup_{1 \leq p, k \leq q} \sum_j |\Gamma_j^n(p, k)| = \left(\sup_{1 \leq j \leq q} |\epsilon_j^n| \right) / \left(\sup_{1 \leq j \leq q} |\epsilon_j^0| \right).$$

В работах [2, 3] были получены оценки, гарантирующие ограниченность скорости роста ошибок округления в l_2 метрике для всех n .

Собственные значения характеристической матрицы A удовлетворяют уравнению

$$\lambda^2 - 2 \frac{1 - 2\gamma^2 \sin^2(\phi/2)}{1 + \delta} \lambda + \frac{1 - \delta}{1 + \delta} = 0.$$

Решая это уравнение, находим

$$\lambda_{1,2} = \frac{1 - 2\gamma^2 \sin^2(\phi/2) \pm \sqrt{(1 - 2\gamma^2 \sin^2(\phi/2))^2 - (1 - \delta^2)}}{1 + \delta}. \quad (2)$$

В работе [2] доказано, что $|\lambda_1| \leq 1$ и $|\lambda_2| \leq 1$ для всех $\gamma \leq 1$ и всех $0 \leq \phi \leq 2\pi$. Это необходимые условия устойчивости рассматриваемой задачи Коши в метрике L_2 , гарантирующие отсутствие неустойчивости экспоненциального типа. Но при наличии кратных собственных значений может возникнуть неустойчивость степенного типа [4]: $\|G^n\| \asymp n^\theta$, $\theta > 0$, G — оператор перехода от слоя к слою. При наличии кратных собственных значений нормальная форма характеристической матрицы имеет блочно-треугольную форму [4]. В [2] проведен детальный спектральный анализ характеристической матрицы системы (1). В рассматриваемом случае нормальной формой характеристической матрицы является треугольная матрица второго порядка. Матрица преобразования подобия

$$T = \begin{bmatrix} \lambda_1 & 1 \\ \lambda_1 - 1 & 1 \end{bmatrix}, \quad \det(T) = 1,$$

приводит матрицу A к треугольному виду:

$$T^{-1}AT = B = \begin{bmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{bmatrix}, \quad B^n = \begin{bmatrix} \lambda_1^n & D_n \\ 0 & \lambda_2^n \end{bmatrix}, \quad D_n = \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2}. \quad (3)$$

Сделав обратное преобразование, получаем

$$A^n = \begin{bmatrix} -\lambda_1\lambda_2 D_n - D_{n+1} & \lambda_1\lambda_2 D_n \\ (\lambda_1 + \lambda_2 - \lambda_1\lambda_2 - 1)D_n & \lambda_1\lambda_2(D_n - D_{n-1}) \end{bmatrix}.$$

Так как $\lambda_1 + \lambda_2 - \lambda_1\lambda_2 - 1 = 2\gamma^2(\cos(\phi) - 1)/(1 + \delta)$, $\gamma \leq 1$, оценка скорости роста ошибок округления в равномерной метрике сводится к оценке роста $\|D_n\|$ в метрике l_1 :

$$\frac{1-\delta}{1+\delta}\|D_n\| \leq \sup_{1 \leq p, k \leq 2} \sum_j |\Gamma_j^n(p, k)| \leq \frac{4}{1+\delta}\|D_n\|.$$

Замечание. Норма $p(e^{i\phi}) = \sum_{k=-n}^n a_k e^{ik\phi}$ в метрике l_1 есть $\sum_{k=-n}^n |a_k|$. При умножении $p(e^{i\phi})$ на $\cos(\phi)$ величина нормы не возрастает: $\|\cos(\phi)p(e^{i\phi})\| \leq \|p(e^{i\phi})\|$.

Алгоритм прямого вычисления норм D_n в метрике l_1 .

В случае четных n , $n = 2nc$, имеем

$$D_n = \frac{p_n(z)}{(1+\delta)^{n-1}}, \quad z = 2(1 - \gamma^2 + \gamma^2 \cos(\phi)), \quad p_n(z) = z \sum_{j=1}^{nc} g_j (\delta^2 - 1)^{nc-j} z^{2(j-1)} = \sum_{j=1}^{nc} b_j z^{2j-1},$$

$$g_j = \sum_{k=1}^j C_n^{2k-1} C_{nc-k}^{j-k}, \quad z^2 = c_1 + c_2 \cos(\phi) + c_3 \cos(2\phi), \quad p_n = \sum_{j=1}^n d_j \cos((j-1)\phi), \quad \|p_n\| = \sum_{j=1}^n |d_j|.$$

Последовательно вычисляем $z^{2k-1} = a_1 + a_2 \cos(\phi) + \dots + a_{2k} \cos((2k-1)\phi)$, $2 \leq k \leq nc$.

Коэффициенты a_j зависят от k и пересчитываются при переходе от z^{2k-1} к z^{2k+1} :

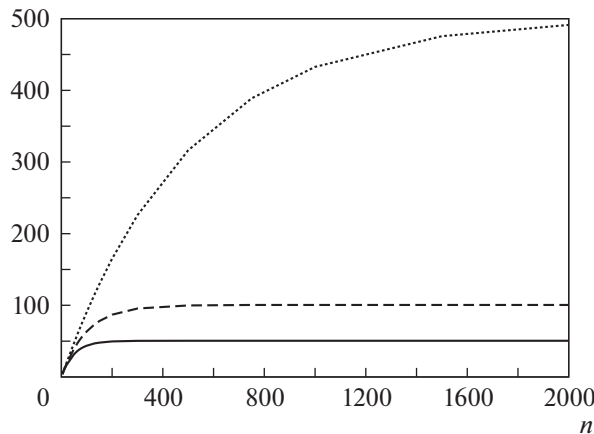
$$z^{2k+1} = z^{2k-1}(c_1 + c_2 \cos(\phi) + c_3 \cos(2\phi)) = a_1(c_1 + c_2 \cos(\phi) + c_3 \cos(2\phi)) + R_k,$$

$$R_k = \sum_{j=2}^{2k} a_j \left[c_1 \cos((j-1)\phi) + c_2 \frac{\cos((j-2)\phi) + \cos(j\phi)}{2} + c_3 \frac{\cos((j-3)\phi) + \cos((j+1)\phi)}{2} \right].$$

В случае нечетных n , $n = 2nc - 1$, имеем

$$D_n = \frac{p_n(z)}{(1+\delta)^{n-1}}, \quad p_n(z) = \sum_{j=1}^{nc} g_j (\delta^2 - 1)^{nc-j} z^{2(j-1)},$$

$$g_j = \sum_{k=1}^j C_n^{2(k-1)} C_{nc-k}^{j-k}, \quad z^2 = c_1 + c_2 \cos(\phi) + c_3 \cos(2\phi), \quad p_n = \sum_{j=1}^n d_j \cos((j-1)\phi), \quad \|p_n\| = \sum_{j=1}^n |d_j|.$$



Фиг. 1. Графики $\|D_n\|$, $n \leq 2000$ в метрике l_1 для уравнения (1). Нижний график относится к случаю $\tau = h = 0.1$. Средний — для $\tau = h = 0.05$, и верхний график — для $h = 0.05$, $\tau = h/5$.

Последовательно вычисляем $z^{2(k-1)} = a_1 + a_2 \cos(\phi) + \dots + a_{2k-1} \cos(2(k-1)\phi)$, $2 \leq k \leq nc$. Коэффициенты a_j зависят от k и пересчитываются при переходе от $z^{2(k-1)}$ к z^{2k} :

$$z^{2k} = z^{2(k-1)}(c_1 + c_2 \cos(\phi) + c_3 \cos(2\phi)) = a_1(c_1 + c_2 \cos(\phi) + c_3 \cos(2\phi)) + R_k,$$

$$R_k = \sum_{j=2}^{2k-1} a_j \left[c_1 \cos((j-1)\phi) + c_2 \frac{\cos((j-2)\phi) + \cos(j\phi)}{2} + c_3 \frac{\cos((j-3)\phi) + \cos((j+1)\phi)}{2} \right].$$

На фиг. 1 представлены графики $\|D_n\|$ в метрике l_1 . Нижний график относится к случаю $\tau = h = 0.1$, средний — к случаю $\tau = h = 0.05$, и верхний — для $h = 0.05$, $\tau = h/5$. Аналогично случаю оценки $\|D_n\|$ в метрике l_2 (см. фиг. 3, [3]), самый быстрый рост наблюдается при $\gamma \neq 1$. Нижний график выходит на плато $\|D_{2000}\| = 50.5$. Средний график также выходит на плато $\|D_{2000}\| = 100.999\dots$. А при $\gamma = 1/5$ имеем $\|D_{2000}\| = 491.333034\dots$, продолжается рост, но видна тенденция выхода на плато:

$$\|D_{1000}\| - \|D_{500}\| = 116.386\dots, \quad \|D_{1500}\| - \|D_{1000}\| = 42.816\dots, \quad \|D_{2000}\| - \|D_{1500}\| = 15.731\dots$$

Чтобы разобраться, мы продолжили счет до $n = 5500$. В результате выяснилось, что верхний график также выходит на плато, см. фиг. 2, где точками отмечены значения $\|D_{n+500}\| - \|D_n\|$. Хорошо видно, что с ростом n значения $\|D_{n+500}\|$ и $\|D_n\|$ сближаются, достигая $\|D_{5500}\| - \|D_{5000}\| = 0.014\dots$

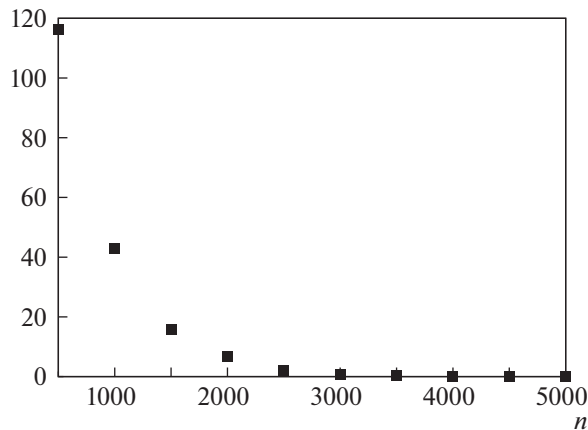
В рассматриваемом случае оценка скорости роста ошибок округления в равномерной метрике свелась к оценке $\|D_n\|$ в метрике l_1 . Но это не всегда так. Рассмотрим еще один пример.

2. УСТОЙЧИВАЯ В L_2 СИСТЕМА, НЕУСТОЙЧИВАЯ В C

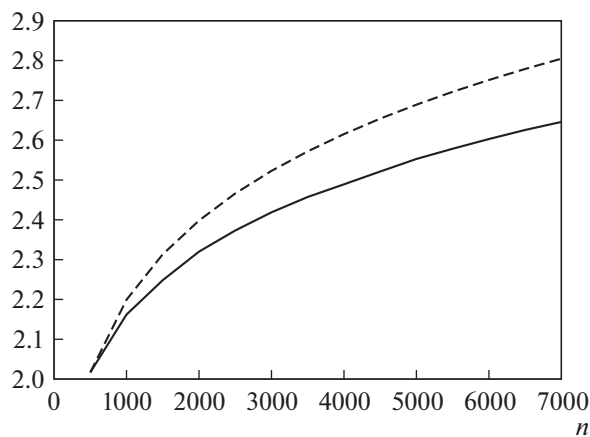
Система разностных уравнений

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\tau}{2h}(v_{j+1}^n - v_{j-1}^n) - \frac{\tau^2}{h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) = 0, \\ v_j^{n+1} &= v_j^n - \frac{\tau}{2h}(u_{j+1}^n - u_{j-1}^n) - \frac{\tau^2}{h^2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n) = 0, \end{aligned} \quad (4)$$

устойчива в L_2 при $\tau \leq h$ (см. [5]).



Фиг. 2. Выход на плато графика $\|D_n\|$ в метрике l_1 (система (1), $h = 0.05$, $\tau = h/5$). Точками отмечены значения $\|D_{n+500}\| - \|D_n\|$, $n = 500, 1000, 1500, 2000, \dots, 5000$.



Фиг. 3. Нижний график представляет $\|A^n\|$, $500 \leq n \leq 7000$ в метрике l_1 для системы (4), $\gamma = 1/2$. Верхний график представляет функцию $\|A^{500}\| * n^{1/8}$.

Характеристическая матрица

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}, \quad a = 1 - \gamma^2(1 - \cos(\phi)), \quad b = i\gamma \sin(\phi),$$

имеет два собственных значения $\lambda_{1,2} = a \pm b$. Матрица преобразования подобия

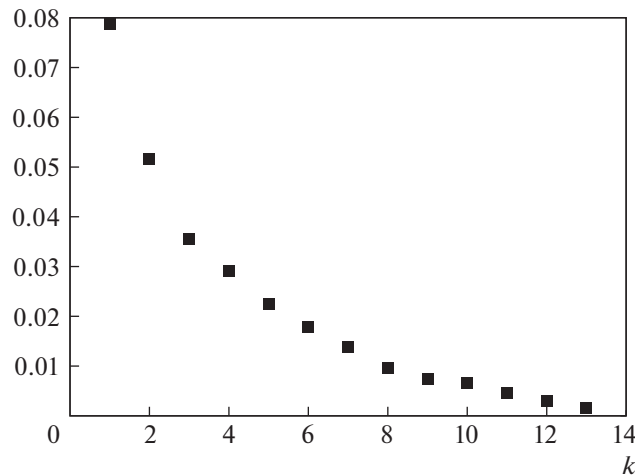
$$T = \begin{bmatrix} \lambda_1 & 1 \\ \lambda_1 - 1 & 1 \end{bmatrix}$$

приводит характеристическую матрицу к треугольному виду

$$TAT^{-1} = \begin{bmatrix} \lambda_2 & 0 \\ b & \lambda_1 \end{bmatrix} = B, \quad B^n = \begin{bmatrix} \lambda_2^n & 0 \\ (\lambda_1^n - \lambda_2^n)/2 & \lambda_1^n \end{bmatrix}.$$

Сделав обратное преобразование, получаем

$$A^n = 0.5 \begin{bmatrix} \lambda_1^n + \lambda_2^n & \lambda_1^n - \lambda_2^n \\ \lambda_1^n - \lambda_2^n & \lambda_1^n + \lambda_2^n \end{bmatrix}.$$



Фиг. 4. На графике точками отмечены значения $r(14)/r(k) - (14/k)^{1/8}$, $k = 1, 2, \dots, 13$, где $r(k) = \|A^{500k}\|$ в метрике l_1 . График демонстрирует скорость роста $\|A^n\|$ порядка $O(n^{1/8})$ для системы (4).

Скорость роста ошибок округления в равномерной метрике определяется скоростью роста норм элементов матрицы A^n в метрике l_1 [5]. В рассматриваемом случае

$$\lambda_1 = 1 - \gamma^2 + \frac{\gamma^2 + \gamma}{2} e^{i\phi} + \frac{\gamma^2 - \gamma}{2} e^{-i\phi} = p(e^{i\phi})e^{-i\phi}, \quad p(x) = d1 + d2x + d3x^2,$$

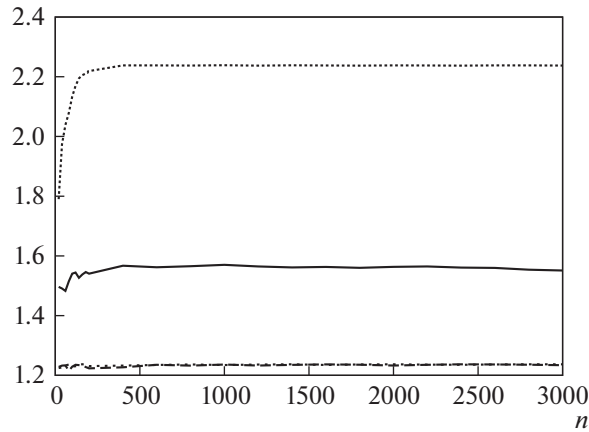
$\lambda_2 = q(e^{i\phi})e^{-i\phi}$, $q(x) = d3 + d2x + d1x^2$. Последовательно вычисляем p^k , q^k , $2 \leq k \leq n$. Сумма модулей коэффициентов многочленов $(p^n + q^n)/2$, $(p^n - q^n)/2$ определяет скорость роста ошибок округления в равномерной метрике.

С использованием системы REDUCE удалось вычислить при $\gamma = 1/2$ на суперкомпьютере “Говорун” $r(n) = \|A^n\|$, $1 \leq n \leq 7000$, в метрике l_1 , которые медленно растут от 2.0166... до 2.645..., что обеспечивает ограниченность роста ошибок округления в равномерной метрике на интервале $[0, 7000\tau]$. См. нижний график на фиг. 3. Верхний график представляет функцию $r(500)n^{1/8}$. В работе [5] аналитически была установлена неустойчивость рассматриваемой системы в C порядка $O(n^{1/8})$. На фиг. 4 точками отмечены значения $\Delta(k) = r(14)/r(k) - (14/k)^{1/8}$, $k = 1, 2, \dots, 13$, где $r(k) = \|A^{500k}\|$. Хорошо видно, что величина $\Delta(k)$ монотонно убывает, достигая в конце значения 0.00159..., что согласуется с результатом, полученным теоретически.

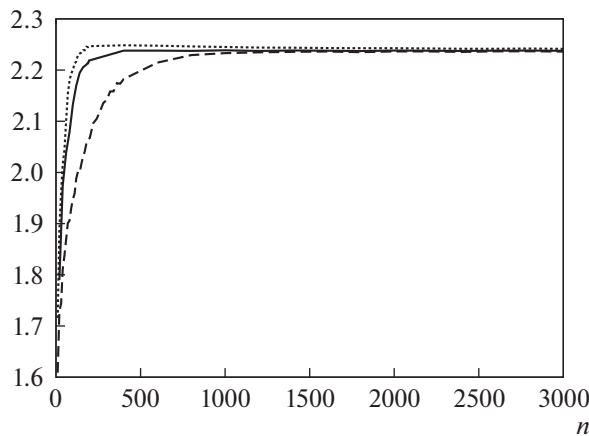
3. ОЦЕНКА ОШИБОК ОКРУГЛЕНИЯ В РАВНОМЕРНОЙ МЕТРИКЕ ДЛЯ СХЕМЫ РУСАНОВА

Был разработан и реализован с использованием системы REDUCE на суперкомпьютере “Говорун” численный метод исследования устойчивости в равномерной метрике схемы Русанова третьего порядка точности, которая используется при решении задач газовой динамики. Для модельного уравнения $u_t = u_x$ рассматривается [6] такая разностная схема:

$$u_v^{n+1} = \left\{ 1 + \frac{\gamma}{12}(-T^2 + 8T - 8T^{-1} + T^{-2}) + \frac{\gamma^2}{8}(T^2 - 2I + T^{-2}) + \right. \\ \left. + \frac{\gamma^3}{12}(T^2 - 2T + 2T^{-1} - T^{-2}) - \frac{\omega}{24}(T^2 - 4T + 6I - 4T^{-1} + T^{-2}) \right\} u_v^n, \quad (5) \\ n \geq 0, \quad v = 0, \pm 1, \pm 2, \dots; \quad x = v h, \quad t = n \tau, \quad \gamma = \tau/h, \quad T u_v^n = u_{v+1}^n.$$



Фиг. 5. Графики $\|A^n\|$ в метрике l_1 — для схемы Рунге-Кутты при $\gamma = 0.5$: верхний график относится к $\omega = 3$, средний — к $\omega = 3/2$, два оставшихся (практически неразличимых при $n > 500$) относятся к $\omega = 0.5, 0.25$.



Фиг. 6. Графики $\|A^n\|$ в метрике l_1 — для схемы Рунге-Кутты при $\omega = 3$: верхний график относится к $\gamma = 0.75$, средний — для $\gamma = 0.5$, нижний — для $\gamma = 0.25$. При $n > 1000$ графики практически неразличимы, выходят на одну и ту же асимптоту.

Характеристическая функция рассматриваемой разностной схемы есть

$$f(e^{i\phi}) = 1 - \frac{\gamma^2}{2} \sin^2(\phi) - \frac{2}{3} \omega \sin^4\left(\frac{\phi}{2}\right) + \gamma i \sin(\phi) \left[1 + \frac{2}{3} (1 - \gamma^2) \sin^2\left(\frac{\phi}{2}\right) \right],$$

$$\begin{aligned} |f(e^{i\phi})|^2 &= 1 + 4P(z)z^2/9, \quad P(z) = [4\gamma^2(1 - \gamma^2)^2 - (\omega - 3\gamma^2)^2]z^2 - \\ &- 2[2\gamma^2(1 - \gamma^2)^2 + 3\gamma^2(\omega - \gamma^2 - 2)]z + 3(\omega - 4\gamma^2 + \gamma^4), \quad z = \sin^2(\phi/2). \end{aligned}$$

Последовательно вычисляем суммы модулей коэффициентов $f, f^2, f^3, \dots$ (сумма модулей коэффициентов f^n есть норма f^n в метрике l_1). На фиг. 5 представлены графики $\|f^n\|$ в метрике l_1 для схемы Рунге-Кутты при $\gamma = 1/2$: верхний график относится к $\omega = 3$, средний — к $\omega = 3/2$, два оставшихся графика (практически неразличимые при $n > 500$) относятся к $\omega = 0.5, 0.25$. Во всех

случаях $\|f^n\|$ ограничены при $n \leq 3000$, что гарантирует ограниченность ошибок округления в равномерной метрике при $t \leq 3000\tau$.

На фиг. 6 представлены графики $\|f^n\|$ в метрике l_1 для схемы Русанова при $\omega = 3$: верхний график относится к $\gamma = 0.75$, средний — для $\gamma = 0.5$, нижний относится к $\gamma = 0.25$. Во всех случаях $\|f^n\|$ ограничены при $n \leq 3000$, что гарантирует ограниченность ошибок округления в равномерной метрике при $t \leq 3000\tau$. Начиная с $n = 1000$, графики практически неразличимы, выходят на одну и ту же асимптоту. Полученные численно результаты согласуются с теоретическими результатами [7].

Авторы благодарны Г.М. Кобелькову за полезные критические замечания и Е.Е. Тыртышникову за поддержку проводимого исследования.

СПИСОК ЛИТЕРАТУРЫ

1. Башашин М.В., Земляная Е.В., Рахмонов И.Р., Шукринов Ю.М., Атанасова П.Х., Волохова А.В. Вычислительная схема и параллельная реализация для моделирования системы длинных джозефсоновских переходов // Компьютерные исследования и моделирование. 2016. Т. 8. № 4. С. 593–604.
2. Сердюкова С.И. Моделирование динамических процессов в длинных джозефсоновских переходах. Проблема вычисления вольт-амперных характеристик. Численный метод оценки скорости роста ошибок округления // Ж. вычисл. матем. и матем. физ. 2020. Т. 60. № 1. С. 3–11.
3. Зув М.И., Сердюкова С.И. Моделирование динамических процессов в длинных джозефсоновских переходах. Проблема вычисления вольт-амперных характеристик. Оценки скорости роста ошибок округления для разностной схемы второго порядка точности // Ж. вычисл. матем. и матем. физ. 2022. Т. 62. № 1. С. 159–166.
4. Урм В.Я. О необходимых и достаточных условиях устойчивости систем разностных уравнений // Докл. АН СССР. 1961. Т. 139. № 1. С. 40–43.
5. Сердюкова С.И. Об устойчивости в равномерной метрике систем разностных уравнений // Ж. вычисл. матем. и матем. физ. 1967. Т. 7. № 3. С. 497–509.
6. Rusanov V.V. On difference schemes of third order accuracy for nonlinear hyperbolic systems // Jour. of Comp. Phys. 1970. V. 5. № 3. P. 507–516.
7. Сердюкова С.И. Схема Русанова. Исследование устойчивости в равномерной метрике. Асимптотика в окрестности изолированного разрыва // Сообщение ОИЯИ. 1977. P5-10708.
8. Adam Gh., Bashashin M., Belyakov D., Kirakosyan M., Matveev M., Podgainy D., Sapozhnikova T., Streltsova O., Torosyan Sh., Vala M., Valova L., Vorontsov A., Zaikina T., Zemlyanaya E., Zuev M. IT-ecosystem of the HybriLIT heterogeneous platform for high-performance computing and training of IT-specialists. Selected Papers of the 8th International Conference “Distributed Computing and Grid-technologies in Science and Education” (GRID 2018), Dubna, Russia, September 10–14, 2018, CEUR-WS.org/Vol-2267”.
9. REDUCE User’s Guide for Unix Systems. Version 3.8 by Winfried Neun ZIB, 2004.

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.612

О ЕДИНСТВЕННОСТИ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ, К КОТОРЫМ РЕДУЦИРУЮТСЯ ОБРАТНЫЕ ЗАДАЧИ ГРАВИМЕТРИИ И МАГНИТОМЕТРИИ: РЕГИОНАЛЬНЫЙ ВАРИАНТ¹⁾

© 2023 г. И. И. Колотов¹, Д. В. Лукьяненко¹, И. Э. Степанова^{1,2,*},
А. В. Щепетилов¹, А. Г. Ягола¹

¹ 119992 Москва, Ленинские горы, 1, МГУ им. М.В. Ломоносова,
физический факультет, кафедра математики, Россия

² 123995 Москва, ул. Б. Грузинская, 10, стр. 1, Институт физики Земли им. О.Ю. Шмидта РАН, Россия

*e-mail: tet@ifz.ru

Поступила в редакцию 06.02.2023 г.
Переработанный вариант 06.02.2023 г.
Принята к публикации 29.05.2023 г.

Рассматриваются условия однозначной разрешимости систем линейных алгебраических уравнений, к решению которых редуцируются многие обратные задачи гравимагниторазведки. Математические постановки обратных задач выполняются с учетом сферичности планеты. Библ. 11. Фиг. 5.

Ключевые слова: разрешимость, вырожденные системы, интегральные представления.

DOI: 10.31857/S0044466923090119, **EDN:** JJFWYJ

ВВЕДЕНИЕ

При решении обратных геофизических задач интерпретационного характера необходимо находить скалярные и векторные функции, описывающие плотности распределения источников поля. В общем случае указанные функции зависят от бесконечного (зачастую, несчетного) числа параметров и, согласно результатам, приведенным, например, в [1], однозначно не могут быть восстановлены. Если перейти к конечномерным аппроксимациям как наблюдаемых элементов аномальных физических полей, так и самих источников, то единственное решение условно-вариационных задач, возникающих при интерпретации данных гравиметрии и магнитометрии, можно получить при выполнении ряда условий. Мы попытаемся сформулировать условия однозначной разрешимости соответствующих систем линейных алгебраических уравнений (СЛАУ) в региональном варианте, т.е. тогда, когда необходимо учитывать сферичность Земли. Постановки такого рода появляются при исследованиях физических полей на больших полигонах, размеры которых превышают 1 градус (приблизительно 111 км для Земли) в двух направлениях.

1. ЕДИНСТВЕННОСТЬ РЕШЕНИЯ СЛАУ В РЕГИОНАЛЬНОМ ВАРИАНТЕ МЕТОДА S-АППРОКСИМАЦИЙ

Основные характеристики метода S-аппроксимаций в двух случаях (локальном и региональном) были изложены в работах первого автора (см., например, [1–5]). Этот метод хорошо зарекомендовал себя при решении самых разнообразных обратных геофизических и геодезических задач и был модифицирован с целью нахождения более адекватных реальности распределений эквивалентных по внешнему полю источников.

Напомним основные положения регионального варианта метода S-аппроксимаций.

¹⁾ Работа выполнена при финансовой поддержке РНФ (код проекта 23-41-00002).

Для нахождения производных гравитационного поля в некоторой совокупности точек M_v с координатами $x^{(v)} = (x_1^{(v)}, x_2^{(v)}, x_3^{(v)})$, из внешности CE некоторой области $E \subset R^3$ в рамках метода линейных интегральных представлений (см. [1]) необходимо действовать следующим образом.

Пусть D — сфера, описываемая уравнением $r = \text{const} = R$.

По заданным плотностям $\rho_1(\xi_1, \xi_2) \in L_2(D)$ и $\rho_2(\xi_1, \xi_2) \in L_2(D)$ простого и двойного слоев мы должны найти совокупность ограниченных линейных функционалов вида

$$p_s = \int_D \rho_1(\xi_1, \xi_2) P_1^{(s)}(\xi_1, \xi_2) \mu(\xi_1, \xi_2) d\xi_1 d\xi_2 + \int_D \rho_2(\xi_1, \xi_2) P_2^{(s)}(\xi_1, \xi_2) \mu(\xi_1, \xi_2) d\xi_1 d\xi_2, \quad (1.1)$$

$$s = 1, 2, \dots, S,$$

где для всех s

$$\begin{aligned} \|P_1^{(s)}\|^2 &= \int_D (P_1^{(s)})^2(\xi_1, \xi_2) d\mu(\xi_1, \xi_2) < +\infty, \\ \|P_2^{(s)}\|^2 &= \int_D (P_2^{(s)})^2(\xi_1, \xi_2) d\mu(\xi_1, \xi_2) < +\infty. \end{aligned} \quad (1.2)$$

В формулах (1.1) и (1.2) $\mu(\xi_1, \xi_2)$ — мера на поверхности D , ξ_1, ξ_2 — координаты на поверхности D . В случае сферы в роли ξ_1, ξ_2 выступают ϑ, φ , $d\mu(\xi_1, \xi_2) = R^2 \sin \vartheta d\vartheta d\varphi$.

В дальнейшем в роли области E у нас будет выступать “идеальная” Земля, а в роли внешности такой идеальной Земли — реальная Земля.

Представим себе идеализированную Землю как внутренность сферы радиуса R_0 . Реальная Земля тогда может трактоваться как область трехмерного пространства, ограниченная замкнутой кусочно-непрерывной поверхностью S , незначительно отклоняющейся от сферы радиуса R_0 . Предполагается, что приближенные значения функции $G(x)$, гармонической вне этой сферы, заданы в произвольном множестве точек $x^{(i)}$, $i = 1, 2, \dots, N$, на поверхности S :

$$f_{i,\delta} = f_i + \delta f_i, \quad f_i = G(x^{(i)}). \quad (1.3)$$

Так как $G(x)$ гармонична при $r > R_0$, она имеет следующее интегральное представление:

$$\begin{aligned} G(x) &= \frac{R_0^2}{4\pi} \int_0^{2\pi} \int_0^\pi \frac{\sigma(\tilde{\vartheta}, \tilde{\varphi}) \sin \tilde{\vartheta} d\tilde{\vartheta} d\tilde{\varphi}}{R(\xi - x)} + \frac{R_0^2}{4\pi} \int_0^{2\pi} \int_0^\pi \frac{w(\tilde{\vartheta}, \tilde{\varphi}) (R_0 - r \cos \vartheta') \sin \tilde{\vartheta} d\tilde{\vartheta} d\tilde{\varphi}}{R^3(\xi - x)}, \\ R(\xi - x) &= (R_0^2 - 2R_0 r \cos \vartheta' + r^2)^{1/2}, \quad \xi = (R_0 \cos \varphi \sin \vartheta, R_0 \sin \varphi \sin \vartheta, R_0 \cos \varphi), \\ x &= (r \cos \varphi \sin \vartheta, r \sin \varphi \sin \vartheta, r \cos \vartheta), \quad \cos \vartheta' = \sin \vartheta \sin \tilde{\vartheta} \cos(\varphi - \tilde{\varphi}) + \cos \vartheta \cos \tilde{\vartheta}. \end{aligned} \quad (1.4)$$

Функция $\sigma(\tilde{\vartheta}, \tilde{\varphi})$ в (1.4) — плотность простого слоя, распределенного на сфере радиуса R_0 , $w(\tilde{\vartheta}, \tilde{\varphi})$ — плотность двойного слоя, распределенного на той же сфере; $R(\xi - x)$ — расстояние между текущей точкой сферы ξ и точкой наблюдения x . Дифференцирование правой части (1.4) — по различным координатам точки наблюдения $G(x)$, даже если эти производные (например, $\partial V / \partial r$ — радиальная производная гравитационного потенциала) не являются гармоническими функциями.

Формула (1.4) является одной из форм интегрального представления гармонической функции вне сферы радиуса R (см. [6]):

$$V(x) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^\pi \left(\frac{1}{R(\xi - x)} \frac{\partial V(\xi)}{\partial \mathbf{n}} - V(\xi) \frac{\partial}{\partial \mathbf{n}} \left(\frac{1}{R(\xi - x)} \right) \right) \sin \tilde{\vartheta} d\tilde{\vartheta} d\tilde{\varphi}, \quad (1.5)$$

где $\mathbf{n}$ — внешняя нормаль к единичной сфере (мы можем написать $\partial / \partial r$, r — радиальная координата радиус-вектора x).

Условно-вариационная задача для σ и w (см. [4, 7]) имеет следующее решение:

$$\sigma(\hat{\xi}) = \sum_{i=1}^N \lambda_i Q_i^{(1)}(\xi), \quad w(\hat{\xi}) = \sum_{i=1}^N \lambda_i Q_i^{(2)}(\xi), \quad \hat{\xi} = (\tilde{\vartheta}, \tilde{\varphi}), \quad \xi = (R, \tilde{\vartheta}, \tilde{\varphi}), \quad (1.6)$$

где $Q_i^{(1)}(\xi)$ и $Q_i^{(2)}(\xi)$ представляются в виде

$$Q_i^{(1)}(\xi) = \frac{1}{R(\xi - x^{(i)})}, \quad Q_i^{(2)}(\xi) = \frac{R_0 - r_i \cos \vartheta_i'}{R^3(\xi - x^{(i)})}. \quad (1.7)$$

Величины λ_i являются компонентами N -вектора λ , дающего решение системы линейных уравнений

$$A\lambda = f_\delta, \quad f_\delta = f + \delta f, \quad (1.8)$$

где f_δ имеет компоненты $f_{i,\delta}$ (см. (1.3)) и элементы матрицы $A = A^T \geq 0$ записываются как

$$a_{ij} = \int_0^{2\pi} \int_0^\pi \left[Q_i^{(1)}(\xi) Q_j^{(1)}(\xi) + Q_i^{(2)}(\xi) Q_j^{(2)}(\xi) \right] \sin \vartheta d\vartheta d\phi = \frac{2\pi}{(h_i h_j)^{1/2} r_i r_j} \left(1 + \frac{0.25}{r_i r_j h_i h_j} \right) \times \\ \times F(2 \arctg(\sqrt{h_i h_j}), \sqrt{0.5(1 + \cos \alpha_{ij})}) - \frac{\pi(3(h_i h_j)^2 - 4h_i h_j \cos \alpha_{ij} + 1)}{r_i^2 r_j^2 h_i h_j (\sqrt{1 - 2h_i h_j \cos \alpha_{ij}} + (h_i h_j)^2)^3}. \quad (1.9)$$

Здесь $h_i = R_0/r_i$, $h_j = R_0/r_j$. α_{ij} — угол между векторами x_i и x_j . Предполагаем, что вектор x_i параллелен оси Oz , а вектор x_j лежит в плоскости zOx (мы всегда так можем выбрать систему координат). Тогда указанные векторы будут иметь координаты

$$x_i = (r_i, 0, 0) \quad \text{и} \quad x_j = (r_j \sin \alpha_{ij}, 0, r_j \cos \alpha_{ij}).$$

Функция $F(\phi, k)$ — это эллиптический интеграл I рода.

В настоящей работе мы ограничимся рассмотрением представления элементов аномального гравитационного поля в региональном и глобальном масштабах в виде потенциала простого слоя. Элементы матрицы системы (1.9) принимают тогда вид

$$a_{ij} = \frac{2\pi}{(h_i h_j)^{1/2} r_i r_j} F(2 \arctg(\sqrt{h_i h_j}), \sqrt{0.5(1 + \cos \alpha_{ij})}). \quad (1.10)$$

Покажем, что в размерности 2 СЛАУ не вырождена.

Для этого выпишем явное выражение для элементов матрицы (1.10):

$$a_{ij} = \frac{2\pi}{(h_i h_j)^{1/2} r_i r_j} \int_0^{\sqrt{h_i h_j}} \frac{dh}{\sqrt{h} \sqrt{1 - 2h \cos \alpha_{ij} + h^2}}. \quad (1.11)$$

Для того чтобы матрица системы (1.8) с элементами (1.10) или (1.11) была вырождена, достаточно, чтобы какие-либо две строки этой матрицы были пропорциональны. Пусть это будут строки с номерами i и j . Покажем, что если соответствующие этим индексам точки наблюдения различны, то пропорциональность не может иметь места. Сначала докажем следующие две леммы.

Лемма 1. Для произвольной непрерывно дифференцируемой на отрезке $[c, d]$, $0 < c < d$, функции $F(y)$, не обращающейся в нуль на этом отрезке вместе со своей производной, строго возрастающей и логарифмически выпуклой вниз, для любых $a, b \in [c, d]$, $0 < a < b$, выполняется неравенство

$$F(a)F(b) \geq [F(\sqrt{ab})]^2. \quad (1.12)$$

Доказательство. Введем вспомогательную переменную x и рассмотрим выражение

$$\Phi(x) = F(a/x)F(bx).$$

Будем считать, что переменная x меняется таким образом, что аргумент функции $F(a/x)$ или $F(bx)$ лежит на отрезке $[c, d]$.

Вычислим производную функции $\Phi(x)$ и приравняем ее нулю:

$$\Phi'(x) = -\frac{a}{x^2} F'(a/x)F(bx) + bF(a/x)F'(bx) = 0 \Rightarrow x^2 = \frac{a F'(a/x)F(bx)}{b F(a/x)F'(bx)}. \quad (1.13)$$

Легко видеть, что при $x^2 = a/b$ аргументы функций в числителе и знаменателе (1.13) совпадают и равны $\sqrt{ab}$. При этом производная функции $\Phi(x)$ в точке $x = \sqrt{ab}$ равна нулю. Следовательно, эта функция достигает в указанной точке экстремума. Для логарифмически выпуклой вниз функции это — минимум.

В самом деле, вторая производная от логарифма такой функции положительна:

$$\ln(F(x))'' = \left(\frac{F'(x)}{F(x)} \right)' > 0.$$

Если $x = \sqrt{a/b}$, $a < b$, то x меньше единицы, а при $x = 1$ будем иметь $\Phi'(1) = -aF'(a)F(b) + bF(a)F'(b) > 0$ для логарифмически выпуклой вниз функции $F(x)$ при $b > a$. На отрезке $[\sqrt{a/b}, 1]$ производная функции $\Phi'(x)$ также положительна вследствие логарифмической выпуклости функции $F(x)$.

Лемма 2. Для любой функции $F(x)$, представляющейся на отрезке $[c, d]$, $0 < c < d$, абсолютно сходящимся степенным рядом с положительными коэффициентами, справедливо неравенство (1.12) при любых $a, b \in [c, d]$, $0 < a < b$.

Доказательство. Выпишем неравенство (1.12) для частичных сумм соответствующих степенных рядов и учтем тот факт, что произведение абсолютно сходящихся степенных рядов есть абсолютно сходящийся степенной ряд:

$$F_n(x) = \sum_{i=0}^n a_i x^i, \quad F(y) = \sum_{k=0}^n a_k y^k,$$

$$F(x)F(y) = \sum_{i=0}^n \sum_{k=0}^n a_i a_k x^i y^k \geq \sum_{i=0}^n \sum_{k=0}^n a_i a_k (\sqrt{xy})^{i+k} = (F(\sqrt{xy}))^2,$$

так как $a_i a_k (x^i y^k + x^k y^i) \geq 2a_i a_k (x^{i+k/2} y^{i+k/2})$, $i, k = 0, \dots, n$.

Устремим теперь $n \rightarrow \infty$ и получим требуемое утверждение при $x = a$ и $y = b$.

Теперь перейдем к доказательству невырожденности системы (1.8)–(1.10). Нам нужно убедиться в том, что

$$a_{ii}a_{jj} \geq (a_{ij})^2, \quad (1.14)$$

причем знак равенства имеет место лишь в случае совпадения двух точек наблюдения. Рассмотрим неравенство (1.14):

$$a_{ii}a_{jj} = \frac{4\pi^2}{h_i^2 h_j^2 r_i^2 r_j^2} \int_0^{h_i} \int_0^{h_j} \frac{dx dy}{\sqrt{xy(1-x)(1-y)}},$$

$$a_{ij}^2 = \frac{4\pi^2}{h_i^2 h_j^2 r_i^2 r_j^2} \int_0^{\sqrt{h_i h_j}} \int_0^{\sqrt{h_i h_j}} \frac{dx dy}{\sqrt{xy} \sqrt{1-2h_i \cos \alpha_{ij} + h_i^2} \sqrt{1-2h_j \cos \alpha_{ij} + h_j^2}}, \quad (1.15)$$

$$a_{ii}a_{jj} \geq a_{ij}^2 \Rightarrow \int_0^{h_i} \int_0^{h_j} \frac{dx dy}{\sqrt{xy(1-x)(1-y)}} \geq \int_0^{\sqrt{h_i h_j}} \int_0^{\sqrt{h_i h_j}} \frac{dx dy}{\sqrt{xy} \sqrt{1-2h_i \cos \alpha_{ij} + h_i^2} \sqrt{1-2h_j \cos \alpha_{ij} + h_j^2}}.$$

Надо доказать, что строгое неравенство в формуле (1.15) выполняется всегда, если $\cos \alpha_{ij} \neq 1$. Подынтегральная функция в левой части (1.15) в каждой точке по (x, y) больше подынтегральной функции в правой части, если точки наблюдения не совпадают ($\cos \alpha_{ij} \neq 1$, $h_i \neq h_j$). По этой причине мы можем оценить интегралы только одной из этих функций (соответствующей произведению диагональных элементов) по двум областям. Но интеграл в левой части (1.15) вычисляется в явном виде:

$$\int_0^{h_i} \int_0^{h_j} \frac{dx dy}{\sqrt{xy(1-x)(1-y)}} = \ln \left(\frac{1+h_i}{1-h_i} \right) \ln \left(\frac{1+h_j}{1-h_j} \right),$$

$$\int_0^{\sqrt{h_i h_j}} \int_0^{\sqrt{h_i h_j}} \frac{dx dy}{\sqrt{xy(1-x)(1-y)}} = \ln \left(\frac{1 + \sqrt{h_i h_j}}{1 - \sqrt{h_i h_j}} \right)^2.$$

Положим теперь $F(x) = \ln \left(\frac{1+x}{1-x} \right)$ и применим лемму 1 или лемму 2. Получим требуемое утверждение. Таким образом, доказана следующая

Теорема 1. Система (1.8)–(1.10) не вырождена в размерности 2, если точки наблюдения различны.

Рассмотрим теперь случай двойного слоя. Элементы матрицы тогда выражаются формулой

$$a_{ij} = \frac{2\pi}{(h_i h_j)^{1/2} r_i r_j} F(2 \operatorname{arctg}(\sqrt{h_i h_j}), \sqrt{0.5(1 + \cos \alpha_{ij})}) - \frac{\pi(3(h_i h_j)^2 - 4h_i h_j \cos \alpha_{ij} + 1)}{r_i^2 r_j^2 h_i h_j (\sqrt{1 - 2h_i h_j \cos \alpha_{ij}} + (h_i h_j)^2)^3}. \quad (1.16)$$

Попробуем оценить второе слагаемое в (1.16). Заметим сначала, что при выполнении условия

$$\frac{1}{3} \leq h_i h_j < 1 \quad (1.17)$$

числитель второго слагаемого в (1.16) всегда неотрицателен.

Числитель $-\frac{\pi(3(h_i h_j)^2 - 4h_i h_j \cos \alpha_{ij} + 1)}{r_i^2 r_j^2 h_i h_j (\sqrt{1 - 2h_i h_j \cos \alpha_{ij}} + (h_i h_j)^2)^3}$ меньше для внедиагональных элементов (когда $\cos \alpha_{ij} \neq 1$ или $h_i \neq h_j$), а знаменатель — соответственно, больше, чем у диагональных: вычитается большее по модулю число. Поскольку для любых четырех положительных чисел a_1, a_2, b_1, b_2 , удовлетворяющих условиям $a_1 a_2 \geq a^2$, $b_1 b_2 \geq b^2$, где a и b — также положительные заданные числа, справедливо неравенство

$$(a_1 + b_1)(a_2 + b_2) \geq (a + b)^2, \quad (1.18)$$

то верна следующая

Теорема 2. Система (1.8), (1.16) не вырождена в размерности 2, если координаты точек наблюдения не совпадают.

Замечание. Точки наблюдения могут иметь одинаковые угловые координаты, но тогда они должны находиться на разных расстояниях от начала координат. Неравенство (1.18) доказывается элементарно. В самом деле,

$$(a_1 + b_1)(a_2 + b_2) \geq a_1 a_2 + b_1 b_2 + a_1 b_2 + a_2 b_1 \geq a^2 + b^2 + 2ab = (a + b)^2,$$

так как $a_1 b_2 + a_2 b_1 \geq 2\sqrt{a_1 a_2 b_1 b_2} \geq 2ab$.

2. ЕДИНСТВЕННОСТЬ РЕШЕНИЯ ОБРАТНЫХ ЗАДАЧ МАГНИТОМЕТРИИ

Обратные задачи магниторазведки (особенно нелинейные, когда требуется определить не только плотность распределения источников аномального магнитного поля, но и локализовать носители масс) являются весьма сложными и требуют применения различных регуляризирующих алгоритмов (см. [8–11]). В настоящей статье мы ограничимся исследованием парамагнитных сред, когда зависимость вектора намагниченности от напряженности внешнего магнитного поля считается линейной. Для ферромагнетиков интегральные уравнения, связывающие результаты наблюдений (компоненты вектора магнитной индукции) и свойства среды, будут нелинейными, и для анализа однозначной разрешимости таких постановок требуется несколько иной математический аппарат.

Мы полагаем, что для магнитной индукции имеет место следующее интегральное представление (см. [8, 9]):

$$\bar{B}(\bar{r}_s) = \frac{\mu_0}{4\pi} \iiint_V K(\bar{r}_s, \bar{r}) M(\bar{r}) dV, \quad \bar{r} = (r, \vartheta, \phi); \quad \bar{r}_s = (r_s, \vartheta_s, \phi_s),$$

$$K(\bar{r}_s, \bar{r}) = \frac{1}{\rho^5} \begin{bmatrix} 2\rho_r^2 - \rho_\vartheta^2 - \rho_\phi^2 & 3\rho_r\rho_\vartheta & 3\rho_r\rho_\phi \\ 3\rho_r\rho_\vartheta & 2\rho_\vartheta^2 - \rho_r^2 - \rho_\phi^2 & 3\rho_\vartheta\rho_\phi \\ 3\rho_r\rho_\phi & 3\rho_\vartheta\rho_\phi & 2\rho_\phi^2 - \rho_r^2 - \rho_\vartheta^2 \end{bmatrix}, \quad (2.1)$$

$$\bar{\rho} = \bar{r} - \bar{r}_s, \quad \rho \equiv |\bar{r} - \bar{r}_s|.$$

Здесь $\bar{r}_s$ — радиус-вектор точки наблюдения P_s в сферической системе координат, $\bar{r}$ — радиус-вектор точки внутри носителя V , $\bar{B}(\bar{r}_s)$ — магнитная индукция в системе СИ, измеренная в точке, характеризующейся радиус-вектором $\bar{r}_s$, μ_0 — магнитная постоянная в системе СИ. Через $K(\bar{r}_s, \bar{r})$ обозначена матричная функция координат точки наблюдения и точки, принадлежащей носителю, $M(\bar{r})$ — вектор-функция, характеризующая намагниченность среды. В отличие от локального варианта, сферическая система координат фиксирована и связана с общеземной, в противном случае результаты наблюдений магнитного поля в разных точках планеты не были бы согласованы. Начало координат расположено в центре Земли. Необходимо также подчеркнуть, всюду в данном разделе статьи предполагается, что компоненты вектора магнитной индукции измеряются в точке наблюдения, с которой связан репер криволинейной системы координат, гладко зависящий от координат этой точки. Фактически, мы имеем дело с некоторым расслоением реперов. Поменять местами магнитные диполи и точки наблюдения нельзя, в отличие от локального случая, когда все измерения выполняются в некоторой декартовой системе координат и базисные векторы зафиксированы в пространстве.

Поле, создаваемое одним магнитным диполем, выражается формулой

$$B(\bar{r}_s) = \frac{\mu_0}{4\pi} \frac{3(\bar{m}, \bar{r} - \bar{r}_s)(\bar{r} - \bar{r}_s) - \rho^2 \bar{m}}{\rho^5}. \quad (2.2)$$

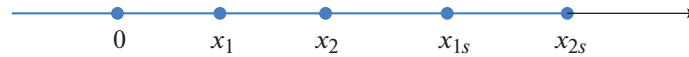
В (2.2) приняты такие же обозначения, что и в (2.1). Через $\bar{m} = (m_r, m_\vartheta, m_\phi)^T$ обозначен магнитный момент точечного диполя в указанной сферической системе координат. Как и в локальном случае, установим в точках наблюдения условные “сенсоры”. В результате дискретизации интегрального представления (2.1) мы приходим к следующей системе линейных алгебраических уравнений:

$$\begin{aligned} \sum_{i=1}^M \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\vartheta} + m_{i\phi}\rho_{ki,\phi}\}\rho_{ki,r}}{\rho_{ki}^5} - \frac{m_{ir}}{\rho_{ki}^3} \right) &= B_{kr}, \\ \sum_{i=1}^M \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\vartheta} + m_{i\phi}\rho_{ki,\phi}\}\rho_{ki,\vartheta}}{\rho_{ki}^5} - \frac{m_{i\vartheta}}{\rho_{ki}^3} \right) &= B_{k\vartheta}, \\ \sum_{i=1}^M \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\vartheta} + m_{i\phi}\rho_{ki,\phi}\}\rho_{ki,\phi}}{\rho_{ki}^5} - \frac{m_{i\phi}}{\rho_{ki}^3} \right) &= B_{k\phi}, \quad k = 1, \dots, N. \end{aligned} \quad (2.3)$$

В общем случае число сенсоров N не совпадает с числом неизвестных дипольных моментов M . В (2.3) через ρ_{ki} обозначено расстояние между k -м сенсором и i -м диполем, а через $\bar{\rho}_{ki} = (\rho_{ki,r}, \rho_{ki,\vartheta}, \rho_{ki,\phi})$ — вектор, соединяющий k -й сенсор с i -м диполем с компонентами в сферической системе координат, $\bar{m}_i = (m_{ir}, m_{i\vartheta}, m_{i\phi})^T$ — магнитный момент i -го диполя в сферической системе координат.

Система (3) имеет блочный вид

$$\begin{bmatrix} K_{11} & K_{12} & \dots & K_{1N} \\ K_{21} & K_{22} & \dots & K_{2N} \\ \dots & \dots & \dots & \dots \\ K_{N1} & K_{N2} & \dots & K_{NN} \end{bmatrix} \begin{bmatrix} \bar{m}_1 \\ \bar{m}_2 \\ \dots \\ \bar{m}_N \end{bmatrix} = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \\ \dots \\ \bar{B}_N \end{bmatrix}, \quad (2.4)$$



Фиг. 1. Расположение двух сенсоров и двух магнитных диполей. Диполи не разделяют сенсоры.

где блок K_{il} соответствует i -му сенсору и l -му диполю, $\bar{m}_i$ – вектор компонент i -го диполя, $\bar{B}_l$ – вектор-функция, измеренная в l -м сенсоре:

$$K_{il} = \begin{bmatrix} \frac{2\rho_{il,r}^2 - \rho_{il,\vartheta}^2 - \rho_{il,\varphi}^2}{\rho_{il}^5} & \frac{3\rho_{il,r}\rho_{il,\vartheta}}{\rho_{il}^5} & \frac{3\rho_{il,r}\rho_{il,\varphi}}{\rho_{il}^5} \\ \frac{3\rho_{il,\vartheta}\rho_{il,r}}{\rho_{il}^5} & \frac{2\rho_{il,\vartheta}^2 - \rho_{il,r}^2 - \rho_{il,\varphi}^2}{\rho_{il}^5} & \frac{3\rho_{il,\vartheta}\rho_{il,\varphi}}{\rho_{il}^5} \\ \frac{3\rho_{il,\varphi}\rho_{il,r}}{\rho_{il}^5} & \frac{3\rho_{il,\varphi}\rho_{il,\vartheta}}{\rho_{il}^5} & \frac{2\rho_{il,\varphi}^2 - \rho_{il,\vartheta}^2 - \rho_{il,r}^2}{\rho_{il}^5} \end{bmatrix}, \quad \rho_{il} = |\bar{r}_i - \bar{r}_l| = \rho_{il,r}. \quad (2.5)$$

Попытаемся выяснить, по аналогии с рассмотренными нами ранее в локальном варианте, условия однозначной разрешимости системы (2.4).

2.1. Два диполя и N сенсоров, расположенных на одной прямой, проходящей через начало координат

Пусть и сенсоры, и диполи располагаются на одной прямой OX , проходящей через центр фиксированной сферической системы координат. Координаты диполей обозначим через x_1, x_2 , а координаты сенсоров – через x_1^s, x_2^s . Предположим сначала, что диполи и сенсоры не разделяют друг друга (см. фиг. 1). В начале координат не могут находиться ни сенсоры, ни диполи ввиду неоднозначности задания координат в указанной точке.

Система (2.4) принимает тогда вид

$$A_{22}\bar{x} = \bar{B}_2, \quad (2.6)$$

$$A_{22} = \begin{bmatrix} \frac{2}{r_{11}^3} & 0 & 0 & \frac{2}{r_{12}^3} & 0 & 0 \\ 0 & -\frac{1}{r_{11}^3} & 0 & 0 & -\frac{1}{r_{12}^3} & 0 \\ 0 & 0 & -\frac{1}{r_{11}^3} & 0 & 0 & -\frac{1}{r_{12}^3} \\ \frac{2}{r_{21}^3} & 0 & 0 & \frac{2}{r_{22}^3} & 0 & 0 \\ 0 & -\frac{1}{r_{21}^3} & 0 & 0 & -\frac{1}{r_{22}^3} & 0 \\ 0 & 0 & -\frac{1}{r_{12}^3} & 0 & 0 & -\frac{1}{r_{22}^3} \end{bmatrix},$$

$$r_{11} = |x_1 - x_{1s}|, \quad r_{12} = |x_1 - x_{2s}|, \quad r_{21} = |x_2 - x_{1s}|, \quad r_{22} = |x_2 - x_{2s}|, \quad (2.7)$$

$$\bar{B}_2 = \begin{pmatrix} B_{1x} \\ B_{1y} \\ B_{1z} \\ B_{2x} \\ B_{2y} \\ B_{2z} \end{pmatrix}, \quad \bar{x} = \begin{pmatrix} m_{1x} \\ m_{1y} \\ m_{1z} \\ m_{2x} \\ m_{2y} \\ m_{2z} \end{pmatrix}.$$



Фиг. 2. Расположение двух сенсоров и двух магнитных диполей. Диполи разделяют сенсоры.

Переобозначим элементы в матрице системы (2.7):

$$\tilde{A}_{22} = \begin{bmatrix} a_{11} & 0 & 0 & a_{12} & 0 & 0 \\ 0 & b_{11} & 0 & 0 & b_{12} & 0 \\ 0 & 0 & b_{11} & 0 & 0 & b_{12} \\ a_{21} & 0 & 0 & a_{22} & 0 & 0 \\ 0 & b_{21} & 0 & 0 & b_{22} & 0 \\ 0 & 0 & b_{21} & 0 & 0 & b_{22} \end{bmatrix}. \quad (2.8)$$

Для того чтобы детерминант матрицы (2.8) был равен нулю, необходимо и достаточно, чтобы координаты сенсоров совпадали (при различных координатах диполей):

$$\begin{aligned} \det A = 0 &\Rightarrow a_{11}a_{22} - a_{12}a_{21} = (b_{11}b_{22} - b_{12}b_{21})^2 = 0 \Rightarrow \\ \Rightarrow a_{ij} &= -2b_{ij} \Rightarrow \det A = 0 \Leftrightarrow b_{11}b_{22} - b_{12}b_{21} = 0 \Leftrightarrow (x_{1s} - x_1)(x_{2s} - x_2) = (x_{1s} - x_2)(x_{2s} - x_1) \Rightarrow \\ &\Rightarrow x_{1s}(x_1 - x_2) = x_{2s}(x_1 - x_2) \Rightarrow (x_{1s} - x_{2s})(x_1 - x_2) = 0. \end{aligned} \quad (2.9)$$

Таким образом, нами доказана следующая

Теорема 3. *Решение системы (2.4) единственно, если два сенсора и два диполя расположены в различных точках на одной и той же прямой, проходящей через начало координат фиксированной сферической системы. Диполи и сенсоры при этом не “разделяют друг друга”.*

Рассмотрим теперь случай, когда в начале координат расположен первый диполь, затем идет второй сенсор, потом – второй диполь, и последний – первый сенсор (см. фиг. 2). В этом случае для равенства нулю детерминанта матрицы получаем равенство

$$x_{1s}x_{2s} + x_1x_2 = 0.5(x_2 + x_1)(x_{1s} + x_{2s}).$$

Если координаты диполей связаны с координатами сенсоров данным соотношением, то получим, что в таком случае система уравнений вырождена, и компоненты двух диполей определить однозначно нельзя.

Замечание. Можно, безусловно, рассмотреть сразу общий случай расположения диполей. Детерминант обращается в нуль при выполнении либо условия

$$(x_{1s} - x_1)(x_{2s} - x_2) = (x_{1s} - x_2)(x_{2s} - x_1),$$

что соответствует случаю “неразделенных” диполей, либо условия

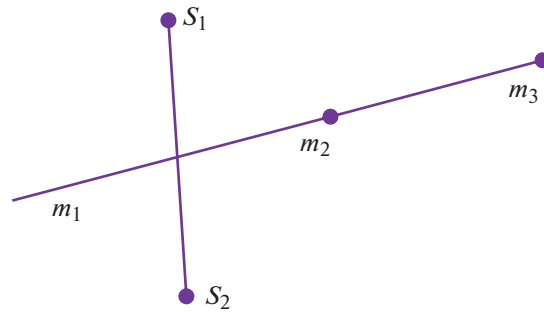
$$(x_{1s} - x_1)(x_{2s} - x_2) = (x_{1s} - x_2)(x_1 - x_{2s}),$$

и в последнем случае система уравнений может быть вырождена.

2.2. N диполей и N сенсоров, расположенных в одной вертикальной плоскости, проходящей через ось OZ

Рассмотрим теперь случай, когда у всех векторов, соединяющих диполи и сенсоры, отсутствует компонента вдоль $\vec{e}_\varphi$.

Пусть требуется определить компоненты N магнитных диполей $\vec{m}_1, \vec{m}_2, \dots, \vec{m}_N$ по измеренным в N произвольных точках некоторой вертикальной плоскости, содержащей ось OZ , трем компонентам вектора магнитной индукции (т.е. известным векторам $\vec{B}_1, \vec{B}_2, \dots, \vec{B}_N$). Принимая во внимание особенности сферической системы координат, здесь и далее будем предполагать, что сенсоры не размещаются вдоль оси OZ , а также что ни сенсоры, ни диполи не находятся в начале координат выбранной нами (фиксированной) сферической системы координат.



Фиг. 3. Расположение сенсоров и магнитных диполей в вертикальной плоскости.

Система уравнений (21) приобретает вид

$$\begin{aligned} \sum_{i=1}^N \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\vartheta}\}\rho_{ki,r}}{\rho_{ki}^5} - \frac{m_{ir}}{\rho_{ki}^3} \right) &= B_{kr}, \\ \sum_{i=1}^N \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\vartheta}\}\rho_{ki,\vartheta}}{\rho_{ki}^5} - \frac{m_{i\vartheta}}{\rho_{ki}^3} \right) &= B_{k\vartheta}, \\ \sum_{i=1}^N \left(-\frac{m_{i\varphi}}{\rho_{ki}^3} \right) &= B_{k\varphi}, \quad k = 1, \dots, N. \end{aligned} \quad (2.10)$$

Если для каких-либо двух индексов k и l справедливы соотношения

$$\rho_{ki} = \rho_{li}, \quad i = 1, \dots, N, \quad (2.11)$$

то две строки в (2.10) будут совпадать (в k -м и l -м блоках для соответствующих сенсоров):

$$\begin{bmatrix} 0 & -\frac{1}{\rho_{k1}^3}; 0 & -\frac{1}{\rho_{k2}^3}; \dots; 0 & -\frac{1}{\rho_{kN}^3} & 0 \\ 0 & -\frac{1}{\rho_{l1}^3}; 0 & -\frac{1}{\rho_{l2}^3}; \dots; 0 & -\frac{1}{\rho_{lN}^3} & 0 \end{bmatrix}. \quad (2.12)$$

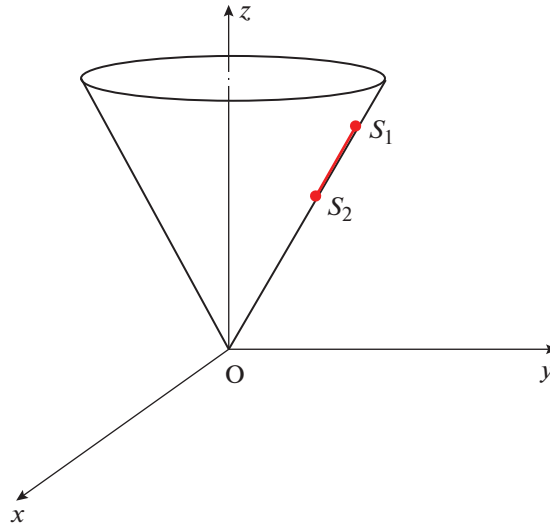
Такая ситуация может возникнуть в том случае, когда какие-либо два сенсора расположены симметрично относительно прямой, на которой лежат магнитные диполи (см. фиг. 3).

Таким образом, нами доказана аналогичная локальному варианту

Теорема 4. Если все сенсоры и диполи расположены в вертикальной плоскости, проходящей через ось OZ , все диполи лежат на одной прямой и для каких-то двух сенсоров имеет место ситуация, изображенная на фиг. 3, то решение системы (2.10) определяется неоднозначно.

2.3. Случай N диполей на оси OZ и N сенсоров, расположенных в произвольных точках

Рассмотрим постановку обратной задачи по поиску компонент N магнитных диполей, которые находятся в точках вертикальной оси OZ (за исключением начала координат), по информации от N сенсоров, расположение которых более-менее произвольно. Из изложенного в п. 2.2 настоящего раздела следует, что если для каких-либо двух сенсоров выполняются условия теоремы 4, то решение системы уравнений (2.4) определяется неоднозначно. В частности, если какие-либо два из N сенсоров лежат на окружности произвольного радиуса с центром на OZ , а сама окружность расположена в горизонтальной плоскости. Тогда радиус-векторы, соединяющие эти два сенсора с каждым из диполей, имеют нулевую компоненту вдоль $\vec{e}_\varphi$, и строки матрицы системы (2.4), соответствующие указанным двум сенсoram, “выглядят” так же, как в (2.12), поскольку расстояния от выделенных сенсоров до диполя с произвольным номером l одинаковы. Таким образом, для рассмотренного варианта взаимного расположения диполей и сенсоров будет справедлив аналог теоремы 4.



Фиг. 4. Расположение сенсоров и магнитных диполей в плоскости, касательной к конусу $\vartheta = \text{const}$.

2.4. Случай N диполей и N сенсоров, расположенных в касательной плоскости к конусу $\vartheta = \text{const}$

Перейдем теперь к изучению случая, когда сенсоры располагаются вдоль образующей конуса $\vartheta = \text{const}$. Диполи при этом должны лежать в касательной плоскости к конусу, содержащей указанную образующую (фиг. 4).

При таком расположении диполей и сенсоров у радиус-векторов, соединяющих диполи и сенсоры, будет отсутствовать компонента $\bar{e}_\vartheta$.

Система уравнений (21) принимает вид

$$\sum_{i=1}^N \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\varphi}\}\rho_{ki,r}}{\rho_{ki}^5} - \frac{m_{ir}}{\rho_{ki}^3} \right) = B_{kr}, \quad \sum_{i=1}^N \left(-\frac{m_{i\vartheta}}{\rho_{ki}^3} \right) = B_{k\vartheta}, \quad (2.13)$$

$$\sum_{i=1}^N \left(\frac{3\{m_{ir}\rho_{ki,r} + m_{i\vartheta}\rho_{ki,\varphi}\}\rho_{ki,\varphi}}{\rho_{ki}^5} - \frac{m_{i\varphi}}{\rho_{ki}^3} \right) = B_{k\varphi}, \quad k = 1, \dots, N.$$

Если диполи лежат на серединном перпендикуляре к отрезку, соединяющему какие-то два сенсора, то опять мы оказываемся в ситуации, когда две строки матрицы системы (2.13) совпадают, поскольку имеют вид (2.12).

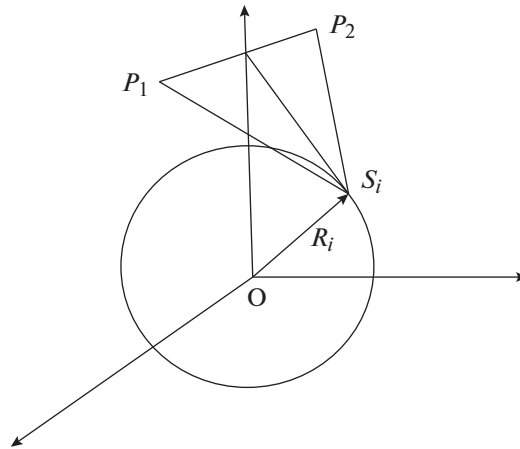
Следовательно, верна следующая

Теорема 5. При расположении N сенсоров и N диполей в касательной плоскости к какому-либо конусу $\vartheta = \text{const}$ и выполнении следующего условия: диполи лежат на серединном перпендикуляре к отрезку, соединяющему какие-либо два сенсора, решение системы (2.13) определяется неоднозначно.

2.5. Случай двух диполей и N сенсоров, произвольным образом расположенных в пространстве

Рассмотрим, наконец, проблему интерпретации магнитометрических данных по информации от N сенсоров, расположенных в пространстве произвольным образом. В каком случае возможно вырождение системы (2.4) из-за обращения в нуль компоненты радиус-векторов $\bar{\rho}_{ki} = (\rho_{ki,r}, \rho_{ki,\vartheta}, \rho_{ki,\varphi})$ вдоль $\bar{e}_r$?

Пусть у нас два каких-либо диполя расположены в точках P_1, P_2 (см. фиг. 5): на некоторой прямой, параллельной плоскости XOY , и симметрично относительно оси OZ . Предположим также, что касательная плоскость в точке сферы, на которой расположен один из сенсоров, проходит через эти два магнитных диполя (фиг. 5). В этом случае радиус-вектор сенсора S_1 , как легко видеть, перпендикулярен указанной касательной плоскости, а векторы, соединяющие два диполя



Фиг. 5. Случай двух диполей и N сенсоров, произвольным образом расположенных в пространстве.

с этим сенсором, не имеют компоненты $\bar{e}_r$ и равны по модулю. Система (2.4) модифицируется и имеет следующий вид:

$$\sum_{i=1}^2 \left(-\frac{m_{ir}}{\rho_{ki}^3} \right) = B_{kr}, \quad \sum_{i=1}^2 \left(\frac{3\{m_{i\vartheta}\rho_{ki,\vartheta} + m_{i\varphi}\rho_{ki,\varphi}\}\rho_{ki,\vartheta}}{\rho_{ki}^5} - \frac{m_{i\vartheta}}{\rho_{ki}^3} \right) = B_{k\vartheta},$$

$$\sum_{i=1}^2 \left(\frac{3\{m_{i\vartheta}\rho_{ki,\vartheta} + m_{i\varphi}\rho_{ki,\varphi}\}\rho_{ki,\varphi}}{\rho_{ki}^5} - \frac{m_{i\varphi}}{\rho_{ki}^3} \right) = B_{k\varphi}, \quad k = 1, \dots, N. \quad (2.14)$$

Если все N сенсоров расположены подобным образом по отношению к выделенным двум диполям, то мы снова оказываемся в ситуации, аналогичной описанной выше, а именно, система (31) вырождается, поскольку, по крайней мере, две ее строки пропорциональны:

$$\begin{bmatrix} 0 & -\frac{1}{\rho_{k1}^3} & 0; & 0 & -\frac{1}{\rho_{k2}^3} & 0 \\ 0 & -\frac{1}{\rho_{l1}^3} & 0; & 0 & -\frac{1}{\rho_{l2}^3} & 0 \end{bmatrix}, \quad \rho_{k1} = \rho_{k2}; \quad \rho_{l1} = \rho_{l2}. \quad (2.15)$$

Таким образом, нами доказана следующая

Теорема 6. Если два диполя расположены в произвольной горизонтальной плоскости симметрично относительно оси OZ и одновременно на всех касательных плоскостях к сферам, на которых находятся N сенсоров, то система (2.14) вырождена.

ЗАКЛЮЧЕНИЕ

Рассмотренные в настоящей работе примеры вырожденных и невырожденных систем линейных алгебраических уравнений, к которым редуцируются многие геофизические задачи интерпретационного характера, могут быть полезны при планировании эксперимента, выборе оптимальной сети пунктов наблюдений при проведении гравиметрической или магнитометрической съемки. Перспективным представляется подход к решению обратных геофизических и геодезических задач, при котором координаты как источников, так и точек наблюдения рассматриваются “единым блоком” — возникает необходимость исследования свойств некоторой расширенной системы нелинейных уравнений, элементы которой являются сечениями некоторого гладкого расслоения.

Метод S -аппроксимаций является весьма простым и эффективным методом решения линейных обратных задач гравиметрии и магнитометрии и может применяться также для поиска нулевого или первого приближения к решению нелинейной обратной задачи по определению геометрии и местоположения источника аномального гравитационного или магнитного поля. При решении такого типа задач необходимо применение регуляризирующих алгоритмов (см. [10, 11]). Постановки, приведенные в разд. 2 настоящей статьи, могут быть рассмотрены и в рамках

метода линейных интегральных представлений: в случае магнитного поля вид линейных функционалов (1) изменится, но не кардинальным образом.

Авторы выражают глубокую благодарность А.С. Леонову за полезные рекомендации и внимание к работе.

Работа выполнена при финансовой поддержке РФФИ (грант № 23-41-00002).

СПИСОК ЛИТЕРАТУРЫ

1. *Страхов В.Н.* Геофизика и математика. М.: ОИФЗ РАН, 1999. 64 с.
2. *Страхов В.Н., Степанова И.Э.* Метод S-аппроксимаций и его использование при решении задач гравиметрии (локальный вариант) // Физика Земли. 2002. № 2. С. 3–19.
3. *Степанова И.Э.* Апробация методики S-аппроксимаций на материалах детальной гравиметрической и магнитометрических съемок // Геофиз. журнал. 2007. Т. 28. № 1. С. 97–105.
4. *Stepanova I.E.* On the S-approximation of the Earth's gravity field. Regional version // Inverse Problem. Sci. and Engineer. 2009. V 17. № 8. P. 1095–1111.
5. *Gudkova T.V., Stepanova I.E., Batov A.V.* Density anomalies in subsurface layers of mars: model estimates for the site of the InSight mission seismometer // Solar System Res. 2020. V. 54. P. 15–19.
6. *Кошляков Н.С., Глинер Э.Б., Смирнов М.М.* Основные дифференциальные уравнения математической физики. М.: Физматгиз, 1962. 767 с.
7. *Лаврентьев М.А., Люстерник Л.А.* Курс вариационного исчисления. М.—Л.: Гостоптехиздат, 1950.
8. *Lukyanenko D.V., Yagola A.G.* Some methods for solving of 3D inverse problem of magnetometry // Eurasian J. Math. and Comp. Appl. 2016. V. 4. № 3. P. 4–14.
9. *Kolotov I.I., Lukyanenko D.V., Stepanova I.E., et al.* Recovering the magnetic image of Mars from satellite observations // J. of Imaging. 2021. V. 7. № 11. P. 234.
10. *Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г.* Численные методы решения некорректных задач. М.: Наука, 1990. 230 с.
11. *Ягола А.Г., Степанова И.Э., Ван Янфей, Титаренко В.Н.* Обратные задачи и методы их решения. Приложения к геофизике. М.: Бином. Лаборатория знаний, 2014. 214 с.

ОПТИМАЛЬНОЕ
УПРАВЛЕНИЕ

УДК 519.85

БЕЗГРАДИЕНТНЫЕ МЕТОДЫ ФЕДЕРАТИВНОГО ОБУЧЕНИЯ
С l_1 И l_2 -РАНДОМИЗАЦИЕЙ ДЛЯ ЗАДАЧ НЕГЛАДКОЙ
ВЫПУКЛОЙ СТОХАСТИЧЕСКОЙ ОПТИМИЗАЦИИ¹⁾© 2023 г. Б. А. Аляшкар¹, А. В. Гасников^{1,2,3}, Д. М. Двинских⁴, А. В. Лобанов^{1,5,6,*},[¶]¹ 141700 Долгопрудный, М.о., Институтский пер., 9, НИУ МФТИ, Россия² 127051 Москва, Б. Каретный пер., 19, стр. 1, ИППИ РАН, Россия³ 385000 Республика Адыгея, Майкоп, ул. Первомайская, 208, КМЦ АГУ, Россия⁴ 109028 Москва, Покровский б-р, 11, НИУ ВШЭ, Россия⁵ 109004 Москва, ул. А. Солженицына, 25, ИСП РАН, Россия⁶ 125993 Москва, Волоколамское шоссе, 4, МАИ (НИУ), Россия

*e-mail: lobbsasha@mail.ru

Поступила в редакцию 18.11.2022 г.

Переработанный вариант 20.05.2023 г.

Принята к публикации 29.05.2023 г.

В статье исследуются негладкие задачи выпуклой стохастической оптимизации. С помощью техники сглаживания, базирующейся на замене значения функции в рассматриваемой точке усредненным значением функции по шару (в l_1 -норме или l_2 -норме) малого радиуса с центром в этой точке, исходная задача сводится к гладкой задаче (константа Липшица градиента которой обратно пропорционально радиусу шара). Важным свойством используемого сглаживания является возможность вычислять несмещенную оценку градиента сглаженной функции на основе только реализаций исходной функции. Полученную гладкую задачу стохастической оптимизации предлагается далее решать в распределенной архитектуре федеративного обучения (задача решается параллельно: узлы делают локальные шаги, например, стохастического градиентного спуска, потом коммуницируют — все со всеми, затем все это повторяется). Цель статьи — построить на базе современных достижений в области безградиентной негладкой оптимизации и в области федеративного обучения безградиентные методы решения задач негладкой стохастической оптимизации в архитектуре федеративного обучения. Библ. 22. Фиг. 3. Табл. 1.

Ключевые слова: безградиентные методы, методы с неточным оракулом, федеративное обучение.

DOI: 10.31857/S0044466923090028, EDN: MLJEUN

1. ВВЕДЕНИЕ

В статье рассматриваются задачи стохастической оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) := \mathbb{E}_{\xi} [f(x, \xi)] \quad (1)$$

и их седловые обобщения. При этом считается, что функция $f(x)$ — выпуклая, негладкая и имеет ограниченную константу Липшица. Также предполагается, что для наблюдения доступна реализация функции $f(x, \xi)$, но не ее градиент (по x) $\nabla f(x, \xi)$. Более того, в работе также рассматривается случай, когда эта реализация доступна с небольшим шумом неслучайной природы. Данная статья основывается на работе [1], в которой предложен оптимальный алгоритм (с точностью до логарифмических по размерности пространства множителей в оценках оракульных вызовов) решения задачи (1) сразу по трем критериям: 1) число оракульных вызовов (вычислений $f(x, \xi)$), 2) число последовательных итераций метода (на одной итерации можно вызывать оракул многократно), 3) максимально допустимый уровень (потенциально враждебного) шума, при котором

¹⁾ Работа выполнена при финансовой поддержке РНФ (проект 23-11-00229), <https://rscf.ru/project/23-11-00229/>.

[¶] Основной вклад в статью принадлежит Александру Лобанову (lobbsasha@mail.ru). Согласно правилам журнала, авторы статьи расставлены в алфавитном порядке.

все еще возможно достичь желаемой точности. В основе подхода [1] лежит довольно старая идея (см., например, [2]) замены исходной негладкой функции $f(x)$ ее сглаженной версией $f_\gamma(x) = \mathbb{E}_{\tilde{e}}[f(x + \gamma\tilde{e})]$, где $\tilde{e}$ – равномерно распределенный случайный вектор на $B_2^d(1)$ – единичном евклидовом шаре с центром в нуле. Для сглаженной функции конечная аппроксимация (с шагом γ) производной по случайному направлению (выбранному равномерно на единичной евклидовой сфере) будет несмещенной (рандомизированной) оценкой градиента, обладающей при использовании симметричной разностной аппроксимации хорошей дисперсией [3] – такой же, как если бы $f(x)$ была гладкой.

В другой работе [4] показано, что если использовать сглаживание не по евклидовому шару, а по шару в l_1 -норме, то в определенных ситуациях можно улучшить оценки [1] на число оракульных вызовов на логарифмический множитель (по размерности пространства). Однако при этом в [4] не обсуждался вопрос о числе последовательных итераций и использовалась другая (более узкая) концепция враждебного шума, для которой вопрос о точных нижних оценках, насколько нам известно, по-прежнему остается открытым.

В настоящей статье показано, как можно в теории улучшить результаты работы [1] за счет схемы сглаживания из [4] (см. также работу [5], в которой данная схема сглаживания также предлагалась, но анализ был проведен менее точно). Как уже отмечалось выше, улучшения носят логарифмический масштаб. Отметим при этом, что проведенные в работе численные эксперименты не смогли явно уловить эффект такого улучшения оценок.

Другим важным направлением развития работы [1] стало обобщение всех результатов этой работы (и ее модификации со сглаживанием на шаре в l_1 -норме) на распределенные алгоритмы в архитектуре федеративного обучения [6]. Отметим, что в федеративном обучении рассматриваются задачи типа (1), во-первых, гладкие, во-вторых, с полноградиентным (стохастическим) оракулом. Задача (1) решается независимо в каждом узле, потом узлы коммуницируют и вычисляют среднее (по узлам), затем снова начинают независимо решать задачу, стартуя с этого среднего. Через некоторое время узлы снова коммуницируют, считают среднее и процесс повторяется... В данной работе за счет использования схемы сглаживания (с введением дополнительной рандомизации) получается свести негладкую задачу стохастической оптимизации с безградиентным оракулом, к гладкой задаче стохастической оптимизации с оракулом, выдающим стохастический градиент, в котором стохастика формируется из изначальной стохастики, присущей исходной постановке задачи, и стохастике, возникшей при сглаживании (рандомизированной). Оказалось, что в существующей сейчас литературе практически отсутствуют методы федеративного обучения для негладких задач. По-видимому, это было связано с тем, что для негладких задач в общем случае невозможно осуществлять батч-параллелизацию (параллелизацию по ξ). Однако для безградиентного оракула батч-параллелизация возможна за счет появления дополнительной (рандомизированной) случайности в виде случайного направления [1]! Собственно за счет этого и удается перенести результаты работы [1] на архитектуру федеративного обучения. Насколько нам известно, результатам, полученным в данном направлении, сейчас нет конкурентов, поэтому здесь не приводится литературный обзор конкурирующих работ.

2. ОСНОВНОЙ ВКЛАД И СТРУКТУРА

Основной вклад статьи состоит в следующем.

- Приводим подробное описание двух схем сглаживания (параллельно): с l_1 и l_2 -рандомизациями. Находим константу Липшица градиента L_{f_γ} для l_1 -рандомизации. Для явного описания оценки второго момента в l_2 -рандомизации с двухточечной обратной связью находим константу c , которая в исходной статье [3], из которой эта константа была взята, так и не была посчитана. Получаем оценку дисперсии (второго момента) для одноточечного случая l_1 -рандомизации. Показываем, что l_1 -рандомизация с одноточечным оракулом также превосходит l_2 -рандомизацию с точностью до $\ln d$, как и с двухточечным оракулом.

- Получаем оптимальные верхние оценки скорости сходимости пробатченных алгоритмов первого порядка Minibatch SMP и Single-Machine SMP для решения задач оптимизации с седловой точкой в архитектуре федеративного обучения.

- Описываем технику генерирования безградиентных алгоритмов (решения седловых задач и задач выпуклой оптимизации), оптимальных по количеству коммуникационных раундов N ,

максимальному значению допустимого шума Δ и оракульной сложности T . Показываем, что одноточечные и двухточечные алгоритмы в архитектуре федеративного обучения, использующие l_1 -рандомизацию, работают лучше алгоритмов, использующих l_2 -рандомизацию, с точностью до логарифма в l_1 -норме. Сравниваем одноточечные алгоритмы с двухточечными и показываем, что для решения с ϵ точностью (по функции) одноточечные алгоритмы требуют в $O(d/\epsilon^2)$ больше обращений к оракулу, чем двухточечные. Анализируем работу алгоритма Minibatch Accelerated SGD, используя разные схемы сглаживания, на практическом эксперименте.

Структура статьи следующая: в разд. 3 и 4 представлено краткое введение в техники сглаживания и федеративное обучение соответственно. В п. 4.2 приводится главный результат работы. В разд. 5 приведены основные идеи доказательств теорем 1, 2. Численные эксперименты представлены в разд. 6.

3. СХЕМЫ СГЛАЖИВАНИЯ

Схема сглаживания позволяет создавать безградиентные методы решения негладких задач, модифицируя одноименные алгоритмы первого порядка, предназначенные для решения гладких задач. Впервые схема сглаживания была описана в книге [2], где представлена идея решения задач методами первого порядка, используя вместо стохастического оракула первого порядка безградиентный стохастический оракул, который получен через теорему Стокса. С тех пор уже придумали различные техники сглаживания, однако основная идея восходит от [2]. В данном разделе представим параллельно две схемы сглаживания: с l_1 -рандомизацией [4, 5] и с l_2 -рандомизацией [1], включающую стохастическую оптимизацию и смещенную оценку безградиентного оракула. Для этого рассматривается стохастическая негладкая выпуклая задача оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} \{f(x) := \mathbb{E}_{\xi} [f(x, \xi)]\}, \quad (2)$$

где Q — выпуклое и компактное множество и $f(x, \xi)$ — выпуклая функция на множестве $Q_{\gamma} := Q + B_p^d(\gamma)$. Здесь предполагается, что безградиентный оракул возвращает значение функции $f(x)$, возможно, с некоторым враждебным шумом $\delta(x)$: $f_{\delta}(x) := f(x) + \delta(x)$.

3.1. Обозначения и предположения

Обозначим через $\langle x, y \rangle := \sum_{i=1}^d x_i y_i$ стандартное скалярное произведение $x, y \in \mathbb{R}^d$. Через $\|x\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ обозначаем l_p -норму ($p \geq 1$) в $\mathbb{R}^d$ и $B_p^d(r) := \{x \in \mathbb{R}^d : \|x\|_p \leq r\}$, и $S_p^d(r) := \{x \in \mathbb{R}^d : \|x\|_p = r\}$ определяют l_p -шар и l_p -сферу соответственно. Объем l_p -шара определяется через

$$V(B_p^d(r)) := c_d r^d = \frac{\left(2\Gamma\left(\frac{1}{p} + 1\right)\right)^d}{\Gamma\left(\frac{d}{p} + 1\right)} r^d,$$

где $\Gamma(\cdot)$ обозначает гамма-функцию. Для обозначения “расстояния” между начальной точкой x^0 и решением исходной задачи x_* вводим $R := \tilde{O}\left(\|x^0 - x_*\|_p\right)$, где $\tilde{O}(\cdot)$ является $O(\cdot)$ с точностью до логарифма $\sqrt{\ln d}$.

Предположение 1 (липшицева непрерывная функция). Функция $f(x, \xi)$ и является M -липшицевой непрерывной функцией в l_p -норме, т.е. для всех $x, y \in Q$ имеем

$$|f(y, \xi) - f(x, \xi)| \leq M(\xi) \|y - x\|_p.$$

Более того, существует положительная константа M , которая определяется следующим образом: $\mathbb{E}[M^2(\xi)] \leq M^2$. В частности, для $p = 2$ используем обозначение M_2 для константы Липшица.

Предположение 2 (ограниченность шума). Для всех $x \in Q$ выполняется $|\delta(x)| \leq \Delta$, где Δ — уровень неточности (шума).

Предположение 3. Для всех $x \in Q$ выполняется $\mathbb{E}_\xi[|f(x, \xi)|^2] \leq G^2$.

3.2. Гладкая аппроксимация

Поскольку задача (2) является негладкой, то вводим следующую гладкую аппроксимацию негладкой функции

$$f_\gamma(x) := \mathbb{E}_{\tilde{e}}[f(x + \gamma\tilde{e})], \quad (3)$$

где $\gamma > 0$, $\tilde{e}$ — случайный вектор, равномерно распределенный на $B_p^d(1)$ (далее ограничимся рассмотрением случаев $p = 1$ и $p = 2$). Здесь $f(x) := \mathbb{E}f(x, \xi)$.

Следующая лемма представляет совокупность свойств аппроксимации функции f в зависимости от распределения вектора $\tilde{e}$. Основываясь на [1, 4, 7] и дополняя найденной константой Липшица градиента L_{f_γ} в случае $p = 1$, выпишем свойства функции f_γ .

Лемма 1. Для всех $x, y \in Q$ с предположением 1 справедливо

$\tilde{e} \in RB_1^d(1)$	$\tilde{e} \in RB_2^d(1)$
$f(x) \leq f_\gamma(x) \leq f(x) + \frac{2}{\sqrt{d}}\gamma M_2$	$f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$
$f_\gamma - M$ -липшицева:	
$ f_\gamma(y) - f_\gamma(x) \leq M \ y - x\ _p$	$ f_\gamma(y) - f_\gamma(x) \leq M \ y - x\ _p$
f_γ имеет L_{f_γ} -липшицевый градиент:	
$\ \nabla f_\gamma(y) - \nabla f_\gamma(x)\ _q \leq \underbrace{\frac{dM}{2\gamma}}_{L_{f_\gamma}} \ y - x\ _p$	$\ \nabla f_\gamma(y) - \nabla f_\gamma(x)\ _q \leq \underbrace{\frac{\sqrt{d}M}{\gamma}}_{L_{f_\gamma}} \ y - x\ _p$

где q такой, что $1/p + 1/q = 1$.

Доказательство приведено в п. 5.1.

3.3. Рандомизация с двухточечной обратной связью

Аппроксимация градиента зашумленной функции $f_\gamma(x, \xi)$ из (3) может быть получена через две точки, близких к x . Для этого определим случайный вектор e , равномерно распределенный на $S_p^d(1)$. Тогда градиент может быть оценен следующей аппроксимацией.

- Оценка градиента для l_1 -рандомизации ($e \in RS_1^d(1)$) [5] (см. также [4]):

$$\nabla f_\gamma(x, \xi, e) = \frac{d}{2\gamma} (f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi)) \text{sign}(e). \quad (4)$$

- Оценка градиента для l_2 -рандомизации ($e \in RS_2^d(1)$) [3]:

$$\nabla f_\gamma(x, \xi, e) = \frac{d}{2\gamma} (f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi))e. \quad (5)$$

Для оценки градиента в (4) и (5) выбрана центральная конечно-разностная схема рандомизации, так как в работе [8] поясняется, что в гладком случае выгоднее оценивать градиент именно центральной конечно-разностной схемой (CFD), а не прямой конечно-разностной схемой (FFD). Отметим, что при $\Delta = 0$ оценки будут несмещенными, т.е. $E_{e, \xi}[\nabla f_\gamma(x, \xi, e)] = \nabla f_\gamma(x)$.

Далее приведем свойства $\nabla f_\gamma(x, \xi, e)$ для двух рандомизаций, используя известные результаты из [1, 3–5, 9, 10]. Во многих работах для l_2 -рандомизации оценка второго момента записывается

с точностью до константы c , поэтому в лемме 2 приводятся оценки второго момента для l_1 и l_2 -рандомизаций с уточнением константы c .

Лемма 2. Для всех $x \in Q$ с предположениями 1 и 2 имеем

i) $\nabla f_\gamma(x, \xi, e)$ с l_1 -рандомизации (4) имеет оценку дисперсии (второй момент):

$$E_e \left[\|\nabla f_\gamma(x, \xi, e)\|_q^2 \right] \leq \kappa(p, d) \left(M_2^2 + \frac{d^2 \Delta^2}{12(1 + \sqrt{2})^2 \gamma^2} \right),$$

где $1/p + 1/q = 1$ и

$$\kappa(p, d) = 48(1 + \sqrt{2})^2 d^{\frac{2-p}{p}}.$$

Если Δ достаточно мала, то

$$E_e \left[\|\nabla f_\gamma(x, \xi, e)\|_q^2 \right] \leq 2\kappa(p, d) M_2^2. \quad (6)$$

ii) $\nabla f_\gamma(x, \xi, e)$ с l_2 -рандомизации (5) имеет оценку дисперсии (второй момент):

$$E_e \left[\|\nabla f_\gamma(x, \xi, e)\|_q^2 \right] \leq \kappa(p, d) \left(M_2^2 + \frac{d^2 \Delta^2}{\sqrt{2} \gamma^2} \right),$$

где $1/p + 1/q = 1$ и

$$\kappa(p, d) = \sqrt{2} \min \{q, \ln d\} d^{\frac{2-p}{p}}.$$

Если Δ достаточно мала, то

$$E_e \left[\|\nabla f_\gamma(x, \xi, e)\|_q^2 \right] \leq 2\kappa(p, d) M_2^2. \quad (7)$$

Доказательство приведено в п. 5.2.

3.4. Рандомизация с односточечной обратной связью

Для случая когда двухточечная обратная связь недоступна, схемы сглаживания могут использовать односточечную обратную связь через следующую несмещенную оценку.

- Оценка градиента для l_1 -рандомизации ($e \in RS_1^d(1)$):

$$\nabla f_\gamma(x, \xi, e) = \frac{d}{\gamma} (f_\delta(x + \gamma e, \xi)) \text{sign}(e). \quad (8)$$

- Оценка градиента для l_2 -рандомизации ($e \in RS_2^d(1)$) [1]:

$$\nabla f_\gamma(x, \xi, e) = \frac{d}{\gamma} (f_\delta(x + \gamma e, \xi)) e. \quad (9)$$

Тогда свойства $\nabla f_\gamma(x, \xi, e)$ для двух рандомизаций будут иметь следующий вид.

Лемма 3. Для всех $x \in Q$ с предположениями 2 и 3 имеем

i) $\nabla f_\gamma(x, \xi, e)$ с l_1 -рандомизации (8) имеет оценку дисперсии (второй момент):

$$E_e \left[\|\nabla f_\gamma(x, \xi, e)\|_q^2 \right] \leq \kappa(p, d) \left(\frac{G^2}{\gamma^2} + \frac{\Delta^2}{\gamma^2} \right),$$

где $1/p + 1/q = 1$ и

$$\kappa(p, d) = d^{\frac{4-p}{p}}.$$

Если Δ достаточно мала, то

$$E_e \left[\left\| \nabla f_\gamma(x, \xi, e) \right\|_q^2 \right] \leq 2\kappa(p, d) \frac{G^2}{\gamma^2}. \quad (10)$$

ii) $\nabla f_\gamma(x, \xi, e)$ с l_2 -рандомизации (9) имеет оценку дисперсии (второй момент):

$$E_e \left[\left\| \nabla f_\gamma(x, \xi, e) \right\|_q^2 \right] \leq \kappa(p, d) \left(\frac{G^2}{\gamma^2} + \frac{\Delta^2}{\gamma^2} \right),$$

где $1/p + 1/q = 1$ и

$$\kappa(p, d) = \min\{q, \ln d\} d^{\frac{3-2}{p}}.$$

Если Δ достаточно мала, то

$$E_e \left[\left\| \nabla f_\gamma(x, \xi, e) \right\|_q^2 \right] \leq 2\kappa(p, d) \frac{G^2}{\gamma^2}. \quad (11)$$

Доказательство приведено в п. 5.3.

3.5. Алгоритм сглаживания

Основываясь на выше представленных элементах, опишем общий подход, именуемый как схема сглаживания. Предположим, что у нас есть некоторый ускоренный пакетный алгоритм $\mathbf{A}(L, \sigma^2)$ архитектуры федеративного обучения, который решает задачу (2) с предположением, что f является гладкой и удовлетворяет

$$\left\| \nabla f(y) - \nabla f(x) \right\|_q \leq L \|y - x\|_p \quad \forall x, y \in Q_\gamma$$

и с помощью стохастического оракула первого порядка, который зависит от случайной величины η и возвращает в точке x смещенный стохастический градиент $\nabla_x f_\gamma(x, \eta)$

$$\mathbb{E}_\eta \left[\left\| \nabla_x f_\gamma(x, \eta) - \nabla f(x) \right\|_q^2 \right] \leq \sigma^2.$$

Тогда общий подход схемы сглаживания состоит в применении $\mathbf{A}(L, \sigma^2)$ к сглаженной задаче

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x) \quad (12)$$

для решения с $\varepsilon/2$ точностью с известными параметрами $\eta = e$, $\nabla_x f_\gamma(x, \eta) = \nabla f_\gamma(x, \xi, e)$, $L = L_{f_\gamma}$ и γ , представленная в следствии 1.

Следствие 1. Согласно лемме 1, чтобы получить ε -точность решения задачи (2), необходимо решить задачу (12) с $\varepsilon/2$ -точностью со следующим параметром

Схема сглаживания с l_1 -рандомизацией	Схема сглаживания с l_2 -рандомизацией
$\gamma = \frac{\sqrt{d\varepsilon}}{4M_2}$	$\gamma = \frac{\varepsilon}{2M_2}$

где $\varepsilon > 0$ является желаемой точностью решения задачи (2) с точки зрения субоптимальности: $\mathbb{E}[f(x^N) - f(x_*)] \leq \varepsilon$.

Следствие 2. Согласно лемме 1, подставляя параметр γ из следствия 1, имеем

Схема сглаживания с l_1 -рандомизацией	Схема сглаживания с l_2 -рандомизацией
$L_{f_\gamma} = \frac{2\sqrt{d}MM_2}{\varepsilon}$	$L_{f_\gamma} = \frac{2\sqrt{d}MM_2}{\varepsilon}$

Следствие 3. Согласно лемме 2 уравнения (6) и (7) для двухточечного оракула примут вид

Схема сглаживания с l_1 -рандомизацией	Схема сглаживания с l_2 -рандомизацией
$\sigma^2 \leq 48(1 + \sqrt{2})^2 d^{\frac{2-p}{p}} M_2^2$	$\sigma^2 \leq 2\sqrt{2} \min\{q, \ln d\} d^{\frac{2-p}{p}} M_2^2$

если Δ достаточно мала.

Следствие 4. Согласно лемме 3 уравнения (10) и (11) для однотоочечного оракула примут вид

Схема сглаживания с l_1 -рандомизацией	Схема сглаживания с l_2 -рандомизацией
$\sigma^2 \leq 32d^{\frac{3-p}{p}} \frac{G^2 M_2^2}{\varepsilon^2}$	$\sigma^2 \leq 8 \min\{q, \ln d\} d^{\frac{3-p}{p}} \frac{G^2 M_2^2}{\varepsilon^2}$

если Δ достаточно мала.

Замечание 1. Если вместо стохастической негладкой выпуклой задачи оптимизации (2) рассматривать стохастическую негладкую выпукло-вогнутую задачу с седловой точкой

$$\min_{x \in Q_x \subseteq \mathbb{R}^{d_x}} \max_{y \in Q_y \subseteq \mathbb{R}^{d_y}} \{f(x, y) := \mathbb{E}[f(x, y, \xi)]\},$$

то применяя схему сглаживания, описанную выше в данном разделе, отдельно к x и y переменным, получим те же результаты, что и для выпуклой оптимизации при соответствующих изменениях $f(x, \xi)$ на $f(z, \xi)$, где $z := (x, y)$, $z \in Q_z := Q_x \times Q_y$, за исключением одного пункта в лемме 1:

- ($\tilde{e} \in RB_1^d(1)$) вместо

$$f(x) \leq f_\gamma(x) \leq f(x) + \frac{2}{\sqrt{d}} \gamma M_2$$

имеем

$$f(x, y) - \frac{2}{\sqrt{d}} \gamma_y M_{2,y} \leq f_\gamma(x, y) \leq f(x, y) + \frac{2}{\sqrt{d}} \gamma_x M_{2,x};$$

- ($\tilde{e} \in RB_2^d(1)$) вместо

$$f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$$

имеем

$$f(x, y) - \gamma_y M_{2,y} \leq f_\gamma(x, y) \leq f(x, y) + \gamma_x M_{2,x},$$

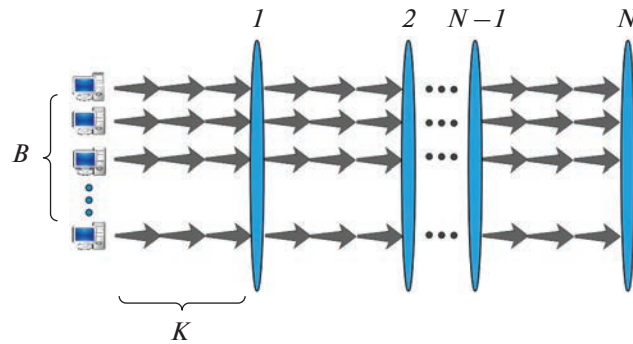
где $\gamma = (\gamma_x, \gamma_y)$, а $M_{2,x}$ и $M_{2,y}$ — соответствующие константы Липшица в l_2 -норме.

4. ФЕДЕРАТИВНОЕ ОБУЧЕНИЕ

Архитектура распределенного обучения выглядит следующим образом: между “компьютерами” распределяется набор данных, каждый компьютер делает одно локальное обновление (локальный шаг, например, шаг стохастического градиентного спуска), после чего происходит глобальное обновление (глобальный шаг, коммуникация всех компьютеров со всеми), далее цепочка локально-глобальных обновлений повторяется. Однако глобальное обновление, например при большом размере данных, тратит много вычислительных ресурсов, в отличие от локального обновления. Тогда вводится архитектура федеративного обучения, представленная на фиг. 1, где в гомогенном случае B компьютеров делают параллельно K локальных обновлений перед каждым коммуникационным раундом, общее число которых составляет N . Таким образом, NK является общим числом итераций алгоритма, а $T = NKVB$ является общим числом вызовов стохастического градиента.

4.1. Оптимальные алгоритмы первого порядка

В данной статье остановимся и рассмотрим класс пробатченных ускоренных методов первого порядка, а именно Minibatch Accelerated SGD (Mb-Ac-SGD) и Single-Machine Accelerated SGD



Фиг. 1. Архитектура федеративного обучения.

(SM-Ac-SGD) из [11], Local-AC-CA из [12] и FedAc из [13], результаты скорости сходимости которых представлены в табл. 1. В таблице были использованы следующие обозначения: R есть $\|x^0 - x_*\|_2$; B – количество работающих компьютеров; K – число локальных обновлений; N – количество коммуникационных раундов; L – гладкость.

Для квадратичной целевой функции было доказано, что $K > 1$ локальное обновление приводит к оптимальным оценкам скорости сходимости (алгоритм Local-AC-CA из [12]). Алгоритм FedAc из [13] обобщает результаты статьи [12] на случай выпуклых функций.

Но уже в 2021 г. были получены оптимальные оценки скорости сходимости в статье [11]. В данной статье утверждается, что оптимальные оценки можно получить только в двух случаях. Первый случай (алгоритм Minibatch Accelerated SGD) предполагает, что каждый компьютер выполняет одно локальное обновление перед коммуникационным раундом. Второй случай (алгоритм Single-Machine Accelerated SGD) предполагает, что работает только один компьютер, который выполняет NK обновлений. Несмотря на доказанные результаты статьи [11], на практике в общем случае выпуклых функций получается использовать $K > 1$ локальных шагов, незначительно теряя при этом в точности, однако существенно выигрывая в вычислительных ресурсах. Именно практические результаты позволяют ожидать в будущем позитивные теоретические результаты.

Существующие алгоритмы первого порядка для архитектуры федеративного обучения зачастую решают задачу выпуклой оптимизации. Что касается задач оптимизации с седловой точкой, то алгоритмы в архитектуре федеративного обучения на данный момент отсутствуют. В данной

Таблица 1. Результаты скорости сходимости

Алгоритм	$\mathbb{E}[f(\cdot)] - f^* \lesssim \dots$	Ссылка
Mb-Ac-SGD	$\frac{LR^2}{N^2} + \frac{\sigma R}{\sqrt{BNK}}$	(Woodworth и др., 2021) [11]
SM-Ac-SGD	$\frac{LR^2}{N^2 K^2} + \frac{\sigma R}{\sqrt{NK}}$	(Woodworth и др., 2021) [11]
Local-AC-CA	$\frac{LR^2}{N^2 K^2} + \frac{\sigma R}{\sqrt{BNK}}$	(Woodworth и др., 2020) [12]
FedAc	$\frac{LR^2}{N^2 K} + \frac{\sigma R}{\sqrt{BNK}} + \min \left\{ \frac{L^{1/3} \sigma^{2/3} R^{4/3}}{NK^{1/3}}, \frac{L^{1/2} \sigma^{1/2} R^{3/2}}{NK^{1/4}} \right\}$	(Yuan, Ma, 2020) [13]
Mb-SMP	$\max \left\{ \frac{LR^2}{N}, \frac{\sigma R}{\sqrt{BNK}} \right\}$	Приложение А
SM-SMP	$\max \left\{ \frac{LR^2}{NK}, \frac{\sigma R}{\sqrt{NK}} \right\}$	Приложение А

статье мы разработали и получили оптимальные оценки скорости сходимости пробатченных методов первого порядка Minibatch SMP (Mb-SMP) и Single-Machine SMP (SM-SMP) для решения задач оптимизации с седловой точкой, используя аналогичный подход, что и в выпуклой оптимизации [11]. Оценки сходимости для алгоритмов решения седловых задач также приведены в табл. 1. Подробное описание получения верхних оценок скорости сходимости для алгоритмов Minibatch SMP (Mb-SMP) и Single-Machine SMP (SM-SMP) приведено в приложении А.

4.2. Оптимальные алгоритмы нулевого порядка

Здесь представлен главный результат данной статьи, который заключается в объединении двух глобальных идей в области оптимизации, представленные в разд. 3 и п. 4.1. А именно решение стохастических негладких выпуклых/выпукло-вогнутых задач оптимизации безградиентными алгоритмами архитектуры федеративного обучения. Для разработки и применения безградиентных методов решения негладких задач предлагается выбрать метод первого порядка, используемый для решения гладких задач в архитектуре федеративного обучения. Далее необходимо модифицировать алгоритм выбранного метода путем замены вычисления стохастического градиента на безградиентную аппроксимацию с l_1 (4) или l_2 (5) рандомизацией. Полученный безградиентный алгоритм будет иметь одноименное название алгоритма первого порядка, однако он не будет требовать информацию о стохастическом градиенте.

Пусть под алгоритмом $\mathbf{A}(L, \sigma^2)$ будем понимать ускоренные пакетные алгоритмы решения задач выпуклой оптимизации: Minibatch и Single-Machine Accelerated SGD, Local-AC-CA и FedAc и ускоренные пакетные алгоритмы решения задач с седловой точкой: Minibatch SMP и Single-Machine SMP в архитектуре федеративного обучения. Тогда в теоремах 1, 2 предполагаем, что выполняется данное свойство: алгоритм $\mathbf{A}(L, \sigma^2)$ (со смещенным безградиентным оракулом $\nabla f_\gamma(x, \xi, e)$) является надежным, если смещение из следствия 5 для l_1 -рандомизации, из следствия 6 для l_2 -рандомизации (рассмотренных в п. 5.4) не накапливается в течение итерации метода. То есть, если для $\mathbf{A}(L, \sigma^2)$ при $\Delta = 0$:

$$E[f_\gamma(x^N) - f(x_*)] \leq \Theta_A(N),$$

тогда

- при $\Delta > 0$ и $d^{4-\frac{2}{p}}\Delta^2\gamma^{-2} \lesssim \kappa(p, d)M_2^2$ из (6):

$$E[f_\gamma(x^N) - f(x_*)] \leq O\left(\Theta_A(N) + d\Delta R\gamma^{-1}\right); \quad (13)$$

- при $\Delta > 0$ и $d^2\Delta^2\gamma^{-2} \lesssim dM_2^2$ из (7):

$$E[f_\gamma(x^N) - f(x_*)] \leq O\left(\Theta_A(N) + \sqrt{d}\Delta R\gamma^{-1}\right). \quad (14)$$

Предположение о выполнении данного свойства базируется на статье [14], где был разработан анализ сходимости методов для смещенных стохастических оракулов. Поэтому в теоремах 1, 2 приводим результаты, предполагая, что можно по аналогии провести анализ сходимости методов федеративного обучения, рассмотренных в этой работе, для смещенных стохастических оракулов. Однако в доказательстве теорем 1, 2 будем рассматривать случай при $\Delta = 0$ и приводить оптимальные оценки параметров разработанных безградиентных алгоритмов.

Теорема 1. *Схема сглаживания из разд. 3, применяемая к задаче (2), обеспечивает сходимость следующих двухточечных безградиентных алгоритмов: Minibatch и Single-Machine Accelerated SGD [11], Local-AC-CA [12] и FedAc [13]. Другими словами, для достижения ε точности решения задачи (2) необходимо проделать NK итераций с максимально допустимым уровнем шума Δ и общим числом вызова безградиентного оракула T в соответствии с выбранным методом и схемой сглаживания:*

- Minibatch Accelerated SGD

i) для l_1 -рандомизации (4):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right);$$

$$N = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)M_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

ii) для l_2 -рандомизации (5):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)dM_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)dM_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

• *Single-Machine Accelerated SGD*

i) для l_1 -рандомизации (4):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right); \quad B = 1;$$

$$T = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

ii) для l_2 -рандомизации (5):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

• *Local-AC-SA*

i) для l_1 -рандомизации (4):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = 1; \quad K = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad B = O\left(\frac{\kappa(p,d)M_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

ii) для l_2 -рандомизации (5):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = 1; \quad K = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad B = O\left(\frac{\kappa(p,d)M_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

• *Federated Accelerated SGD (FedAc)*

i) для l_1 -рандомизации (4):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = O\left(\frac{d^{1/6}(\kappa(p,d)M)^{1/3}M_2R^{4/3}}{K^{1/3}\varepsilon^{4/3}}\right); \quad K = O\left(\frac{\kappa(p,d)M_2^2R^2}{BN\varepsilon^2}\right); \quad B = 1;$$

$$T = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1; \end{cases}$$

ii) для l_2 -рандомизации (5):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = O\left(\frac{d^{1/2}(\kappa(p,d)M)^{1/3}M_2R^{4/3}}{K^{1/3}\varepsilon^{4/3}}\right); \quad K = O\left(\frac{\kappa(p,d)M_2^2R^2}{BN\varepsilon^2}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2, \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1. \end{cases}$$

Подробное доказательство теоремы 1 приведено в приложении В.

По результатам теоремы 1 стоит обратить внимание, что количество локальных обновлений $K > 1$ с использованием $B > 1$ компьютеров при оптимальных значениях N и T получилось только для одного алгоритма, а именно Local-AC-SA, который применяется для частного квадратичного случая. Это подтверждает, что в теории на данный момент существует только два оптимальных (по Δ , N и T) алгоритма: Minibatch и Single-Machine Accelerated SGD. В работе [13] показана

но, что на практике можно получить результат, в котором алгоритм совершает параллельно $K > 1$ локальных обновлений на каждом $B > 1$ компьютерах.

В теореме 2 также получены оптимальные оценки параметров, рассмотренных в п. 4.1 алгоритмов для решения стохастических негладких задач с седловой точкой, оптимальных по Δ , N и T .

Теорема 2. *Схема сглаживания из разд. 3, применяемая к седловой задаче (см. замечание 1), обеспечивает сходимость следующих двухточечных безградиентных алгоритмов: Minibatch SMP и Single-Machine SMP из приложения А. Другими словами, для достижения ϵ точности решения седловой задачи (см. замечание 1) необходимо проделать NK итераций с максимально допустимым уровнем шума Δ и общим числом вызова безградиентного оракула T в соответствии с выбранным методом и схемой сглаживания:*

• *Minibatch SMP*

i) для l_1 -рандомизации (4):

$$\begin{aligned}\Delta &= O\left(\frac{\epsilon^2}{\sqrt{d}M_2R}\right); \\ N &= 1; \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right); \\ T &= O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\epsilon^2}\right), & p = 2, \\ O\left(\frac{M_2^2R^2}{\epsilon^2}\right), & p = 1; \end{cases}\end{aligned}$$

ii) для l_2 -рандомизации (5):

$$\begin{aligned}\Delta &= O\left(\frac{\epsilon^2}{\sqrt{d}M_2R}\right); \\ N &= 1; \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right); \\ T &= \tilde{O}\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\epsilon^2}\right), & p = 2, \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\epsilon^2}\right), & p = 1; \end{cases}\end{aligned}$$

• *Single-Machine SMP*

i) для l_1 -рандомизации (4):

$$\begin{aligned}\Delta &= O\left(\frac{\epsilon^2}{\sqrt{d}M_2R}\right); \\ N &= 1; \quad K = O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right); \quad B = 1; \\ T &= O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\epsilon^2}\right), & p = 2, \\ O\left(\frac{M_2^2R^2}{\epsilon^2}\right), & p = 1; \end{cases}\end{aligned}$$

ii) для l_2 -рандомизации (5):

$$\Delta = O\left(\frac{\epsilon^2}{\sqrt{d}M_2R}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2, \\ \tilde{O}\left(\frac{(\ln d)M_2^2 R^2}{\varepsilon^2}\right), & p = 1. \end{cases}$$

Подробное доказательство теоремы 2 приведено в приложении С.

По результатам теоремы 1 и теоремы 2 не трудно увидеть, что для всех алгоритмов оптимальное число вызовов безградиентного оракула в l_1 -норме равняется $\tilde{O}\left(\frac{(\ln d)M_2^2 R^2}{\varepsilon^2}\right)$ с l_2 -рандомизацией, в то время как для l_1 -рандомизации равняется $O\left(\frac{M_2^2 R^2}{\varepsilon^2}\right)$, где ε является точностью решения негладкой задачи. Данный результат подтверждает, что схема сглаживания с l_1 -рандомизацией лучше работает в архитектуре федеративного обучения, чем схема сглаживания с l_2 -рандомизацией. В разд. 6 проверим этот результат на практическом эксперименте.

Замечание 2. Для получения результатов теорем 1 и 2 использовалось предположение о доступности двухточечной обратной связи (см. п. 3.3). Если вместо двухточечной обратной связи использовать односточечную (см. п. 3.4), то это приведет к такой же итерационной сложности (количеству коммуникационных раундов) N и максимальному уровню неточности Δ , однако оракульная сложность возрастет в $O(d / \varepsilon^2)$ раз. Данный случай подробно рассмотрен в приложении D.

Замечание 3. Стоит отметить, что идея объединения двух техник, а именно федеративного обучения со схемами сглаживания, не ограничивается рассмотренными алгоритмами первого порядка и может быть использована для решения негладких задач другими алгоритмами, используемыми в архитектуре федеративного обучения.

5. СХЕМЫ ДОКАЗАТЕЛЬСТВ

Приведем схемы для доказательств теоремы 1 и теоремы 2. Подробные доказательства этих теорем можно найти в приложении В, С. Здесь сфокусируемся на доказательстве леммы 1-3 и на нахождении оценки уровня неточности (враждебного шума).

5.1. Доказательство леммы 1

В данном п. рассмотрим неевклидовы случай, когда случайный вектор $\tilde{\epsilon}$ равномерно распределен на l_1 -шаре. Доказательство евклидового случая можно найти в теореме 2.1 из [1].

Для всех $x, y \in Q$ верно следующее:

$$1) f(x) \leq f_{\gamma}(x) \leq f(x) + \frac{2}{\sqrt{d}} \gamma M_2;$$

$$2) f_{\gamma} - M\text{-липшицева:}$$

$$|f_{\gamma}(y) - f_{\gamma}(x)| \leq M \|y - x\|_p,$$

$$3) f_{\gamma} \text{ имеет } L_{f_{\gamma}} = \frac{dM}{2\gamma}\text{-липшицевый градиент:}$$

$$\|\nabla f_{\gamma}(y) - \nabla f_{\gamma}(x)\|_q \leq L_{f_{\gamma}} \|y - x\|_2 \leq L_{f_{\gamma}} \|y - x\|_p,$$

где q такой, что $1/p + 1/q = 1$.

Доказательство. Для первого неравенства первого свойства воспользуемся выпуклостью функции $f(x)$

$$f_{\gamma}(x) = \mathbb{E}_{\tilde{e}} [f(x + \gamma\tilde{e})] \geq \mathbb{E}_{\tilde{e}} [f(x) + \langle \nabla f(x), \gamma\tilde{e} \rangle] = \mathbb{E}_{\tilde{e}} [f(x)] = f(x).$$

Для второго неравенства первого свойства, применяя лемму 1 из [4], имеем:

$$|f_{\gamma}(x) - f(x)| = |\mathbb{E}_{\tilde{e}} [f(x + \gamma\tilde{e})] - f(x)| \leq \mathbb{E}_{\tilde{e}} [|f(x + \gamma\tilde{e}) - f(x)|] \leq \gamma M_2 \mathbb{E}_{\tilde{e}} [\|\tilde{e}\|_2] \leq \frac{2}{\sqrt{d}} \gamma M_2,$$

используя тот факт, что f является M_2 -липшицевой.

Для второго свойства:

$$|f_{\gamma}(y) - f_{\gamma}(x)| = |\mathbb{E}_{\tilde{e}} [f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})]| \leq \mathbb{E}_{\tilde{e}} [|f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})|] \leq M \|y - x\|_p.$$

И для третьего свойства, применяя лемму 11 из [15] имеем для любых $x, y \in Q$,

$$\begin{aligned} \|\nabla f_{\gamma}(y) - \nabla f_{\gamma}(x)\|_q &= \left\| \int_{B_1^d(\gamma)} \nabla f(y + \tilde{e}) \mu(\tilde{e}) d\tilde{e} - \int_{B_1^d(\gamma)} \nabla f(x + \tilde{e}) \mu(\tilde{e}) d\tilde{e} \right\|_q \leq \\ &\leq \left\| \int_{Q_{\gamma}} \nabla f(z) \mu(z - y) dz - \int_{Q_{\gamma}} \nabla f(z) \mu(z - x) dz \right\|_q \leq M \underbrace{\int_{Q_{\gamma}} |\mu(z - y) - \mu(z - x)| dz}_{I_1}, \end{aligned}$$

где второе неравенство следует из $z = x + \tilde{e}$ и $\mu(x) = \begin{cases} \frac{1}{V(B_1^d(\gamma))}, & x \in B_1^d(\gamma), \\ 0 & \text{иначе.} \end{cases}$ Далее для оценки интеграла I_1 применим тот же подход, что и в доказательстве леммы 8 из [16] и рассмотрим случаи

где $\|y - x\|_1 > 2\gamma$ и $\|y - x\|_1 \leq 2\gamma$.

Случай 1 ($\|y - x\|_1 > 2\gamma$). Для всех $\tilde{e}$ пусть $\|\tilde{e} - x\|_1 \leq \gamma$, тогда имеем $\|\tilde{e} - y\|_1 > \gamma \Rightarrow \mu(\tilde{e} - y) = 0$, поэтому

$$\int_{\|\tilde{e} - x\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = 1.$$

Аналогично для всех $\tilde{e}$ пусть $\|\tilde{e} - y\|_1 \leq \gamma$, тогда имеем $\mu(\tilde{e} - x) = 0$, поэтому

$$\int_{\|\tilde{e} - y\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = 1.$$

Следовательно,

$$\int_{Q_{\gamma}} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = \int_{\|\tilde{e} - x\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} + \int_{\|\tilde{e} - y\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = 2.$$

Поскольку $2 < \frac{\|y - x\|_1}{\gamma} \leq \frac{d^{1-1/p} \|y - x\|_p}{\gamma}$, учитывая тот факт, что для $p \in [1, 2]$ выполнено

$\|y - x\|_1 \leq d^{\frac{1-1}{p}} \|y - x\|_p$, тогда получим следующее неравенство:

$$\int_{Q_{\gamma}} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} \leq \frac{d^{1-1/p} \|y - x\|_p}{\gamma}.$$

Случай 2 ($\|y - x\|_1 \leq 2\gamma$). Разложим интеграл I_1 на 4 интеграла

$$\begin{aligned} & \int_{Q_\gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = \\ &= \int_{\|\tilde{e}-x\|_1 \leq \gamma \ \& \ \|\tilde{e}-y\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} + \int_{\|\tilde{e}-x\|_1 \leq \gamma \ \& \ \|\tilde{e}-y\|_1 \geq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} + \\ &+ \int_{\|\tilde{e}-x\|_1 \geq \gamma \ \& \ \|\tilde{e}-y\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} + \int_{\|\tilde{e}-x\|_1 \geq \gamma \ \& \ \|\tilde{e}-y\|_1 \geq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e}. \end{aligned}$$

Теперь рассмотрим каждый интеграл из разложения по отдельности.

1. Для первого и четвертого интегралов справедливо следующее:

$$\mu(\tilde{e} - x) = \mu(\tilde{e} - y).$$

Тогда, подставляя в первый интеграл, получим

$$\int_{\|\tilde{e}-x\|_1 \leq \gamma \ \& \ \|\tilde{e}-y\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = 0. \quad (15)$$

Аналогично, подставляя в четвертый интеграл, получим

$$\int_{\|\tilde{e}-x\|_1 \geq \gamma \ \& \ \|\tilde{e}-y\|_1 \geq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = 0. \quad (16)$$

2. Когда $\|\tilde{e} - x\|_1 \leq \gamma$ & $\|\tilde{e} - y\|_1 \geq \gamma$ или $\|\tilde{e} - x\|_1 \geq \gamma$ & $\|\tilde{e} - y\|_1 \leq \gamma$, имеем

$$\int_{\|\tilde{e}-x\|_1 \leq \gamma \ \& \ \|\tilde{e}-y\|_1 \geq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = \int_{\|\tilde{e}-x\|_1 \geq \gamma \ \& \ \|\tilde{e}-y\|_1 \leq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e},$$

где $\mu(\tilde{e} - x)$ и $\mu(\tilde{e} - y)$ не пересекаются, поэтому, используя это и симметрию интегралов, и, определив множество $S = \{\tilde{e} \in \mathbb{R}^d \mid \|\tilde{e} - x\|_1 \leq \gamma \text{ и } \|\tilde{e} - y\|_1 \geq \gamma\}$, получим сумму второго и третьего интегралов

$$2 \int_{\|\tilde{e}-x\|_1 \leq \gamma \ \& \ \|\tilde{e}-y\|_1 \geq \gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = \frac{2}{c_d \gamma^d} V_S, \quad (17)$$

где V_S — объем на множестве S .

Суммируя значения четырех интегралов (15)–(17), получим

$$\int_{Q_\gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} = \frac{2}{c_d \gamma^d} V_S, \quad (18)$$

где V_S — объем на множестве S .

Теперь найдем верхнюю границу V_S через $\|y - x\|_p$. Пусть $V_{\text{cap}}(r)$ — объем сферического колпака с расстоянием r до центра l_1 -сферы. Тогда

$$V_S = c_d \gamma^d - 2V_{\text{cap}}\left(\frac{\|y - x\|_p}{4}\right). \quad (19)$$

Объем d -мерного сферического колпака V_{cap} может быть вычислен через $(d-1)$ -мерную l_1 -сферу следующим образом:

$$V_{\text{cap}}(r) = \int_r^\gamma c_{d-1} (\gamma - \rho)^{d-1} d\rho \quad \text{для} \quad r \in [0, \gamma],$$

где $c_d = \frac{2^d}{d!}$, $d \geq 1$. Для $r \in [0, \gamma]$ имеем

$$\begin{aligned} V'_{\text{cap}}(r) &= -c_{d-1}(\gamma - r)^{d-1} \leq 0, \\ V''_{\text{cap}}(r) &= (d-1)c_{d-1}(\gamma - r)^{d-2} \geq 0, \end{aligned}$$

где V'_{cap} , V''_{cap} — первая и вторая производные по r соответственно. Следовательно, V_{cap} выпукла на $[0, \gamma]$, и по определению субградиента имеем

$$V_{\text{cap}}(0) + V'_{\text{cap}}(0)r \leq V_{\text{cap}}(r) \quad \text{для } r \in [0, \gamma].$$

Так как $V_{\text{cap}}(0) = \frac{1}{2}c_d\gamma^d$ и $V'_{\text{cap}}(0) = -c_{d-1}\gamma^{d-1}$, отсюда следует, что

$$\frac{1}{2}c_d\gamma^d - c_{d-1}\gamma^{d-1}r \leq V_{\text{cap}}(r) \quad \text{для } r \in [0, \gamma]. \quad (20)$$

Поскольку $\|y - x\|_1/2 \leq \gamma$ и отмечая, что $\|y - x\|_p/4 \leq \|y - x\|_p/2 \leq \|y - x\|_1/2$ выполняется для $p \in [1, 2]$, можем задать $r = \|y - x\|_p/4 \leq \gamma$ и подставить в (20). Сделав это и используя (19), получим

$$V_S = c_d\gamma^d - 2V_{\text{cap}}\left(\frac{\|y - x\|_p}{4}\right) \leq 2c_{d-1}\gamma^{d-1} \frac{\|y - x\|_p}{4}.$$

Теперь, подставляя полученную оценку V_S в (18), имеем

$$\int_{Q_\gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} \leq \frac{c_{d-1}}{c_d} \frac{\|y - x\|_p}{\gamma}.$$

Поскольку $c_d = \frac{2^d}{d!}$, можно увидеть, что

$$\int_{Q_\gamma} |\mu(\tilde{e} - y) - \mu(\tilde{e} - x)| d\tilde{e} \leq \frac{d}{2\gamma} \|y - x\|_p.$$

Теперь, получив оценку для интеграла I_1 , имеем, что

$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq \frac{dM}{2\gamma} \|y - x\|_p.$$

5.2. Доказательство леммы 2

В данном подразделе сосредоточимся на доказательство леммы 2 для l_2 -рандомизации. Во многих работах, например в [1, 3, 9], оценка второго момента для безградиентной аппроксимации (5) приведена и доказана с точностью до некоторой числовой константы c . В данном доказательстве разберемся, чему же равна эта числовая константа c . А подробное доказательство леммы 2 для l_1 -рандомизации приведено в лемме 4 [4].

Для всех $x \in Q$ с предположениями 1 и 2 $\nabla f_\gamma(x, \xi, e)$ из (5) имеет нижнюю оценку (второй момент):

$$E_{\xi, e} \left[\|\nabla f_\gamma(x, \xi, e)\|_q^2 \right] \leq \kappa(p, d) \left(dM_2^2 + \frac{d^2 \Delta^2}{\sqrt{2}\gamma^2} \right),$$

где $1/p + 1/q = 1$ и

$$\kappa(p, d) = \sqrt{2} \min \{q, \ln d\} d^{2/q-1}.$$

Доказательство. Давайте рассмотрим

$$\begin{aligned} \mathbb{E}_{\xi, e} \left[\left\| \nabla f_{\gamma}(x, \xi, e) \right\|_q^2 \right] &= \mathbb{E}_{\xi, e} \left[\left\| \frac{d}{2\gamma} (f_{\delta}(x + \gamma e, \xi) - f_{\delta}(x - \gamma e, \xi)) e \right\|_q^2 \right] = \\ &= \frac{d^2}{4\gamma^2} \mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) + \delta(x + \gamma e) - f(x - \gamma e, \xi) - \delta(x - \gamma e))^2 \right] \leq \\ &\leq \frac{d^2}{2\gamma^2} \left(\mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) - f(x - \gamma e, \xi))^2 \right] + \mathbb{E}_e \left[\|e\|_q^2 (\delta(x + \gamma e) - \delta(x - \gamma e))^2 \right] \right), \end{aligned} \quad (21)$$

где использовался тот факт, что для всех $a, b, (a + b)^2 \leq 2a^2 + 2b^2$. Для первого слагаемого (21) выполняется следующее с произвольным параметром α с учетом симметричного распределения e

$$\begin{aligned} &\frac{d^2}{2\gamma^2} \mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) - f(x - \gamma e, \xi))^2 \right] = \\ &= \frac{d^2}{2\gamma^2} \mathbb{E}_{\xi, e} \left[\|e\|_q^2 ((f(x + \gamma e, \xi) - \alpha) - (f(x - \gamma e, \xi) - \alpha))^2 \right] \leq \\ &\leq \frac{d^2}{\gamma^2} \mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) - \alpha)^2 + (f(x - \gamma e, \xi) - \alpha)^2 \right] = \\ &= \frac{d^2}{\gamma^2} \left(\mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) - \alpha)^2 \right] + \mathbb{E}_{\xi, e} \left[(f(x - \gamma e, \xi) - \alpha)^2 \right] \right) = \\ &= \frac{2d^2}{\gamma^2} \mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) - \alpha)^2 \right]. \end{aligned} \quad (22)$$

Применяя неравенство Коши–Шварца для (22) и используя $\sqrt{\mathbb{E}[\|e\|_q^4]} \leq \kappa'(p, d)$, где $\kappa'(p, d) = \min\{q, \ln d\} d^{2/q-1}$, получим

$$\begin{aligned} \frac{2d^2}{\gamma^2} \mathbb{E}_{\xi, e} \left[\|e\|_q^2 (f(x + \gamma e, \xi) - \alpha)^2 \right] &\leq \frac{2d^2}{\gamma^2} \mathbb{E}_{\xi} \left[\sqrt{\mathbb{E}[\|e\|_q^4]} \sqrt{\mathbb{E}_e[(f(x + \gamma e, \xi) - \alpha)^4]} \right] \leq \\ &\leq \frac{2d^2 \kappa'(p, d)}{\gamma^2} \mathbb{E}_{\xi} \left[\sqrt{\mathbb{E}_e[(f(x + \gamma e, \xi) - \alpha)^4]} \right]. \end{aligned} \quad (23)$$

Следующая лемма, которую рассмотрим, является уточнением леммы 9 [3] с указанием числовой константы c .

Лемма 4. Для любой функции $f(e)$, которая является M -липшицевой относительно l_2 -нормы, имеет место, что если e равномерно распределен на евклидовой сфере, тогда

$$\sqrt{\mathbb{E}[(f(e) - \mathbb{E}[f(e)])^4]} \leq \frac{M_2^2}{\sqrt{2d}}.$$

Доказательство. Пусть неравенство концентрации меры, использующее математическое ожидание, имеет вид

$$\Pr(|f - \mathbb{E}f| > t) \leq K \exp(-\eta t^2),$$

где K и η — неизвестные параметры. Тогда, чтобы найти параметры K и η , выпишем неравенство, использующую медиану функции, используя параметры K и η (так как в литературе, например в [10], обычно записывается неравенство концентрации меры на сфере, используя медиану функции M_f , а не математическое ожидание):

$$\Pr(|f - M_f| > t) \leq 2K \exp(-\eta t^2/4) = 4 \exp(-dt^2/2M_2^2). \quad (24)$$

Подставляя параметры $K = 2$ и $\eta = 2d/M_2^2$ из неравенства (24), запишем стандартный результат о концентрации липшицевых функций на евклидовой единичной сфере

$$\Pr(|f(e) - \mathbb{E}[f(e)]| > t) \leq 2 \exp(-2dt^2/M_2^2).$$

Следовательно,

$$\begin{aligned} \sqrt{\mathbb{E}[(f(e) - \mathbb{E}[f(e)])^4]} &= \sqrt{\int_{t=0}^{\infty} \Pr((f(e) - \mathbb{E}[f(e)])^4 > t) dt} = \\ &= \sqrt{\int_{t=0}^{\infty} \Pr(|f(e) - \mathbb{E}[f(e)]| > \sqrt[4]{t}) dt} \leq \sqrt{\int_{t=0}^{\infty} 2 \exp\left(-\frac{2d\sqrt[4]{t}}{M_2^2}\right) dt} = \sqrt{2 \frac{M_2^4}{(2d)^2}}, \end{aligned}$$

где в последнем шаге использовался факт, что $\int_{t=0}^{\infty} \exp(-\sqrt{x}) dx = 2$. Таким образом, найдена числовая константа $c = \frac{1}{\sqrt{2}}$ из леммы 9 [3].

Затем используем лемму 4 вместе с тем фактом, что $f(x + \gamma e, \xi)$ является $\gamma M_2(\xi)$ -липшицевой относительно e с точки зрения l_2 -нормы. Таким образом, для (23) и $\alpha := \mathbb{E}[f(x + \gamma e, \xi)]$ выполняется

$$\frac{2d^2 \kappa'(p, d)}{\gamma^2} \mathbb{E}_{\xi} \left[\sqrt{\mathbb{E}_e \left[(f(x + \gamma e, \xi) - \alpha)^4 \right]} \right] \leq \frac{2d^2 \kappa'(p, d)}{\gamma^2} \cdot \frac{\gamma^2 \mathbb{E} [M_2^2(\xi)]}{\sqrt{2d}} = \sqrt{2} \kappa'(p, d) d M_2^2. \quad (25)$$

Для второго слагаемого (21) выполняется следующее:

$$\frac{d^2}{2\gamma^2} \mathbb{E}_e \left[\|e\|_q^2 (\delta(x + \gamma e) - \delta(x - \gamma e))^2 \right] \leq \frac{d^2 \Delta^2}{\gamma^2} \mathbb{E}_e \left[\|e\|_q^2 \right] \leq \frac{\kappa'(p, d) d^2 \Delta^2}{\gamma^2}. \quad (26)$$

Подставляя (25) и (26) в неравенство (21) и занося коэффициент $\sqrt{2}$ в $\kappa'(p, d)$, получим утверждение леммы 2 для l_2 -рандомизации с $\kappa(p, d) = \sqrt{2} \kappa'(p, d) = \sqrt{2} \min\{q, \ln d\} d^{2/q-1}$.

5.3. Доказательство леммы 3

Рассмотрим краткое доказательство леммы 3 для случая с l_1 -рандомизацией. С евклидовым случаем можно познакомиться в следующих работах [1, 17].

Для всех $x \in Q$ с предположениями 2 и 3 $\nabla f_{\gamma}(x, \xi, e)$ из (8) имеет нижнюю оценку (второй момент):

$$\mathbb{E}_e \left[\|\nabla f_{\gamma}(x, \xi, e)\|_q^2 \right] \leq \frac{d^{4-\frac{2}{p}}}{\gamma^2} (G^2 + \Delta^2),$$

где $1/p + 1/q = 1$.

Доказательство. Доказательство этой леммы будет базироваться на доказательстве леммы 4 из [4]. Используя определение $\nabla f_{\gamma}(x, \xi, e)$, получим

$$\mathbb{E} \left[\|\nabla f_{\gamma}(x, \xi, e)\|_q^2 \right] = \frac{d^2}{\gamma^2} \mathbb{E} \left[(f(x + \gamma e) + \delta(x))^2 \|\text{sign}(e)\|_q^2 \right] = \frac{d^{4-\frac{2}{p}}}{\gamma^2} \mathbb{E} \left[(f(x + \gamma e) + \delta(x))^2 \right].$$

Далее с предположениями 2 и 3 получим утверждение леммы:

$$E \left[\|\nabla f_{\gamma}(x, \xi, e)\|_q^2 \right] = \frac{d^{4-\frac{2}{p}}}{\gamma^2} E \left[(f(x + \gamma e) + \delta(x))^2 \right] \leq \frac{d^{4-\frac{2}{p}}}{\gamma^2} (G^2 + \Delta^2).$$

5.4. Оценки уровня неточности

Приведем необходимые леммы и следствия для нахождения двух оценок уровня неточности (враждебного шума): для l_1 -рандомизация (4) и для l_2 -рандомизация (5). Для этого воспользуемся известными результатами и теми же суждениями, что и в [18].

Лемма 5 (см. [4]). Функция $f_\gamma(x)$ дифференцируема со следующим градиентом с l_1 -рандомизацией:

$$\nabla f_\gamma(x) = \mathbb{E}_e \left[\frac{d}{\gamma} f(x + \gamma e) \text{sign}(e) \right].$$

Лемма 6 (см. [19]). Функция $f_\gamma(x)$ дифференцируема со следующим градиентом с l_2 -рандомизацией:

$$\nabla f_\gamma(x) = \mathbb{E}_e \left[\frac{d}{\gamma} f(x + \gamma e) e \right].$$

Лемма 7 (см. [20]). Пусть вектор e — случайный единичный вектор из евклидовой единичной сферы $\{e : \|e\|_2 = 1\}$. Тогда для всех $r \in \mathbb{R}^d$ следует

$$\mathbb{E}[\langle e, r \rangle] \leq \frac{\|r\|_2}{\sqrt{d}}.$$

Лемма 8. Для $\nabla f_\gamma(x, \xi, e)$ и $\nabla f_\gamma(x)$ с предположением 2, выполняется следующее:

$$\mathbb{E}_{\xi, e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] \geq \langle \nabla f_\gamma(x), r \rangle - \frac{d \Delta \mathbb{E}_e[\langle \text{sign}(e), r \rangle]}{\gamma},$$

где ∇f_γ с l_1 -рандомизацией;

$$\mathbb{E}_{\xi, e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] \geq \langle \nabla f_\gamma(x), r \rangle - \frac{d \Delta \mathbb{E}_e[\langle e, r \rangle]}{\gamma},$$

где ∇f_γ с l_2 -рандомизацией.

Доказательство. Рассмотрим следующие утверждения.

i) для l_1 -рандомизации (4):

$$\begin{aligned} \nabla f_\gamma(x, \xi, e) &= \frac{d}{2\gamma} (f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi)) \text{sign}(e) = \\ &= \frac{d}{2\gamma} (f(x + \gamma e, \xi) + \delta(x + \gamma e) - f(x - \gamma e, \xi) - \delta(x - \gamma e)) \text{sign}(e) = \\ &= \frac{d}{2\gamma} ((f(x + \gamma e, \xi) - f(x - \gamma e, \xi)) \text{sign}(e) + (\delta(x + \gamma e) - \delta(x - \gamma e)) \text{sign}(e)). \end{aligned}$$

Из данного равенства следует

$$\begin{aligned} \mathbb{E}_{\xi, e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] &= \frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle (f(x + \gamma e, \xi) - f(x - \gamma e, \xi)) \text{sign}(e), r \rangle] + \\ &+ \frac{d}{2\gamma} \mathbb{E}_e[\langle (\delta(x + \gamma e) - \delta(x - \gamma e)) \text{sign}(e), r \rangle]. \end{aligned} \quad (27)$$

Применяя лемму 5 к первому слагаемому (27), получим

$$\begin{aligned} &\frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle (f(x + \gamma e, \xi) - f(x - \gamma e, \xi)) \text{sign}(e), r \rangle] = \\ &= \frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle f(x + \gamma e, \xi) \text{sign}(e), r \rangle] + \frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle f(x - \gamma e, \xi) \text{sign}(e), r \rangle] = \\ &= \frac{d}{\gamma} \mathbb{E}_e[\langle \mathbb{E}_\xi[f(x + \gamma e, \xi)] \text{sign}(e), r \rangle] = \frac{d}{\gamma} \mathbb{E}_e[\langle f(x + \gamma e) \text{sign}(e), r \rangle] = \langle \nabla f_\gamma(x), r \rangle. \end{aligned} \quad (28)$$

Для второго слагаемого (27) с учетом $|\delta(x)| \leq \Delta$ получим

$$\frac{d}{\gamma} \mathbb{E}_e[\langle (\delta(x + \gamma e) - \delta(x - \gamma e)) \text{sign}(e), r \rangle] \geq -\frac{d}{2\gamma} 2\Delta \mathbb{E}_e[|\langle \text{sign}(e), r \rangle|] = -\frac{d}{\gamma} \Delta \mathbb{E}_e[|\langle \text{sign}(e), r \rangle|]. \quad (29)$$

Используя уравнения (28) и (29) для уравнения (27), получим утверждение леммы для l_1 -рандомизации.

ii) для l_2 -рандомизации (5):

$$\begin{aligned} \nabla f_\gamma(x, \xi, e) &= \frac{d}{2\gamma} (f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi))e = \\ &= \frac{d}{2\gamma} (f(x + \gamma e, \xi) + \delta(x + \gamma e) - f(x - \gamma e, \xi) - \delta(x - \gamma e))e = \\ &= \frac{d}{2\gamma} ((f(x + \gamma e, \xi) - f(x - \gamma e, \xi))e + (\delta(x + \gamma e) - \delta(x - \gamma e))e). \end{aligned}$$

Из данного равенства следует

$$\begin{aligned} E_{\xi, e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] &= \frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle (f(x + \gamma e, \xi) - f(x - \gamma e, \xi))e, r \rangle] + \\ &+ \frac{d}{2\gamma} \mathbb{E}_e[\langle (\delta(x + \gamma e) - \delta(x - \gamma e))e, r \rangle]. \end{aligned} \quad (30)$$

Применяя лемму 6 к первому слагаемому (30), получим

$$\begin{aligned} &\frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle (f(x + \gamma e, \xi) - f(x - \gamma e, \xi))e, r \rangle] = \\ &= \frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle f(x + \gamma e, \xi)e, r \rangle] + \frac{d}{2\gamma} \mathbb{E}_{\xi, e}[\langle f(x - \gamma e, \xi)e, r \rangle] = \\ &= \frac{d}{\gamma} \mathbb{E}_e[\langle \mathbb{E}_\xi[f(x + \gamma e, \xi)]e, r \rangle] = \frac{d}{\gamma} \mathbb{E}_e[\langle f(x + \gamma e)e, r \rangle] = \langle \nabla f_\gamma(x), r \rangle. \end{aligned} \quad (31)$$

Для второго слагаемого (30) с учетом $|\delta(x)| \leq \Delta$ получим

$$\frac{d}{\gamma} \mathbb{E}_e[\langle (\delta(x + \gamma e) - \delta(x - \gamma e))e, r \rangle] \geq -\frac{d}{2\gamma} 2\Delta \mathbb{E}_e[|\langle e, r \rangle|] = -\frac{d}{\gamma} \Delta \mathbb{E}_e[|\langle e, r \rangle|]. \quad (32)$$

Используя уравнения (31) и (32), для уравнения (30) получим утверждение леммы для l_2 -рандомизации.

Следствие 5. Заметим, что вектор $\text{sign}(e) = (\text{sign}(e_1), \dots, \text{sign}(e_d))$ в l_1 -норме при большой размерности пространства d ведет себя схожим образом, как вектор из независимых радемахеровских случайных величин. Тогда из неравенства Хинчина следует $E[|\langle \text{sign}(e), r \rangle|] \leq \|r\|_2$, где $r \in \mathbb{R}^d$. Применяя данное неравенство к лемме 8, получим

$$\mathbb{E}_e[\langle \nabla f_\gamma(x, e) - \nabla f_\gamma(x), r \rangle] \leq \frac{d\Delta \|r\|_2}{\gamma}.$$

Следствие 6. Применяя утверждение леммы 7 к лемме 8, получим то же неравенство, что и в работе [18] для всех $r \in \mathbb{R}^d$

$$\mathbb{E}_e[\langle \nabla f_\gamma(x, e) - \nabla f_\gamma(x), r \rangle] \leq \frac{\sqrt{d}\Delta \|r\|_2}{\gamma}.$$

Теперь рассмотрим, как получить оценки уровня неточности (шума).

Из (13) и (14), подставляя значение параметра γ из следствия 1 ($\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ для l_1 -рандомизации, $\gamma = \frac{\varepsilon}{2M_2}$ для l_2 -рандомизации) и $R = \|x^0 - x_*\|_2$, получим ограничительные условия на уровень неточности (шума) для l_1 и l_2 -рандомизаций:

Схема сглаживания с l_1 -рандомизацией	Схема сглаживания с l_2 -рандомизацией
$\Delta \leq \frac{\varepsilon^2}{\sqrt{d}M_2R}$	$\Delta \leq \frac{\varepsilon^2}{\sqrt{d}M_2R}$

Замечание 4. Стоит отметить, что результаты леммы 8 и следствий 5 и 6 для седловых задач такие же и доказываются аналогично. Следовательно, оценки уровня неточности для задачи с седловой точкой совпадают с оценками для выпуклой оптимизации. С целью избежания повторений будем ссылаться на это замечание как на результат получения оценок уровня неточности для седловых задач.

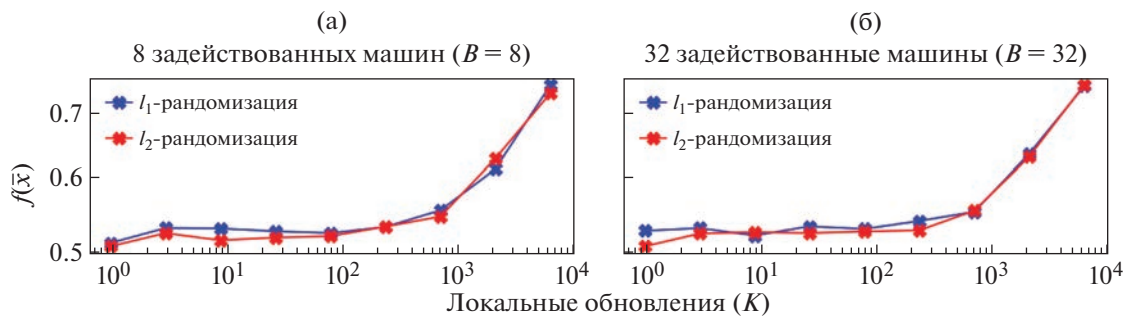
6. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Приведем численное сравнение двух техник сглаживания в архитектуре федеративного обучения. Рассматриваем стохастическую негладкую задачу оптимизации на множестве симплекса $Q = \{x \in \mathbb{R}^d : \|x\|_1 = 1, x \geq 0\}$ с функцией $f : \mathbb{R}^d \rightarrow \mathbb{R}$, определенную следующим образом:

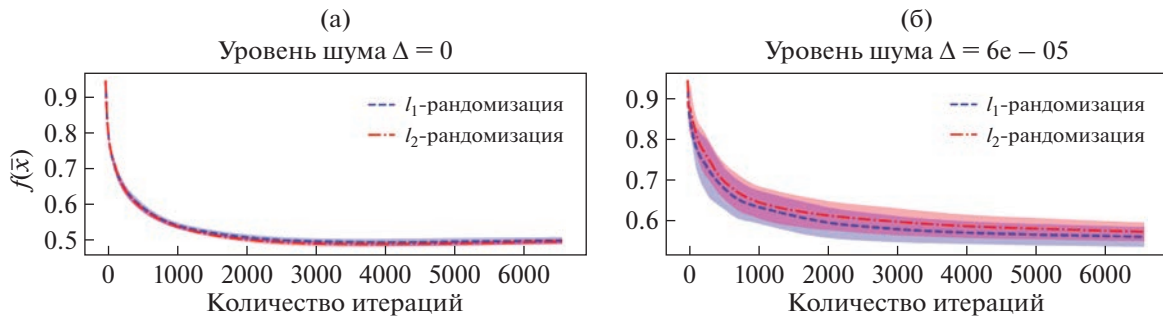
$$f(x) = \langle b, x \rangle + \|x\|_\infty,$$

где $b \in \mathbb{R}^d$ — случайный вектор, равномерно распределенный на отрезке $[0, 1]$. В качестве метода оптимизации используется безградиентный двухточечный алгоритм Minibatch Accelerated SGD, рассмотренный в разд. 4. На фиг. 2 представлена зависимость ошибки $f(\bar{x})$ на последней итерации от изменения количества локальных вызовов безградиентного оракула $K = 3^0, \dots, 3^8$ при разном количестве машин $B = 8$ и $B = 32$, работающих параллельно, где $\bar{x} = \frac{1}{t} \sum_{i=1}^t x_i$, а t — номер итерации. Количество итераций было выбрано $KN = 3^8$. Чем больше K , тем меньше количество коммуникационных раундов N , обратное также верно. Уровень неточности $\Delta = 0$ (без враждебного шума), размерность задачи $d = 100$.

Не трудно заметить из фиг. 2, что при увеличении числа локальных обновлений K (уменьшении раундов связи N), точность ухудшается (тем самым подтверждая теоретические результаты), однако не критично. То есть при решении практических задач, несмотря на теорию, можно брать число локальных обновлений $K \leq 3^6$, чтобы получить достаточно хороший результат. Также стоит обратить внимание, что на практике схема сглаживания с l_2 -рандомизацией не только не уступает схеме сглаживания с l_1 -рандомизацией, но иногда и превосходит ее. Однако возникает интерес узнать, в каких случаях схема сглаживания с l_1 -рандомизацией будет превосходить



Фиг. 2. Зависимость ошибки от числа машин B и различных локальных обновлений K .



Фиг. 3. Зависимость ошибки от количества итераций при различных значениях уровня неточности при 20 запусках.

схему сглаживания с l_2 -рандомизацией. Для этого рассмотрим два случая вычисления безградиентного оракула: с враждебным шумом и без враждебного шума. На фиг. 3 представлена зависимость ошибки $f(\bar{x})$ от количества итераций и уровня неточности (наличия враждебного шума) $\Delta = 0$ и $\Delta = 6 \times 10^{-5}$. Число работающих параллельно машин выбрано $B = 8$. На каждой машине количество локальных вызовов безградиентного оракула равняется $K = 3^2$, а количество коммуникационных раундов 3^6 . Таким образом, общее число итераций $NK = 3^8$. Размерность задачи $d = 100$, количество запусков равняется 20.

По фиг. 3 можно сделать вывод, что при добавления враждебного шума схема сглаживания с l_1 рандомизацией работает лучше, чем схема сглаживания с l_2 -рандомизацией в архитектуре федеративного обучения.

7. ЗАКЛЮЧЕНИЕ

В данной работе были получены верхние оценки оптимального алгоритма решения седловых задач в архитектуре федеративного обучения и нашли константу Липшица для градиента в схеме сглаживания с l_1 -рандомизацией. Применяя сглаживающие схемы, создали оптимальные безградиентные двухточечные и одноточечные алгоритмы с l_1 и l_2 -рандомизацией, благодаря которым можно решать стохастические негладкие выпуклые задачи оптимизации и выпукло-вогнутые задачи оптимизации в настройке федеративного обучения с неточным безградиентным оракулом. Показали, при каких условиях схема сглаживания с l_1 -рандомизацией работает лучше, чем с l_2 -рандомизацией в архитектуре федеративного обучения. Было показано на практике, что число локальных обновлений может быть увеличено за счет уменьшения количества коммуникационных раундов связи, при этом общее количество итераций остается прежним.

ПРИЛОЖЕНИЕ

А. ВЕРХНИЕ ОЦЕНКИ

Рассмотрим гладкую выпукло-вогнутую седловую задачу

$$\min_{x \in Q_x \subseteq \mathbb{R}^{d_x}} \max_{y \in Q_y \subseteq \mathbb{R}^{d_y}} f(x, y), \quad (\text{П.33})$$

где $f : Q_x \times Q_y \rightarrow \mathbb{R}$ — выпукло-вогнутая липшицева непрерывная функция, Q_x и Q_y — выпуклые множества. Для простоты изложения введем множество $Q_z = Q_x \times Q_y$, $z = (x, y)$ и монотонный оператор F :

$$F(z) = F(x, y) = \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix}. \quad (\text{П.34})$$

Тогда $z_* \in Q_z$ — решение вариационного неравенства (ВН) выглядит следующим образом:

$$\langle F(z), z_* - z \rangle \leq 0 \quad \forall z \in Q_z. \quad (\text{П.35})$$

Оценивать неточность возможного решения $z \in Q_z$ будем по ошибке

$$\text{Err}_{v_i}(z) = \max_{u \in Q_z} \langle F(u), z - u \rangle. \quad (\text{П.36})$$

В дальнейшем накладываем на F , помимо монотонности, требование

$$\forall (z, z' \in Q_z) : \|F(z) - F(z')\|_* \leq L \|z - z'\| + V, \quad (\text{П.37})$$

где $L \geq 0$, $V \geq 0$ — известные константы, $\|\cdot\|_* = \max_{z: \|z\| \leq 1} \langle \cdot, z \rangle$. И предполагаем, что $F(z, \xi_t)$ является несмещенным и имеет ограниченную дисперсию, т.е. $\forall z \in Q_z$ справедливо

$$\mathbb{E}[F(z, \xi_t)] = F(z), \quad \mathbb{E}[\|F(z, \xi_t) - F(z)\|_*^2] \leq \sigma^2.$$

Для распространения основного результата статьи [11] на седловые задачи и ВН в качестве алгоритма выберем Stochastic Mirror Prox (SMP), описанного в работе [21]. В дальнейшем рассмотрим два метода архитектуры FL, предложенные в [11] и назовем их Minibatch SMP и Single-Machine SMP. Для этого воспользуемся известными результатами и предположениями.

Algorithm 1. Алгоритм SMP

1: Инициализация. Выберите $r_0 \in Z^0$ и размер шага η_t , $1 \leq \tau \leq t$.

2: Шаг τ , $\tau = 1, 2, \dots, t$: По известному $r_{\tau-1} \in Z^0$, вычислите

$$w_\tau = P_{r_{\tau-1}}(\eta_\tau \hat{F}(r_{\tau-1})), \quad r_\tau = P_{r_{\tau-1}}(\eta_\tau \hat{F}(w_\tau)). \quad (\text{П.38})$$

Когда $\tau < t$, выполнять цикл до шага $t + 1$.

3: На шаге t выведите

$$\hat{z}_t = \left[\sum_{\tau=1}^t \eta_\tau \right]^{-1} \sum_{\tau=1}^t \eta_\tau w_\tau.$$

Предположение 4. Для всех $z \in Q_z$ с $\mu \in [0, \infty)$ имеем

$$\| \mathbb{E}[F(z, \xi_t) - F(z)] \|_* \leq \mu, \quad \mathbb{E}[\|F(z, \xi_t) - F(z)\|_*^2] \leq \sigma^2.$$

Предположение 5. Для всех $z \in Q_z$ и всех t имеем

$$\mathbb{E} \left[\exp \left\{ \|F(z, \xi_t) - F(z)\|_*^2 / \sigma^2 \right\} \right] \leq \exp \{1\}.$$

Лемма 9 (см. [21]). Пусть ВН (П.35) с монотонным оператором F (П.34), удовлетворяющим требованию (П.37) решается с помощью t шагового алгоритма 8, используя стохастический оракул ($\hat{F} = F(z, \xi_t)$), и пусть размеры шага $\eta_t \equiv \eta$ удовлетворяют $0 \leq \eta \leq \frac{1}{\sqrt{3}L}$. Тогда

(i) с предположением 4 имеем

$$\mathbb{E} \{ \text{Err}_{v_i}(\hat{z}_t) \} \leq K_0(t) \equiv \left[\frac{R^2}{t\eta} + \frac{7\eta}{2} [V^2 + 2\sigma^2] \right] + 2\mu R;$$

(ii) с предположениями 4, 5 для любого $\Lambda > 0$

$$\text{Prob} \{ \text{Err}_{v_i}(\hat{z}_t) > K_0(t) + \Lambda K_1(t) \} \leq \exp \{ -\Lambda^2 / 3 \} + \exp \{ -\Lambda t \},$$

где

$$K_1(t) = \frac{7\sigma^2\eta}{2} + \frac{2\sigma R}{\sqrt{t}}.$$

Следствие 7. Используя результаты леммы 9, распространим идею оптимального алгоритма [11] и получим верхние оценки.

• Minibatch SMP

Данный алгоритм выполняет N итераций SMP, используя пробатченные градиенты размера BK . Во время каждого раунда связи каждая машина вычисляет K стохастического оракула, затем машины посылают свои минибатчи, усредняя в один большой минибатч размером BK , затем они обновляют w_t и r_t в соответствии с (П.38). Так как вычисление пробатченного стохастического оракула уменьшает дисперсию в BK раз, то обозначим $\sigma_{BK}^2 = \frac{\sigma^2}{BK}$. Предположим, что размер шага $\eta_t \equiv \eta$ алгоритма 8 равен

$$\eta_t = \min \left[\frac{1}{\sqrt{3}L}, 7R \sqrt{\frac{2BK}{7N(V^2 + 2\sigma^2)}} \right], \quad (\text{П.39})$$

где N — заданное число итераций. Тогда с предположением 4

$$\mathbb{E}\{\text{Err}_{vi}(\hat{z}_N)\} \leq K_0^*(N) \equiv \max \left[\frac{7LR^2}{4N}, 7R \sqrt{\frac{V^2 + 2\sigma^2}{3BKN}} \right] + 2\mu R. \quad (\text{П.40})$$

Так как верхняя граница (П.40) для ошибки алгоритма 8 со стратегией шага (П.39) зависит схожим образом от σ и V , то V можем заменить на σ . Тогда

$$\mathbb{E}\{\text{Err}_{vi}(\hat{z}_N)\} \leq K_0^*(N) \equiv \max \left[\frac{7LR^2}{4N}, 7 \frac{\sigma R}{\sqrt{BKN}} \right] + 2\mu R,$$

если выполняются предположения 4 и 5, то

$$\text{Prob}\{\text{Err}_{vi}(\hat{z}_N) > K_0^*(N) + \Lambda K_1^*(N)\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda N\},$$

где

$$K_1^*(N) = \frac{7\sigma_{BK}R}{2\sqrt{N}} = \frac{7}{2} \frac{\sigma R}{\sqrt{BKN}}.$$

• Single-Machine SMP

Данный алгоритм в отличие от Mini-batch SMP игнорирует $B-1$ машин и выполняет KN шагов алгоритма SMP. Тогда предположим, что размер шага η_t алгоритма SMP равен

$$\eta_t = \min \left[\frac{1}{\sqrt{3}L}, 7R \sqrt{\frac{2}{7KN(V^2 + 2\sigma^2)}} \right], \quad (\text{П.41})$$

где KN — заданное число итераций. Тогда с предположением 4 имеем

$$\mathbb{E}\{\text{Err}_{vi}(\hat{z}_{KN})\} \leq K_0^*(KN) \equiv \max \left[\frac{7LR^2}{4KN}, 7R \sqrt{\frac{V^2 + 2\sigma^2}{3KN}} \right] + 2\mu R. \quad (\text{П.42})$$

Так как верхняя граница (П.42) для ошибки алгоритма 8 со стратегией шага (П.41) зависит схожим образом от σ и V , то V можем заменить на σ . Тогда

$$\mathbb{E}\{\text{Err}_{vi}(\hat{z}_{KN})\} \leq K_0^*(KN) \equiv \max \left[\frac{7LR^2}{4KN}, 7 \frac{\sigma R}{\sqrt{KN}} \right] + 2\mu R,$$

если выполняются предположения 4 и 5, то

$$\text{Prob}\{\text{Err}_{vi}(\hat{z}_{KN}) > K_0^*(KN) + \Lambda K_1^*(KN)\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda KN\},$$

где

$$K_1^*(KN) = \frac{7}{2} \frac{\sigma R}{\sqrt{KN}}.$$

Запишем лемму, используя такой же подход, что и Woodworth [11], чтобы объединить два алгоритма (Mini-batch SMP и Single-Machine SMP) в один оптимальный алгоритм.

Следствие 8. Для любых $L, R, \sigma, K, N, M > 0$ алгоритм использует Minibatch SMP, когда $K \leq \frac{\sigma^2 N}{L^2 R^2}$, и использует Single-Machine SMP, когда $K > \frac{\sigma^2 N}{L^2 R^2}$, тогда с предположением 4 имеем

$$\mathbb{E}\{\text{Err}_{v_i}(\hat{z})\} \leq c \left(\frac{LR^2}{KN} + \frac{\sigma R}{\sqrt{BKN}} + \min \left[\frac{LR^2}{N}, \frac{\sigma R}{\sqrt{KN}} \right] \right) + 2\mu R,$$

где c — некоторая числовая константа.

В. ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 1

Приведем полное доказательство теоремы 1. Для этого разобьем доказательство на две части: доказательство для l_1 -рандомизации и доказательство для l_2 -рандомизации.

Для l_1 -рандомизации имеем следующие алгоритмы:

• Minibatch Accelerated SGD

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{d\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$ нужно

$$\frac{d\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.43})$$

$$\frac{4L_{f_\gamma}R^2}{N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.44})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.45})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из следствия 3) в неравенства (П.43)–(П.45), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6dR} \Rightarrow \Delta \leq \frac{\varepsilon^2}{24\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2 \geq \frac{96\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow N \geq \frac{4\sqrt{6}d^{1/4}\sqrt{MM_2}R}{\varepsilon} \Rightarrow N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right)$$

количество коммуникационных раундов,

$$B \geq \frac{576\sigma^2R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{1152\kappa(p, d)M_2^2R^2}{KN\varepsilon^2} \Rightarrow B = O\left(\frac{\kappa(p, d)M_2^2R^2}{KN\varepsilon^2}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{1152\kappa(p,d)M_2^2R^2}{\epsilon^2} \Rightarrow$$

$$\Rightarrow T = O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right) \text{ следующее: } = \begin{cases} O\left(\frac{dM_2^2R^2}{\epsilon^2}\right), & p = 2 \quad (q = 2), \\ O\left(\frac{M_2^2R^2}{\epsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• **Single-Machine Accelerated SGD**

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{NK+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{N^2K^2} + \frac{4\sigma R}{\sqrt{NK}} + \frac{d\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\epsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\epsilon}{4M_2}$ (из следствия 1), то имеем ϵ -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Чтобы была $\frac{\epsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{d\Delta R}{\gamma} \leq \frac{\epsilon}{6}, \quad (\text{П.46})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\epsilon}{6}, \quad (\text{П.47})$$

$$\frac{4\sigma R}{\sqrt{KN}} \leq \frac{\epsilon}{6}. \quad (\text{П.48})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\epsilon}{4M_2}$ и $\sigma^2 = 2\kappa(p,d)M_2^2$ (из следствия 3), в неравенства (П.46)–(П.48), получим

$$\Delta \leq \frac{\gamma\epsilon}{6dR} \Rightarrow \Delta \leq \frac{\epsilon^2}{24\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\epsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{576\sigma^2R^2}{\epsilon^2} \Rightarrow NK \geq \frac{1152\kappa(p,d)M_2^2R^2}{\epsilon^2}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{\kappa(p,d)M_2^2R^2}{\epsilon^2}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{1152\kappa(p,d)M_2^2R^2}{\epsilon^2} \Rightarrow$$

$$\Rightarrow T = O\left(\frac{\kappa(p,d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• **Local-AC-SA**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [12, 22]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{K^2N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{d\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{d\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.49})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.50})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.51})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = 2\kappa(p,d)M_2^2$ (из следствия 3), в неравенства (П.49)–(П.51), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6dR} \Rightarrow \Delta \leq \frac{\varepsilon^2}{24\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2K^2 \geq \frac{96\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{4\sqrt{6}d^{1/4}\sqrt{MM_2}R}{\varepsilon}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right)$$

количество локальных вызовов безградиентного оракула,

$$B \geq \frac{576\sigma^2R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{1152\kappa(p,d)M_2^2R^2}{KN\varepsilon^2} \Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2R^2}{KN\varepsilon^2}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{1152\kappa(p,d)M_2^2R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = O\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ O\left(\frac{M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• **Federated Accelerated SGD (FedAc)**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [13]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{L_{f_\gamma} R^2}{KN^2} + \frac{\sigma R}{\sqrt{BKN}} + \min\left\{\frac{L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} N}, \frac{L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{K^{1/4} N}\right\} + \frac{d\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d\varepsilon}}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}} \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$ нужно

$$\frac{d\Delta R}{\gamma} \leq \frac{\varepsilon}{8}, \quad (\text{П.52})$$

$$\frac{L_{f_\gamma} R^2}{KN^2} \leq \frac{\varepsilon}{8}, \quad (\text{П.53})$$

$$\frac{\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{8}, \quad (\text{П.54})$$

$$\min\left\{\frac{L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} N}, \frac{L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{K^{1/4} N}\right\} \leq \frac{\varepsilon}{8}. \quad (\text{П.55})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d\varepsilon}}{4M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из следствия 3), в неравенства (П.52)–(П.55), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6dR} \Rightarrow \Delta \leq \frac{\varepsilon^2}{24\sqrt{d}M_2 R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2 R}\right)$$

уровень неточности,

$$NK \geq \frac{8^{1/2} K^{1/2} L_{f_\gamma}^{1/2} R}{\varepsilon^{1/2}} \Rightarrow NK \geq \frac{16^{1/2} K^{1/2} d^{1/4} M^{1/2} M_2^{1/2} R}{\varepsilon} \Rightarrow NK \geq \frac{K^{1/2}}{\varepsilon},$$

$$NK \geq \frac{64\sigma^2 R^2}{B\varepsilon^2} \Rightarrow NK \geq \frac{72\kappa(p, d)dM_2^2 R^2}{B\varepsilon^2} \Rightarrow NK \geq \frac{1}{B\varepsilon^2}$$

и

$$NK \geq \min\left\{\frac{K^{2/3} L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{\varepsilon}, \frac{K^{3/4} L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{\varepsilon}\right\} \Rightarrow$$

$$\Rightarrow NK \geq \min \left\{ \frac{2^{1/3} K^{2/3} d^{1/6} (MM_2)^{1/3} \sigma^{2/3} R^{4/3}}{\varepsilon^{4/3}}, \frac{2^{1/2} K^{3/4} d^{1/4} (MM_2)^{1/2} \sigma^{1/2} R^{3/2}}{\varepsilon^{3/2}} \right\} \Rightarrow$$

$$\Rightarrow NK \geq \frac{2^{1/3} K^{2/3} d^{1/6} (MM_2)^{1/3} \sigma^{2/3} R^{4/3}}{\varepsilon^{4/3}} \Rightarrow NK \gtrsim \frac{K^{2/3}}{\varepsilon^{4/3}}.$$

Таким образом, наименьшее число коммуникационных раундов при условии, что $NK \in [1, \varepsilon^{-2}]$ и $T \in [1, \varepsilon^{-2}]$, будет иметь вид

$$N \sim \frac{1}{\varepsilon}: \quad N \geq \frac{8L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} \varepsilon} \Rightarrow N = O\left(\frac{d^{1/6} (\kappa(p, d)M)^{1/3} M_2 R^{4/3}}{K^{1/3} \varepsilon^{4/3}}\right),$$

тогда

$$K \sim \frac{1}{\varepsilon}: \quad K \geq \frac{\sigma^2 R^2}{BN \varepsilon^2} \Rightarrow K = O\left(\frac{\kappa(p, d) M_2^2 R^2}{BN \varepsilon^2}\right)$$

количество локальных вызовов безградиентного оракула, $B = 1$ — количество работающих параллельно машин и

$$T \sim \frac{1}{\varepsilon^2}: \quad T = NKB = O\left(\frac{\kappa(p, d) M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{d M_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ O\left(\frac{M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула.

Для l_2 -рандомизации имеем следующие алгоритмы:

- **Minibatch Accelerated SGD**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma} R^2}{N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.56})$$

$$\frac{4L_{f_\gamma} R^2}{N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.57})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.58})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из следствия 3), в неравенства (П.56)–(П.58), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2 \geq \frac{48\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow N \geq \frac{4\sqrt{3}d^{1/4}\sqrt{MM_2}R}{\varepsilon} \Rightarrow N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right)$$

количество коммуникационных раундов,

$$B \geq \frac{576\sigma^2R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{1152\kappa(p, d)M_2^2R^2}{KN\varepsilon^2} \Rightarrow B = O\left(\frac{\kappa(p, d)M_2^2R^2}{KN\varepsilon^2}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{1152\kappa(p, d)M_2^2R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• **Single-Machine Accelerated SGD**

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{NK+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{N^2K^2} + \frac{4\sigma R}{\sqrt{NK}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.59})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.60})$$

$$\frac{4\sigma R}{\sqrt{KN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.61})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из следствия 3), в неравенства (П.59)–(П.61), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{576\sigma^2 R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{1152\kappa(p, d)M_2^2 R^2}{\varepsilon^2}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{1152\kappa(p, d)M_2^2 R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{(\ln d)M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• **Local-AC-SA**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [12, 22]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{K^2N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.62})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.63})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.64})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из следствия 3), в неравенства (П.62)–(П.64), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2 K^2 \geq \frac{48\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{4\sqrt{3}d^{1/4}\sqrt{MM_2R}}{\varepsilon}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right)$$

количество локальных вызовов безградиентного оракула,

$$B \geq \frac{576\sigma^2 R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{1152\kappa(p,d)M_2^2 R^2}{KN\varepsilon^2} \Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2 R^2}{KN\varepsilon^2}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{1152\kappa(p,d)M_2^2 R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{(\ln d)M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• **Federated Accelerated SGD (FedAc)**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [13]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{L_{f_\gamma} R^2}{KN^2} + \frac{\sigma R}{\sqrt{BKN}} + \min\left\{\frac{L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} N}, \frac{L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{K^{1/4} N}\right\} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f(x_{ag}^{N+1}) - f(x_*) \leq f(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{8}, \quad (\text{П.65})$$

$$\frac{L_{f_\gamma} R^2}{KN^2} \leq \frac{\varepsilon}{8}, \quad (\text{П.66})$$

$$\frac{\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{8}, \quad (\text{П.67})$$

$$\min\left\{\frac{L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} N}, \frac{L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{K^{1/4} N}\right\} \leq \frac{\varepsilon}{8}. \quad (\text{П.68})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = 2\kappa(p, d)dM_2^2$ (из следствия 3), в неравенства (П.65)–(П.68), получим

$$\Delta \leq \frac{\gamma\varepsilon}{8\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{16\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{8^{1/2}K^{1/2}L_{f_\gamma}^{1/2}R}{\varepsilon^{1/2}} \Rightarrow NK \geq \frac{16^{1/2}K^{1/2}d^{1/4}M^{1/2}M_2^{1/2}R}{\varepsilon} \Rightarrow NK \geq \frac{K^{1/2}}{\varepsilon},$$

$$NK \geq \frac{64\sigma^2R^2}{B\varepsilon^2} \Rightarrow NK \geq \frac{72\kappa(p, d)dM_2^2R^2}{B\varepsilon^2} \Rightarrow NK \geq \frac{1}{B\varepsilon^2}$$

и

$$NK \geq \min\left\{\frac{K^{2/3}L_{f_\gamma}^{1/3}\sigma^{2/3}R^{4/3}}{\varepsilon}, \frac{K^{3/4}L_{f_\gamma}^{1/2}\sigma^{1/2}R^{3/2}}{\varepsilon}\right\} \Rightarrow$$

$$\Rightarrow NK \geq \min\left\{\frac{2^{1/3}K^{2/3}d^{1/6}(MM_2)^{1/3}\sigma^{2/3}R^{4/3}}{\varepsilon^{4/3}}, \frac{2^{1/2}K^{3/4}d^{1/4}(MM_2)^{1/2}\sigma^{1/2}R^{3/2}}{\varepsilon^{3/2}}\right\} \Rightarrow$$

$$\Rightarrow NK \geq \frac{2^{1/3}K^{2/3}d^{1/6}(MM_2)^{1/3}\sigma^{2/3}R^{4/3}}{\varepsilon^{4/3}} \Rightarrow NK \geq \frac{K^{2/3}}{\varepsilon^{4/3}}.$$

Таким образом, наименьшее число коммуникационных раундов при условии, что $NK \in [1, \varepsilon^{-2}]$ и $T \in [1, \varepsilon^{-2}]$, будет иметь вид

$$N \sim \frac{1}{\varepsilon}: \quad N \geq \frac{8L_{f_\gamma}^{1/3}\sigma^{2/3}R^{4/3}}{K^{1/3}\varepsilon} \Rightarrow N = O\left(\frac{d^{1/2}(\kappa(p, d)M)^{1/3}M_2R^{4/3}}{K^{1/3}\varepsilon^{4/3}}\right),$$

тогда

$$K \sim \frac{1}{\varepsilon}: \quad K \geq \frac{\sigma^2R^2}{BN\varepsilon^2} \Rightarrow K = O\left(\frac{\kappa(p, d)dM_2^2R^2}{BN\varepsilon^2}\right)$$

количество локальных вызовов безградиентного оракула, $B = 1$ – количество работающих параллельно машин и

$$T \sim \frac{1}{\varepsilon^2}: \quad T = NKB = \tilde{O}\left(\frac{\kappa(p, d)dM_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула.

С. ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 2

Приведем полное доказательство теоремы 2. Для этого разобьем доказательство на две части: доказательство для l_1 -рандомизации и доказательство для l_2 -рандомизации.

Для l_1 -рандомизации имеем следующие алгоритмы:

• Minibatch SMP

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$\mathbb{E}[f_\gamma(z_N)] \leq \max\left\{\frac{7}{4}\frac{LR^2}{N}, 7\frac{\sigma R}{\sqrt{BKN}}\right\} + \frac{d\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{d\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.69})$$

$$\max \left\{ \frac{7}{4} \frac{LR^2}{N}, 7 \frac{\sigma R}{\sqrt{BKN}} \right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.70})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из следствия 3), в неравенства (П.69) и (П.70), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6dR} \Rightarrow \Delta \leq \frac{\varepsilon^2}{24\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2 R^2}{BK\varepsilon^2} \Rightarrow N \geq \frac{1568\kappa(p, d)M_2^2 R^2}{BK\varepsilon^2}.$$

Так как $K = 1$, а N напрямую зависит от B , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow B = O\left(\frac{\kappa(p, d)M_2^2 R^2}{KN\varepsilon^2}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{1568\kappa(p, d)M_2^2 R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = O\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ O\left(\frac{M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• Single-Machine SMP

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$\mathbb{E}[f_\gamma(z_{NK})] \leq \max \left\{ \frac{7}{4} \frac{LR^2}{KN}, 7 \frac{\sigma R}{\sqrt{KN}} \right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.71})$$

$$\max \left\{ \frac{7}{4} \frac{LR^2}{KN}, 7 \frac{\sigma R}{\sqrt{KN}} \right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.72})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из 3), в неравенства (П.71) и (П.72), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6dR} \Rightarrow \Delta \leq \frac{\varepsilon^2}{24\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2 R^2}{K\varepsilon^2} \Rightarrow N \geq \frac{1568\kappa(p, d)M_2^2 R^2}{K\varepsilon^2}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow K = O\left(\frac{\kappa(p, d)M_2^2 R^2}{KN\varepsilon^2}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{1568\kappa(p, d)M_2^2 R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = O\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ O\left(\frac{M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула.

Для l_2 -рандомизации имеем следующие алгоритмы:

- **Minibatch SMP**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$\mathbb{E}[f_\gamma(z_N)] \leq \max \left\{ \frac{7}{4} \frac{LR^2}{N}, 7 \frac{\sigma R}{\sqrt{BKN}} \right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.73})$$

$$\max \left\{ \frac{7}{4} \frac{LR^2}{N}, 7 \frac{\sigma R}{\sqrt{BKN}} \right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.74})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из 3), в неравенства (П.73) и (П.74), получим

$$\Delta \leq \frac{\gamma\varepsilon}{4\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{8\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2 R^2}{BK\varepsilon^2} \Rightarrow N \geq \frac{1568\kappa(p, d)M_2^2 R^2}{BK\varepsilon^2}.$$

Так как $K = 1$, а N напрямую зависит от B , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow B = O\left(\frac{\kappa(p, d)M_2^2 R^2}{KN\varepsilon^2}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{1568\kappa(p, d)M_2^2 R^2}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2 R^2}{\varepsilon^2}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{(\ln d)M_2^2 R^2}{\varepsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула;

• Single-Machine SMP

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$\mathbb{E}[f_\gamma(z_{NK})] \leq \max\left\{\frac{7}{4}\frac{LR^2}{KN}, 7\frac{\sigma R}{\sqrt{KN}}\right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$ нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.75})$$

$$\max\left\{\frac{7}{4}\frac{LR^2}{KN}, 7\frac{\sigma R}{\sqrt{KN}}\right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.76})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = 2\kappa(p, d)M_2^2$ (из 3), в неравенства (П.75) и (П.76), получим

$$\Delta \leq \frac{\gamma\varepsilon}{4\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{8\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2 R^2}{K\epsilon^2} \Rightarrow N \geq \frac{1568\kappa(p,d)M_2^2 R^2}{K\epsilon^2}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow K = O\left(\frac{\kappa(p,d)M_2^2 R^2}{KN\epsilon^2}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{1568\kappa(p,d)M_2^2 R^2}{\epsilon^2} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2 R^2}{\epsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2 R^2}{\epsilon^2}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{(\ln d)M_2^2 R^2}{\epsilon^2}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов двухточечного безградиентного оракула.

Д. ОДНОТОЧЕЧНЫЕ БЕЗГРАДИЕНТНЫЕ АЛГОРИТМЫ

Прделаем такую же процедуру, как и в п. 4.2 для создания безградиентных однотоочечных алгоритмов.

Для начала на примере леммы 10 покажем, что лемма 8 также выполняется и для однотоочечного оракула.

Лемма 10. Для $\nabla f_\gamma(x, \xi, e)$ и $\nabla f_\gamma(x)$ с предположением 2, выполняется следующее:

$$\mathbb{E}_{\xi,e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] \geq \langle \nabla f_\gamma(x), r \rangle - \frac{d\Delta E_e[\langle \text{sign}(e), r \rangle]}{\gamma},$$

где ∇f_γ с l_1 -рандомизацией;

$$\mathbb{E}_{\xi,e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] \geq \langle \nabla f_\gamma(x), r \rangle - \frac{d\Delta E_e[\langle e, r \rangle]}{\gamma},$$

где ∇f_γ с l_2 -рандомизацией.

Доказательство. Рассмотрим следующие утверждения:

i) для l_1 -рандомизации (8):

$$\begin{aligned} \nabla f_\gamma(x, \xi, e) &= \frac{d}{\gamma} f_\delta(x + \gamma e, \xi) \text{sign}(e) = \frac{d}{\gamma} (f(x + \gamma e, \xi) + \delta(x + \gamma e)) \text{sign}(e) = \\ &= \frac{d}{\gamma} (f(x + \gamma e, \xi) \text{sign}(e) + \delta(x + \gamma e) \text{sign}(e)). \end{aligned}$$

Из данного равенства следует

$$\mathbb{E}_{\xi,e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] = \frac{d}{\gamma} \mathbb{E}_{\xi,e}[\langle f(x + \gamma e, \xi) \text{sign}(e), r \rangle] + \frac{d}{\gamma} \mathbb{E}_e[\langle \delta(x + \gamma e) \text{sign}(e), r \rangle]. \quad (\text{П.77})$$

Применяя лемму 5 к первому слагаемому (П.77), получим

$$\begin{aligned} \frac{d}{\gamma} \mathbb{E}_{\xi,e}[\langle f(x + \gamma e, \xi) \text{sign}(e), r \rangle] &= \frac{d}{\gamma} \mathbb{E}_e[\langle \mathbb{E}_\xi[f(x + \gamma e, \xi)] \text{sign}(e), r \rangle] = \\ &= \frac{d}{\gamma} \mathbb{E}_e[\langle f(x + \gamma e) \text{sign}(e), r \rangle] = \langle \nabla f_\gamma(x), r \rangle. \end{aligned} \quad (\text{П.78})$$

Для второго слагаемого (П.77) с предположением 2 получим

$$\frac{d}{\gamma} \mathbb{E}_e[\langle \delta(x + \gamma e) \text{sign}(e), r \rangle] \geq -\frac{d}{\gamma} \Delta \mathbb{E}_e[\langle \text{sign}(e), r \rangle]. \quad (\text{П.79})$$

Используя уравнения (П.78) и (П.79), для уравнения (П.77), получим утверждение леммы для l_1 -рандомизации.

ii) для l_2 -рандомизации (9):

$$\begin{aligned} \nabla f_\gamma(x, \xi, e) &= \frac{d}{\gamma} f_\delta(x + \gamma e, \xi) e = \frac{d}{\gamma} (f(x + \gamma e, \xi) + \delta(x + \gamma e)) e = \\ &= \frac{d}{\gamma} (f(x + \gamma e, \xi) e + \delta(x + \gamma e) e). \end{aligned}$$

Из данного равенства следует

$$\mathbb{E}_{\xi, e}[\langle \nabla f_\gamma(x, \xi, e), r \rangle] = \frac{d}{\gamma} \mathbb{E}_{\xi, e}[\langle f(x + \gamma e, \xi) e, r \rangle] + \frac{d}{\gamma} \mathbb{E}_e[\langle \delta(x + \gamma e) e, r \rangle]. \quad (\text{П.80})$$

Применяя лемму 6 к первому слагаемому (П.80), получим

$$\begin{aligned} \frac{d}{\gamma} \mathbb{E}_{\xi, e}[\langle f(x + \gamma e, \xi) e, r \rangle] &= \frac{d}{\gamma} \mathbb{E}_e[\langle \mathbb{E}_\xi[f(x + \gamma e, \xi)] e, r \rangle] = \\ &= \frac{d}{\gamma} \mathbb{E}_e[\langle f(x + \gamma e) e, r \rangle] = \langle \nabla f_\gamma(x), r \rangle. \end{aligned} \quad (\text{П.81})$$

Для второго слагаемого (П.80) с предположением 2 получим

$$\frac{d}{\gamma} \mathbb{E}_e[\langle \delta(x + \gamma e) e, r \rangle] \geq -\frac{d}{\gamma} \Delta \mathbb{E}_e[\langle e, r \rangle]. \quad (\text{П.82})$$

Используя уравнения (П.81) и (П.82), для уравнения (П.80) получим утверждение леммы для l_2 -рандомизации.

Так как лемма 8 выполняется для одноточечного оракула, то следствия 5, 6 и замечание 4 также выполняются для одноточечного оракула. Таким образом, теперь можем получить оценки параметров одноточечных безградиентных методов для l_1 и l_2 -рандомизации, аналогичным образом, как и в п. 4.2, используя п. 3.4. В теореме 3 представлены оценки безградиентных методов выпуклой оптимизации, а в теореме 4 оценки для седловых задач.

Теорема 3. *Схема сглаживания из разд. 3, применяемая к задаче (2), обеспечивает сходимость следующих одноточечных безградиентных алгоритмов: Minibatch и Single-Machine Accelerated SGD [11], Local-AC-SA [12] и FedAc [13]. Другими словами, для достижения ϵ точности решения задачи (2) необходимо проделать NK итераций с максимально допустимым уровнем шума Δ и общим числом вызова безградиентного оракула T в соответствии с выбранным методом и схемой сглаживания:*

• *Minibatch Accelerated SGD*

i) для l_1 -рандомизации (8):

$$\begin{aligned} \Delta &= O\left(\frac{\epsilon^2}{\sqrt{d} M_2 R}\right); \\ N &= O\left(\frac{d^{1/4} \sqrt{M M_2} R}{\epsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{K N \epsilon^4}\right); \\ T &= \tilde{O}\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{\epsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases} \end{aligned}$$

ii) для l_2 -рандомизации (9):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right);$$

$$N = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)M_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

• *Single-Machine Accelerated SGD*

i) для l_1 -рандомизации (8):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dG^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

ii) для l_2 -рандомизации (9):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

• *Local-AC-SA*

i) для l_1 -рандомизации (8):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right);$$

$$N = 1; \quad K = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad B = O\left(\frac{\kappa(q,d)d^2G^2R^2}{KN\varepsilon^4}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dG^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

ii) для l_2 -рандомизации (9):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = 1; \quad K = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right); \quad B = O\left(\frac{\kappa(q,d)d^2G^2R^2}{KN\varepsilon^4}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

• *Federated Accelerated SGD (FedAc)*

i) для l_1 -рандомизации (8):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = O\left(\frac{d^{1/6}(\kappa(q,d)MM_2)^{1/3}G^{2/3}R^{4/3}}{K^{1/3}\varepsilon^2}\right); \quad K = O\left(\frac{\kappa(q,d)d^2G^2R^2}{BN\varepsilon^4}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dG^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

ii) для l_2 -рандомизации (9):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{dM_2R}}\right);$$

$$N = O\left(\frac{d^{1/6}(\kappa(q,d)MM_2)^{1/3}G^{2/3}R^{4/3}}{K^{1/3}\varepsilon^2}\right); \quad K = O\left(\frac{\kappa(q,d)d^2G^2R^2}{BN\varepsilon^4}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(q,d)d^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty). \end{cases}$$

Доказательство. Рассмотрим доказательство для каждой рандомизации и каждого метода отдельно.

Для l_1 -рандомизации имеем следующие алгоритмы:

• **Minibatch Accelerated SGD**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.83})$$

$$\frac{4L_{f_\gamma}R^2}{N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.84})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.85})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.83)–(П.85), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2 \geq \frac{48\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow N \geq \frac{4\sqrt{3}d^{1/4}\sqrt{MM_2}R}{\varepsilon} \Rightarrow N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right)$$

количество коммуникационных раундов,

$$B \geq \frac{576\sigma^2R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{2304\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4} \Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dM_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов односточечного безградиентного оракула;

• Single-Machine Accelerated SGD

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{NK+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{N^2K^2} + \frac{4\sigma R}{\sqrt{NK}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.86})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.87})$$

$$\frac{4\sigma R}{\sqrt{KN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.88})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.86)–(П.88), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{576\sigma^2R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dM_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов односточного безградиентного оракула;

• Local-AC-SA

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [12, 22]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{K^2N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.89})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.90})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.91})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.89)–(П.91), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2K^2 \geq \frac{48\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{4\sqrt{3}d^{1/4}\sqrt{MM_2R}}{\varepsilon}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{d^{1/4}\sqrt{MM_2R}}{\varepsilon}\right)$$

количество локальных вызовов безградиентного оракула,

$$B \geq \frac{576\sigma^2R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{2304\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4} \Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dM_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула;

• Federated Accelerated SGD (FedAc)

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [13]) в соответствии со следствием 5:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{L_{f_\gamma}R^2}{KN^2} + \frac{\sigma R}{\sqrt{BKN}} + \min\left\{\frac{L_{f_\gamma}^{1/3}\sigma^{2/3}R^{4/3}}{K^{1/3}N}, \frac{L_{f_\gamma}^{1/2}\sigma^{1/2}R^{3/2}}{K^{1/4}N}\right\} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$ нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{8}, \quad (\text{П.92})$$

$$\frac{L_{f_\gamma} R^2}{KN^2} \leq \frac{\varepsilon}{8}, \quad (\text{П.93})$$

$$\frac{\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{8}, \quad (\text{П.94})$$

$$\min \left\{ \frac{L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} N}, \frac{L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{K^{1/4} N} \right\} \leq \frac{\varepsilon}{8}. \quad (\text{П.95})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2 G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.92)–(П.95), получим

$$\Delta \leq \frac{\gamma\varepsilon}{8\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{16\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{8^{1/2} K^{1/2} L_{f_\gamma}^{1/2} R}{\varepsilon^{1/2}} \Rightarrow NK \geq \frac{16^{1/2} K^{1/2} d^{1/4} M^{1/2} M_2^{1/2} R}{\varepsilon} \Rightarrow NK \geq \frac{K^{1/2}}{\varepsilon},$$

$$NK \geq \frac{64\sigma^2 R^2}{B\varepsilon^2} \Rightarrow NK \geq \frac{256\kappa(p,d)M_2^2 G^2 R^2}{B\varepsilon^4} \Rightarrow NK \geq \frac{1}{B\varepsilon^4}$$

и

$$\begin{aligned} NK &\geq \min \left\{ \frac{K^{2/3} L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{\varepsilon}, \frac{K^{3/4} L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{\varepsilon} \right\} \Rightarrow \\ &\Rightarrow NK \geq \min \left\{ \frac{16K^{2/3} d^{1/6} (MM_2)^{1/3} \kappa(p,d)^{1/3} G^{2/3} R^{4/3}}{\varepsilon^2}, \frac{16K^{3/4} d^{1/4} (MM_2)^{1/2} \kappa(p,d)^{1/4} G^{1/2} R^{3/2}}{\varepsilon^2} \right\} \Rightarrow \\ &\Rightarrow NK \geq \frac{16K^{2/3} d^{1/6} (MM_2)^{1/3} \kappa(p,d)^{1/3} G^{2/3} R^{4/3}}{\varepsilon^2} \Rightarrow NK \geq \frac{K^{2/3}}{\varepsilon^2}. \end{aligned}$$

Таким образом, наименьшее число коммуникационных раундов при условии, что $NK \in [1, \varepsilon^{-4}]$ и $T \in [1, \varepsilon^{-4}]$, будет выглядеть следующим образом:

$$N \sim \frac{1}{\varepsilon} \quad N \geq \frac{8L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} \varepsilon} \Rightarrow N = O\left(\frac{d^{1/6} (\kappa(p,d)MM_2)^{1/3} G^{2/3} R^{4/3}}{K^{1/3} \varepsilon^2}\right),$$

тогда

$$K \sim \frac{1}{\varepsilon^3}: \quad K \geq \frac{\sigma^2 R^2}{BN\varepsilon^2} \Rightarrow K = O\left(\frac{\kappa(p,d)M_2^2 G^2 R^2}{BN\varepsilon^4}\right)$$

количество локальных вызовов безградиентного оракула, $B = 1$ – количество работающих параллельно машин и

$$T \sim \frac{1}{\varepsilon^4}: \quad T = NKB = \tilde{O}\left(\frac{\kappa(p,d)M_2^2 G^2 R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула.

Для I_2 -рандомизации имеем следующие алгоритмы:

• **Minibatch Accelerated SGD**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma} R^2}{N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.96})$$

$$\frac{4L_{f_\gamma} R^2}{N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.97})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.98})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2 G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.96)–(П.98), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2 R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2 R}\right)$$

уровень неточности,

$$N^2 \geq \frac{48\sqrt{d}MM_2 R^2}{\varepsilon^2} \Rightarrow N \geq \frac{4\sqrt{3}d^{1/4}\sqrt{MM_2}R}{\varepsilon} \Rightarrow N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right)$$

количество коммуникационных раундов,

$$B \geq \frac{576\sigma^2 R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{2304\kappa(p,d)M_2^2 G^2 R^2}{KN\varepsilon^4} \Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2 G^2 R^2}{KN\varepsilon^4}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)M_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула;

• **Single-Machine Accelerated SGD**

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. [11, 22]) в соответствии со следствием 6:

$$E[f_\gamma(x_{ag}^{NK+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{N^2K^2} + \frac{4\sigma R}{\sqrt{NK}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.99})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.100})$$

$$\frac{4\sigma R}{\sqrt{KN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.101})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.99)–(П.101), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{576\sigma^2R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)M_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула;

• **Local-AC-SA**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [12, 22]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{4L_{f_\gamma}R^2}{K^2N^2} + \frac{4\sigma R}{\sqrt{BKN}} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{6}, \quad (\text{П.102})$$

$$\frac{4L_{f_\gamma}R^2}{K^2N^2} \leq \frac{\varepsilon}{6}, \quad (\text{П.103})$$

$$\frac{4\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{6}. \quad (\text{П.104})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.102)–(П.104), получим

$$\Delta \leq \frac{\gamma\varepsilon}{6\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{12\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N^2K^2 \geq \frac{48\sqrt{d}MM_2R^2}{\varepsilon^2} \Rightarrow NK \geq \frac{4\sqrt{3}d^{1/4}\sqrt{MM_2}R}{\varepsilon}.$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$NK = K = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right)$$

количество локальных вызовов безградиентного оракула,

$$B \geq \frac{576\sigma^2R^2}{KN\varepsilon^2} \Rightarrow B \geq \frac{2304\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4} \Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)M_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула;

• **Federated Accelerated SGD (FedAc)**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. [13]) в соответствии со следствием 6:

$$\mathbb{E}[f_\gamma(x_{ag}^{N+1}) - f(x_*)] \leq \frac{L_{f_\gamma}R^2}{KN^2} + \frac{\sigma R}{\sqrt{BKN}} + \min\left\{\frac{L_{f_\gamma}^{1/3}\sigma^{2/3}R^{4/3}}{K^{1/3}N}, \frac{L_{f_\gamma}^{1/2}\sigma^{1/2}R^{3/2}}{K^{1/4}N}\right\} + \frac{\sqrt{d}\Delta R}{\gamma},$$

где $x_*(\gamma) = \operatorname{argmin}_{x \in Q_\gamma} f_\gamma(x)$.

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(x)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(x)$:

$$f_\gamma(x_{ag}^{N+1}) - f(x_*) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x_{ag}^{N+1}) - f(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(x)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{8}, \quad (\text{П.105})$$

$$\frac{L_{f_\gamma}R^2}{KN^2} \leq \frac{\varepsilon}{8}, \quad (\text{П.106})$$

$$\frac{\sigma R}{\sqrt{BKN}} \leq \frac{\varepsilon}{8}, \quad (\text{П.107})$$

$$\min\left\{\frac{L_{f_\gamma}^{1/3}\sigma^{2/3}R^{4/3}}{K^{1/3}N}, \frac{L_{f_\gamma}^{1/2}\sigma^{1/2}R^{3/2}}{K^{1/4}N}\right\} \leq \frac{\varepsilon}{8}. \quad (\text{П.108})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.105)–(П.108), получим

$$\Delta \leq \frac{\gamma\varepsilon}{8\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{16\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$NK \geq \frac{8^{1/2}K^{1/2}L_{f_\gamma}^{1/2}R}{\varepsilon^{1/2}} \Rightarrow NK \geq \frac{16^{1/2}K^{1/2}d^{1/4}M^{1/2}M_2^{1/2}R}{\varepsilon} \Rightarrow NK \geq \frac{K^{1/2}}{\varepsilon},$$

$$NK \geq \frac{64\sigma^2R^2}{B\varepsilon^2} \Rightarrow NK \geq \frac{256\kappa(p,d)M_2^2G^2R^2}{B\varepsilon^4} \Rightarrow NK \geq \frac{1}{B\varepsilon^4},$$

и

$$\begin{aligned}
NK &\geq \min \left\{ \frac{K^{2/3} L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{\epsilon}, \frac{K^{3/4} L_{f_\gamma}^{1/2} \sigma^{1/2} R^{3/2}}{\epsilon} \right\} \Rightarrow \\
&\Rightarrow NK \geq \min \left\{ \frac{16K^{2/3} d^{1/6} (MM_2)^{1/3} \kappa(p, d)^{1/3} G^{2/3} R^{4/3}}{\epsilon^2}, \frac{16K^{3/4} d^{1/4} (MM_2)^{1/2} \kappa(p, d)^{1/4} G^{1/2} R^{3/2}}{\epsilon^2} \right\} \Rightarrow \\
&\Rightarrow NK \geq \frac{16K^{2/3} d^{1/6} (MM_2)^{1/3} \kappa(p, d)^{1/3} G^{2/3} R^{4/3}}{\epsilon^2} \Rightarrow NK \geq \frac{K^{2/3}}{\epsilon^2}.
\end{aligned}$$

Таким образом, наименьшее число коммуникационных раундов при условии, что $NK \in [1, \epsilon^{-4}]$ и $T \in [1, \epsilon^{-4}]$, будет иметь вид

$$N \sim \frac{1}{\epsilon}: \quad N \geq \frac{8L_{f_\gamma}^{1/3} \sigma^{2/3} R^{4/3}}{K^{1/3} \epsilon} \Rightarrow N = O\left(\frac{d^{1/6} (\kappa(p, d) MM_2)^{1/3} G^{2/3} R^{4/3}}{K^{1/3} \epsilon^2}\right),$$

тогда

$$K \sim \frac{1}{\epsilon^3}: \quad K \geq \frac{\sigma^2 R^2}{BN \epsilon^2} \Rightarrow K = O\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{BN \epsilon^4}\right)$$

количество локальных вызовов безградиентного оракула, $B = 1$ — количество работающих параллельно машин и

$$T \sim \frac{1}{\epsilon^4}: \quad T = NKB = \tilde{O}\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{\epsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d) M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула.

Теорема 4. Схема сглаживания из разд. 3, применяемая к седловой задаче (см. замечание 1), обеспечивает сходимость следующих одноточечных безградиентных алгоритмов: *Minibatch SMP* и *Single-Machine SMP* из Приложения 8. Другими словами, для достижения ϵ точности решения седловой задачи (см. замечание 1) необходимо проделать NK итераций с максимально допустимым уровнем шума Δ и общим числом вызова безградиентного оракула T в соответствии с выбранным методом и следующей схемой сглаживания:

• *Minibatch SMP*

i) для l_1 -рандомизации (8):

$$\begin{aligned}
\Delta &= O\left(\frac{\epsilon^2}{\sqrt{d} M_2 R}\right); \\
N &= 1; \quad K = 1; \quad B = O\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{\epsilon^4}\right); \\
T &= \tilde{O}\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{\epsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}
\end{aligned}$$

ii) для l_2 -рандомизации (9):

$$\begin{aligned}
\Delta &= O\left(\frac{\epsilon^2}{\sqrt{d} M_2 R}\right); \\
N &= 1; \quad K = 1; \quad B = O\left(\frac{\kappa(p, d) M_2^2 G^2 R^2}{\epsilon^4}\right);
\end{aligned}$$

$$T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

• *Single-Machine SMP*

i) для l_1 -рандомизации (8):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2 R}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{\varepsilon^4}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dM_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

ii) для l_2 -рандомизации (9):

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2 R}\right);$$

$$N = 1; \quad K = O\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{\varepsilon^4}\right); \quad B = 1;$$

$$T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d)M_2^2 G^2 R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty). \end{cases}$$

Доказательство. Рассмотрим доказательство для каждой рандомизации и каждого метода отдельно.

Для l_1 -рандомизации имеем следующие алгоритмы:

• **Minibatch SMP**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$E[f_\gamma(z_N)] \leq \max\left\{\frac{7}{4} \frac{LR^2}{N}, 7 \frac{\sigma R}{\sqrt{BKN}}\right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \frac{2}{\sqrt{d}} \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.109})$$

$$\max \left\{ \frac{7LR^2}{4N}, 7\frac{\sigma R}{\sqrt{BKN}} \right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.110})$$

Подставляя $L_{f_\gamma} = \frac{dM}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{4M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.109) и (П.110), получим

$$\Delta \leq \frac{\gamma\varepsilon}{4\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{8\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2R^2}{BK\varepsilon^2} \Rightarrow N \geq \frac{3136\kappa(p,d)M_2^2G^2R^2}{BK\varepsilon^4} \Rightarrow .$$

Так как $K = 1$, а N напрямую зависит от B , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dM_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула;

• Single-Machine SMP

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$\mathbb{E}[f_\gamma(z_{NK})] \leq \max \left\{ \frac{7LR^2}{4KN}, 7\frac{\sigma R}{\sqrt{KN}} \right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \frac{2}{\sqrt{d}}\gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.111})$$

$$\max \left\{ \frac{7LR^2}{4KN}, 7\frac{\sigma R}{\sqrt{KN}} \right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.112})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{4M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.111) и (П.112), получим

$$\Delta \leq \frac{\gamma\varepsilon}{4\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{8\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2R^2}{K\varepsilon^2} \Rightarrow N \geq \frac{3136\kappa(p,d)M_2^2G^2R^2}{K\varepsilon^4} \Rightarrow .$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow K = O\left(\frac{\kappa(p,d)M_2^2G^2R^2}{KN\varepsilon^4}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2G^2R^2}{\varepsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2M_2^2G^2R^2}{\varepsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{dM_2^2G^2R^2}{\varepsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула.

Для l_2 -рандомизации имеем следующие алгоритмы:

• **Minibatch SMP**

Данный алгоритм после N раундов связи дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$E[f_\gamma(z_N)] \leq \max\left\{\frac{7}{4}\frac{LR^2}{N}, 7\frac{\sigma R}{\sqrt{BKN}}\right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\varepsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\varepsilon}{2M_2}$ (из следствия 1), то имеем ε -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Чтобы была $\frac{\varepsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\varepsilon}{4}, \quad (\text{П.113})$$

$$\max\left\{\frac{7}{4}\frac{LR^2}{N}, 7\frac{\sigma R}{\sqrt{BKN}}\right\} \leq \frac{\varepsilon}{4}. \quad (\text{П.114})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\varepsilon}{2M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2G^2}{\varepsilon^2}$ (из следствия 4), в неравенства (П.113) и (П.114), получим

$$\Delta \leq \frac{\gamma\varepsilon}{4\sqrt{d}R} \Rightarrow \Delta \leq \frac{\varepsilon^2}{8\sqrt{d}M_2R} \Rightarrow \Delta = O\left(\frac{\varepsilon^2}{\sqrt{d}M_2R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2 R^2}{BK\epsilon^2} \Rightarrow N \geq \frac{3136\kappa(p,d)M_2^2 G^2 R^2}{BK\epsilon^4} \Rightarrow .$$

Так как $K = 1$, а N напрямую зависит от B , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow B = O\left(\frac{\kappa(p,d)M_2^2 G^2 R^2}{KN\epsilon^4}\right)$$

количество работающих параллельно машин и

$$T = NKB = \frac{2304\kappa(p,d)M_2^2 G^2 R^2}{\epsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p,d)M_2^2 G^2 R^2}{\epsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d) M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула;

• **Single-Machine SMP**

Данный алгоритм после NK итераций дает скорость сходимости для $f_\gamma(x)$ (см. следствие 7) в соответствии с замечанием 4:

$$\mathbb{E}[f_\gamma(z_{NK})] \leq \max\left\{\frac{7 LR^2}{4 KN}, 7 \frac{\sigma R}{\sqrt{KN}}\right\} + \frac{\sqrt{d}\Delta R}{\gamma}.$$

Если имеем $\frac{\epsilon}{2}$ -точность для функции $f_\gamma(z)$ с $\gamma = \frac{\epsilon}{2M_2}$ (из следствия 1), то имеем ϵ -точность для функции $f(z)$:

$$f(z_N) \leq f_\gamma(z_N) + \gamma M_2 \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Чтобы была $\frac{\epsilon}{2}$ -точность для $f_\gamma(z)$, нужно

$$\frac{\sqrt{d}\Delta R}{\gamma} \leq \frac{\epsilon}{4}, \quad (\text{П.115})$$

$$\max\left\{\frac{7 LR^2}{4 KN}, 7 \frac{\sigma R}{\sqrt{KN}}\right\} \leq \frac{\epsilon}{4}. \quad (\text{П.116})$$

Подставляя $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$ (из следствия 2), где $\gamma = \frac{\epsilon}{2M_2}$ и $\sigma^2 = \frac{4\kappa(p,d)M_2^2 G^2}{\epsilon^2}$ (из следствия 4), в неравенства (П.115) и (П.116), получим

$$\Delta \leq \frac{\gamma\epsilon}{4\sqrt{d}R} \Rightarrow \Delta \leq \frac{\epsilon^2}{8\sqrt{d}M_2 R} \Rightarrow \Delta = O\left(\frac{\epsilon^2}{\sqrt{d}M_2 R}\right)$$

уровень неточности,

$$N \geq \frac{784\sigma^2 R^2}{K\epsilon^2} \Rightarrow N \geq \frac{3136\kappa(p,d)M_2^2 G^2 R^2}{K\epsilon^4} \Rightarrow$$

Так как N напрямую зависит от K , то количество коммуникационных раундов можно взять $N = 1$, тогда

$$\Rightarrow K = O\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{KN\epsilon^4}\right)$$

количество локальных вызовов безградиентного оракула и

$$T = NKB = \frac{2304\kappa(p, d)M_2^2 G^2 R^2}{\epsilon^4} \Rightarrow$$

$$\Rightarrow T = \tilde{O}\left(\frac{\kappa(p, d)M_2^2 G^2 R^2}{\epsilon^4}\right) = \begin{cases} \tilde{O}\left(\frac{d^2 M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 2 \quad (q = 2), \\ \tilde{O}\left(\frac{d(\ln d) M_2^2 G^2 R^2}{\epsilon^4}\right), & p = 1 \quad (q = \infty), \end{cases}$$

общее количество вызовов одноточечного безградиентного оракула.

СПИСОК ЛИТЕРАТУРЫ

1. The power of first-order smooth optimization for black-box non-smooth problems / A. Gasnikov et al. // arXiv preprint arXiv:2201.12289. 2022.
2. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. 1979.
3. *Shamir O.* An optimal algorithm for bandit and zero-order convex optimization with twopoint feedback // The Journal of Machine Learning Research. 2017. Т. 18. № 1. С. 1703–1713.
4. A gradient estimator via L1-randomization for online zero-order optimization with two point feedback / A. Akhavan et al. // arXiv preprint arXiv:2205.13910. 2022.
5. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex / A.V. Gasnikov et al. // Automation and Remote Control. 2016. Т. 77. № 11. С. 2018–2034.
6. Advances and open problems in federated learning / P. Kairouz [и др.] // Foundations and Trends® in Machine Learning. 2021. Т. 14, № 1/2. С. 1–210.
7. Optimal rates for zero-order convex optimization: The power of two function evaluations / J.C. Duchi et al. // IEEE 5. С. 2788–2806. Transactions on Information Theory. 2015. Т. 61.
8. *Scheinberg K.* Finite Difference Gradient Approximation: To Randomize or Not? // INFORMS Journal on Computing. 2022. Т. 34. № 5. С. 2384–2388.
9. *Beznosikov A., Gorbunov E., Gasnikov A.* Derivative-free method for composite optimization with applications to decentralized distributed optimization // IFAC-PapersOnLine. 2020. Т. 53. № 2. С. 4038–4043.
10. *Ledoux M.* The concentration of measure phenomenon// American Mathematical Soc., 2001.
11. The min-max complexity of distributed stochastic convex optimization with intermittent communication / B.E. Woodworth et al. // Conference on Learning Theory. PMLR. 2021. С. 4386–4437.
12. Is local SGD better than minibatch SGD? / B. Woodworth et al. // Internat. Conference on Machine Learning. PMLR. 2020. С. 10334–10343.
13. *Yuan H., Ma T.* Federated accelerated stochastic gradient descent // Advances in Neural Information Processing Systems. 2020. Т. 33. С. 5332–5344.
14. *Gorbunov E., Dvinskikh D., Gasnikov A.* Optimal decentralized distributed algorithms for stochastic convex optimization // arXiv preprint arXiv:1911.07363. 2019.
15. *Duchi J.C., Bartlett P.L., Wainwright M.J.* Randomized smoothing for stochastic optimization // SIAM Journal on Optimization. 2012. Т. 22. № 2. С. 674–701.
16. *Yousefian F., Nedic' A., Shanbhag U.V.* On stochastic gradient and subgradient methods with adaptive steplength sequences // Automatica. 2012. Т. 48. № 1. С. 56–67.

17. Стохастическая онлайн оптимизация. Одноточечные и двухточечные нелинейные многорукие бандиты. Выпуклый и сильно выпуклый случаи / А.В. Гасников [и др.] // Автоматика и телемеханика. 2017. № 2. С. 36–49.
18. Gradient-Free Optimization for Non-Smooth Minimax Problems with Maximum Value of Adversarial Noise / D. Dvinskikh et al. // arXiv preprint arXiv:2202.06114. 2022.
19. *Flaxman A.D., Kalai A.T., McMahan H.B.* Online convex optimization in the bandit setting: gradient descent without a gradient // Proc. of the sixteenth annual ACM/IEEE symposium on Discrete algorithms. 2005. С. 385–394.
20. *Dvurechensky P., Gorbunov E., Gasnikov A.* An accelerated directional derivative method for smooth stochastic convex optimization // European Journal of Operational Research. 2021. Т. 290. № 2. С. 601–621.
21. *Juditsky A., Nemirovski A., Tauvel C.* Solving variational inequalities with stochastic mirrorprox algorithm // Stochastic Systems. 2011. Т. 1. № 1. С. 17–58.
22. *Lan G.* An optimal method for stochastic composite optimization // Math. Programming. 2012. Т. 133. № 1. С. 365–397.

ОПТИМАЛЬНОЕ
УПРАВЛЕНИЕ

УДК 519.63

ЧИСЛЕННЫЙ АЛГОРИТМ ОПРЕДЕЛЕНИЯ ИСТОЧНИКА
ДИФФУЗИОННО-ЛОГИСТИЧЕСКОЙ МОДЕЛИ ПО ДАННЫМ
ИНТЕГРАЛЬНОГО ТИПА, ОСНОВАННЫЙ
НА ТЕНЗОРНОЙ ОПТИМИЗАЦИИ¹⁾© 2023 г. Т. А. Звонарева^{1,2}, С. И. Кабанихин^{2,3}, О. И. Криворотко^{1,2,3,*}¹ 630090 Новосибирск, пр-кт Акад. Лаврентьева, 6, ИВМиМГ СО РАН, Россия² 630090 Новосибирск, ул. Пирогова, 2, НГУ, Россия³ 630090 Новосибирск, пр-кт Акад. Коптюга, 4, ИМ им. С.Л. Соболева СО РАН, Россия

*e-mail: krivorotko.olya@mail.ru

Поступила в редакцию 09.12.2022 г.
Переработанный вариант 09.12.2022 г.
Принята к публикации 29.05.2023 г.

Разработан алгоритм численного решения задачи определения источника в модели распространения информации в синтетических онлайн социальных сетях, описываемой уравнениями типа “реакции–диффузии”, по дополнительной информации о процессе в фиксированные моменты времени. Исследована степень некорректности задачи определения источника в параболическом уравнении, основанная на анализе сингулярных чисел линеаризованного оператора обратной задачи. Разработанный алгоритм основан на комбинации метода тензорной оптимизации и градиентного спуска с учетом регуляризации А.Н. Тихонова. Численные расчеты демонстрируют наименьшую относительную погрешность восстановленного источника, полученную разработанным алгоритмом в сравнении с классическими подходами. Библ. 26. Фиг. 4. Табл. 2.

Ключевые слова: задача об источнике, модель “реакции–диффузии”, обратная задача, тензорная оптимизация, регуляризация, градиентные методы.

DOI: 10.31857/S0044466923090193, EDN: DFESUX

ВВЕДЕНИЕ

Оценка последствий принимаемых решений в социальных и экономических сферах опирается на создание и исследование моделей управляемых систем. В этом направлении существенную помощь могут оказать аналогии с процессами из различных предметных областей (математическая физика, механика сплошной среды, геофизика и др.) с хорошо развитым математическим аппаратом (см. [1]).

В настоящей работе рассматривается начально-краевая задача для диффузионно-логистической модели

$$\begin{aligned} u_t &= du_{xx} + \left(1 - \frac{u}{K}\right)r(t)u, \quad l_1 \leq x \leq l_2, \quad t \geq 1, \\ u(x, 1) &= Q(x), \quad l_1 \leq x \leq l_2, \\ u_x(l_1, t) &= u_x(l_2, t) = 0, \quad t \geq 1, \end{aligned} \quad (1)$$

которая описывает процесс распространения информации в онлайн социальной сети (см. [2]). Здесь $u(x, t)$ – плотность пользователей, вовлеченных в процесс распространения информации, K – константа, описывающая максимальную пропускную способность сети, $r(t)$ [1/час] – скорость роста числа активных пользователей, $Q(x)$ – гладкая функция источника распространения

¹⁾Работа выполнена при финансовой поддержке РНФ (проект № 18-71-10044-П) и Математического Центра в Академгородке, соглашение с Минобрнауки РФ 075-15-2022-281.

информации в социальной сети в зависимости от количества связей $x = 1, 2, \dots$, интерполированная с помощью сплайнов по дискретному набору связей, удовлетворяющая ограничениям

$$0 \leq Q(x) \leq Q_{\max}, \quad Q(x) = 0 \quad \forall x < l_1 \quad \text{и} \quad Q(x) = Q_{\max} \quad \forall x > l_2 \geq l_1.$$

Переменная x характеризует расстояние от источника, т.е. $x = 0$ соответствует источнику, $x = i$ – минимальное количество связей от источника до вовлеченного пользователя. Например, вовлеченный пользователь отреагировал на новость “друга”, который в свою очередь отреагировал (репост, реакция) на новость своего “друга”-источника. В таком случае вовлеченный пользователь находится на расстоянии $x = 2$ от источника.

Подобное уравнение “реакции–диффузии” было рассмотрено в статье А.Н. Колмогорова, И.Г. Петровского и Н.С. Пискунова [3], в которой $u(x, t)$ означала концентрацию исследуемого вида в популяции при $Q_{\max} = 1$, $K = 1$. В этом случае предельная скорость перемещения популяции равна $v(t) = 2\sqrt{dr(t)}$.

Если ввести функцию потока $\Psi(x, t)$, характеризующую количество вовлеченных пользователей, проходящую через точку с координатой x в единицу времени в момент t , то закон сохранения для модели (1) примет вид

$$\frac{\partial u(x, t)}{\partial t} + \frac{\partial \Psi(x, t)}{\partial x} = \left(1 - \frac{u}{K}\right) r(t) u.$$

В случае, когда функции потока $\Psi(x, t)$ и плотности $u(x, t)$ связаны соотношением Фика

$$\Psi(x, t) = -v(u) \frac{\partial u}{\partial x},$$

уравнение из модели (1) примет вид

$$\frac{\partial u}{\partial t} = \left(1 - \frac{u}{K}\right) r(t) u + \frac{\partial}{\partial x} \left(v(u) \frac{\partial u}{\partial x} \right).$$

Здесь $v(u)$ – неотрицательная функция скорости потока.

Замечание. Начально-краевая задача об определении источника $\varphi(x)$ может быть записана в виде

$$\begin{aligned} \frac{\partial \tilde{u}}{\partial t} &= \left(1 - \frac{\tilde{u}}{K}\right) r(t) \tilde{u} + d \frac{\partial^2 \tilde{u}}{\partial x^2} + \varphi(x) k(x, t), \quad t \in (1, T), x \in (l_1, l_2), \\ \tilde{u}(x, 1) &< 0, \quad x \in (l_1, l_2), \\ \tilde{u}_x(l_1, t) &= \tilde{u}_x(l_2, t) = 0, \quad t \in (1, T). \end{aligned} \quad (2)$$

В случае $k(x, t) = \delta(x) \delta(t)$, где δ – дельта-функция Дирака, задача определения функции $\varphi(x)$ из (2) сводится к задаче определения функции $Q(x)$ из (1) (см. [4]).

Определение источника $\varphi(x)$ крайне важно при моделировании динамики популяции, распространения информации в социальных сетях, приложениях математической физики, так как решение $\tilde{u}(x, t)$ сильно зависит от линейной части члена, отвечающего за реакцию. В физических приложениях линейная часть часто неизвестна или известна частично, и $\varphi(x)$ нельзя измерить напрямую, поскольку он является результатом смешанного влияния нескольких факторов. Таким образом, в большинстве случаев $\tilde{u}(x, t)$ не доступны одновременно для всех x и всех моментов времени t . В работе разработан эффективный алгоритм определения функции источника $Q(x)$ в начально-краевой задаче (1) по дополнительной информации о процессе в фиксированные моменты времени.

В Институте системного программирования РАН разрабатываются подходы анализа социальных процессов, основанные на онтологии, графовых структурах, методах машинного обучения и искусственного интеллекта (см. [5]). В настоящей работе предложен упрощенный подход описания процесса распространения информации в онлайн социальных сетях, основанный на описании процессов посредством дифференциальных уравнений.

Статья организована следующим образом. В разд. 1 сформулирована обратная задача (задача об источнике) для модели (1) в классической и вариационной постановках и исследован линеаризованный оператор обратной задачи на степень некорректности. В разд. 2 сформулирован комбинированный алгоритм численного решения задачи об источнике, основанный на методах

тензорной оптимизации и градиентного типа. В разд. 3 приведены результаты и анализ численных расчетов определения функции источника распространения информации в онлайн социальных сетях по синтетическим данным о количестве вовлеченных пользователей в фиксированные моменты времени. В заключении сформулированы выводы по статье и направления дальнейших исследований.

1. ПОСТАНОВКА ОБРАТНОЙ ЗАДАЧИ

В обратной задаче, в отличие от прямой, помимо функции $u(x, t)$ неизвестной является и начальная функция плотности $Q(x)$. Рассматривается дополнительная информация следующего вида:

$$\sum_{i=1}^{N_1} u(x_i, t_k) = f_k, \quad k = 1, \dots, N_2. \quad (3)$$

Таким образом, в обратной задаче (1), (3) требуется восстановить функцию $Q(x)$ модели (1) по данным f_k вида (3). Так как данные обратной задачи не полны, такая обратная задача некорректна (см. [6]) (ее решение может быть неединственным и/или неустойчивым).

Благодаря неравенствам Карлемана, метод, введенный Бухгеймом и Клибановым (см. [7]), позволил установить важные результаты, включающие неравенства устойчивости, связывающие восстанавливаемую функцию источника с наблюдениями на части области $\omega \times (1, T)$, $\omega \subset (l_1, l_2)$, и наблюдения на всей области (l_1, l_2) в фиксированный момент времени t^* (см. [8, 9]). В [10] был предложен подход определения функции источника по дополнительной информации о решении в фиксированный момент времени и точки пространства в одномерном случае, получены результаты единственности и устойчивости решения. В [11] В. Исаков получил теорему условной устойчивости решения задачи об источнике для модели (2) с необходимостью дополнительных граничных наблюдений. В [12] А. Хасанов использовал подход слабого решения для минимизации функционала невязки. Больше теоретических и практических результатов приведены в современных работах и ссылках в них (см. [13, 14]).

Для описания быстротекущих процессов (движения плазмы, модели теплопроводности, распространение информации в социальных сетях) в [15, 16] предложен подход регуляризации параболической системы (2) путем добавления второй производной по времени с малым параметром ε в качестве коэффициента. В нашем случае

$$\begin{aligned} \varepsilon u_{tt} + u_t &= du_{xx} + \left(1 - \frac{u}{K}\right) r(t)u, \quad l_1 \leq x \leq l_2, \quad t \geq 1, \\ u(x, 1) &= Q(x), \quad u_t(x, 1) = 0, \quad l_1 \leq x \leq l_2, \\ u_x(l_1, t) &= u_x(l_2, t) = 0, \quad t \geq 1. \end{aligned} \quad (4)$$

Гиперболическая модель имеет заметные преимущества при использовании явных схем численного решения прямых и сопряженных задач. Особенно ярко эти преимущества по сравнению с параболической моделью (2) проявляются при использовании подробных пространственных сеток, применение которых стало возможным с появлением вычислительных систем сверхвысокой производительности. В то же время добавленный член имеет сингулярный характер. Вопрос о близости решений сингулярно возмущенных и невозмущенных уравнений впервые исследован в работах А.Н. Тихонова [17, 18]. В [16] показано, что оптимальным представляется выбор параметра ε в виде:

$$\varepsilon < \frac{h_x}{V},$$

где h_x — шаг по пространственной сетке, V — характерная скорость диффузионного процесса. При этом, с одной стороны, обеспечивается близость решений параболической и гиперболической моделей (см. [15]), а с другой — обеспечивается заметный вычислительный эффект при использовании явных схем.

1.1. Исследование сингулярных чисел оператора линеаризованной обратной задачи

Используя дискретизацию начально-краевой задачи (1), обратную задачу определения $q = (q_0, \dots, q_5) = (Q(1), \dots, Q(6))$ по дополнительной информации вида (3) можно представить в виде системы линейных алгебраических уравнений

$$Aq = f,$$

где A — матрица размера $N_2 \times N$, а q и $f = (f_1, \dots, f_{N_2})$ — векторы размерности N и N_2 соответственно. Обратную задачу (4), (3) аналогично можно представить в виде $A_\epsilon q = f$.

Зависимость решения от возмущения правой части f в случае невырожденной матрицы A для возмущенной задачи

$$A(q + \delta q) = f + \delta f$$

имеет вид $A\delta q = \delta f$, откуда $\delta q = A^{-1}\delta f$, $\|\delta q\| \leq \|A^{-1}\|\|\delta f\|$. Кроме того, $\|A\|\|q\| \geq \|f\|$.

Из этих соотношений мы имеем неулучшаемую оценку для относительной ошибки решения:

$$\frac{\|\delta q\|}{\|q\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta f\|}{\|f\|}. \quad (5)$$

Таким образом, погрешность определяется константой $\text{cond}(A) = \|A\|\|A^{-1}\|$, которая называется *числом обусловленности* системы (матрицы). Системы с плохо обусловленными матрицами можно считать практически неустойчивыми, хотя формально задача корректна, и выполнено условие устойчивости $\|A^{-1}\| < \infty$.

В случае возмущения матрицы оценка (5) принимает вид

$$\frac{\|\delta q\|}{\|q\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|} / \left(1 - \text{cond}(A) \frac{\|\delta A\|}{\|A\|}\right)$$

(при $\|A^{-1}\|\|\delta A\| < 1$).

Известно, что число обусловленности выражается через сингулярные числа матрицы следующим образом:

$$\text{cond}(A) = \frac{\sigma_1(A)}{\sigma_p(A)},$$

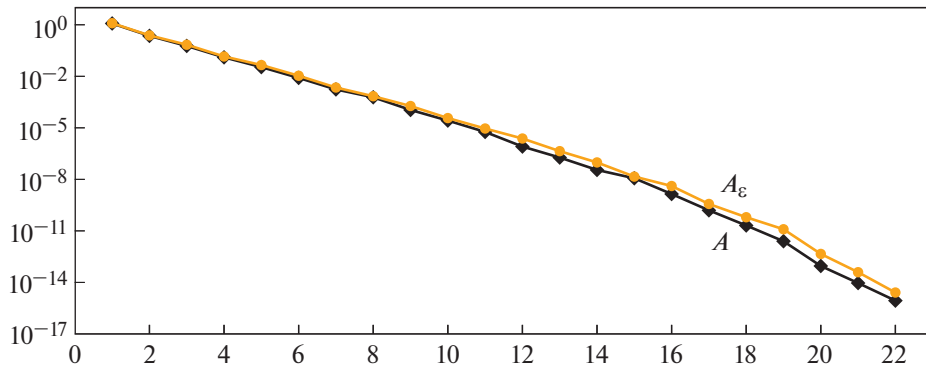
где $\sigma_1(A)$ и $\sigma_p(A)$ — максимальное и минимальное сингулярные числа матрицы A , $p = \min(N, N_2)$.

После линеаризации и дискретизации обратных задач (1), (3) (см. [19]) и (4), (3) при $N_1 = 5$, $N_2 = 22$ получим дискретные аналоги операторов обратных задач. В численных расчетах $K = 25$, $d = 0.01$, и функция $r(t)$ имеет вид:

$$r(t) = \frac{\beta_2}{\beta_1} - e^{-\beta_1(t-1)} \left(\frac{\beta_2}{\beta_1} - \beta_3 \right), \quad \text{где } \beta_1 = 1.5, \quad \beta_2 = 0.375, \quad \beta_3 = 1.65.$$

Сингулярные числа матриц A и A_ϵ представлены на фиг. 1.

Таким образом, числа обусловленности матриц A и A_ϵ имеют порядок 10^{16} и 10^{15} соответственно. Это означает, что решение линеаризованной обратной задачи неустойчиво. Добавление малого параметра ϵ со второй производной по времени не значительно уменьшает число обусловленности матрицы A_ϵ . В следующем разделе обратная задача сведена к задаче минимизации функционала с учетом регуляризирующего слагаемого.



Фиг. 1. Графики убывания сингулярных чисел операторов A (черная линия с ромбами) и A_ε при $\varepsilon = 0.1$ (оранжевая линия с кругами) линеаризованных обратных задач (1), (3) и (4), (3) в логарифмической шкале.

1.2. Вариационная постановка обратной задачи

Пусть $u(x, t) \in L^2((l_1, l_2) \times (1, +\infty))$. Тогда обратную задачу можно свести к соответствующей задаче минимизации следующего целевого функционала:

$$J(q) = \frac{T-1}{N_2} \sum_{k=1}^{N_2} \left| \sum_{i=1}^{N_1} u(x_i, t_k; q) - f_k \right|^2, \quad (6)$$

где $u(x, t; q)$ — решение прямой задачи для начальной функции плотности $Q(x)$, определяемой из набора параметров $q = (q_0, \dots, q_{N-1})$, т.е. $Q(x_i) = q_{i-1}$, $i = 1, \dots, N$, $N = 6$ — число параметров.

Для данной обратной задачи также был сформулирован регуляризирующий функционал А.Н. Тихонова (см. [17, 20])

$$J_T(q) = \frac{T-1}{N_2} \sum_{k=1}^{N_2} \left| \sum_{i=1}^{N_1} u(x_i, t_k; q) - f_k \right|^2 + \alpha \sum_{j=0}^{N-1} |q_j - q_j^0|^2 \quad (7)$$

с параметром регуляризации $\alpha = 10^{-2}$ и линейно убывающим q^0 .

2. МЕТОДЫ ОПТИМИЗАЦИИ

Задача определения вектора $q \in R^6$ является некорректной, а именно, как показано в п. 1.1, решение неустойчиво (число обусловленности достигает порядка 10^{16}). Более того, функция $Q(x)$ определяется на всем интервале $x \in [l_1, l_2]$ с помощью интерполяции вектора q кубическими сплайнами, что порождает неединственность решения.

Дополнительная информация (3) обратной задачи задана в дискретном виде.

В [19] были проведены расчеты для более полных данных $u(x_i, t_k) = f_{ik}$, $i = 1, \dots, N_1$, $k = 1, \dots, N_2$, с применением комбинации методов глобальной оптимизации (метод роя частиц) и градиентных подходов. Для решения поставленной обратной задачи был применен метод оптимизации тензорного поезда, который сводит функцию нескольких переменных к тензору (см. [21]) и является в сравнении с природоподобными алгоритмами эффективнее без потери точности для многомасштабных задач.

2.1. Метод тензорного разложения функционала

Метод глобальной оптимизации тензорного поезда (Tensor Train global optimization, ТТ) (см. [21]) заменяет исходную задачу оптимизации на эквивалентную задачу поиска максимума по модулю. Для этого исходная функция непрерывно и монотонно отображается на часть интервала $[0; +\infty)$ и после дискретизации преобразуется в тензор.

Однако число элементов в таком тензоре растет экспоненциально с ростом числа искомых параметров. Поэтому идеей метода является аппроксимация тензора методом TT-CROSS (см. [22]), который получает приближение тензора в TT-формате. Этот формат представляет собой разложение, в котором исходный тензор размерности N сводится к N тензорам размерности 3.

Для вычислительных экспериментов был адаптирован пакет программ ttpu на языке программирования Python. В данном случае исходный функционал $J(q)$ отображается с помощью функции $g(J(q) - \alpha)$ и вводится сетка с n узлами по каждому из N направлений, а тензор значений функции $g(q)$ обозначается через G . Тогда алгоритм метода имеет следующий вид:

0. Для каждого $i = 1, \dots, N - 1$:

- Случайным образом генерируется тензор $\hat{G}_i(j_1; j_2; j_3)$ и считается QR -разложение матрицы развертки $\hat{G}_i(j_1 j_2; j_3) = Q_i R_i$.

- Процедура $\text{maxvol}(Q_i)$ находит набор номеров строк $\{p\}$, соответствующий квадратной подматрице максимального объема (см. [23]).

- Используя значения сетки и полученный на предыдущем шаге $\hat{q}_{i-1}$, генерируется массив $\hat{q}_i$, из которого выбираются строки с номерами из $\{p\}$.

1. Пока число итераций меньше n_{swp} (параметр функции `min_func` в пакете ttpu), чередуются следующие действия:

(а) Для $i = N - 1, \dots, 1$:

- На основе $\hat{q}_i$ и $\hat{q}_{i+1}$ генерируется множество потенциальных решений M и обновляется сдвиг α .

- Массив значений функционала $J(\tilde{q})$, $\tilde{q} \in M$, представляется в виде тензора $A_i(j_1; j_2; j_3)$ и считается сингулярное разложение матрицы развертки $A_i(j_2 j_3; j_1) = U_i S_i (V_i)^*$.

- Процедура $\text{gest_maxvol}(U_i)$ находит набор номеров строк $\{p\}$, соответствующий прямоугольной подматрице максимального объема (см. [24]).

- Используя значения сетки и $\hat{q}_{i+1}$, генерируется массив $\hat{q}_i$, из которого выбираются строки с номерами из $\{p\}$.

(б) Для $i = 0, \dots, N - 2$:

- На основе $\hat{q}_i$ и $\hat{q}_{i+1}$ генерируется множество потенциальных решений M и обновляется сдвиг α .

- Массив значений функционала $J(\tilde{q})$, $\tilde{q} \in M$, представляется в виде тензора $A_i(j_1; j_2; j_3)$ и считается QR -разложение матрицы развертки $A_i(j_1 j_2; j_3) = Q_i R_i$.

- Процедура $\text{gest_maxvol}(Q_i)$ находит набор номеров строк $\{p\}$, соответствующий прямоугольной подматрице максимального объема.

- Используя значения сетки и $\hat{q}_i$, генерируется массив $\hat{q}_{i+1}$, из которого выбираются строки с номерами из $\{p\}$.

На нечетных итерациях выполняется вариант (а), на четных итерациях – (б).

Данный метод использует только некоторые строки и столбцы изначального тензора значений функции $g(q)$, из-за чего может не найти точку глобального оптимума, а лишь обнаружить близкое к оптимальному значение. Поэтому для увеличения точности решения обратной задачи мы рассмотрели комбинированный метод.

2.2. Комбинированный метод

Данный метод состоит в последовательном применении метода глобальной оптимизации ТТ и локального многоуровневого градиентного метода (МГМ) (см. [25]). Ранее уже были проведены численные эксперименты для подобной обратной задачи, решаемой с помощью МГМ (см. [26]). Алгоритм МГМ имеет следующий вид:

$$q^{m+1} = \theta_{m+1} z^m + (1 - \theta_{m+1}) y^m, \quad \theta_{m+1} \in \arg \min_{\theta \in [0;1]} J(\theta z^m + (1 - \theta) y^m),$$

$$y^{m+1} = q^{m+1} - \zeta_{m+1} J'(q^{m+1}), \quad \zeta_{m+1} \in \arg \min_{\zeta \geq 0} J(q^{m+1} - \zeta J'(q^{m+1})),$$

$$z^{m+1} = z^m - \eta_{m+1} J'(q^{m+1}), \quad \eta_{m+1} = \frac{1}{2L_{m+1}} + \sqrt{\frac{1}{4L_{m+1}^2} + \eta_m^2},$$

$$L_{m+1} = \frac{\|J'(q^{m+1})\|^2}{2(J(q^{m+1}) - J(y^{m+1}))}, \quad \eta_0 = 0.$$

В качестве начального приближения для МГМ был взят результат метода ТТ. Скорость сходимости по функционалу метода МГМ имеет порядок $O(m^{-2})$.

3. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Решение задач минимизации (6) и (7) проводилось с помощью методов ТТ, ТТ в комбинации с МГМ (ТТ + МГМ). Сравнение результатов проведено с результатом комбинации методов роя частиц и МГМ для синтетических данных обратной задачи. Вычисления проводились на вычислительном узле с двумя процессорами Intel Xeon Gold 6248R, количество ядер в каждом процессоре – 24, частота – 3 Гц, 384 Гб оперативной памяти.

3.1. Входные параметры

Прямая задача решается конечно-разностным методом на равномерной сетке в замкнутой области $\bar{D} = \{(x, t) \mid l_1 \leq x \leq l_2, 1 \leq t \leq T\}$:

$$\bar{\omega} = \{(x_j, t_n) \mid x_j = l_1 + jh, t_n = 1 + n\tau, j = 0, \dots, N_x, n = 0, \dots, N_t\},$$

где $h = \frac{l_2 - l_1}{N_x}$ и $\tau = \frac{T - 1}{N_t}$.

Для применения классического непрерывного подхода функция начальной плотности $Q(x)$ определяется из вектора q_{ex} (см. табл. 2) интерполяцией кубическими сплайнами (см. [26]).

Прямая задача (1) решается с помощью явной конечно-разностной схемы с порядком аппроксимации $O(\tau + h^2)$.

Положим в численных расчетах $l_1 = 1, l_2 = 6, T = 24, N_x = 50$ и $N_t = 575$ (см. [26]). И в случае функционала А.Н. Тихонова $q_j^0 \in [0, 6]$ расположены равномерно, т.е. $q^0 = (6, 4.8, 3.6, 2.4, 1.2, 0)$.

В качестве синтетических данных f_k были взяты значения решения прямой задачи в каждой десятой точке по x и в каждой двадцать пятой точке по $t \geq 50$, т.е. $N_1 = 6$ и $N_2 = 22$. Значения параметров q_{ex} в табл. 2 эмитируют реальные данные новостного сайта Digg.com, представленные в [2].

Относительная ошибка вычислялась по формуле

$$\delta_m = \frac{\|q_{\text{ex}} - q^m\|^2}{\|q_{\text{ex}}\|^2}.$$

Здесь q_{ex} – точные значения дискретизованной функции $Q(x)$, а q^m – аппроксимация решения обратной задачи, которая соответствует минимуму функционала $J(q^m)$ (6).

3.1.1. Параметры ТТ. Исходный функционал отображался с помощью функции $g(q) = \pi/2 - \arctg(J(q) - \alpha)$. Для всех N направлений выбирался интервал изменения параметров $[0, 6]$, а максимально возможный ранг тензоров в разложении $r_{\text{max}} = 7$ и количество итераций во всех численных экспериментах $n_{\text{swp}} = 10$.

При увеличении количества узлов n по каждому из направлений растет вычислительная сложность (см. п. 3.3), поэтому в расчетах было выбрано оптимальное по времени и значению целевого функционала число узлов $n = 51$.

3.1.2. Параметры МГМ. На каждой итерации параметр спуска определяется на отрезке $[0, \zeta_{\text{max}}]$ с шагом ζ_n . А градиент вычислялся через решение сопряженной задачи (см. [26]).

Таблица 1. Результаты, $t_{ЭВМ}$ — время работы программы

Метод	Функционал	δ	$t_{ЭВМ}$, ч
ТТ	$J(q) = 6.16 \times 10^0$	2.04×10^{-2}	1.09
ТТ + МГМ	$J(q) = 7.43 \times 10^{-3}$	1.89×10^{-2}	1.1
ТТ + МГМ + рег.	$J_T(q) = 1.56 \times 10^{-1}$	1.29×10^{-2}	2.55
МРЧ + МГМ	$J(q) = 6.67 \times 10^{-2}$	1.03×10^0	0.64

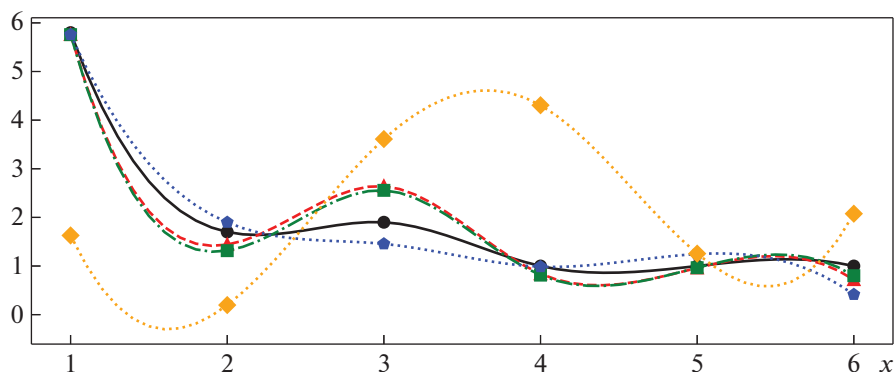
Таблица 2. Значения точного q_{ex} и найденных решений обратной задачи (1), (3)

Параметр	q_0	q_1	q_2	q_3	q_4	q_5
q_{ex}	5.8	1.7	1.9	1	0.95	0.7
$q_{ТТ}$	5.76	1.44	2.639	0.84	0.96	0.72
$q_{ТТ+МГМ}$	5.758	1.316	2.556	0.813	0.965	0.801
$q_{ТТ+МГМ+рег.}$	5.759	1.899	1.456	0.986	1.246	0.415
$q_{МРЧ+МГМ}$	1.629	0.197	3.611	4.307	1.255	2.076

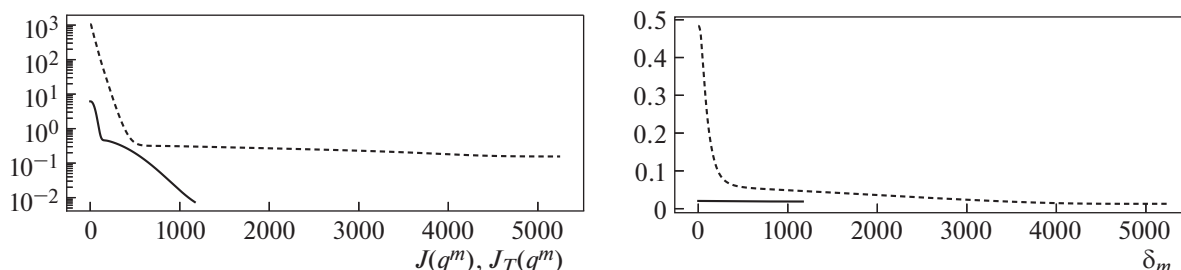
3.2. Результаты численных экспериментов

Было выявлено, что при значениях $n \leq 50$ изменение параметров $r_{\max}$ и n_{swp} никак не влияет на результат работы программы, лишь увеличивая время ее работы. А минимальные значения функционала $J(q) = 6.16 \times 10^0$ и погрешности $\delta = 2.04 \times 10^{-2}$ достигались при $n = 51$, $r_{\max} = 7$ и $n_{\text{swp}} = 10$. Однако решение обратной задачи еще далеко от точного, поэтому для уточнения применялся МГМ.

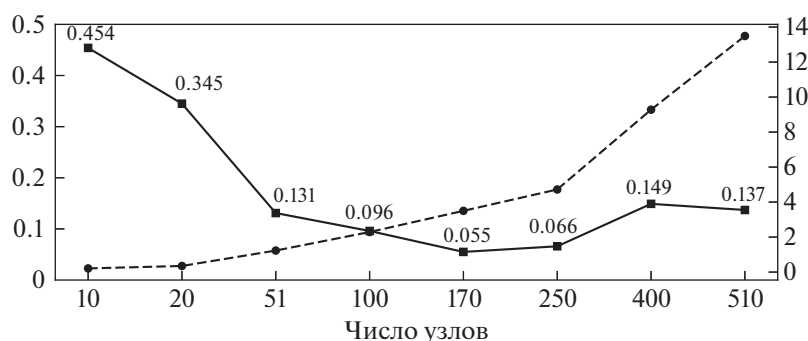
В табл. 1 и 2 и на фиг. 2 представлены результаты решения задачи минимизации целевого функционала методом ТТ, комбинацией ТТ + МГМ, ТТ + МГМ с регуляризацией А.Н. Тихонова и комбинацией стохастического глобального метода роя частиц (МРЧ) + МГМ. Можно заметить, что метод ТТ работает почти в 2 раза дольше комбинации МРЧ + МГМ, но находит решение лучше в смысле значений погрешности. Применение МГМ к методу ТТ приводит к уменьшению функционала, но незначительно влияет на погрешность, а применение регуляризации А.Н. Тихонова уменьшает погрешность, но увеличивает время работы программы более чем в



Фиг. 2. Восстановленная начальная функция плотности активных пользователей $Q(x)$. Непрерывной черной линией с кругами изображено точное решение обратной задачи, красной штриховой линией с треугольниками — решение, полученное методом ТТ, зеленой пунктирной линией с квадратами — решение, полученное комбинацией ТТ + МГМ, синей пунктирной линией с пятиугольниками — решение, полученное с помощью комбинации ТТ + МГМ с регуляризацией А.Н. Тихонова, и оранжевой пунктирной линией с ромбами изображено решение, полученное комбинацией МРЧ + МГМ.



Фиг. 3. Графики зависимости функционалов (в логарифмической шкале) и относительной погрешности МГМ от номера итерации m с начальным приближением в виде решения, полученного методом ТТ для случаев без (непрерывная линия) и с регуляризацией А.Н. Тихонова (пунктирная линия).



Фиг. 4. Непрерывная линия с квадратами демонстрирует убывание функционала с увеличением числа узлов, а штриховая линия с кругами — график возрастания времени работы программы с увеличением числа узлов.

2 раза. Таким образом, наименьшее значение погрешности $\delta = 1.29 \times 10^{-2}$ достигается при работе регуляризованного комбинированного метода ТТ + МГМ.

На фиг. 3 непрерывными линиями изображены кривые, полученные комбинацией ТТ + МГМ, а пунктирной линией — кривые, полученные комбинированным методом с регуляризацией А.Н. Тихонова. Из фиг. 3 следует, что функционал (6) убывает быстрее, а функционал (7) не достигает минимального значения. Это связано с выбором q^0 , что не позволяет функционалу достигнуть похожих значений. А погрешность регуляризованного комбинированного метода хоть и имеет сравнительно большие значения на первых итерациях, но достигает значения меньшего, чем в случае отсутствия регуляризации.

3.3. Вычислительная сложность

Был проведен ряд численных экспериментов для фиксированных параметров метода ТТ $r_{\max} = 7$ и $n_{\text{swp}} = 10$ и набора $n = \{10, 20, 51, 100, 170, 250, 400, 510\}$, который показал, что при увеличении числа узлов n метода ТТ уменьшаются значения функционала и погрешности, но критически увеличивается время работы программы, что продемонстрировано на фиг. 4. В результате было выбрано значение $n = 51$ с учетом оптимального времени работы и приемлемого значения функционала.

4. ЗАКЛЮЧЕНИЕ

Математическая модель типа “реакции–диффузии”, возникающая в биологии, экологии и других приложениях, была адаптирована для описания процесса распространения информации в онлайн социальных сетях. Особенность моделирования состояла в описании социальных процессов в дискретном пространстве с неизвестным источником. В работе численно исследована и решена задача восстановления источника распространения информации по синтетическим

данным интегрального типа в фиксированные моменты времени. Анализ степени убывания сингулярных чисел линеаризованного оператора показал необходимость применения методов регуляризации. Разработанный алгоритм решения обратной задачи основан на комбинации вариации метода покрытий и градиентного метода с применением регуляризации А.Н. Тихонова. Было проведено сравнение с природоподобным алгоритмом и показано преимущество использования комбинации метода тензорной оптимизации и градиентного метода с регуляризацией А.Н. Тихонова.

СПИСОК ЛИТЕРАТУРЫ

1. Четверушкин Б.Н., Осипов В.П., Балута В.И. Подходы к моделированию последствий принятия решений в условиях противодействия // Препринт ИПМ им. М.В. Келдыша. 2018. № 43. 15 с.
2. Wang F., Wang H., Xu K., Wu J., Jia X. Characterizing Information Diffusion in Online Social Networks with Linear Diffusive Model // Proceed. of ICDSCS. 2013. P. 307–316.
3. Колмогоров А.Н., Петровский И.Г., Пискунов Н.С. Исследование уравнения диффузии, соединенной с возрастанием вещества, и его применение к одной биологической проблеме // Бюлл. МГУ. Сер. А. Математика и механика. 1937. Т. 1. № 6. С. 1–26.
4. Владимиров В.С. Уравнения математической физики. М.: Наука, 1981.
5. Самохвалов Д.И. Определение аккаунтов злоумышленников в социальной сети ВКонтакте при помощи методов машинного обучения // Тр. Института системного программирования РАН. 2020. Т. 32. № 3. С. 109–117.
6. Kabanikhin S. Definitions and examples of inverse and ill-posed problems // J. Inverse Ill-Posed Probl. 2009. V. 16. № 4. P. 317–357.
7. Бухгейм А.Л., Клибанов М.В. Единственность в целом одного класса многомерных обратных задач // Докл. АН СССР. 1981. Т. 260. № 2. С. 269–272.
8. Yamamoto M., Zou J. Simultaneous reconstruction of the initial temperature and heat radiative coefficient // Inverse Problems. 2001. V. 17. P. 1181.
9. Bellassoued M., Yamamoto M. Inverse source problem for a transmission problem for a parabolic equation // J. Inverse Ill-Posed Probl. 2006. V. 14(1). P. 47–56.
10. Cristofol M., Garnier J., Hamel F., Roques L. Uniqueness from pointwise observations in a multi-parameter inverse problem // Commun. Pure and Appl. Analys. 2012. V. 11(1). P. 173–188.
11. Isakov V. Inverse Problems for Partial Differential Equations. New York: Springer, 2017.
12. Hasanov A. Simultaneous determination of source terms in a linear parabolic problem from the final overdetermination: Weak solution approach // J. Math. Anal. Appl. 2007. V. 330. Iss. 2. P. 766–779.
13. Penenko A., Mukatova Z. Inverse modeling of diffusion-reaction processes with image-type measurement data // 2018 11th Inter. Multiconference Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB). 2018. P. 39–43.
14. Kaltenbacher B., Rundell W. The inverse problem of reconstructing reaction-diffusion systems // Inverse Problems. 2020. V. 36. P. 065011.
15. Моисеев Т.Е., Мышечная Е.Е., Тишкин В.Ф. О близости решений невозмущенных гиперболизированных уравнений теплопроводности для разрывных начальных данных // Докл. АН. Математика. 2018. Т. 481. № 6. С. 605–609.
16. Четверушкин Б.Н., Ольховская О.Г. Моделирование процесса лучистой теплопроводности на высокопроизводительных вычислительных системах // Докл. АН. Математика, информатика, процессы управления. 2020. Т. 491. № 1. С. 111–114.
17. Тихонов А.Н. Об устойчивости обратных задач // Докл. АН СССР. 1943. Т. 39. № 5. С. 195–198.
18. Тихонов А.Н. О зависимости решений дифференциальных уравнений от малого параметра // Матем. сб. 1948. Т. 22(64). № 2. С. 193–204.
19. Krivorotko O., Zvonareva T., Zyatkov N. Numerical solution of the inverse problem for diffusion-logistic model arising in online social networks // Commun. Comput. Info. Sci. 2021. V. 1476. P. 444–459.
20. Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г. Регуляризирующие алгоритмы и априорная информация. М.: Наука, 1983.

21. Zheltkova V.V., Zheltkov D.A., Grossman Z., Bocharov G.A., Tyrtysnikov E.E. Tensor based approach to the numerical treatment of the parameter estimation problems in mathematical immunology // J. Inverse Ill-Posed Probl. 2018. V. 26. № 1. P. 51–66.
22. Oseledets I.V., Tyrtysnikov E.E. TT-cross approximation for multidimensional arrays // Linear Algebra Appl. 2010. V. 432. № 1. P. 70–88.
23. Goreinov S.A., Oseledets I.V., Savostyanov D.V., Tyrtysnikov E.E., Zamarashkin N.L. How to Find a Good Submatrix // Matrix Methods: Theory, Algorithms and Appl. 2010. P. 247–256.
24. Mikhalev A., Oseledets I. Rectangular maximum-volume submatrices and their applications // Linear Algebra Appl. 2018. V. 538. P. 187–211.
25. Gasnikov A.V., Nesterov Y.E. Universal method for stochastic composite optimization problems // Comput. Math. Math. Phys. 2018. V. 58. № 1. P. 48–64.
26. Звонарева Т.А., Криворотько О.И. Сравнительный анализ градиентных методов определения источника диффузионно-логистической модели // Ж. вычисл. матем. и матем. физ. 2022. Т. 62. № 4. С. 694–704.

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

О ВЗАИМОДЕЙСТВИИ ГРАНИЧНЫХ ОСОБЫХ ТОЧЕК В ОДНОЙ ЭЛЛИПТИЧЕСКОЙ КРАЕВОЙ ЗАДАЧЕ

© 2023 г. А. М. Боговский^{1,*}

¹ 119991 Москва, Ленинские горы, МГУ, ф-т ВМК, Россия

*e-mail: abogovski@gmail.com

Поступила в редакцию 20.02.2023 г.
Переработанный вариант 03.05.2023 г.
Принята к публикации 29.05.2023 г.

Статья продолжает построение L_p -теории эллиптических краевых задач Дирихле и Неймана с разрывными кусочно-постоянными коэффициентами в дивергентной форме для неограниченной области $\Omega \subset \mathbb{R}^2$ с кусочно- C^1 -некомпактной липшицевой границей и C^1 -гладкими линиями разрыва коэффициентов. Ранее построенная L_p -теория обобщается на случай несовпадения наименьших собственных значений, соответствующих конечной и бесконечной особым точкам, продолжая исследование эффекта их взаимодействия в функциональном классе с первыми производными из $L_p(\Omega)$ во всей шкале значений показателя $p \in (1, \infty)$. Библ. 3. Фиг. 1.

Ключевые слова: эллиптическое уравнение в дивергентной форме, разрывный кусочно-постоянный коэффициент, неограниченная область, кусочно-гладкая некомпактная липшицева граница, гладкие линии разрыва коэффициента, задача Дирихле, задача Неймана, слабое решение с первыми производными из L_p , L_p -теория, взаимодействие особенностей,

DOI: 10.31857/S0044466923090041, EDN: RJMDHQ

ВВЕДЕНИЕ

В неограниченной плоской области $\Omega \subset \mathbb{R}^2$ с одним выходом на бесконечность и кусочно-гладкой некомпактной границей $\partial\Omega$, состоящей из конечного числа гладких кривых, рассматривается краевая задача Дирихле для эллиптического уравнения в дивергентной форме с разрывным скалярным кусочно-постоянным коэффициентом $\kappa > 0$, имеющим конечное число подобластей непрерывности в Ω , ограниченных кусочно- C^1 гладкими линиями разрыва. Задача Дирихле рассматривается в слабой постановке, соответствующей функциональному классу решений с первыми производными из L_p во всей шкале значений показателя $p \in (1, \infty)$. Статья продолжает начатое в [1] исследование вопросов существования и единственности решений для случая одной конечной и одной бесконечной граничных особых точек.

Конечной граничной особой точкой будем называть точку негладкости $\partial\Omega$ и/или вхождения линий разрыва коэффициента κ в $\partial\Omega$. Углы, образуемые входящими в особую точку кривыми границы и линиями разрыва коэффициента, предполагаются ненулевыми.

Бесконечной граничной особой точкой будем называть связную компоненту окрестности бесконечности в Ω с образующими ненулевой угол асимптотами $\partial\Omega$ и, возможно, асимптотами линий разрыва κ в этой компоненте. Такое определение допускает области с несколькими выходами на бесконечность. В рамках данной статьи ограничимся единственным выходом Ω на бесконечность.

Ограничимся также рассмотрением особых точек, допускающих локальное разделение переменных в полярных координатах в некоторой своей окрестности. Переход к общему случаю производится с помощью известных L_p -оценок и теоремы об устойчивости индекса при малых возмущениях в операторной норме.

Полученный результат является обобщением полученного ранее в [1]. Теперь допускается неравенство наименьших собственных значений $\lambda_1, \tilde{\lambda}_1$, возникающих в задаче Штурма–Лиувилля по полярному углу. При этом возникают интервалы значений p , на которых эллиптический оператор $L : \mathring{L}_p^1(\Omega) \rightarrow \mathring{L}_p^{-1}(\Omega)$ имеет конечные ненулевые размерности ядра и коядра.

Отметим, что приведенные здесь и полученные ранее результаты переносятся и на задачу Неймана в смысле установленной в [2] взаимосвязи разрешающих операторов.

1. ПОСТАНОВКА ЗАДАЧИ

Определим пространство $\mathring{L}_p^1(\Omega)$ как замыкание подпространства $\mathring{C}^\infty(\Omega)$ в пространстве Соболева $L_p^1(\Omega)$ с нормой

$$\|u\|_{\mathring{L}_p^1(\Omega)} \stackrel{\text{def}}{=} \|\nabla u\|_{L_p(\Omega)} + \|u\|_{L_p(\omega)},$$

где ω — какая-либо фиксированная ограниченная подобласть $\omega \subset \Omega$, и введем эквивалентную норму $\|u\|_{\mathring{L}_p^1(\Omega)} \stackrel{\text{def}}{=} \|\nabla u\|_{L_p(\Omega)}$. Пространство $\mathring{L}_p^{-1}(\Omega)$ определим как двойственное к $\mathring{L}_p^1(\Omega)$ пространство линейных непрерывных функционалов, т.е.

$$\mathring{L}_p^{-1}(\Omega) \stackrel{\text{def}}{=} (L_q^1(\Omega))^*, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Понимая обобщенную дивергенцию в смысле $\mathcal{D}'(\Omega)$, рассмотрим дифференциальный оператор

$$L = \text{div}(\kappa \nabla \cdot) : \mathring{L}_p^1(\Omega) \rightarrow \mathring{L}_p^{-1}(\Omega), \quad (1.1)$$

позволяющий записать в операторной форме $Lu = f$ слабую постановку задачи Дирихле

$$\text{div}(\kappa \nabla u) = f, \quad u|_{\partial\Omega} = 0.$$

Операторная постановка проясняет связь размерностей $\dim \ker$ и $\dim \text{coker}$ ядра и коядра

$$\ker L \stackrel{\text{def}}{=} \{u \in \mathring{L}_p^1(\Omega) \mid Lu = 0\},$$

$$\text{coker } L \stackrel{\text{def}}{=} \mathring{L}_p^{-1}(\Omega) / \{Lu \mid u \in \mathring{L}_p^1(\Omega)\}$$

линейного оператора (1.1) в случае, когда эти размерности конечны. В свою очередь, операторная постановка эквивалентна слабой L_p -постановке той же задачи Дирихле

$$\int_{\Omega} (\kappa \nabla u, \nabla \psi) dx = -f(\psi) \quad \forall \psi \in \mathring{L}_p^1(\Omega), \quad \frac{1}{p} + \frac{1}{q} = 1,$$

для класса слабых решений $u \in \mathring{L}_p^1(\Omega)$. При этом, как хорошо известно, для всякого функционала $f \in \mathring{L}_p^{-1}(\Omega)$ найдется векторное поле $\mathbf{F} \in \mathbf{L}_p(\Omega)$ такое, что $f = \text{div } \mathbf{F}$ в смысле $\mathcal{D}'(\Omega)$ с оценкой

$$\|\mathbf{F}\|_{\mathbf{L}_p(\Omega)} \leq \|f\|_{\mathring{L}_p^{-1}(\Omega)}.$$

Определение 1. Для произвольной области $\Omega \subset \mathbb{R}^2$ элемент $u \in \mathring{L}_p^1(\Omega)$ будем называть *слабым решением* задачи Дирихле, если выполняется интегральное тождество

$$\int_{\Omega} \kappa(x) (\nabla u, \nabla \psi) dx = \int_{\Omega} (\mathbf{F}, \nabla \psi) dx \quad \forall \psi \in \mathring{L}_q^1(\Omega) \quad (1.2)$$

при заданной вектор-функции $\mathbf{F} \in \mathbf{L}_p(\Omega)$, где $(\cdot, \cdot)$ — скалярное произведение в $\mathbb{R}^2$, $q = p/(p-1)$ — показатель, сопряженный к $p \in (1, \infty)$.

Отметим, что в случае ограниченной области $\Omega \subset \mathbb{R}^2$ класс пробных функций $\mathring{L}_q^1(\Omega)$ может быть заменен на пространство Соболева $\mathring{W}_q^1(\Omega)$.

Замечание 1. При построении примеров несуществования слабых решений можно ограничиться рассмотрением правых частей $\operatorname{div} \mathbf{F} = f \in \mathring{C}^\infty(\Omega)$. В таком случае тождество (1.2) примет более простой вид:

$$\int_{\Omega} \kappa(x)(\nabla u, \nabla \psi) dx = - \int_{\Omega} f \psi dx \quad \forall \psi \in \mathring{L}_q^1(\Omega). \quad (1.3)$$

2. ЗАДАЧА ШТУРМА–ЛИУВИЛЛЯ

При разделении переменных в полярных координатах по полярному углу возникает задача Штурма–Лиувилля. Напомним постановку этой задачи и приведем здесь без доказательства некоторые ее свойства, установленные в [1].

Взяв в качестве веса кусочно-постоянный коэффициент $\kappa = \kappa(\varphi) > 0$ с конечным числом линий разрыва, введем пространство Лебега $L_{2,\kappa}(0, \alpha)$ со скалярным произведением

$$(\Phi, \Psi)_{\kappa} = \int_0^{\alpha} \kappa(\varphi) \Phi(\varphi) \Psi(\varphi) d\varphi,$$

порождающим соответствующую весовую норму $\|\Phi\|_{2,\kappa}$.

На $L_{2,\kappa}(0, \alpha)$ введем операцию слабого дифференцирования

$$\mathcal{L} = \frac{1}{\kappa} (\kappa \Phi)', \quad (2.1)$$

где штрих означает слабое дифференцирование на $L_{1,\text{loc}}(\mathbb{R})$ или, в более широком понимании, дифференцирование в смысле $\mathcal{D}'(\mathbb{R})$ с требованием регулярности обобщенной производной.

Рассматривая операцию дифференцирования (2.1) на подпространстве

$$D_{\mathcal{L}} \stackrel{\text{def}}{=} \{\Phi \in \mathring{W}_2^1(0, \alpha) : \kappa \Phi' \in \mathring{W}_2^1(0, \alpha)\} \subset L_{2,\kappa}(0, \alpha),$$

приходим к замкнутому неограниченному дифференциальному оператору

$$\mathcal{L} = \frac{1}{\kappa} (\kappa \Phi)' : D_{\mathcal{L}} \subset L_{2,\kappa}(0, \alpha) \rightarrow L_{2,\kappa}(0, \alpha). \quad (2.2)$$

Справедливы следующие утверждения.

Теорема 1. Геометрическая кратность всех собственных значений задачи Штурма–Лиувилля $\mathcal{L}u + \lambda u = 0$ равна единице.

Теорема 2. Система собственных функций задачи Штурма–Лиувилля является ортогональным базисом в $\mathring{W}_{2,\kappa}^1(0, \alpha)$.

Занумеруем собственные функции $\{\Phi_k\}_{k=1}^{\infty}$ по возрастанию чисел $\lambda_k > 0$. Верны утверждения следующих трех теорем и леммы.

Теорема 3. Существует зависящая только от чисел α , $\kappa_{\min}$, $\kappa_{\max}$ постоянная $M > 0$ такая, что

$$\sum_{k=1}^{\infty} \frac{1}{\lambda_k^2} \leq M.$$

Теорема 4. Собственная функция Φ_1 задачи Штурма–Лиувилля, соответствующая λ_1 , не имеет нулей на $(0, \alpha)$.

Теорема 5. Собственные функции Φ_k , $k \geq 2$, задачи Штурма–Лиувилля меняют знак на $(0, \alpha)$ в силу $\lambda_k > \lambda_1$.

Лемма 1. Для собственных функций Φ_k задачи Штурма–Лиувилля верны неравенства

$$\exists C > 0: \left| \frac{\Phi_k(\varphi)}{\Phi_1(\varphi)} \right| \leq C \lambda_k \quad \forall k \geq 2.$$

3. ОСОБЫЕ ТОЧКИ

Напомним, что рассматриваются уже описанные во введении конечная и бесконечная граничные особые точки, допускающие локальное разделение переменных в полярных координатах. При этом ключевым вопросом в вычислении $\dim \ker L$ является количество значений $\lambda_k \in (0, 1)$ в соответствующих особым точкам задачах Штурма—Лиувилля по φ . Особые точки, имеющие значения $\lambda_k \in (0, 1)$, будем называть *сингулярными*, а не имеющие таковых — *регулярными*. Простой пример сингулярной особой точки можно найти в [1].

Дополнительно потребовав $\lambda_1 < 1 < \lambda_2$ и $\tilde{\lambda}_1 < 1 < \tilde{\lambda}_2$, представим решение $u \in \mathring{L}_p^1(\Omega)$ в окрестностях конечной и бесконечной особых точек в виде

$$u(r, \varphi) = \sum_{k=1}^{\infty} A_k r^{\mu_k} \Phi_k(\varphi) + B_1 r^{-\mu_1} \Phi_1(\varphi), \quad 0 < r < \delta, \quad \mu_k = \sqrt{\lambda_k}, \quad (3.1)$$

$$u(\tilde{r}, \tilde{\varphi}) = \tilde{A}_1 \tilde{r}^{\tilde{\mu}_1} \tilde{\Phi}_1(\tilde{\varphi}) + \sum_{k=1}^{\infty} \tilde{B}_k \tilde{r}^{-\tilde{\mu}_k} \tilde{\Phi}_k(\tilde{\varphi}), \quad \tilde{r} > R > \delta, \quad \tilde{\mu}_k = \sqrt{\tilde{\lambda}_k}. \quad (3.2)$$

При исследовании взаимодействия конечной и бесконечной особых точек понадобятся следующие свойства локальной знакоопределенности/знакопеременности решения.

Лемма 2 (знакоопределенность в окрестности особой точки).

1. Пусть $p < 2/(1 + \mu_1)$ и, соответственно, в окрестности конечной особой точки решение $u(r, \varphi)$ имеет вид

$$u(r, \varphi) = \sum_{k=1}^{\infty} A_k r^{\mu_k} \Phi_k(\varphi) + B_1 r^{-\mu_1} \Phi_1(\varphi), \quad 0 < r < \delta < 1.$$

Если $B_1 > 0$, то $u(r, \varphi) \geq 0$ в некоторой окрестности этой особой точки.

2. Пусть $p > 2/(1 - \tilde{\mu}_1)$ и, соответственно, в окрестности бесконечной особой точки решение $u(\tilde{r}, \tilde{\varphi})$ имеет вид

$$u(\tilde{r}, \tilde{\varphi}) = \tilde{A}_1 \tilde{r}^{\tilde{\mu}_1} \tilde{\Phi}_1(\tilde{\varphi}) + \sum_{k=1}^{\infty} \tilde{B}_k \tilde{r}^{-\tilde{\mu}_k} \tilde{\Phi}_k(\tilde{\varphi}), \quad \tilde{r} > R > 1.$$

Если $\tilde{A}_1 > 0$, то $u(\tilde{r}, \tilde{\varphi}) \geq 0$ в некоторой окрестности этой особой точки.

Доказательство. Доказательства 1 и 2 аналогичны. Приведем доказательство второго утверждения леммы, для удобства убрав тильды из обозначений.

Оценим решение, переписав его в виде

$$u(r, \varphi) = \Phi_1(\varphi) \left[A_1 r^{\mu_1} + B_1 r^{-\mu_1} + \sum_{k=2}^{\infty} B_k r^{-\mu_k} \frac{\Phi_k(\varphi)}{\Phi_1(\varphi)} \right].$$

Заметим, что коэффициенты B_k разложения в ряд Фурье по ортонормированному базису в $L_{2,\infty}(0, \alpha)$ ограничены в силу, например,

$$|B_k| = \left| \int_0^\alpha u(R, \varphi) \Phi_k(\varphi) d\varphi \right| \leq \|u\|_{r=R} \| \Phi_k \|_{L_{2,\infty}(0, \alpha)}.$$

Используя оценку $|\Phi_k(\varphi)/\Phi_1(\varphi)| \leq C \mu_k^2 \quad \forall k \geq 1$ из леммы 1, получаем

$$\left| B_1 r^{-\mu_1} + \sum_{k=2}^{\infty} B_k r^{-\mu_k} \frac{\Phi_k}{\Phi_1} \right| \leq |B_1| r^{-\mu_1} + M \sum_{k=2}^{\infty} \mu_k^2 r^{-\mu_k} \quad \forall r > R.$$

При этом ряд

$$\sum_{k=2}^{\infty} \mu_k^2 r^{-\mu_k} = r^{-\mu_1} \sum_{k=2}^{\infty} \mu_k^{-4} \mu_k^6 r^{\mu_1 - \mu_k}$$

сходится и имеет сумму порядка $O(r^{-\mu_1})$ при $r \rightarrow \infty$ в силу теоремы 3 о сходимости ряда $\sum \mu_k^{-4}$ и очевидного неравенства

$$\sup_{s>1} (s^6 r^{\mu_1-s}) < m(R) < \infty \quad \forall r > R$$

с некоторой постоянной $m(R) > 0$.

Таким образом, $u(r, \varphi) = \Phi_1(\varphi)[A_1 r^{\mu_1} + O(r^{-\mu_1})]$, $r \rightarrow \infty$. А по теореме 4 функция Φ_1 знакопостоянна на $(0, \alpha)$. Лемма доказана.

Лемма 3 (смена знака в окрестности особой точки).

1. Пусть в окрестности конечной особой точки решение $u(r, \varphi)$ имеет вид

$$u(r, \varphi) = \sum_{k=2}^{\infty} A_k r^{\mu_k} \Phi_k(\varphi), \quad 0 < r < \delta < 1.$$

Тогда $u(r, \varphi)$ меняет знак в любой окрестности этой особой точки.

2. Пусть в окрестности бесконечной особой точки решение $u(\tilde{r}, \tilde{\varphi})$ имеет вид

$$u(\tilde{r}, \tilde{\varphi}) = \sum_{k=2}^{\infty} \tilde{B}_k \tilde{r}^{-\tilde{\mu}_k} \tilde{\Phi}_k(\tilde{\varphi}), \quad \tilde{r} > R > 1.$$

Тогда $u(\tilde{r}, \tilde{\varphi})$ меняет знак в любой окрестности этой особой точки.

Доказательство. Доказательства 1 и 2 аналогичны. Приведем доказательство первого утверждения леммы.

Предположим, что найдется такой номер $m \geq 2$, что $A_m \neq 0$, тогда как все $A_k = 0$ при $k < m$. Согласно теореме 5, собственная функция Φ_m меняет на Ω_R знак, а тогда на любой дуге $\{r = \delta, 0 < \varphi < \alpha\}$ с достаточно малым δ происходит смена знака рассматриваемого решения u . Действительно, на дуге $r = \delta, 0 < \varphi < \alpha$ функция Φ_m в малой окрестности своего нуля строго монотонна и принимает значения $(-\sigma, \sigma)$ с некоторым $\sigma > 0$. Оценивая при достаточно малых δ оставшуюся часть ряда

$$\sum_{k=m+1}^{\infty} A_k r^{\mu_k} \Phi_k(\varphi),$$

устанавливаем смену знака.

Лемма доказана.

4. ВЗАИМОДЕЙСТВИЕ ОСОБЕННОСТЕЙ

Напомним, что рассматривается задача Дирихле с одной конечной и одной бесконечной граничными особыми точками с локально-разделяющимися переменными. При этом в соответствующих особым точкам задачах Штурма–Лиувилля наложены ограничения $\mu_1 < 1 < \mu_2$ и $\tilde{\mu}_1 < 1 < \tilde{\mu}_2$.

Лемма 4 (нетривиальное решение однородной задачи).

1. Пусть $\mu_1 < \tilde{\mu}_1$, $p \in (2/(1 + \tilde{\mu}_1), 2/(1 + \mu_1))$, тогда существует нетривиальное решение однородной задачи Дирихле, имеющее особенность в конечной точке.

2. Пусть $\mu_1 > \tilde{\mu}_1$, $p \in (2/(1 - \mu_1), 2/(1 - \tilde{\mu}_1))$, тогда существует нетривиальное решение однородной задачи Дирихле, имеющее особенность на бесконечности.

Доказательство. Доказательства 1 и 2 аналогичны. Приведем конструктивное доказательство первого утверждения.

Рассмотрим $u(x) = \eta_\sigma(x) u_{\mu_1}(x) - v(x)$, где $u_{\mu_1} = r^{-\mu_1} \Phi_1(\varphi)$, срезающая функция $\eta_\sigma \in C^\infty(\mathbb{R}^2)$ равна единице в σ -окрестности конечной особой точки и нулю — вне 2σ -окрестности.

В силу выполнения тождеств

$$\int_{\Omega} \kappa(\nabla(\eta_{\sigma} u_{\mu}), \nabla \psi) dx = \int_{\Omega} \kappa(\eta_{\sigma} \nabla u_{\mu}, \nabla \psi) dx + \int_{\Omega} \kappa(u_{\mu} \nabla \eta_{\sigma}, \nabla \psi) \quad \forall \psi \in \dot{C}^{\infty}(\Omega),$$

$$0 = \int_{\Omega} \kappa(\nabla u_{\mu}, \nabla(\psi \eta_{\sigma})) dx = \int_{\Omega} \kappa(\eta_{\sigma} \nabla u_{\mu}, \nabla \psi) dx + \int_{\Omega} \kappa(\nabla u_{\mu}, \nabla \eta_{\sigma}) \psi dx \quad \forall \psi \in \dot{C}^{\infty}(\Omega),$$

определяющее решение тождество

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

равносильно

$$\int_{\Omega} \kappa(\nabla v, \nabla \psi) dx = - \int_{\Omega} \kappa(\nabla u_{\mu_1}, \nabla \eta_{\sigma}) \psi dx + \int_{\Omega} \kappa(u_{\mu_1} \eta_{\sigma}, \nabla \psi) dx \quad \forall \psi \in \dot{C}^{\infty}(\Omega).$$

Применяя теорему Рисса в гильбертовом пространстве $\dot{L}_{2,\kappa}^1(\Omega)$ со скалярным произведением $\int_{\Omega} \kappa(x)(\cdot, \cdot) dx$, заключаем, что существует $v \in \dot{L}_2^1(\Omega)$. А значит, не только $\eta_{\sigma}(x)u_{\mu_1}(x)$, но и u попадает в $L_p^1(\Omega)$ в силу общего вида решения в окрестностях особых точек при рассматриваемых здесь значениях p .

Лемма доказана.

Теорема 6 (тривиальное ядро).

1. Если $\mu_1 < \tilde{\mu}_1$, то $\dim \ker L = 0$ для $p \in (1, 2/(1 + \tilde{\mu}_1)] \cup [2/(1 + \mu_1), \infty)$.

2. Если $\mu_1 > \tilde{\mu}_1$, то $\dim \ker L = 0$ для $p \in (1, 2/(1 - \tilde{\mu}_1)] \cup [2/(1 - \mu_1), \infty)$.

Доказательство. Доказательства 1 и 2 аналогичны. Приведем здесь доказательство второго утверждения, т.е. для $\mu_1 > \tilde{\mu}_1$.

Выбор $2/(1 + \tilde{\mu}_1) \leq p \leq 2/(1 - \tilde{\mu}_1)$ исключает из рядов (3.1) и (3.2), представляющих решение в окрестностях особых точек, члены $\tilde{A}_1 \tilde{r}^{\tilde{\mu}_1} \tilde{\Phi}_1(\tilde{\varphi})$, $B_1 r^{-\mu_1} \Phi_1(\varphi)$. Решение попадает в $L_2^1(\Omega)$, а значит, оно тривиальное.

При $p \geq 2/(1 - \mu_1) > 2/(1 - \tilde{\mu}_1)$ сразу получаем $A_1 = B_1 = 0$. Докажем от противного, что и $\tilde{A}_1 = 0$. Не ограничивая общности, предположим $\tilde{A}_1 > 0$. Выберем дугу $\tilde{r} = R > 1$ в такой окрестности бесконечной особой точки, что

- переменные разделяются в полярных координатах,
- по лемме 2 решение $u \geq 0$.

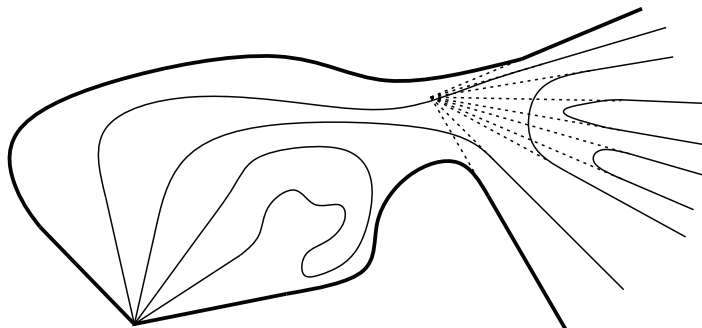
Применяя принцип существенного максимума (см. теорему 9.27 в [3]) на $\Omega_R = \Omega \cap \{x | \tilde{r} < R\}$, получаем $u|_{\Omega_R} \geq 0$. Однако в окрестности конечной особой точки $A_1 = B_1 = 0$ и применима лемма 3 о знакопеременности решения. Приходим к противоречию.

При $p < 2/(1 + \tilde{\mu}_1)$ сразу получаем $\tilde{A}_1 = \tilde{B}_1 = 0$. Докажем от противного, что и $B_1 = 0$. Не ограничивая общности, возьмем $B_1 > 0$. Выберем дугу $r = \delta < 1$ в такой окрестности конечной особой точки, что

- переменные разделяются в полярных координатах,
- по лемме 2 решение $u \geq 0$.

Применяя принцип существенного максимума на $\Omega_{\delta} = \Omega \cap \{x | r > \delta\}$, получаем $u|_{\Omega_{\delta}} \geq 0$. Однако в окрестности бесконечной особой точки $\tilde{A}_1 = \tilde{B}_1 = 0$ и применима лемма 3 о знакопеременности решения. Приходим к противоречию.

Теорема доказана.



Фиг. 1. Пример допустимой области.

Теорема 7 (нетривиальное ядро).

1. Если $\mu_1 < \tilde{\mu}_1$, то $\dim \ker L = 1$ для $p \in (2/(1 + \tilde{\mu}_1), 2/(1 + \mu_1))$.

2. Если $\mu_1 > \tilde{\mu}_1$, то $\dim \ker L = 1$ для $p \in (2/(1 - \mu_1), 2/(1 - \tilde{\mu}_1))$.

Доказательство. Доказательства 1 и 2 аналогичны. Приведем здесь доказательство второго утверждения, т.е. для $\mu_1 > \tilde{\mu}_1$.

Согласно лемме 4 существует нетривиальное решение u_1 однородной задачи. Введем обозначение

$$u_1 = \sum_{k=1}^{\infty} A_k' r^{\mu_k} \Phi_k(\varphi) + B_1' r^{-\mu_1} \Phi_1(\varphi), \quad 0 < r < \delta < 1.$$

Покажем, что произвольное решение u_2 линейно зависимо от u_1 . Введем обозначение

$$u_2 = \sum_{k=1}^{\infty} A_k'' r^{\mu_k} \Phi_k(\varphi) + B_1'' r^{-\mu_1} \Phi_1(\varphi), \quad 0 < r < \delta < 1.$$

Линейная комбинация $u = B_1'' u_1 - B_1' u_2$ в окрестности конечной особой точки будет по построению иметь коэффициент $B_1 = 0$ при $r^{-\mu_1} \Phi_1(\varphi)$. Применяя схему доказательства случая $p \geq 2/(1 - \mu_1)$ предыдущей теоремы, показываем, что $u = 0$.

Теорема доказана.

Следствие 1 (размерность коядра).

1. Если $\mu_1 < \tilde{\mu}_1$, то $\dim \operatorname{coker} L = 1$ для $p \in (2/(1 + \tilde{\mu}_1), 2/(1 + \mu_1))$ и $\dim \operatorname{coker} L = 0$ для остальных $1 < p < \infty$ кроме $2/(1 \pm \tilde{\mu}_1)$, $2/(1 \pm \mu_1)$.

2. Если $\mu_1 > \tilde{\mu}_1$, то $\dim \operatorname{coker} L = 1$ для $p \in (2/(1 - \mu_1), 2/(1 - \tilde{\mu}_1))$ и $\dim \operatorname{coker} L = 0$ для остальных $1 < p < \infty$ кроме $2/(1 \pm \tilde{\mu}_1)$, $2/(1 \pm \mu_1)$.

Замечание 2. Не требуется совпадения числа линий разрыва в конечной и бесконечной особых точках $\partial\Omega$. Так, линия разрыва может замыкаться через особую точку, т.е. начало и конец линии разрыва могут совпадать с этой особой точкой. На фиг. 1 приводится пример допустимой области: показаны линии разрыва коэффициента и граница $\partial\Omega$, а пунктиром — их асимптоты в окрестности бесконечности.

СПИСОК ЛИТЕРАТУРЫ

1. Денисов В.Н., Боговский А.М. О взаимодействии граничных особых точек в задаче Дирихле для эллиптического уравнения с кусочно-постоянными коэффициентами в плоской области // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 12. С. 2155–2174.
2. Денисов В.Н., Боговский А.М. О взаимосвязи слабых решений эллиптических краевых задач Дирихле и Неймана для плоской односвязной области // Матем. заметки. 2020. Т. 107. № 1. С. 32–48.
3. Brezis H. Functional analysis, Sobolev spaces and partial differential equations. Universitext. New York: Springer, 2011.

**УРАВНЕНИЯ
В ЧАСТНЫХ ПРОИЗВОДНЫХ**

УДК 519.642

**CONSTRUCTING SOLUTIONS OF CAUCHY TYPE INTEGRAL
EQUATIONS BY USING FOUR KINDS OF BASIS¹⁾**

© 2023 г. М. Yaghobifar^{1,*}, F. Hosseini Shekarabi^{1,**}

¹ *Department of Mathematics, Shahid Rajaei Teacher Training University, Tehran, Iran*

**e-mail: yaghobifar@pnu.ac.ir,*

***e-mail: f_hosseini@sru.ac.ir*

Поступила в редакцию 29.01.2023 г.

Переработанный вариант 17.03.2023 г.

Принята к публикации 29.05.2023 г.

Построение решений интегральных уравнений типа Коши с использованием четырех видов базиса. Решение интегрального уравнения типа Коши выписано с помощью разложения по специальным функциям и в виде ряда Маклорена. Комбинируя эти представления, решение записано в виде удобных многочленов. Численная реализация предложенного алгоритма показала высокую точность и эффективность получаемого решения.

Ключевые слова: интегральные уравнения типа Коши, специальные функции, высокоточный алгоритм.

DOI: 10.31857/S0044466923090156, **EDN:** MOKTIA

¹⁾ Полный текст статьи печатается в английской версии журнала.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 517.958

ДОБАВЛЕНИЕ К КЛАССИЧЕСКОМУ РЕЗУЛЬТАТУ А.Н. ТИХОНОВА
ОБ ЭЛЕКТРОМАГНИТНЫХ ЗОНДИРОВАНИЯХ
ДЛЯ СРЕДЫ С ТОНКИМИ СЛОЯМИ© 2023 г. А. С. Барашков^{1,*}¹ 111250 Москва, ул. Красноказарменная, 14, НИУ МЭИ, Россия

*e-mail: BarashkovAS@mpei.ru

Поступила в редакцию 20.02.2023 г.
Переработанный вариант 12.05.2023 г.
Принята к публикации 29.05.2023 г.

Результат А.Н. Тихонова использует значения функции на отрезке, т.е. для восстановления среды требуется бесконечно много значений функции. В настоящей работе ставится вопрос: пусть известны только несколько значений этой функции, какую информацию о среде можно тогда получить? Ответ оказался максимально благоприятным. Если массив данных содержит k значений функции, то среду можно характеризовать таким же числом параметров. Библ. 6.

Ключевые слова: обратная задача, конечный набор частот, одномерное уравнение Гельмгольца, теорема единственности.

DOI: 10.31857/S004446692309003X, EDN: EFXRCI

1. ВВЕДЕНИЕ

В работе [1] А.Н. Тихонов рассмотрел задачу определения коэффициента $\sigma(x)$ в одномерном уравнении Гельмгольца

$$\frac{d^2 u}{dx^2} + \lambda \sigma(x) u = 0 \quad (1)$$

по функции $\varphi(\lambda)$ (так называемый импеданс), которая выражается через решение уравнения при $x = 0$. При этом существенно использовалось, что функция $\varphi(\lambda)$ известна на некотором отрезке, а значит (так как функция является аналитической), при всех допустимых значениях λ . Похожий подход использовали и другие авторы (см., например, [2]).

Есть ситуации, когда функция $\varphi(\lambda)$ известна на конечном наборе частот:

$\lambda \in \{\lambda_1, \dots, \lambda_k\}$. Более того, пусть функция $\varphi(\lambda)$ известна на некотором отрезке, но ее можно аппроксимировать в пределах точности измерений по значениям $\varphi(\lambda_i)$, $i = 1, 2, \dots, k$.

Понятно, что эти k значений функции $\varphi(\lambda)$ несут такую же информацию о коэффициенте $\sigma(x)$, что и вся функция $\varphi(\lambda)$ на отрезке.

В настоящей работе выясняется, какие сведения о коэффициенте $\sigma(x)$ можно получить по конечному набору значений функции $\varphi(\lambda)$. Естественно предположить, что если функция $\varphi(\lambda)$ известна при k значениях, то и коэффициент $\sigma(x)$ можно характеризовать k параметрами. Проверим эту гипотезу для среды с “тонкими” слоями. Слой для кусочно-постоянной среды характеризуется своими границами: $b_1 < x < b_2$, проводимостью $\sigma(x) = \gamma = \text{const}$ или суммарной проводимостью $\alpha = \gamma(b_2 - b_1)$ (см. [2]). Для тонкого слоя есть две характеристики: $b_1 \approx b_2 = b$ (глубина залегания слоя) и суммарная проводимость α .

2. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим непроводящий слой (т.е. $\sigma = 0$) толщины 1, лежащий на идеальном проводнике (т.е. $\sigma = \infty$). В этом непроводящем слое расположены проводящие тонкие слои с параметрами b_i, α_i . Электромагнитные поля в случае E -поляризации описываются соотношениями

$$\frac{d^2 u}{dx^2} = 0; \quad u|_{x=1} = 0; \quad \frac{du}{dx}\bigg|_{x=0} = -1, \quad [u]|_{x=b_i} = 0, \quad \left[\frac{du}{dx}\right]_{x=b_i} = -\lambda \alpha_i u(b_i), \quad (2)$$

$$i = 1, 2, \dots, n; \quad 0 < b_1 < b_2 < \dots < b_n < 1.$$

Определение 1. *Прямой задачей* для системы (2) назовем задачу нахождения функции

$$\varphi(\lambda) = u(x=0, \lambda). \quad (3)$$

Определение 2. *Обратной задачей* для системы (2) назовем задачу определения чисел $b_i, \alpha_i, i = 1, 2, \dots, n$, по функции $\varphi(\lambda)$ из (3).

Замечание 1. Функция $\varphi(\lambda)$ (импеданс) определена через решение краевой задачи. Можно определить ее через решение задачи Коши. Для этого вместо граничного условия $\frac{du}{dx}\bigg|_{x=0} = -1$ ставим второе начальное условие $\frac{du}{dx}\bigg|_{x=1} = 1$. Тогда $\varphi(\lambda) = -\frac{u(x=0, \lambda)}{u'(x=0, \lambda)}$.

3. РЕШЕНИЕ ПРЯМОЙ ЗАДАЧИ

Рассмотрим вспомогательную задачу, которую можно интерпретировать как задачу (2) для одного слоя, или, точнее, как пересчет импеданса с подошвы одного тонкого слоя на подошву следующего тонкого слоя.

Ищем функцию $v(x)$ и определяем характеристику $B = \frac{v}{\frac{\partial v}{\partial x}\big|_{x=d-0}}$ (импеданс):

$$\frac{d^2 v}{dx^2} = 0 \quad \text{для} \quad 0 < x < d, \quad \frac{v}{\frac{\partial v}{\partial x}\big|_{x=0-0}} = A, \quad [v]|_{x=0} = 0, \quad \left[\frac{dv}{dx}\right]_{x=0} = -\lambda \alpha v(0).$$

Получаем

$$v(x) = C(A + (1 - \lambda \alpha A)x), \quad C = \text{const}, \quad B = d + \frac{1}{-\lambda \alpha + 1/A}.$$

Вводим числа

$$d_1 = 1 - b_n, \quad d_2 = b_n - b_{n-1}, \dots, d_i = b_{n-i+2} - b_{n-i+1}, \dots, d_{n+1} = b_1, \quad (4)$$

$$a_{2i-1} = d_i, \quad i = 1, \dots, n+1; \quad a_{2i} = -\lambda \alpha_{n-i+1}, \quad i = 1, \dots, n.$$

Тогда решение задачи (2) выражается цепной дробью (см. [3])

$$\varphi(\lambda) = \Phi_{2n+1}(a_1, a_2(\lambda), \dots, a_{2n}(\lambda), a_{2n+1}) = \Phi_{2n+1}(\lambda) = a_{2n+1} + \frac{1}{a_{2n} + \frac{1}{a_{2n-1} + \dots + \frac{1}{a_2 + \frac{1}{a_1}}}}. \quad (5)$$

4. РЕШЕНИЕ ОБРАТНОЙ ЗАДАЧИ

Теорема. Пусть известно $2n$ значений функции $\varphi(\lambda)$ из (3): $\varphi(\lambda_1), \varphi(\lambda_2), \dots, \varphi(\lambda_{2n}), \lambda_i \neq 0, n \geq 1$. Тогда обратная задача имеет единственное решение.

Доказательство. Преобразуем соотношение (5), используя результат из [3].

Вводим последовательность $P_k: P_{-2} = 0, P_{-1} = 1, P_k = a_k P_{k-1} + P_{k-2}, k \geq 0$. Тогда

$$\varphi(\lambda) = \Phi_{2n+1}(\lambda) = \frac{P_{2n+1}}{P_{2n}}. \quad (6)$$

Возвращаясь к обозначениям d_i, α_i из (4), замечаем, что $\varphi(\lambda)$ является рациональной дробью относительно λ вида

$$\varphi(\lambda) = \Phi_{2n+1}(\lambda) = \frac{k_n(-\lambda)^n + k_{n-1}(-\lambda)^{n-1} + \dots + k_0}{m_n(-\lambda)^n + m_{n-1}(-\lambda)^{n-1} + \dots + m_0}, \quad (7)$$

где

$$\begin{aligned} m_n &= d_1 d_2 \dots d_n \alpha_1 \alpha_2 \dots \alpha_n, & m_{n-1} &= d_1 d_2 \dots d_{n-1} \alpha_2 \alpha_3 \dots \alpha_n + \dots + d_2 d_3 \dots d_n \alpha_1 \alpha_2 \dots \alpha_{n-1}, & \dots, & & m_0 &= 1, \\ k_n &= m_n d_{n+1}, & k_{n-1} &= d_{n+1} m_{n-1} + d_1 d_2 \dots d_n \alpha_2 \alpha_3 \dots \alpha_{n-1}, & \dots, & & k_0 &= d_1 + d_2 + \dots + d_n + d_{n+1} = 1. \end{aligned} \quad (8)$$

По условию теоремы известны $2n$ значений функции $\varphi(\lambda)$. Учтем, что $\varphi(0) = 1$. Таким образом, для восстановления функции $\varphi(\lambda) = \Phi_{2n+1}$ есть $2n+1$ значений этой функции.

Можно рассматривать эту задачу как задачу интерполяции с помощью рациональной функции. Известно (см. [4, 5]), что такая задача интерполяции имеет не более одного решения. Так как в рассматриваемом случае одно решение заведомо есть (решение прямой задачи), то рациональная функция восстанавливается однозначно. Тем самым функция $\varphi(\lambda)$ известна при всех значениях λ . Однако использовать результат работы [1] о единственности решения обратной задачи не получается. А.Н. Тихонов в работе [1] ограничился случаем кусочно-аналитического коэффициента $\sigma(x)$ в (1). Среда с тонкими слоями при описании уравнением (1) задается коэффициентом $\sigma(x)$, который содержит обобщенные дельта-функции. Поэтому продолжим самостоятельный анализ функции $\varphi(\lambda) = \Phi_{2n+1}(\lambda)$ из (7).

При условиях (8) не только функция (7) определяется однозначно, но и коэффициенты k_j, m_j восстанавливаются однозначно. В самом деле, дробь (7) является несократимой в силу построения многочленов P_k из (6), и коэффициент $m_0 = 1$. Зная коэффициенты k_j, m_j функции $\Phi_{2n+1}(\lambda)$ из (7), можно определить параметры первого слоя

$$(b_1, \alpha_1): \quad b_1 = d_{n+1} = \frac{k_n}{m_n}, \quad \alpha_1 = \frac{m_n}{k_{n-1} - d_{n+1} m_{n-1}}.$$

Теперь можно пересчитать исходные значения $\varphi(\lambda_i)$ на уровень $x = b_1 + 0$. Получаем аналогичную задачу с $n - 1$ слоем. Отличие функции $\Phi_{2n-1}(\lambda)$ от функции в (7) будет в том, что степени многочленов в числителе и знаменателе дроби будут равны $n - 1$, и коэффициент $k_0 = d_1 + d_2 + \dots + d_n = 1 - b_1$. Значений функции $\Phi_{2n-1}(\lambda)$ будет уже с избытком. Зная коэффициенты k_j, m_j функции $\Phi_{2n-1}(\lambda)$, определяем параметры второго слоя (b_2, α_2) и т.д. Теорема доказана.

5. ПРИМЕР

Рассмотрим двухслойную среду с параметрами $b_1 = 1/3, \alpha_1 = 1, b_2 = 2/3, \alpha_2 = 3$. Решим для этой среды прямую задачу. Полученные данные будем считать информацией для решения обратной задачи и по ним восстановим двухслойную среду.

Введем вспомогательные величины

$$d_1 = 1 - b_2 = 1 - 2/3 = 1/3, \quad d_2 = b_2 - b_1 = 2/3 - 1/3 = 1/3, \quad d_3 = b_1 = 1/3.$$

Согласно (5),

$$\varphi(\lambda) = a_5 + \frac{1}{a_4 + \frac{1}{a_3 + \frac{1}{a_2 + \frac{1}{a_1}}}} = d_3 + \frac{1}{-\lambda\alpha_1 + \frac{1}{d_2 + \frac{1}{-\lambda\alpha_2 + \frac{1}{d_1}}}} = 1/3 + \frac{1}{-\lambda + \frac{1}{1/3 + \frac{1}{-\lambda 3 + \frac{1}{1/3}}}}.$$

Вычисляем $\varphi(\lambda)$ для четырех значений $\lambda \in \{-i, -2i, -3i, -4i\}$:

$$\varphi(-i) = \frac{56 - 24i}{87}, \quad \varphi(-2i) = \frac{155 - 66i}{303}, \quad \varphi(-3i) = \frac{40 - 16i}{87}, \quad \varphi(-4i) = \frac{731 - 276i}{1707}.$$

По полученным четырем числам восстанавливаем двухслойную среду.

Записываем функцию $\varphi(\lambda)$ в виде рациональной дроби (7):

$$\varphi(\lambda) = \Phi_5(\lambda) = \frac{k_2(-\lambda)^2 + k_1(-\lambda) + k_0}{m_2(-\lambda)^2 + m_1(-\lambda) + m_0} = \frac{k_2\lambda^2 - k_1\lambda + k_0}{m_2\lambda^2 - m_1\lambda + m_0}. \quad (9)$$

При записи (9) учтено, что $d_1 + d_2 + d_3 = k_0 = 1$.

Для определения коэффициентов k_2, k_1, m_2, m_1 получаем систему линейных уравнений

$$\begin{aligned} -k_2 + ik_1 + 1 &= \frac{56 - 24i}{87}(-m_2 + im_1 + 1), \\ -4k_2 + 2ik_1 + 1 &= \frac{155 - 66i}{303}(-4m_2 + 2im_1 + 1), \\ -9k_2 + 3ik_1 + 1 &= \frac{40 - 16i}{87}(-9m_2 + 3im_1 + 1), \\ -16k_2 + 4ik_1 + 1 &= \frac{731 - 276i}{1707}(-16m_2 + 4im_1 + 1). \end{aligned}$$

Решение системы есть $k_2 = 1/9, k_1 = 8/9, m_2 = 1/3, m_1 = 5/3$.

Коэффициенты дроби (9) связаны с параметрами среды следующими соотношениями:

$$\begin{aligned} k_2 &= \alpha_1\alpha_2d_1d_2d_3 = 1/9, \quad k_1 = d_3(\alpha_1(d_1 + d_2) + \alpha_2d_1) + d_2d_1\alpha_2 = 8/9, \quad m_2 = \alpha_1\alpha_2d_1d_2 = 1/3, \\ m_1 &= \alpha_1(d_1 + d_2) + \alpha_2d_1 = 5/3. \end{aligned}$$

Решаем возникшую нелинейную систему относительно $\alpha_1, \alpha_2, d_1, d_2, d_3$, учтя, что

$$d_1 + d_2 + d_3 = 1:$$

$$\begin{aligned} \frac{k_2}{m_2} = d_3 = \frac{1/9}{1/3} = 1/3; \quad \frac{m_2}{k_1 - d_3m_1} = \alpha_1 = \frac{1/3}{8/9 - 1/3 \times 5/3} = 1; \\ \frac{k_1 - d_3m_1}{m_1 - \alpha_1(d_1 + d_2)} = d_2 = \frac{1/3}{5/3 - 1 \times 5/3} = 1/3; \\ \frac{m_1 - \alpha_1(d_1 + d_2)}{\alpha_1} = \alpha_2 = \frac{1}{1/3} = 3. \end{aligned}$$

Ответ: $b_1 = d_3 = 1/3, \alpha_1 = 1, b_2 = 1 - d_3 = 1/3, \alpha_1 = 3$.

Замечание 2. В приведенном примере значения λ выбраны чисто мнимыми, так как в работе [1] λ тоже чисто мнимые. В свою очередь, в [1] выбор вида λ вызван тем, что при зондировании на низких частотах λ именно такие.

6. ЗАКЛЮЧЕНИЕ

Получен следующий результат: если известно n значений импеданса, то можно определить n параметров, задающих одномерную среду с тонкими слоями. Такое бывает не всегда. В похожей задаче для двумерной среды с тонкими слоями обнаружился случай неоднозначности решения

(см. [6]). При этом массив данных для восстановления среды может быть насколько угодно больше числа параметров, задающих среду.

СПИСОК ЛИТЕРАТУРЫ

1. *Тихонов А.Н.* К математическому обоснованию теории электромагнитных зондирований // Ж. вычисл. матем. и матем. физ. 1965. Т. 5. № 1. С. 545–554.
2. *Дмитриев В.И.* Обратные задачи геофизики. М.: МАКС Пресс, 2012.
3. *Хинчин А.Я.* Цепные дроби. М.: Наука, 1978.
4. *Уолш Дж.* Интерполяция и аппроксимация рациональными функциями в комплексной области. Изд-во иностр. лит., 1961.
5. *Чередниченко В.Г.* Рациональная интерполяция, аналитическое решение // Сиб. матем. журн. 2002. Т. 41. № 1. С. 188–193.
6. *Барашков А.С.* О возможности обнаружения тонких проводящих слоев по измерениям полей на поверхности среды // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 12. С. 2127–2138.

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.63

ОБРАТНЫЕ ЗАДАЧИ ДЛЯ ДИФФУЗИОННО-ДРЕЙФОВОЙ МОДЕЛИ ЗАРЯДКИ НЕОДНОРОДНОГО ПОЛЯРНОГО ДИЭЛЕКТРИКА¹⁾

© 2023 г. Р. В. Бризицкий^{1,*}, Н. Н. Максимова^{2,**}, А. Г. Масловская^{2,***}

¹ 690041 Владивосток, ул. Радио, 7, ИПМ ДВО РАН, Россия

² 675000 Благовещенск, ул. Игнатьевское шоссе, 21, АмГУ, Россия

*e-mail: mlnwizard@mail.ru

**e-mail: maksimova.nn@amursu.ru

***e-mail: maslovskaya.ag@amursu.ru

Поступила в редакцию 12.02.2023 г.

Переработанный вариант 12.02.2023 г.

Принята к публикации 29.03.2023 г.

Исследуются задачи восстановления неизвестных параметров модели электронно-индуцированной зарядки неоднородного полярного диэлектрика по дополнительной информации об объемной плотности распределения заряда и напряженности электрического поля. В рамках оптимизационного подхода указанные обратные задачи сводятся к задачам управления и доказываются их разрешимость. Для экстремальных задач выводятся системы оптимальности и на основе их анализа доказываются локальная единственность решения одной из рассматриваемых задач. С учетом введенной характеристики неоднородности диэлектрика корректируются вспомогательные результаты о разрешимости и свойствах решений краевой задачи, полученные ранее для модели зарядки однородного диэлектрика. Библи. 31.

Ключевые слова: модель дрейфа–диффузии электронов, модель зарядки неоднородного полярного диэлектрика, глобальная разрешимость, локальная единственность, принцип максимума, обратная задача, задача управления, система оптимальности.

DOI: 10.31857/S0044466923090053, **EDN:** DNND SQ

1. ВВЕДЕНИЕ. ПОСТАНОВКА КРАЕВОЙ ЗАДАЧИ

Диффузионно-дрейфовое приближение часто используется в практике физико-математического моделирования процессов зарядки диэлектриков в неравновесных внешних условиях. Развитию данного подхода посвящены как ранние работы [1–6], так и современные исследования [7–14]. В частности, применение детерминированной модели обусловлено необходимостью прогнозирования состояния функциональных диэлектрических материалов при диагностике и модификации их свойств методами растровой электронной микроскопии. В числе важнейших аспектов, подлежащих исследованию, в указанных работах рассмотрены вопросы разработки фундаментальных основ, развития математических моделей, создания математического и программного обеспечения для исследования процессов стимулированной зарядки. Отдельно отметим статьи [12, 13], в которых предложена авторская модификация классической нестационарной модели процесса зарядки сегнетоэлектрических материалов при электронном облучении с учетом собственной радиационно-стимулированной проводимости объекта, диффузии и дрейфа инжектированных зарядов.

Априорный анализ показывает, что если облучение диэлектрического материала поддерживается достаточно длительное время (на практике достаточно и доли секунды), то это время намного превышает временной диапазон, который необходим для перехода динамической системы в стационарное состояние (менее микросекунды). В связи с этим особую актуальность для практики приобретает детальное исследование математических моделей, описывающих стационарные режимы процессов зарядки. Корректность одной из таких моделей была впервые обоснована в [15] в предположении свойства однородности поляризационных характеристик объекта.

¹⁾ Работа выполнена в рамках госзадания ИПМ ДВО РАН (№ 075-01290-23-00) и при поддержке Минобрнауки РФ (проект № 122082400001-8 и проект № 075-02-2023-946).

Моделирование процессов зарядки в неоднородных диэлектриках представляет сложную и многоаспектную задачу, поскольку требует рассмотрения покомпонентного представления распределения поляризации в объекте. В качестве приближения к общей проблеме в настоящей работе рассмотрим упрощенную модель зарядки неоднородного полярного диэлектрика, в которой “неоднородность” описывается с помощью нормализованной функции. Такой подход позволяет задать пространственную характеристику изменения уровня заряда, обусловленного начальным (неиндуцированным) состоянием самого диэлектрика. Учитывая введенное предположение и модельную характеристику, для рассматриваемого объекта далее будем использовать понятие “неоднородный диэлектрик”. Будем считать, что функция $\beta(\mathbf{x})$ описывает неравномерность потерь заряда в пространстве, что на практике может быть обусловлено анизотропией кристалла, наличием дефектной структуры, композитным составом материала или присутствием примеси, предварительной ионизирующей обработкой и др.

Математическая модель процесса зарядки неоднородного полярного диэлектрика может быть представлена следующей краевой задачей, рассматриваемой в ограниченной области $\Omega \subset \mathbb{R}^3$ с границей Γ :

$$-\operatorname{div}(d\nabla\rho) + \mu_n \mathbf{E} \cdot \nabla\rho + \frac{\mu_n}{\varepsilon\varepsilon_0} \beta|\rho|\rho = f \quad \text{в } \Omega, \quad (1)$$

$$\operatorname{rot} \mathbf{E} = \mathbf{0}, \quad \operatorname{div} \mathbf{E} = \frac{1}{\varepsilon\varepsilon_0} \rho \quad \text{в } \Omega, \quad (2)$$

$$\rho = 0, \quad \mathbf{E} \times \mathbf{n} = \mathbf{0} \quad \text{на } \Gamma. \quad (3)$$

Здесь ρ — объемная плотность заряда, $\mathbf{E}$ — вектор-функция напряженности электрического поля, $d(\mathbf{x}) > 0$ — коэффициент диффузии электронов, μ_n — дрейфовая подвижность электронов, ε — диэлектрическая проницаемость материала, ε_0 — электрическая постоянная, f — генерационное слагаемое, отвечающее за действие объемного источника зарядов в объекте, $\beta(\mathbf{x}) > 0$ — нормализованный коэффициент потери заряда. Ниже на задачу (1)–(3) будем ссылаться как на задачу 1.

В [15] доказаны глобальная разрешимость и локальная единственность решения соответствующей краевой задачи в случае, когда $\beta \equiv 1$, а также установлен принцип минимума и максимума для плотности заряда. Хорошо известно, что принцип максимума применяется для контроля вычислительных экспериментов. В указанной статье один из вычислительных экспериментов по решению прямой задачи посвящен проверке установленного теоретически принципа максимума.

В настоящей работе результаты [15] по исследованию краевой задачи обобщены на случай, когда коэффициент β не является константой. По возможности полученные в настоящей статье результаты о разрешимости краевой задачи представлены в виде коррекции результатов [15]. Отметим, что функция $\beta(\mathbf{x})$ влияет на достаточные условия локальной единственности решения краевой задачи и входит в оценку принципа максимума.

Основные результаты статьи связаны с исследованием задачи восстановления неизвестных функций d , β и f по измеренным в некоторой подобласти области Ω плотности заряда ρ и напряженности электрического поля $\mathbf{E}$. В частности, восстановление функции β позволит обнаружить вкрапления в кристалле диэлектрика. В рамках оптимизационного подхода указанная обратная задача сводится к задаче управления (о корректности применяемого подхода см. [16, 17]). В результате формулируется задача управления, содержащая два мультипликативных управления d и β и одно распределенное управление f , и доказывается ее разрешимость. Далее для задачи управления выводится система оптимальности и обосновывается ее локальная регулярность. Наконец, на основе анализа системы оптимальности доказывается локальная единственность решения задачи распределенного управления. В заключение отметим работу [18], в которой доказана разрешимость задачи мультипликативного управления при более жестких условиях на функцию d .

2. РАЗРЕШИМОСТЬ КРАЕВОЙ ЗАДАЧИ

При анализе краевой задачи будем использовать функциональные пространства Соболева $H^s(D)$, $s \in \mathbb{R}$. Здесь D обозначает область Ω , либо некоторую подобласть $Q \subset \Omega$, либо границу Γ . Через $\|\cdot\|_{s,Q}$, $|\cdot|_{s,Q}$ и $(\cdot, \cdot)_{s,Q}$ будем обозначать норму, полунорму и скалярное произведение в $H^s(Q)$.

Нормы и скалярные произведения в $L^2(Q)$ и $L^2(\Omega)$ будем обозначать соответственно через $\|\cdot\|_Q$ и $(\cdot, \cdot)_Q$, $\|\cdot\|_\Omega$ и $(\cdot, \cdot)_\Omega$.

Введем функциональные пространства $H^1(\Delta, \Omega) = \{h \in H^1(\Omega) : \Delta h \in L^2(\Omega)\}$,

$$H_N^1(\Omega) = \{\mathbf{h} \in H^1(\Omega)^3 : \mathbf{h} \times \mathbf{n}|_\Gamma = \mathbf{0}\}, \tilde{H}_N^1(\Omega) = H_N^1(\Omega) \cap \ker(\operatorname{rot}),$$

функциональные множества $L_{d_0}^p(\Omega) = \{h \in L^p(\Omega) : h \geq d_0 > 0 \text{ п.в. в } \Omega\}$, $1 \leq p \leq \infty$, $H_{d_0}^s(\Omega) = \{h \in H^s(\Omega) : h \geq d_0 > 0 \text{ п.в. в } \Omega\}$, $s \geq 0$, где d_0 — положительная константа и произведение пространств $X = H_0^1(\Omega) \times \tilde{H}_N^1(\Omega)$.

Предположим, что выполняются следующие условия:

(i) Ω — ограниченная область в $\mathbb{R}^3$ с границей $\Gamma \in C^{0,1}$;

(ii) $f \in L^2(\Omega)$, $d \in L_{d_0}^\infty(\Omega)$;

(iii) $\beta \in L^4(\Omega)$ и $\beta \geq \beta_0 \geq 1/2$ п.в. в Ω .

Напомним, что в силу теоремы вложения Соболева пространство $H^1(\Omega)$ вкладывается в пространство $L^s(\Omega)$ непрерывно при $s \leq 6$, компактно при $s < 6$ и с некоторой константой C_s , зависящей от s и Ω , справедлива оценка

$$\|h\|_{L^s(\Omega)} \leq C_s \|h\|_{1,\Omega} \quad \forall h \in H^1(\Omega). \quad (4)$$

При $s = 2$ мы полагаем $C_2 = 1$.

Ниже будем использовать формулы Грина (см. [19, 20])

$$-(\Delta u, v) = (\nabla u, \nabla v) - (\partial u / \partial n, v)_\Gamma \quad \forall u \in H^1(\Delta, \Omega), \quad v \in H^1(\Omega), \quad (5)$$

$$(\mathbf{u}, \nabla v) + (\operatorname{div} \mathbf{u}, v) = \langle \mathbf{u} \cdot \mathbf{n}, v \rangle_\Gamma \quad \forall \mathbf{u} \in L^2(\Omega)^3 \quad \text{с} \quad \operatorname{div} \mathbf{u} \in L^{3/2}(\Omega), \quad v \in H^1(\Omega). \quad (6)$$

Справедливы следующие леммы (см. [15, 20]).

Лемма 2.1. При выполнении условий (i), $\mathbf{E} \in H^1(\Omega)^3$ существуют положительные константы C_0 , δ_1 , γ_1' и γ_1 , зависящие соответственно от Ω , такие что

$$|(\nabla h, \nabla \eta)| \leq C_0 \|h\|_{1,\Omega} \|\eta\|_{1,\Omega}, \quad (7)$$

$$|(\mathbf{E} \cdot \nabla h, \eta)| \leq \gamma_1' \|\mathbf{E}\|_{L^4(\Omega)^3} \|h\|_{1,\Omega} \|\eta\|_{1,\Omega} \leq \gamma_1 \|\mathbf{E}\|_{1,\Omega} \|h\|_{1,\Omega} \|\eta\|_{1,\Omega} \quad \forall h, \eta \in H^1(\Omega),$$

$$(\nabla h, \nabla h) \geq \delta_1 \|h\|_{1,\Omega}^2 \quad \forall h \in H_0^1(\Omega). \quad (8)$$

Если функции $\mathbf{E} \in H^1(\Omega)^3$ и $\rho \in H_0^1(\Omega)$ связаны вторым соотношением в (2), то справедливо равенство

$$(\mathbf{E} \cdot \nabla \rho, h) = -(\nabla h \cdot \mathbf{E}, \rho) - (1/\varepsilon \varepsilon_0)(h, \rho^2) \quad \forall h \in H_0^1(\Omega), \quad (9)$$

принимая при $h = \rho$ следующий вид:

$$\mu_n(\mathbf{E} \cdot \nabla \rho, \rho) = -(\mu_n/2\varepsilon \varepsilon_0)(\rho, \rho^2). \quad (10)$$

Лемма 2.2. При выполнении условия (i) для любой функции $\sigma \in L^2(\Omega)$ существует единственное решение $\mathbf{E} \in \tilde{H}_N^1(\Omega)$ задачи $\operatorname{rot} \mathbf{E} = \mathbf{0}$, $\operatorname{div} \mathbf{E} = \sigma$ в Ω , $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ на Γ , для которого справедлива оценка $\|\mathbf{E}\|_{1,\Omega} \leq C_N \|\sigma\|_\Omega$, где C_N — положительная константа, зависящая от Ω .

Пусть $(\rho, \mathbf{E}) \in (C^2(\Omega) \cap C^0(\bar{\Omega})) \times (C^1(\Omega)^3 \cap \tilde{H}_N^1(\Omega))$ – классическое решение задачи 1. Умножим уравнение в (1) на функцию $h \in H_0^1(\Omega)$ и проинтегрируем по Ω , применяя формулу Грина (6). Приходим к слабой формулировке задачи 1

$$(d\nabla\rho, \nabla h) + \mu_n(\mathbf{E} \cdot \nabla\rho, h) + (\mu_n/\varepsilon\varepsilon_0)(\beta|\rho|, h) = (f, h) \quad \forall h \in H_0^1(\Omega), \quad (11)$$

$$\operatorname{div} \mathbf{E} = (1/\varepsilon\varepsilon_0)\rho \quad \text{в } \Omega. \quad (12)$$

Разрешимость задачи (11), (12) докажем с помощью теоремы Шаудера. Для этого построим отображение $G: H_0^1(\Omega) \rightarrow H_0^1(\Omega)$, действующее по формуле $\rho = G(r)$.

Здесь функция $\rho \in H_0^1(\Omega)$ является решением линейной задачи

$$a(\rho, h) \equiv (d\nabla\rho, \nabla h) + \mu_n(\mathbf{E}(r) \cdot \nabla\rho, h) + (\mu_n/\varepsilon\varepsilon_0)(\beta|r|, h) = (f, h) \quad \forall h \in H_0^1(\Omega), \quad (13)$$

$$\operatorname{div} \mathbf{E}(r) = (1/\varepsilon\varepsilon_0)r \quad \text{в } \Omega. \quad (14)$$

Из леммы 2.2 вытекает, что для любой функции $r \in H_0^1(\Omega)$ существует единственное решение $\mathbf{E} \in \tilde{H}_N^1(\Omega)$ задачи (14), для которого справедлива оценка

$$\|\mathbf{E}\|_{1,\Omega} \leq (C_N/\varepsilon\varepsilon_0)\|r\|_{1,\Omega}. \quad (15)$$

Из леммы 2.1 вытекает, что форма $a(\rho, h)$ непрерывна. Так же из леммы 2.1 и равенства (14) следует, что

$$(\mathbf{E} \cdot \nabla\rho, h) = -(\nabla h \cdot \mathbf{E}, \rho) - \frac{1}{\varepsilon\varepsilon_0}(h, r\rho) \quad \forall h \in H_0^1(\Omega). \quad (16)$$

При $h = \rho$ равенство (16) принимает вид

$$(\mathbf{E} \cdot \nabla\rho, \rho) = -\frac{1}{2\varepsilon\varepsilon_0}(r, \rho^2). \quad (17)$$

Положим $h = \rho$ в (13). С учетом (17) приходим к равенству

$$a(\rho, \rho) = (d\nabla\rho, \nabla\rho) + \frac{\mu_n}{\varepsilon\varepsilon_0}(\beta|r| - (1/2)r, \rho^2) = (f, \rho). \quad (18)$$

Поскольку $\beta \geq 1/2$ в силу (iii), то $(\beta|r| - (1/2)r, \rho^2) \geq 0$.

Тогда из (18) и леммы 2.1 вытекает, что форма a коэрцитивна на пространстве $H_0^1(\Omega)$ с константой $d_0\delta_1$. Из теоремы Лакса–Мильграма следует, что при любом $r \in H_0^1(\Omega)$ существует единственное решение $\rho \in H_0^1(\Omega)$ задачи (13), для которого справедлива оценка

$$\|\rho\|_{1,\Omega} \leq C_* \|f\|_{\Omega}, \quad C_* = (d_0\delta_1)^{-1}. \quad (19)$$

Из вышесказанного вытекает, что отображение $G: H_0^1(\Omega) \rightarrow H_0^1(\Omega)$ определено корректно и переводит шар B_R радиуса $R = C_* \|f\|$ в себя. Компактность и непрерывность оператора G на B_R показывается в точности, как в [15].

Тогда из теоремы Шаудера вытекает, что оператор G имеет неподвижную точку $\rho = G(\rho)$, которая является решением уравнения (11). В свою очередь, пара $(\rho, \mathbf{E})$ является искомым решением задачи (11), (12). При этом для функции ρ справедлива оценка (19). Тогда для электрического поля $\mathbf{E}$ из (15) в силу леммы 2.2 и с учетом (19) получаем оценку

$$\|\mathbf{E}\|_{1,\Omega} \leq (1/\varepsilon\varepsilon_0)C_N C_* \|f\|_{\Omega}. \quad (20)$$

Установим достаточные условия единственности решения задачи (11), (12). Обозначим через $(\rho_1, \mathbf{E}_1) \in X$ и $(\rho_2, \mathbf{E}_2) \in X$ любые два ее решения.

Несложно проверить, что разности $\rho = \rho_1 - \rho_2$ и $\mathbf{E} = \mathbf{E}_1 - \mathbf{E}_2$ удовлетворяют соотношениям

$$\begin{aligned} (d\nabla\rho, \nabla h) + \mu_n(\mathbf{E}_1 \cdot \nabla\rho, h) + \frac{\mu_n}{\varepsilon\varepsilon_0}(\beta|\rho_1|\rho, h) = \\ = -\mu_n(\mathbf{E} \cdot \nabla\rho_2, h) - \frac{\mu_n}{\varepsilon\varepsilon_0}(\beta(|\rho_1| - |\rho_2|)\rho_2, h) \quad \forall h \in H_0^1(\Omega), \end{aligned} \quad (21)$$

$$\operatorname{div} \mathbf{E} = \frac{1}{\varepsilon\varepsilon_0} \rho \quad \text{в } \Omega. \quad (22)$$

Поскольку $\operatorname{div} \mathbf{E}_i = (1/\varepsilon\varepsilon_0)\rho_i$ в Ω , $i = 1, 2$, то с учетом леммы 2.1 получаем, что

$$\mu_n(\mathbf{E}_1 \cdot \nabla\rho, \rho) = -\frac{\mu_n}{2\varepsilon\varepsilon_0}(\rho_1\rho, \rho). \quad (23)$$

Тогда равенство (21) при $h = \rho$ с учетом (23) принимает вид

$$(d\nabla\rho, \nabla\rho) + \frac{\mu_n}{\varepsilon\varepsilon_0}((\beta|\rho_1| - (1/2)\rho_1)\rho, \rho) = -\mu_n(\mathbf{E} \cdot \nabla\rho_2, \rho) - \frac{\mu_n}{\varepsilon\varepsilon_0}(\beta(|\rho_1| - |\rho_2|)\rho_2, \rho). \quad (24)$$

Для левой части (24) при $\beta \geq 1/2$ в силу (8) справедлива следующая оценка

$$\lambda_* \|\rho\|_{1,\Omega}^2 \leq (d\nabla\rho, \nabla\rho) + \frac{\mu_n}{\varepsilon\varepsilon_0}((\beta|\rho_1| - (1/2)\rho_1)\rho, \rho), \quad \lambda_* = d_0\delta_1.$$

Оценки для правой части (24) вытекают из неравенства Гёльдера, соотношения (7), и с учетом (4), (19) и (20) принимают следующий вид:

$$|\mu_n(\mathbf{E} \cdot \nabla\rho_2, \rho)| \leq \frac{\mu_n\gamma_1}{\varepsilon\varepsilon_0\lambda_*} C_N \|f\|_{\Omega} \|\rho\|_{1,\Omega}^2,$$

$$\frac{\mu_n}{\varepsilon\varepsilon_0} |(\beta(|\rho_1| - |\rho_2|)\rho_2, \rho)| \leq \frac{\mu_n}{\varepsilon\varepsilon_0\lambda_*} C_4^3 \|\beta\|_{L^4(\Omega)} \|f\|_{\Omega} \|\rho\|_{1,\Omega}^2.$$

Окончательно приходим к неравенству

$$\lambda_* \|\rho\|_{1,\Omega}^2 \leq \frac{\mu_n}{\varepsilon\varepsilon_0\lambda_*} (\gamma_1 C_N + C_4^3 \|\beta\|_{L^4(\Omega)}) \|f\|_{\Omega} \|\rho\|_{1,\Omega}^2. \quad (25)$$

Пусть исходные данные задачи 1 таковы, что выполняется условие

$$(\mu_n/\varepsilon\varepsilon_0) (\gamma_1 C_N + C_4^3 \|\beta\|_{L^4(\Omega)}) \|f\|_{\Omega} < \lambda_*^2. \quad (26)$$

Тогда из (25) вытекает, что $\|\rho\|_{1,\Omega} = 0$ или $\rho_1 = \rho_2$ в Ω . В таком случае из (22) получаем, что $\operatorname{div} \mathbf{E} = 0$ в Ω . Последнее в силу леммы 2.2 означает, что $\mathbf{E} = \mathbf{0}$ или $\mathbf{E}_1 = \mathbf{E}_2$ в Ω .

Сформулируем полученные результаты в виде следующей теоремы.

Теорема 2.1. При выполнении условий (i)–(iii) существует слабое решение $(\rho, \mathbf{E}) \in X$ задачи 1 и справедливы оценки (19) и (20). Если к тому же выполняется условие (26), то слабое решение задачи 1 единственно.

3. ПРИНЦИП МИНИМУМА И МАКСИМУМА

Положим $f_{\max}$ — положительная константа.

Пусть в дополнение к (i)–(iii) выполнено условие

(iv) $0 \leq f \leq f_{\max}$ п.в. в Ω .

Рассуждая, как в [15], установим принцип минимума и максимума для плотности заряда ρ .

Лемма 3.1. При выполнении условий (i)–(iv) для слабого решения $\rho \in H_0^1(\Omega)$ задачи 1 справедлив следующий принцип минимума и максимума:

$$0 \leq \rho \leq M \text{ п.в. в } \Omega, \quad M = \left(\frac{f_{\max} \varepsilon \varepsilon_0}{\mu_n \beta_0} \right)^{1/2}. \quad (27)$$

Доказательство. Докажем, что $\rho \geq 0$ п.в. в Ω . Введем функцию $w = \min\{\rho, 0\}$. Ясно, что оценка $\rho \geq 0$ выполняется тогда и только тогда, когда $w = 0$ п.в. в Ω .

Через $Q \subset \Omega$ обозначим измеримое подмножество Ω , в котором $w < 0$. Очевидно, что $w \in H^1(\Omega)$, а из [21, 22] вытекает, что $w|_{\Gamma} = \min\{\rho|_{\Gamma}, 0\} = 0$. Тогда $w \in H_0^1(\Omega)$.

Справедливо следующее равенство: $\nabla w = \nabla \rho$ п.в. в Q (см. [21]).

Из вышесказанного вытекает, что

$$(d\nabla \rho, \nabla w) = (d\nabla w, \nabla w)_Q = (d\nabla w, \nabla w), \quad (28)$$

$$(\mathbf{E} \cdot \nabla \rho, w) = (\mathbf{E} \cdot \nabla w, w)_Q = (\mathbf{E} \cdot \nabla w, w). \quad (29)$$

С учетом (10) из второго равенства в (29) получаем

$$\mu_n(\mathbf{E} \cdot \nabla w, w) = -\frac{\mu_n}{2\varepsilon \varepsilon_0}(\rho, w^2) = -\frac{\mu_n}{2\varepsilon \varepsilon_0}(w, w^2)_Q. \quad (30)$$

При этом

$$\frac{\mu_n}{\varepsilon \varepsilon_0}(\beta|\rho|, w) = \frac{\mu_n}{\varepsilon \varepsilon_0}(\beta|w|, w)_Q. \quad (31)$$

Полагая в (11) $h = w$, будем иметь

$$(d\nabla \rho, \nabla w) + \mu_n(\mathbf{E} \cdot \nabla \rho, w)_Q + \frac{\mu_n}{\varepsilon \varepsilon_0}(\beta|\rho|, w)_Q = (f, w)_Q. \quad (32)$$

С учетом (28)–(31) последнее соотношение (32) принимает вид

$$(d\nabla w, \nabla w) + \frac{\mu_n}{\varepsilon \varepsilon_0}(\beta|w| - (1/2)w, w^2)_Q = (f, w)_Q. \quad (33)$$

Поскольку $w < 0$ в Q и $f \geq 0$ п.в. в Ω , то $(f, w)_Q \leq 0$. С учетом того, что $\beta \geq 1/2$, в силу неравенства Фридрихса–Пуанкаре (8) из (33) приходим к оценке

$$\lambda_* \|w\|_{1,\Omega}^2 \leq (f, w)_Q \leq 0,$$

из которой вытекает, что $\|w\|_{1,\Omega} = 0$, а следовательно, $w = 0$ п.в. в Ω . Тогда $\rho \geq 0$ п.в. в Ω .

Для доказательства принципа максимума введем функцию $\tilde{\rho} = \max\{\rho - M, 0\}$, которая, как и вспомогательная функция w , принадлежит $H^1(\Omega)$. Ясно, оценка $\rho \leq M$ п.в. в Ω выполняется тогда и только тогда, когда $\tilde{\rho} = 0$ п.в. в Ω . Из [21, 22] вытекает, что $\tilde{\rho}|_{\Gamma} = \max\{\rho|_{\Gamma} - M, 0\} = 0$. В таком случае $\tilde{\rho} \in H_0^1(\Omega)$.

Через $Q_M \subset \Omega$ обозначим открытое измеримое подмножество Ω , в котором $\rho > M$. Как и выше, из [21] вытекает, что

$$(d\nabla \rho, \nabla \tilde{\rho}) = (d\nabla \tilde{\rho}, \nabla \tilde{\rho})_{Q_M} = (d\nabla \tilde{\rho}, \nabla \tilde{\rho}). \quad (34)$$

Из сказанного выше вытекает, что

$$\mu_n(\mathbf{E} \cdot \nabla \rho, \tilde{\rho}) = \mu_n(\mathbf{E} \cdot \nabla \tilde{\rho}, \tilde{\rho}). \quad (35)$$

Тогда с учетом леммы 2.1 приходим к равенству

$$\mu_n(\mathbf{E} \cdot \nabla \tilde{\rho}, \tilde{\rho}) = -\frac{\mu_n}{2\varepsilon \varepsilon_0}(\rho, \tilde{\rho}^2)_{Q_M} = -\frac{\mu_n}{2\varepsilon \varepsilon_0}(\tilde{\rho} + M, \tilde{\rho}^2)_{Q_M} = -\frac{\mu_n}{2\varepsilon \varepsilon_0}(\tilde{\rho}^2 + M\tilde{\rho}, \tilde{\rho})_{Q_M}. \quad (36)$$

Далее

$$(\beta|\rho|\rho, \tilde{\rho})_{Q_M} = (\beta\rho^2, \tilde{\rho})_{Q_M} = (\beta(\tilde{\rho} + M)^2, \tilde{\rho}) = (\beta(\tilde{\rho}^2 + 2M\tilde{\rho} + M^2), \tilde{\rho})_{Q_M}. \quad (37)$$

Подставляя $h = \tilde{\rho}$ в (11), с учетом (36), (37) получим

$$(d\nabla\tilde{\rho}, \nabla\tilde{\rho}) + \frac{\mu_n}{2\epsilon\epsilon_0}((2\beta - 1)\tilde{\rho}^2 + (4\beta - 1)M\tilde{\rho}, \tilde{\rho})_{Q_M} = (f - (1/\epsilon\epsilon_0)\mu_n\beta M^2, \tilde{\rho})_{Q_M}. \quad (38)$$

Поскольку $\beta \geq 1/2$, то, применив к левой части (38) оценку (8), приходим к оценке

$$d_0\delta_1\|\tilde{\rho}\|_{1,\Omega}^2 \leq (f_{\max} - (1/\epsilon\epsilon_0)\mu_n\beta_0M^2, \tilde{\rho})_{Q_M},$$

из которой выводим, что $\tilde{\rho} = 0$ п.в. в Ω , если параметр M определен в (27). Лемма 3.1 доказана.

4. ПОСТАНОВКА И РАЗРЕШИМОСТЬ ЗАДАЧИ УПРАВЛЕНИЯ

В данном разделе мы исследуем задачу управления для задачи 1 с двумя мультипликативными управлениями: d и β , и одним распределенным управлением f .

Будем считать, что функции d , β и f могут изменяться соответственно во множествах K_1 , K_2 и K_3 , удовлетворяющих условию

(j) $K_1 \subset H_{d_0}^s(\Omega) \cap L^\infty(\Omega)$, $s \geq 1/2$, $K_2 \subset L_{\beta_0}^4(\Omega)$ и $K_3 \subset L^2(\Omega)$ — непустые выпуклые замкнутые множества.

Введем функциональные пространства $X = H_0^1(\Omega) \times \tilde{H}_N^1(\Omega)$, $Y = H^{-1}(\Omega) \times L^2(\Omega)$ и положим $\mathbf{x} = (\rho, \mathbf{E}) \in X$, $u = (d, \beta, f) \in K$, где $K = K_1 \times K_2 \times K_3$. Далее введем оператор $F = (F_1, F_2) : X \times K \rightarrow Y$ по формулам

$$\langle F_1(\mathbf{x}, u), h \rangle = (d\nabla\rho, \nabla h) + \mu_n(\mathbf{E} \cdot \nabla\rho, h) + \frac{\mu_n}{\epsilon\epsilon_0}(\beta|\rho|\rho, h) - (f, h) \quad \forall h \in H_0^1(\Omega),$$

$$F_2(\mathbf{x}) = \operatorname{div} \mathbf{E} - \frac{1}{\epsilon\epsilon_0}\rho \quad \text{в } \Omega$$

и перепишем слабую формулировку задачи 1 в виде операторного уравнения $F(\mathbf{x}, u) = 0$.

Пусть $I : X \rightarrow \mathbb{R}$ — слабо полунепрерывный снизу функционал. Рассмотрим следующую задачу управления:

$$J(\mathbf{x}, u) \equiv \frac{\mu_0}{2}I(\mathbf{x}) + \frac{\mu_1}{2}\|d\|_{s,\Omega}^2 + \frac{\mu_2}{2}\|f\|_\Omega^2 \rightarrow \inf, \quad F(\mathbf{x}, u) = 0, \quad (\mathbf{x}, u) \in X \times K. \quad (39)$$

Через $Z_{ad} = \{(\mathbf{x}, u) \in X \times K : F(\mathbf{x}, u) = 0, J(\mathbf{x}, u) < \infty\}$ обозначим множество допустимых пар для задачи (39).

Пусть, в дополнение к (j), выполняются следующие условия:

(jj) множество K_1 ограничено по норме $L^\infty(\Omega)$, множество K_2 ограничено по норме $L^4(\Omega)$;

(jjj) $\mu_0 > 0$, $\mu_i \geq 0$, $i = 1, 2$, множество K_1 ограничено в $H^s(\Omega)$, $s > 1/2$, множество K_3 ограничено в $L^2(\Omega)$ или $\mu_i > 0$, $i = 0, 1, 2$, и функционал I ограничен снизу.

Будем использовать следующие функционалы качества:

$$I_1(\rho) = \|\rho - \rho^d\|_Q^2, \quad I_2(\mathbf{E}) = \|\mathbf{E} - \mathbf{E}^d\|_Q^2. \quad (40)$$

Здесь $\rho^d \in L^2(Q)$ обозначает заданное поле концентрации в подобласти $Q \subset \Omega$. Функция $\mathbf{E}^d$ имеет аналогичный смысл для электрического поля.

Теорема 4.1. Пусть выполняются условия (i)–(iii) и (j)–(jjj), функционал $I : X \rightarrow \mathbb{R}$ слабо полунепрерывен снизу и $Z_{ad} \neq \emptyset$. Тогда существует по крайней мере одно решение $(\mathbf{x}, u) \in X \times K$ задачи управления (39).

Доказательство. Через $(\mathbf{x}_m, u_m) = (\rho_m, \mathbf{E}_m, d_m, \beta_m, f_m) \in Z_{ad}$ обозначим минимизирующую последовательность, для которой

$$\lim_{m \rightarrow \infty} J(\mathbf{x}_m, u_m) = \inf_{(\mathbf{x}, u) \in Z_{ad}} J(\mathbf{x}, u) \equiv J^*.$$

Из условий (jj) и (jjj) и теоремы 2.1 вытекают следующие оценки:

$$\|d_m\|_{s, \Omega} + \|d_m\|_{L^\infty(\Omega)} \leq c_1, \quad \|\beta\|_{L^4(\Omega)} \leq c_2, \quad \|f\|_\Omega \leq c_3, \quad \|\rho_m\|_{1, \Omega} \leq c_4, \quad \|\mathbf{E}_m\|_{1, \Omega} \leq c_5, \quad (41)$$

где константы c_1, c_2, c_5 не зависят от m . Из оценок (41) и условия (j) вытекает, что существуют слабые пределы $d^* \in K_1, \beta^* \in K_2, f^* \in K_3, \rho^* \in H_0^1(\Omega)$ и $\mathbf{E}^* \in \tilde{H}_N^1(\Omega)$ некоторых подпоследовательностей последовательностей $\{d_m\}, \{\beta_m\}, \{f_m\}, \{\rho_m\}$ и $\{\mathbf{E}_m\}$ соответственно.

С учетом этого можно считать, что при $m \rightarrow \infty$ имеем

$$\begin{aligned} \rho_m &\rightarrow \rho^* \text{ слабо в } H^1(\Omega) \text{ и сильно в } L^s(\Omega), \quad s < 6, \\ \mathbf{E}_m &\rightarrow \mathbf{E}^* \text{ слабо в } H^1(\Omega)^3 \text{ и сильно в } L^p(\Omega)^3, \quad p < 6, \\ d_m &\rightarrow d^* \text{ слабо в } H^{1/2}(\Omega), \text{ слабо в } L^\infty(\Omega) \text{ и сильно в } L^q(\Omega), \quad q < 3, \\ \beta_m &\rightarrow \beta^* \text{ слабо в } L^4(\Omega), \quad f_m \rightarrow f^* \text{ слабо в } L^2(\Omega). \end{aligned} \quad (42)$$

Ясно, что $F_2(\mathbf{x}^*) = 0$. Покажем теперь, что $F_1(\mathbf{x}^*, u^*) = 0$, т.е. что

$$(d^* \nabla \rho^*, \nabla h) + \mu_n(\mathbf{E}^* \cdot \nabla \rho^*, h) + \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta^* |\rho^*| \rho^*, h) = (f^*, h) \quad \forall h \in H_0^1(\Omega). \quad (43)$$

При этом пара $(\mathbf{x}_m, u_m)$ удовлетворяет соотношению

$$(d_m \nabla \rho_m, \nabla h) + \mu_n(\mathbf{E}_m \cdot \nabla \rho_m, h) + \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta_m |\rho_m| \rho_m, h) = (f, h) \quad \forall h \in H_0^1(\Omega). \quad (44)$$

Перейдем в (44) к пределу при $m \rightarrow \infty$, начав со слагаемого $(d_m \nabla \rho_m, \nabla h)$:

$$(d_m \nabla \rho_m, \nabla h) - (d^* \nabla \rho^*, \nabla h) = ((d_m - d^*) \nabla \rho_m, \nabla h) + (\nabla(\rho_m - \rho^*), d^* \nabla h). \quad (45)$$

Поскольку $d^* \nabla h \in L^2(\Omega)^3$, то согласно (42) для второго слагаемого в (45) имеем

$$(\nabla(\rho_m - \rho^*), d^* \nabla h) \rightarrow 0 \quad \text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega).$$

Для работы с первым слагаемым в (45) введем последовательность $\{h_n\} \in \mathcal{D}(\Omega)$ такую, что $h_n \rightarrow h$ при $n \rightarrow \infty$ по норме $H^1(\Omega)$. Существование такой последовательности следует из плотности вложения $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$. Справедливо следующее равенство:

$$((d_m - d^*) \nabla \rho_m, \nabla h) = ((d_m - d^*) \nabla \rho_m, \nabla h_n) + ((d_m - d^*) \nabla \rho_m, \nabla(h - h_n)). \quad (46)$$

В силу равномерной ограниченности величин $\|d_m - d^*\|_{L^\infty(\Omega)}$ и $\|\nabla \rho_m\|_\Omega$ по m существует такой номер $N = N(\varepsilon, h)$, что для второго слагаемого в (46) справедливо неравенство

$$|((d_m - d^*) \nabla \rho_m, \nabla(h - h_n))| \leq \|d_m - d^*\|_{L^\infty(\Omega)} \|\nabla \rho_m\|_\Omega \|\nabla(h - h_n)\|_\Omega \leq \varepsilon/2, \quad n \geq N, \quad m \in \mathbb{N}. \quad (47)$$

Из равномерной ограниченности величин $\|\nabla h_n\|_{L^\infty(\Omega)}$ по n и $\|\nabla \rho_m\|_\Omega$ по m соответственно и из (42) следует существование такого номера $M = M(\varepsilon, h)$, что для первого слагаемого в (46) справедливо неравенство

$$|((d_m - d^*) \nabla \rho_m, \nabla h_n)| \leq \|d_m - d^*\|_{L^q(\Omega)} \|\nabla \rho_m\|_\Omega \|\nabla h_n\|_{L^\infty(\Omega)} \leq \varepsilon/2, \quad m \geq M, \quad n \in \mathbb{N}, \quad q \geq 3. \quad (48)$$

Тогда из (46)–(48) вытекает, что

$$((d_m - d^*) \nabla \rho_m, \nabla h) \rightarrow 0 \quad \text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega). \quad (49)$$

В таком случае $(d_m \nabla \rho_m, \nabla h) \rightarrow (d^* \nabla \rho^*, \nabla h)$ при $m \rightarrow \infty$ для всех $h \in H_0^1(\Omega)$.

Для второго слагаемого имеем

$$(\mathbf{E}_m \cdot \nabla \rho_m, h) - (\mathbf{E}^* \cdot \nabla \rho^*, h) = ((\mathbf{E}_m - \mathbf{E}^*) \cdot \nabla \rho_m, h) + (\mathbf{E}^* \cdot \nabla (\rho_m - \rho^*), h).$$

В силу (42), используя лемму 2.1 и (41), получаем для первого слагаемого

$$|((\mathbf{E}_m - \mathbf{E}^*) \cdot \nabla \rho_m, h)| \leq \gamma_1 \|\mathbf{E}_m - \mathbf{E}^*\|_{L^4(\Omega)^3} \|\rho_m\|_{1,\Omega} \|h\|_{1,\Omega} \rightarrow 0 \quad \text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega).$$

Поскольку $\mathbf{E}^* h \in L^2(\Omega)^3$, то для второго слагаемого в силу (42) имеем

$$(\mathbf{E}^* \cdot \nabla (\rho_m - \rho^*), h) = (\nabla (\rho_m - \rho^*), \mathbf{E}^* h) \rightarrow 0 \quad \text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega).$$

Тем самым, $(\mathbf{E}_m \cdot \nabla \rho_m, h) \rightarrow (\mathbf{E}^* \cdot \nabla \rho^*, h)$ при $m \rightarrow \infty$ для всех $h \in H_0^1(\Omega)$.

Далее рассмотрим третье слагаемое:

$$(\beta_m |\rho_m| \rho_m, h) - (\beta^* |\rho^*| \rho^*, h) = (\beta_m (|\rho_m| \rho_m - |\rho^*| \rho^*), h) + (\beta_m - \beta^*, |\rho^*| \rho^*, h). \quad (50)$$

Так как $|\rho^*| \rho^* h \in L^2(\Omega)$, то согласно (42) для второго слагаемого в (50) имеем

$$(\beta_m - \beta^*, |\rho^*| \rho^*, h) \rightarrow 0 \quad \text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega).$$

Первое слагаемое в (50) перепишем в виде

$$(\beta_m (|\rho_m| \rho_m - |\rho^*| \rho^*), h) = (\beta_m |\rho_m| (\rho_m - \rho^*), h) + (\beta_m \rho^* (|\rho_m| - |\rho^*|), h). \quad (51)$$

В силу (42) и равномерной ограниченности по m величин $\|\beta_m\|_{L^4(\Omega)}$, $\|\rho_m\|_{L^4(\Omega)}$ для каждого слагаемого в правой части (51) получаем

$$|(\beta_m |\rho_m| (\rho_m - \rho^*), h)| \leq \|\beta_m\|_{L^4(\Omega)} \|\rho_m\|_{L^4(\Omega)} \|\rho_m - \rho^*\|_{L^4(\Omega)} \|h\|_{L^4(\Omega)} \rightarrow 0$$

$$\text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega),$$

$$|(\beta_m \rho^* (|\rho_m| - |\rho^*|), h)| \leq \|\beta_m\|_{L^4(\Omega)} \|\rho^*\|_{L^4(\Omega)} \|\rho_m - \rho^*\|_{L^4(\Omega)} \|h\|_{L^4(\Omega)} \rightarrow 0$$

$$\text{при } m \rightarrow \infty \quad \forall h \in H_0^1(\Omega).$$

Следовательно, $(\beta_m |\rho_m| \rho_m, h) \rightarrow (\beta^* |\rho^*| \rho^*, h)$ при $m \rightarrow \infty$ для всех $h \in H_0^1(\Omega)$.

Поскольку функционал J слабо полунепрерывен снизу на $X \times K$, то из (41) получаем, что $J(\mathbf{x}^*, u^*) = J^*$. Теорема 4.1 доказана.

Замечание 4.1. Ясно, что все функционалы качества из (40) удовлетворяют условиям теоремы 4.1.

5. ВЫВОД СИСТЕМЫ ОПТИМАЛЬНОСТИ

Через

$$X^* = H^{-1}(\Omega) \times \tilde{H}_N^1(\Omega)^*, \quad Y^* = H_0^1(\Omega) \times L^2(\Omega)$$

обозначим пространства, двойственные к пространствам X и Y , введенным в разд. 4.

Несложно показать, что производная Фреше от оператора $F : X \times K \rightarrow Y$ по состоянию $\mathbf{x}$ в любой точке $(\hat{\mathbf{x}}, \hat{u}) = (\hat{\rho}, \hat{\mathbf{E}}, \hat{d}, \hat{\beta}, \hat{f})$ является линейным непрерывным оператором $F'_x(\hat{\mathbf{x}}, \hat{u}) : X \rightarrow Y$, который каждому элементу $(\tau, \mathbf{e}) \in X$ ставит в соответствие элемент

$$F'_x(\hat{\mathbf{x}}, \hat{u})(\tau, \mathbf{e}) = (\hat{y}_1, \hat{y}_2) \in Y.$$

Здесь $\hat{y}_1 \in H^{-1}(\Omega)$ и $\hat{y}_2 \in L_0^2(\Omega)$ определяются по паре $(\hat{\rho}, \hat{\mathbf{E}})$ и $(\tau, \mathbf{e})$ с помощью следующих соотношений:

$$\begin{aligned} \langle \hat{y}_1, (\tau, \mathbf{e}) \rangle &= (\hat{d} \nabla \tau, \nabla h) + \mu_n (\hat{\mathbf{E}} \cdot \nabla \tau, h) + \mu_n (\mathbf{e} \cdot \nabla \hat{\rho}, h) + (2/\varepsilon \varepsilon_0) \mu_n (\hat{\beta} |\hat{\rho}| \tau, h) \quad \forall (\tau, \mathbf{e}) \in X, \\ \hat{y}_2 &= \operatorname{div} \mathbf{e} - (1/\varepsilon \varepsilon_0) \tau. \end{aligned} \quad (52)$$

Через $F'_x(\hat{\mathbf{x}}, \hat{u})^* : Y^* \rightarrow X^*$ обозначим сопряженный к $F'_x(\hat{\mathbf{x}}, \hat{u})$ оператор.

Следуя общей теории гладко-выпуклых экстремальных задач (см. [23]), введем элемент $\mathbf{y}^* = (\theta, \sigma) \in Y^* = H_0^1(\Omega) \times L^2(\Omega)$, на который будем ссылаться, как на сопряженное состояние, и введем Лагранжиан $\mathcal{L} : X \times K \times \mathbb{R} \times Y^* \rightarrow \mathbb{R}$ по формуле

$$\mathcal{L}(\mathbf{x}, u, \lambda_0, \mathbf{y}^*) = \lambda_0 J(\mathbf{x}, u) + \langle \mathbf{y}^*, F(\mathbf{x}, u) \rangle_{Y^* \times Y} \equiv \lambda_0 J(\mathbf{x}, u) + \langle F_1(\mathbf{x}, u), \theta \rangle + (F_2(\mathbf{x}), \sigma). \quad (53)$$

Доказательство следующей теоремы проведем по схеме, предложенной в гл. 6 из [19].

Теорема 5.1. Пусть выполняются условия (i)–(iii) и (j)–(jjj) и элемент $(\hat{\mathbf{x}}, \hat{u}) \in X \times K$ является точкой локального минимума для задачи (39). Предположим, что функционал качества $I : X \rightarrow \mathbb{R}$ непрерывно дифференцируем по Фреше по состоянию $\mathbf{x}$ в точке $\hat{\mathbf{x}}$. Тогда:

1) существует ненулевой множитель Лагранжа $(\lambda_0, \mathbf{y}^*) = (\lambda_0, \theta, \sigma) \in \mathbb{R}^+ \times Y^*$, с которым выполняется уравнение Эйлера–Лагранжа $F'_x(\hat{\mathbf{x}}, \hat{u})^* \mathbf{y}^* = -\lambda_0 J'_x(\hat{\mathbf{x}}, \hat{u})$ в X^* , эквивалентное соотношениям

$$(\hat{d}\nabla \tau, \nabla \theta) + \mu_n(\hat{\mathbf{E}} \cdot \nabla \tau, \theta) + \frac{2\mu_n}{\varepsilon \varepsilon_0}(\hat{\beta}|\hat{\rho}|\tau, \theta) - \frac{1}{\varepsilon \varepsilon_0}(\tau, \sigma) = -\lambda_0 \frac{\mu_0}{2} \langle I'_p(\hat{\mathbf{x}}), \tau \rangle \quad \forall \tau \in H_0^1(\Omega), \quad (54)$$

$$\mu_n(\mathbf{E} \cdot \nabla \hat{\rho}, \theta) + (\operatorname{div} \mathbf{e}, \sigma) = -\lambda_0 \frac{\mu_0}{2} \langle I'_E(\hat{\mathbf{x}}), \mathbf{e} \rangle \quad \forall \mathbf{e} \in \tilde{H}_N^1(\Omega), \quad (55)$$

и справедлив принцип минимума $\mathcal{L}(\hat{\mathbf{x}}, \hat{u}, \lambda_0, \mathbf{y}^*) \leq \mathcal{L}(\hat{\mathbf{x}}, u, \lambda_0, \mathbf{y}^*)$ для всех $u \in K$, эквивалентный неравенствам

$$\lambda_0 \mu_1(\hat{d}, d - \hat{d})_{s, \Omega} + ((d - \hat{d})\nabla \hat{\rho}, \nabla \theta) \geq 0 \quad \forall d \in K_1, \quad s \geq 1/2, \quad (56)$$

$$((\beta - \hat{\beta})|\hat{\rho}|\hat{\rho}, \theta) \geq 0 \quad \forall \beta \in K_2, \quad (57)$$

$$\lambda_0 \mu_2(\hat{f}, f - \hat{f}) - (f - \hat{f}, \theta) \geq 0 \quad \forall f \in K_3; \quad (58)$$

2) если, к тому же, выполняется условие

$$\frac{\gamma \mu_n C_N}{\varepsilon \varepsilon_0} \|f\|_{\Omega} \leq \lambda_*^2, \quad (59)$$

то любой нетривиальный множитель Лагранжа $(\lambda_0, \mathbf{y}^*)$, удовлетворяющий (54)–(58), является регулярным, т.е. имеет вид $(1, \mathbf{y}^*)$ и определяется единственным образом по заданной паре $(\hat{\mathbf{x}}, \hat{u})$.

Доказательство. Согласно гл. 2 из [23] для доказательства существования множителя Лагранжа $(\lambda_0, \mathbf{y}^*)$ достаточно показать, что $F'_x(\hat{\mathbf{x}}, \hat{u}) : X \rightarrow Y$ – фредгольмов оператор. В силу (52) оператор $F'_x(\hat{\mathbf{x}}, \hat{u}) : X \rightarrow Y$ можно представить в следующем виде:

$$F'_x = \Phi + \hat{\Phi} \equiv (\Phi_1, \Phi_2) + (\hat{\Phi}_1, 0).$$

Здесь $\Phi_2(\mathbf{x}) = \operatorname{div} \mathbf{E} - (1/\varepsilon \varepsilon_0)\tau$ для всех $\mathbf{x} = (\tau, \mathbf{E}) \in X$ и операторы Φ_1 и $\hat{\Phi}_1 : X \rightarrow Y$ определяются формулами

$$\begin{aligned} \langle \Phi_1(\tau), h \rangle &= (\hat{d}\nabla \tau, \nabla h) + \mu_n(\hat{\mathbf{E}} \cdot \nabla \tau, h) + (2\mu_n/\varepsilon \varepsilon_0)(\hat{\beta}|\hat{\rho}|\tau, h), \\ \langle \hat{\Phi}_1(\mathbf{e}), h \rangle &= \mu_n(\mathbf{e} \cdot \nabla \hat{\rho}, h). \end{aligned} \quad (60)$$

Покажем, что оператор $\Phi = (\Phi_1, \Phi_2) : X \rightarrow Y$ является изоморфизмом. Для этого достаточно показать, что для любой пары $(f, s) \in Y$ существует единственное решение $(\tau, \mathbf{e}) \in X$ линейной задачи

$$a_1(\tau, h) = (\hat{d}\nabla \tau, \nabla h) + \mu_n(\hat{\mathbf{E}} \cdot \nabla \tau, h) + (2\mu_n/\varepsilon \varepsilon_0)(\hat{\beta}|\hat{\rho}|\tau, h) = \langle f, h \rangle \quad \forall h \in H_0^1(\Omega), \quad (61)$$

$$\operatorname{div} \mathbf{e} - (1/\varepsilon \varepsilon_0)\tau = s \quad \text{в } \Omega. \quad (62)$$

Из леммы 2.1 и условия (iii) вытекают непрерывность и коэрцитивность билинейной формы a_1 с константой $\lambda_* = d_0 \delta_1$. Тогда по теореме Лакса–Мильграма существует единственное решение $\tau \in H_0^1(\Omega)$ задачи (61), для которого справедлива оценка

$$\|\tau\|_{1,\Omega} \leq C_* \|f\|_{-1,\Omega}, \quad C_* = \lambda_*^{-1}. \quad (63)$$

Тогда в силу леммы 2.2 для любой функции $s \in L^2(\Omega)$ существует единственное решение $e \in \tilde{H}_N^1(\Omega)$ задачи (62) и справедлива оценка

$$\|e\|_{1,\Omega} \leq C_N \left((C_*/\varepsilon \varepsilon_0) \|f\|_{-1,\Omega} + \|s\|_{\Omega} \right). \quad (64)$$

Докажем, что оператор $\hat{\Phi} = (\hat{\Phi}_1, 0) : X \rightarrow Y$, определенный формулой (60), является непрерывным и компактным. Поскольку пространство $H^1(\Omega)^3$ непрерывно и компактно вложено в $L^4(\Omega)^3$, то данное утверждение вытекает из оценки

$$|(e \cdot \hat{\rho}, h)| \leq \gamma'_2 \|e\|_{L^4(\Omega)^3} \|\hat{\rho}\|_{1,\Omega} \|h\|_{1,\Omega}.$$

В таком случае оператор $F'_x(x, u) : X \rightarrow Y$ является фредгольмовым, как сумма изоморфизма $\Phi : X \rightarrow Y$ и компактного оператора $\hat{\Phi} : X \rightarrow Y$.

Для доказательства второго утверждения теоремы 5.1 достаточно доказать, что однородная система (54), (55) (при $\lambda_0 = 0$) имеет только тривиальное решение $y^* = (\theta, \sigma) \equiv 0$.

Предположим противное, т.е., что существует по крайней мере одно нетривиальное решение $y^* = (\theta, \sigma) \in Y^*$ системы (54), (55) при $\lambda_0 = 0$, в которой элементы $\hat{x} = (\hat{\rho}, \hat{E})$ и $\hat{u} = (\hat{d}, \hat{\beta}, \hat{f})$ связаны соотношением $F(\hat{x}, \hat{u}) = 0$.

Подставим $\tau = \theta$ и $e = \tilde{e}$, такое что $\operatorname{div} \tilde{e} = \sigma$ в Ω , в (54), (55). Существование указанной функции $\tilde{e}$ вытекает из леммы 2.2, причем справедлива оценка $\|\tilde{e}\|_{1,\Omega} \leq C_N \|\sigma\|_{\Omega}$. В результате приходим к соотношениям

$$(\hat{d} \nabla \theta, \nabla \theta) + \mu_n (\hat{E} \cdot \nabla \theta, \theta) + \frac{2\mu_n}{\varepsilon \varepsilon_0} (\hat{\beta} |\hat{\rho}| \theta, \theta) - \frac{1}{\varepsilon \varepsilon_0} (\theta, \sigma) = 0, \quad (65)$$

$$\mu_n (\tilde{e} \cdot \nabla \hat{\rho}, \theta) + (\sigma, \sigma) = 0. \quad (66)$$

Из (66) с учетом (7) и (19) выводим оценку

$$\|\sigma\|_{\Omega} \leq \gamma_1 \mu_n C_* C_N \|f\|_{\Omega} \|\theta\|_{1,\Omega}. \quad (67)$$

Используя (67) и коэрцитивность формы a_1 , из (65) приходим к неравенству

$$\lambda_* \|\theta\|_{1,\Omega}^2 \leq (1/\varepsilon \varepsilon_0) \gamma_1 \mu_n C_* C_N \|f\|_{\Omega} \|\theta\|_{1,\Omega}^2. \quad (68)$$

Из (68) вытекает, что если выполняется (59), то $\|\theta\|_{1,\Omega} = 0$ или $\theta = 0$ в Ω . Тогда из (67) получаем, что $\sigma = 0$, но это противоречит предполагаемой нетривиальности множителя Лагранжа (θ, σ) . Единственность и регулярность множителя Лагранжа $(1, y^*)$ при условии (59) вытекают из фредгольмовости оператора $F'_x(\hat{x}, \hat{u}) : X \rightarrow Y$. Теорема 5.1 доказана.

6. ЕДИНСТВЕННОСТЬ РЕШЕНИЯ ЭКСТРЕМАЛЬНОЙ ЗАДАЧИ

В данном разделе докажем локальную единственность решения задачи управления

$$J(x, f) \equiv \frac{\mu_0}{2} \|\rho - \rho^d\|_Q^2 + \frac{\mu_1}{2} \|f\|_{\Omega}^2 \rightarrow \inf, \quad F(x, f) = 0, \quad (x, f) \in X \times K, \quad (69)$$

отвечающую функционалу качества I_1 в (40), роль управления в которой играет только одна функция f .

Пусть выполняются следующие условия:

(j)' $K \subset L^2(\Omega)$ — непустое выпуклое замкнутое множество;

(jj)' $\mu_0 > 0$, $\mu_i \geq 0$, множество K ограничено или $\mu_i > 0$, $i = 0, 1$.

Из теоремы 2.1 вытекают оценки

$$\|\rho\|_{1,\Omega} \leq M_\rho \equiv \sup_{f \in K} C_* \|f\|_\Omega, \quad \|\mathbf{E}\|_{1,\Omega} \leq M_E \equiv \sup_{f \in K} (1/\varepsilon \varepsilon_0) C_N C_* \|f\|_\Omega, \quad C_* = \lambda_*^{-1}. \quad (70)$$

Положим

$$\rho = \rho_1 - \rho_2, \quad \mathbf{E} = \mathbf{E}_1 - \mathbf{E}_2, \quad f = f_1 - f_2. \quad (71)$$

Для удобства разобьем доказательство единственности решения задачи (69) на отдельные этапы.

1. Вывод оценок норм разностей ρ и $\mathbf{E}$ через норму разности управлений f .

Для этого вычтем уравнения (11), (12), записанные при $(\rho_2, \mathbf{E}_2, f_2)$, из уравнений (11), (12) для $(\rho_1, \mathbf{E}_1, f_1)$. Рассуждая, как при выводе (21), (22), получаем

$$\begin{aligned} (d\nabla \rho, \nabla h) + \mu_n(\mathbf{E}_1 \cdot \nabla \rho, h) + \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta |\rho_1| \rho, h) = \\ = -\mu_n(\mathbf{E} \cdot \nabla \rho_2, h) - \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta (|\rho_1| - |\rho_2|) \rho_2, h) + (f, h) \quad \forall h \in H_0^1(\Omega), \end{aligned} \quad (72)$$

$$\operatorname{div} \mathbf{E} = \frac{1}{\varepsilon \varepsilon_0} \rho \quad \text{в } \Omega. \quad (73)$$

Из (73) в силу леммы 2.2 вытекает оценка

$$\|\mathbf{E}\|_{1,\Omega} \leq (1/\varepsilon \varepsilon_0) C_N \|\rho\|_{1,\Omega}. \quad (74)$$

С учетом (74), рассуждая как при доказательстве локальной единственности решения задачи 1, при выполнении условия

$$\frac{\mu_n}{\varepsilon \varepsilon_0} (\gamma_1 C_N + C_4^3 \|\beta\|_{L^4(\Omega)}) M_\rho \leq \frac{\lambda_*}{2} \quad (75)$$

из неравенства

$$\lambda_* \|\rho\|_{1,\Omega}^2 \leq (\mu_n/\varepsilon \varepsilon_0) (\gamma_1 C_N + C_4^3 \|\beta\|_{L^4(\Omega)}) M_\rho \|\rho\|_{1,\Omega}^2 + \|f\|_\Omega \|\rho\|_{1,\Omega} \quad (76)$$

приходим к следующей оценке:

$$\|\rho\|_{1,\Omega} \leq 2C_* \|f\|_\Omega. \quad (77)$$

С учетом (77) оценка (74) примет вид

$$\|\mathbf{E}\|_{1,\Omega} \leq (2/\varepsilon \varepsilon_0) C_N C_* \|f\|_\Omega. \quad (78)$$

2. Вывод оценок для множителей Лагранжа θ_i и σ_i , $i = 1, 2$.

Предполагая, что выполняются условия теоремы 5.1, запишем (54), (55) при $I = I_1(\rho) = \|\rho - \rho^d\|_Q^2$, $\lambda_0 = 1$ для $(\rho_i, \mathbf{E}_i, \theta_i, \sigma_i)$. Будем иметь

$$\begin{aligned} (d\nabla \tau, \nabla \theta_i) + \mu_n(\mathbf{E}_i \cdot \nabla \tau, \theta_i) + \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta |\rho_i| \tau, \theta_i) - \frac{1}{\varepsilon \varepsilon_0} (\tau, \sigma_i) = \\ = -\mu_0(\rho_i - \rho^d, \tau)_Q \quad \forall \tau \in H_0^1(\Omega), \end{aligned} \quad (79)$$

$$\mu_n(\mathbf{e} \cdot \nabla \rho_i, \theta_i) + (\operatorname{div} \mathbf{e}, \sigma_i) = 0 \quad \forall \mathbf{e} \in \tilde{H}_N^1(\Omega). \quad (80)$$

В силу леммы 2.2 существуют функции $\tilde{\mathbf{e}}_i$ такие, что $\operatorname{div} \tilde{\mathbf{e}}_i = \sigma_i$ и $\|\tilde{\mathbf{e}}_i\|_{1,\Omega} \leq C_N \|\sigma_i\|_\Omega$, $i = 1, 2$. Полагая $\mathbf{e} = \tilde{\mathbf{e}}_i$ в (80), приходим к неравенству

$$\|\sigma_i\|_\Omega \leq \mu_n \gamma_1 C_N M_\rho \|\theta_i\|_{1,\Omega}, \quad i = 1, 2. \quad (81)$$

Полагая $\tau = \theta_i$ в (79), с учетом лемм 2.1 и 2.2 и оценки (81) при выполнении условия

$$\frac{\mu_n \gamma_1 C_N}{\varepsilon \varepsilon_0} M_\rho \leq \frac{\lambda_*}{2} \quad (82)$$

получаем оценку для θ_i :

$$\|\theta_i\|_{1,\Omega} \leq \mu_0 M_\theta \equiv 2\mu_0 C_* \left(M_\rho + \|\rho^d\|_Q \right), \quad i = 1, 2. \quad (83)$$

С учетом (83) из (81) получаем оценку для σ_i :

$$\|\sigma_i\|_\Omega \leq \mu_0 M_\sigma \equiv 2\mu_0 \mu_n \gamma_1 C_N M_\rho M_\theta, \quad i = 1, 2. \quad (84)$$

3. Вывод оценок норм θ и σ через норму f .

В дополнение к (71) положим $\theta = \theta_1 - \theta_2$ и $\sigma = \sigma_1 - \sigma_2$ и вычтем равенства (79), (80), записанные при $i = 2$, из (79), (80) при $i = 1$. Будем иметь

$$\begin{aligned} & (d\nabla \tau, \nabla \theta) + \mu_n (\mathbf{E}_1 \cdot \nabla \tau, \theta) + \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta |\rho_1| \tau, \theta) = \\ & = -\mu_n (\mathbf{E} \cdot \nabla \tau, \theta_2) - \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta (|\rho_1| - |\rho_2|) \tau, \theta_2) + \frac{1}{\varepsilon \varepsilon_0} (\tau, \sigma) - \mu_0 (\rho, \tau)_Q \quad \forall \tau \in H_0^1(\Omega), \end{aligned} \quad (85)$$

$$\mu_n (\mathbf{e} \cdot \nabla \rho, \theta_1) + \mu_n (\mathbf{e} \cdot \nabla \rho_2, \theta) + (\operatorname{div} \mathbf{e}, \sigma) = 0 \quad \forall \mathbf{e} \in \tilde{H}_N^1(\Omega). \quad (86)$$

Рассуждая, как при выводе (84), из (86) с учетом (77) приходим к неравенству

$$\|\sigma\|_\Omega \leq \mu_0 \mu_n \gamma_1 M_\theta C_N \|\rho\|_{1,\Omega} + \mu_n \gamma_1 M_\rho C_N \|\theta\|_{1,\Omega} \leq 2\mu_0 \mu_n \gamma_1 C_* M_\theta C_N \|f\|_\Omega + \mu_n \gamma_1 M_\rho C_N \|\theta\|_{1,\Omega}. \quad (87)$$

Подставим $\tau = \theta$ в (85). Применяя неравенство Гёльдера и оценки леммы 2.1, с учетом (77), (78) будем иметь

$$\begin{aligned} \lambda_* \|\theta\|_{1,\Omega} & \leq \mu_0 \mu_n \gamma_1 M_\theta \|\mathbf{E}\|_{1,\Omega} + \mu_0 (2/\varepsilon \varepsilon_0) \mu_n C_4^3 M_\theta \|\beta\|_{L^4(\Omega)} \|\rho\|_{1,\Omega} + \\ & + (1/\varepsilon \varepsilon_0) \|\sigma\|_\Omega + \mu_0 \|\rho\|_{1,\Omega} \leq \mu_0 \omega_1 \|f\|_\Omega + (1/\varepsilon \varepsilon_0) \mu_n \gamma_1 C_N M_\rho \|\theta\|_{1,\Omega}, \\ \omega_1 & = (4/\varepsilon \varepsilon_0) \mu_n M_\theta C_* \left(\gamma_1 C_N + C_4^3 \|\beta\|_{L^4(\Omega)} \right) + 2C_*. \end{aligned} \quad (88)$$

При выполнении условия (82) из (88) выводим оценку для разности θ :

$$\|\theta\|_{1,\Omega} \leq 2\mu_0 C_* \omega_1 \|f\|_\Omega. \quad (89)$$

С учетом (89) из (87) получаем оценку для разности σ :

$$\|\sigma\|_\Omega \leq \mu_0 \omega_2 \|f\|_\Omega, \quad \omega_2 = 2\mu_n \gamma_1 C_* C_N (M_\theta + \omega_1 M_\rho). \quad (90)$$

4. Вывод основного неравенства.

Полагая в (85) $\tau = \rho$, получим

$$\begin{aligned} & (d\nabla \rho, \nabla \theta) + \mu_n (\mathbf{E}_1 \cdot \nabla \rho, \theta) + \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta |\rho_1| \rho, \theta) = \\ & = -\mu_n (\mathbf{E} \cdot \nabla \rho, \theta_2) - \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta (|\rho_1| - |\rho_2|) \rho, \theta_2) + \frac{1}{\varepsilon \varepsilon_0} (\rho, \sigma) - \mu_0 (\rho, \rho)_Q. \end{aligned} \quad (91)$$

Положим теперь $h = \theta$ в (72). Будем иметь

$$(d\nabla \rho, \nabla \theta) + \mu_n (\mathbf{E}_1 \cdot \nabla \rho, \theta) + \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta |\rho_1| \rho, \theta) = -\mu_n (\mathbf{E} \cdot \nabla \rho_2, \theta) - \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta (|\rho_1| - |\rho_2|) \rho_2, \theta) + (f, \theta). \quad (92)$$

Вычитая (92) из (91), получим

$$\begin{aligned} & \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta |\rho_1| \rho, \theta) + \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta (|\rho_1| - |\rho_2|) \rho, \theta_2) - \frac{1}{\varepsilon \varepsilon_0} (\rho, \sigma) + \mu_0 \|\rho\|_Q^2 + (f, \theta) = \\ & = \mu_n (\mathbf{E} \cdot \nabla \rho_2, \theta) - \mu_n (\mathbf{E} \cdot \nabla \rho, \theta_2) + \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta (|\rho_1| - |\rho_2|) \rho_2, \theta). \end{aligned} \quad (93)$$

Положим $f = f_1$ в неравенстве (58) (при $\lambda_0 = 1$ и при $\mu_2 = \mu_1$ согласно (70)), записанном для (f_2, θ_2) , и положим $f = f_2$ в (58) при (f_1, θ_1) . Будем иметь

$$\mu_1(f_2, f) - (f, \theta_2) \geq 0, \quad -\mu_1(f_1, f) + (f, \theta_1) \geq 0. \quad (94)$$

Складывая эти неравенства, приходим к оценке

$$(f, \theta) \geq \mu_1 \|f\|_{\Omega}. \quad (95)$$

С учетом (95) из (93) получаем основное неравенство

$$\begin{aligned} & \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta |p_1| p, \theta) + \frac{2\mu_n}{\varepsilon \varepsilon_0} (\beta (|p_1| - |p_2|) p, \theta_2) - \frac{\mu_n}{\varepsilon \varepsilon_0} (\beta (|p_1| - |p_2|) p_2, \theta) - \\ & - \mu_n (\mathbf{E} \cdot \nabla p_2, \theta) + \mu_n (\mathbf{E} \cdot \nabla p, \theta_2) - \frac{1}{\varepsilon \varepsilon_0} (p, \sigma) + \mu_0 \|p\|_Q^2 + \mu_1 \|f\|_{\Omega}^2 \leq 0. \end{aligned} \quad (96)$$

5. Оценка слагаемых в (96) через норму разности f .

Применяя неравенство Гёльдера, оценки леммы 2.1 и используя (70), (77), (78), (83), (89) и (90), оценим слагаемые в левой части последнего неравенства:

$$\begin{aligned} & \frac{\mu_n}{\varepsilon \varepsilon_0} |\beta |p_1| p, \theta| \leq \frac{\mu_n}{\varepsilon \varepsilon_0} C_4^3 M_p \|\beta\|_{L^4(\Omega)} \|p\|_{1,\Omega} \|\theta\|_{1,\Omega} \leq 4\mu_0 \frac{\mu_n}{\varepsilon \varepsilon_0} C_4^3 M_p \|\beta\|_{L^4(\Omega)} C_*^2 \omega_1 \|f\|_{\Omega}^2; \\ & \frac{2\mu_n}{\varepsilon \varepsilon_0} |\beta (|p_1| - |p_2|) p, \theta_2| \leq \frac{2\mu_n}{\varepsilon \varepsilon_0} C_4^3 \|\beta\|_{L^4(\Omega)} \|p\|_{1,\Omega}^2 \|\theta_2\|_{1,\Omega} \leq 4\mu_0 \frac{2\mu_n}{\varepsilon \varepsilon_0} C_4^3 M_{\theta} \|\beta\|_{L^4(\Omega)} C_*^2 \|f\|_{\Omega}^2; \\ & \frac{\mu_n}{\varepsilon \varepsilon_0} |\beta (|p_1| - |p_2|) p_2, \theta| \leq \frac{\mu_n}{\varepsilon \varepsilon_0} C_4^3 M_p \|\beta\|_{L^4(\Omega)} \|p\|_{1,\Omega} \|\theta\|_{1,\Omega} \leq 4\mu_0 \frac{\mu_n}{\varepsilon \varepsilon_0} C_4^3 M_p \|\beta\|_{L^4(\Omega)} C_*^2 \omega_1 \|f\|_{\Omega}^2; \\ & \mu_n |(\mathbf{E} \cdot \nabla p_2, \theta)| \leq 4\mu_0 \frac{\mu_n}{\varepsilon \varepsilon_0} \gamma_1 M_p C_N C_*^2 \omega_1 \|f\|_{\Omega}^2; \\ & \mu_n |(\mathbf{E} \cdot \nabla p, \theta_2)| \leq 4\mu_0 \frac{\mu_n}{\varepsilon \varepsilon_0} \gamma_1 M_{\theta} C_N C_*^2 \|f\|_{\Omega}^2; \\ & \frac{1}{\varepsilon \varepsilon_0} |(p, \sigma)| \leq \frac{1}{\varepsilon \varepsilon_0} \|p\|_{1,\Omega} \|\sigma\|_{1,\Omega} \leq 2\mu_0 \frac{1}{\varepsilon \varepsilon_0} C_* \omega_2 \|f\|_{\Omega}^2 = 4\mu_0 \frac{\mu_n}{\varepsilon \varepsilon_0} \gamma_1 (M_{\theta} + \omega_1 M_p) C_N C_*^2 \|f\|_{\Omega}^2. \end{aligned}$$

Пусть выполняются условия

$$8(\mu_n/\varepsilon \varepsilon_0) C_*^2 (M_{\theta} + \omega_1 M_p) (C_4^3 \|\beta\|_{L^4(\Omega)} + \gamma_1 C_N) < \mu_1/\mu_0, \quad (97)$$

где параметры M_p , M_{θ} и ω_1 введены соответственно в (70), (83) и (88).

Тогда из (96), (77) и (78) вытекает, что $f = 0$, $p = 0$ и $\mathbf{E} = \mathbf{0}$ в Ω .

Сформулируем полученный результат в виде теоремы.

Теорема 6.1. Пусть в дополнение к условиям (i)–(iii) и (j)', (jj)' выполняются условия (75) и (97).

Тогда задача управления (69) имеет единственное решение $(p, \mathbf{E}, f) \in H_0^1(\Omega) \times \tilde{H}_N^1(\Omega) \times K$.

7. ЗАКЛЮЧЕНИЕ

В настоящей работе обоснована корректность задачи моделирования зарядки неоднородного полярного диэлектрика и проведен качественный анализ свойств ее решения. При этом обобщены результаты [15] по исследованию разрешимости краевой задачи (1)–(3), полученные при $\beta \equiv 1$. Здесь отметим работы [24–26] по исследованию схожих реакционно-диффузионных моделей, также учитывающих зависимость аналогичной нелинейности от пространственных переменных, которая описывает неоднородность протекания химической реакции в рассматриваемой области.

Одним из основных результатов настоящей работы является доказательство разрешимости задачи управления при минимальной гладкости используемых в ней мультипликативных управлений и вывод для нее системы оптимальности. На основе анализа данной системы установлены достаточные условия локальной единственности решения задачи распределенного управления.

Отметим также аналогичные результаты для близких моделей тепломассопереноса, сложного теплообмена и поляризационных процессов, полученные в работах [27–31].

СПИСОК ЛИТЕРАТУРЫ

1. Chan D.S.H., Sim K.S., Phang J.C.H. A simulation model for electron irradiation induced specimen charging in a scanning electron microscope // *Scanning Spectroscopy*. 1993. V. 7. № 31. P. 847–859.
2. Sessler G.M., Yang G.M. Charge dynamics in electron-irradiated polymers // *Braz. J. Phys.* 1999. V. 29. № 2. P. 233–240.
3. Suga H., Tadokoro H., Kotera M. A simulation of electron beam induced charging-up of insulators // *Electron Microscopy*. 1998. V. 1. P. 177–178.
4. Cazaux J. About the mechanisms of charging in EPMA, SEM, and ESEM with their time evolution // *Microscopy and Microanalysis*. 2004. V. 10. № 6. P. 670–680.
5. Борисов С.С., Грачев Е.А., Зайцев С.И. Моделирование поляризации диэлектрика в процессе облучения электронным пучком // *Прикладная физика*. 2004. № 1. С. 118–124.
6. Kotera M., Yamaguchi K., Suga H. Dynamic simulation of electron-beam-induced charging up of insulators // *Japan J. Appl. Phys.* 1999. V. 38. № 12 B. P. 7176–7179.
7. Ohya K., Inai K., Kuwada H., Hauashi T., Saito M. Dynamic simulation of secondary electron emission and charging up of an insulting material // *Surface and Coating Technology*. 2008. V. 202. P. 5310–5313.
8. Maslovskaya A.G. Physical and mathematical modeling of the electron-beam-induced charging of ferroelectrics during the process of domain structure switching // *J. of Surface Investigation*. 2013. V. 7. № 4. P. 680–684.
9. Pavelchuk A.V., Maslovskaya A.G. Approach to numerical implementation of the drift-diffusion model of field effects induced by a moving source // *Russ. Phys. J.* 2020. V. 63. P. 105–112.
10. Raftari B., Budko N.V., Vuik C. Self-consistence drift-diffusion-reaction model for the electron beam interaction with dielectric samples // *J. Appl. Phys.* 2015. V. 118. P. 204101 (17).
11. Chezganov D.S., Kuznetsov D.K., Shur V.Ya. Simulation of spatial distribution of electric field after electron beam irradiation of MgO -doped $LiNbO_3$ covered by resist layer // *Ferroelectrics*. 2016. V. 496. P. 70–78.
12. Maslovskaya A., Pavelchuk A. Simulation of dynamic charging processes in ferroelectrics irradiated with SEM // *Ferroelectrics*. 2015. V. 476. P. 157–167.
13. Maslovskaya A., Sivunov A.V. Simulation of electron injection and charging processes in ferroelectrics modified with SEM-techniques // *Solid State Phenomena*. 2014. V. 213. P. 119–124.
14. Arat K.T., Klimpel T., Hagen C.W. Model improvements to simulate charging in scanning electron microscope // *J. of Micro/ Nanolithography, MEMS, and MOEMS*, 2019. V. 18. № 4. P. 04403 (13).
15. Бризицкий Р.В., Максимов Н.Н., Масловская А.Г. Теоретический анализ и численная реализация стационарной диффузионно-дрейфовой модели зарядки полярных диэлектриков // *Ж. вычисл. матем. и матем. физ.* 2022. Т. 62. № 10. С. 1696–1706.
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. Brizitskii R.V., Saritskaya Zh. Yu. Optimization analysis of the inverse coefficient problem for the nonlinear convection-diffusion-reaction equation // *J. Inverse Ill-Posed Probl.* 2018. V. 26. № 6. P. 821–833.
18. Maksimova N.N., Brizitskii R.V. Inverse problem of recovering the electron diffusion coefficient // *Дальневосточный матем. журн.* 2022. Т. 22. № 2. С. 201–206.
19. Алексеев Г.В. Оптимизация в стационарных задачах тепломассопереноса и магнитной гидродинамики. М.: Научный мир, 2010. 412 с.
20. Buffa A. Some numerical and theoretical problems in computational electromagnetism. Thesis. 2000.
21. Гилбарг Д., Трудингер М. Эллиптические дифференциальные уравнения с частными производными второго порядка. М.: Наука, 1989. 463 с.
22. Berninger H. Non-overlapping domain decomposition for the Richards equation via superposition operators // *Domain Decomposition Methods in Science and Engineering XVIII*. Springer, 2009. P. 169–176.
23. Фурсиков А.В. Оптимальное управление распределенными системами. Теория и приложения. Новосибирск: Научн. книга, 1999. 352 с.
24. Алексеев Г.В., Бризицкий Р.В., Сарицкая Ж.Ю. Оценки устойчивости решений экстремальных задач для нелинейного уравнения конвекции–диффузии–реакции // *Сиб. журн. индустр. матем.* 2016. Т. 19. № 2. С. 3–16.
25. Бризицкий Р.В., Сарицкая Ж.Ю. Обратные коэффициентные задачи для нелинейного уравнения конвекции–диффузии–реакции // *Изв. РАН. Сер. матем.* 2018. Т. 82. Вып. 1. С. 17–33.

26. Бризицкий Р.В., Сарицкая Ж.Ю. Задача граничного управления для нелинейного уравнения конвекции–диффузии–реакции // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 12. С. 2139–2152.
27. Алексеев Г.В. Коэффициентные обратные экстремальные задачи для стационарных уравнений тепло-массопереноса // Ж. вычисл. матем. и матем. физ. 2007. Т. 47. № 6. С. 1055–1076.
28. 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. of Math. Analys. and Appl. 2018. V. 460. № 2. P. 737–744.
29. Chebotarev A. Yu., 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 // Commun. Nonlinear Sci. Numer. Simulat. 2018. V. 57. P. 290–298.
30. Chebotarev A. Y., Grenkin G. V., Kovtanyuk A. E. Inhomogeneous steady-state problem of complex heat transfer // ESAIM: Math. Model. and Numeric. Analys. 2017. V. 51. № 6. P. 2511–2519.
31. Maslovskaya A. G., Moroz L. I., Chebotarev A. Y., Kovtanyuk A. E. Theoretical and numerical analysis of the Landau-Khalatnikov model of ferroelectric hysteresis // Commun. Nonlinear Sci. Numer. Simulat. 2021. V. 93. P. 105524.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 519.635

THREE-DIMENSIONAL NUMERICAL TRANSFER MODEL
USING A MONOTONIZED Z-SCHEME¹⁾

© 2023 г. N. M. Kashchenko^{1,*}, S. A. Ishanov^{1,**}, E. V. Zubkov^{1,***}, L. V. Zinin^{1,****}

¹ Immanuel Kant Baltic Federal University, Kaliningrad 236016, Russia

*e-mail: kaschtschenko@mail.ru

**e-mail: SIshanov@kantiana.ru

***e-mail: evgenyzubkov@mail.ru,

****e-mail: leonid.zinin@gmail.com

Поступила в редакцию 27.06.2022 г.

Переработанный вариант 24.04.2023 г.

Принята к публикации 29.05.2023 г.

Трехмерная численная модель переноса с использованием монотонизированной Z-схем. Целью работы является исследование трехмерной конечно-разностной монотонизированной схемы второго порядка точности для уравнения переноса. Исследование проведено для модельных трехмерных уравнений переноса несжимаемой среды. В работе исследуются свойства трехмерного обобщения Z-схемы с нелинейной коррекцией. Данная работа является развитием предыдущих работ автора, в которых построена нелинейная коррекция одномерных уравнений переноса. При коррекции предложенная схема использует вместо потоков их аналог — “косые разности”, содержащие значения из разных временных слоев. Численно исследована монотонность полученной нелинейной схемы для семейства функций-ограничителей, как для гладких решений, так и для негладких непрерывных решений. Построенная схема является абсолютно устойчивой, но теряет свойство монотонности при превышении шага Куранта. Отличительной особенностью предложенной конечно-разностной схемы является минимальность ее шаблона. Построенная численная схема предназначена для моделей плазменных неустойчивостей различных масштабов в низкоширотной ионосферной плазме Земли. Одна из реальных задач, при решении которых возникают подобные уравнения, — это численное моделирование сильно нестационарных среднемасштабных и мелкомасштабных процессов в низкоширотной ионосфере Земли в условиях возникновения неустойчивости Рэлея–Тейлора и неустойчивостей других типов, что приводит к явлению F-рассеяния. Вследствие того, что процессы переноса в ионосферной плазме контролируются магнитным полем, в поперечном к магнитному полю направлении предполагается выполнение условия несжимаемости плазмы.

Ключевые слова: нелинейная конечно-разностная схема, Z-схема, математическое моделирование, численное моделирование, уравнения переноса, ионосфера, неустойчивость Рэлея–Тейлора, плазменные неустойчивости.

DOI: 10.31857/S004446692309017X, **EDN:** XEONIP

¹⁾ Полный текст статьи печатается в английской версии журнала.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 519.635

УНИВЕРСАЛЬНЫЙ ПОДХОД К АНАЛИЗУ ДИССИПАТИВНЫХ СВОЙСТВ ЧИСЛЕННОГО МЕТОДА РЕШЕНИЯ УРАВНЕНИЙ ГАЗОДИНАМИКИ¹⁾

© 2023 г. П. В. Чувахов^{1,2,*}, И. О. Погорелов^{1,2,**}, К. В. Шубин^{2,***}

¹ 140180 Жуковский, М.о., ул. Жуковского, 1, Центральный аэрогидродинамический институт
им. Н.Е. Жуковского, Россия

² 141701 Долгопрудный, М.о., Институтский пер., 9, Московский физико-технический институт
(национальный исследовательский университет), Россия

*e-mail: pavel_chuvahov@mail.ru

**e-mail: ilya.pogorelov@phystech.edu

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

Поступила в редакцию 27.10.2022 г.
Переработанный вариант 26.03.2023 г.
Принята к публикации 28.04.2023 г.

Предложен универсальный подход к верификации численного метода решения уравнений Навье–Стокса, который позволяет надежно оценивать его диссипативные свойства. Подход основан на явлении вязкого затухания слабых элементарных возмущений, распространяющихся в однородном потоке. Соответствующее теоретическое решение позволяет определить порядок сходимости численного метода. Библ. 8. Фиг. 5.

Ключевые слова: численная диссипация, численная вязкость, численный метод, уравнения Навье–Стокса, уравнения Эйлера, сходимость.

DOI: 10.31857/S0044466923090065, **EDN:** LRQWBQ

1. ВВЕДЕНИЕ

Численный метод неизбежно привносит в нестационарное решение дополнительную диссипацию и дисперсию, обусловленную аппроксимацией исходного непрерывного решения на дискретной пространственно-временной сетке. Численная диссипация — или схемная вязкость, или численная вязкость, или диссипативные свойства численного метода и т. д. — зависит от используемого численного метода и обязана уменьшаться по мере уменьшения шага сетки (свойство сходимости численного решения) [1]. При моделировании нестационарных газодинамических явлений численная диссипация проявляется в виде избыточного затухания возмущений, развивающихся в расчетной области. Поскольку это может приводить к искажению исследуемых физических явлений, проводятся исследования диссипативных и дисперсионных свойств конкретных разностных схем (см., например, [2–6]). В качестве примера детального анализа дисперсионных и диссипативных свойств разностных схем для уравнения переноса в линейном случае можно указать монографию [7].

В настоящей работе предлагается универсальный подход к верификации произвольного численного метода решения уравнений газодинамики путем оценки его диссипативных свойств. Рассматривается общий случай полных уравнений Навье–Стокса, а применимость подхода тестируется на примере конкретной реализации метода Годунова. Оценка проводится с помощью методического исследования простейшего процесса — вязкого затухания слабой монохроматической акустической волны или волны энтропии в однородном газовом потоке без газодинамических разрывов. Предполагается, что такая оценка останется справедливой для широкого класса задач.

Для выбранной задачи можно получить точное теоретическое решение.

¹⁾ Работа выполнена при финансовой поддержке РНФ в МФТИ (проект № 19-79-10132).

2. ТЕОРЕТИЧЕСКОЕ РЕШЕНИЕ

Для простоты рассмотрим задачу развития малых возмущений в одномерной постановке, когда вектор направления энергии возмущений направлен вдоль оси Ox . После простых алгебраических преобразований уравнения Навье—Стокса, приведенные в [8] в консервативном виде в трехмерной безразмерной постановке, сводятся к виду (добавлено уравнение состояния):

$$\begin{aligned}\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} &= 0, \\ \rho \frac{\partial u}{\partial t} + \rho u \frac{\partial u}{\partial x} + \frac{\partial p}{\partial x} - \frac{4}{3\text{Re}_{\infty,L}} \frac{\partial}{\partial x} \left(\mu \frac{\partial u}{\partial x} \right) &= 0, \\ \frac{1}{\gamma} \frac{\partial \rho T}{\partial t} + \frac{\partial \rho u T}{\partial x} - (\gamma - 1) M_{\infty}^2 u \frac{\partial p}{\partial x} - \frac{1}{\text{Re}_{\infty,L} \text{Pr}} \frac{\partial}{\partial x} \left(\mu \frac{\partial T}{\partial x} \right) - \frac{4\mu}{3\text{Re}_{\infty,L}} \frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x} \right)^2 (\gamma - 1) M_{\infty}^2 &= 0, \\ \gamma M_{\infty}^2 p &= \rho T.\end{aligned}$$

В этих уравнениях единичная длина, скорость и температура соответствуют характерному размеру задачи L^* и параметрам набегающего потока V_{∞}^* , T_{∞}^* . Звездочка в верхнем индексе обозначает размерную величину. Единичное давление соответствует удвоенному скоростному напору $\rho_{\infty}^* V_{\infty}^{*2}$. Число Маха набегающего потока вычисляется как $M_{\infty} = V_{\infty}^*/a$, где a — скорость звука, $\text{Re}_{\infty,L} = \rho_{\infty}^* V_{\infty}^* L^*/\mu_{\infty}^*$ — число Рейнольдса, $\text{Pr} = \mu_{\infty}^* c_p^*/\lambda_{\infty}^* = 0.72$ — число Прандтля, $\gamma = c_p^*/c_v^* = 1.4$ — показатель адиабаты.

Линейное приближение этой системы относительно однородного набегающего потока — $(u, \rho, p, T) = (1, 1, 1/\gamma M_{\infty}^2, 1) + (u', \rho', p', T')$ — имеет вид

$$\begin{aligned}\frac{\partial \rho'}{\partial t} + \frac{\partial u'}{\partial x} + \frac{\partial p'}{\partial x} &= 0, \quad \frac{\partial u'}{\partial t} + \frac{\partial u'}{\partial x} + \frac{\partial p'}{\partial x} - \frac{4}{3\text{Re}_{\infty,L}} \frac{\partial^2 u'}{\partial x^2} = 0, \\ M_{\infty}^2 \left(\frac{\partial p'}{\partial t} + \frac{\partial p'}{\partial x} \right) + \frac{\partial u'}{\partial x} - \frac{1}{\text{Re}_{\infty,L} \text{Pr}} \frac{\partial^2 T'}{\partial x^2} &= 0, \quad \gamma M_{\infty}^2 p' = \rho' + T' .\end{aligned}$$

Будем искать решение в виде бегущей волны $(u', \rho', p', T') = (\hat{u}, \hat{\rho}, \hat{p}, \hat{T}) \exp(i\alpha x - i\omega t)$, где ω — действительная частота, а α — комплексное волновое число. После такой подстановки последняя система дифференциальных уравнений сводится к системе алгебраических уравнений с малым параметром $\epsilon = 1/\text{Re}_{\infty,L} \ll 1$

$$\begin{pmatrix} i\alpha & \gamma M_{\infty}^2 (i\alpha - i\omega) & -i\alpha + i\omega \\ i\alpha - \omega + \frac{4}{3}\alpha^2 \epsilon & i\alpha & 0 \\ i\alpha & M_{\infty}^2 (i\alpha - i\omega) & \frac{1}{\text{Pr}} \alpha^2 \epsilon \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{\rho} \\ \hat{T} \end{pmatrix} \equiv \mathbf{A} \begin{pmatrix} \hat{u} \\ \hat{\rho} \\ \hat{T} \end{pmatrix} = \mathbf{0}.$$

В главном приближении по ϵ (при $\epsilon = 0$) условие разрешимости $\det \mathbf{A} = 0$ дает известные дисперсионные соотношения для энтропийной и акустической волн: $\alpha_0 = \omega$ — для энтропийной волны; $\alpha_0 (1 \pm 1/M_{\infty}) = \omega$ — для акустических волн (верхний знак соответствует быстрой волне, нижний — медленной). Первое приближение получим, раскладывая искомое собственное значение α по формуле Тейлора до $o(\epsilon)$ при $\epsilon \rightarrow 0$

$$\alpha(\omega) \equiv \alpha_0 + \alpha_1 + o(\epsilon) = \alpha_0 + \left(\frac{d\alpha}{d\epsilon} \right)_{\epsilon=0} \epsilon + o(\epsilon).$$

Производную $\left(\frac{d\alpha}{d\epsilon}\right)_{\epsilon=0}$ найдем, дифференцируя систему по параметру ϵ :

$$\left(\frac{\partial \mathbf{A}}{\partial \epsilon}\right)_{\epsilon=0} \begin{pmatrix} \hat{u} \\ \hat{p} \\ \hat{T} \end{pmatrix} = \mathbf{0}.$$

После ряда алгебраических манипуляций получим

$$\left(\frac{d\alpha}{d\epsilon}\right)_{\epsilon=0} = i \frac{\alpha_0^2}{\text{Pr}} \frac{M_\infty^2 (\alpha - \omega)^2 (4\text{Pr} + 3\gamma) - 3\alpha_0^2}{9M_\infty^2 (\alpha - \omega)^2 + \omega^2 - (3\alpha_0 - \omega)^2}.$$

Так как $\alpha_0 \in \mathbb{R}$, элементарные волны не затухают в отсутствие вязкости. В линейном приближении вязкость приводит исключительно к демпфированию колебаний. Декремент вязкого затухания для энтропийной волны ($\alpha_0 = \omega$) есть

$$\alpha_1 = \alpha_{1,i} = \frac{\omega^2}{\text{Re}_{\infty,L} \text{Pr}},$$

а для акустической волны ($\alpha_0 c = \alpha_0 (1 \pm 1/M_\infty) = \omega$)

$$\alpha_1 = \alpha_{1,i} = \frac{2}{3} \frac{\omega^2}{c^3 \text{Re}_{\infty,L}} \left(1 + \frac{3}{4} \frac{\gamma - 1}{\text{Pr}}\right). \quad (1)$$

Аналогичный результат для акустической волны получен с помощью общих энергетических соображений в классической литературе [9]. В случае нулевой объемной вязкости (гипотеза Стокса) он согласуется с выражением, полученным выше.

Из полученных выражений следует, что в предельном случае $\text{Re}_{\infty,L} \rightarrow \infty$ вязкое затухание отсутствует и любое затухание, наблюдаемое в численном моделировании нестационарных течений в рамках уравнений Эйлера, есть проявление диссипативных свойств численного метода. Исследуем этот случай численно, определив декремент численного затухания как

$$\alpha_{i,\text{числ}} = \alpha_i - \alpha_{i,\text{теор}}, \quad (2)$$

где α_i — декремент, восстановленный из численного решения, а $\alpha_{i,\text{теор}}$ — его теоретическое значение, которое в невязком случае равно нулю.

3. ЧИСЛЕННЫЙ МЕТОД

Уравнения Навье—Стокса в двумерной постановке интегрировались численно с помощью авторского пакета расчетных программ [8]. В пакете реализован квазимоноотонный численный метод конечных объемов второго порядка аппроксимации по времени и пространству. Интерполяция параметров потока на границы ячеек выполняется методом WENO третьего порядка, решение задачи Римана о распаде разрыва осуществляется методом Roe с малой энтропийной коррекцией. Полученная при аппроксимации система нелинейных уравнений решается модифицированным методом Ньютона, на каждой итерации по нелинейности используется метод GMRes для решения линейной системы уравнений. Расширенное описание численного метода дано в [8].

Расчеты ведутся в квазиодномерной постановке. Для этого используется структурированная сетка в прямоугольной области с $N_x \times 5$ равномерно распределенных узлов, величина N_x варьируется. Направление y моделируется формально на ширину шаблона аппроксимации. На левой (входной) границе $x = 0$ накладываются условия I рода $\mathbf{q} \equiv (u, v, p, T) = \mathbf{q}_b + \mathbf{q}'$, задающие базовый однородный поток $\mathbf{q}_b = (1, 0, 1/\gamma M_\infty^2, 1)$ и его нестационарное возмущение $\mathbf{q}' = \delta \mathbf{q} \cdot \sin \omega t$ в ви-

де акустической $\delta \mathbf{q} = \varepsilon \cdot (\pm M_\infty, 0, 1, (\gamma - 1) M_\infty^2)$ или энтропийной $\delta \mathbf{q} = (0, 0, 0, \varepsilon)$ волны с различными амплитудами ε . На правой (выходной) границе накладывается условие линейной экстраполяции всех газодинамических величин изнутри расчетной области: $\partial \mathbf{q} / \partial x = \mathbf{0}$. На границах по координате y накладывается условие симметрии потока. В качестве начального условия рассматривается однородный поток $\mathbf{q} = \mathbf{q}_b$.

4. РЕЗУЛЬТАТЫ ЧИСЛЕННОГО МОДЕЛИРОВАНИЯ

4.1. Невязкий случай

Расчеты, воспроизводящие невязкий случай $Re_{\infty, L} \rightarrow \infty$, выполнены в рамках полных двумерных уравнений Навье—Стокса при числе Рейнольдса $Re_{\infty, L} = 10^{16}$ в квазиодномерной постановке: плоская монохроматическая волна распространяется в однородном потоке в направлении x . В базовом случае рассматривается медленная акустическая волна с частотой $\omega = 131$ и амплитудой давления $\delta p = 10^{-5} p_\infty$, число Маха $M_\infty = 6$. Такие параметры характерны для задач восприимчивости сверхзвуковых пограничных слоев к акустическим возмущениям.

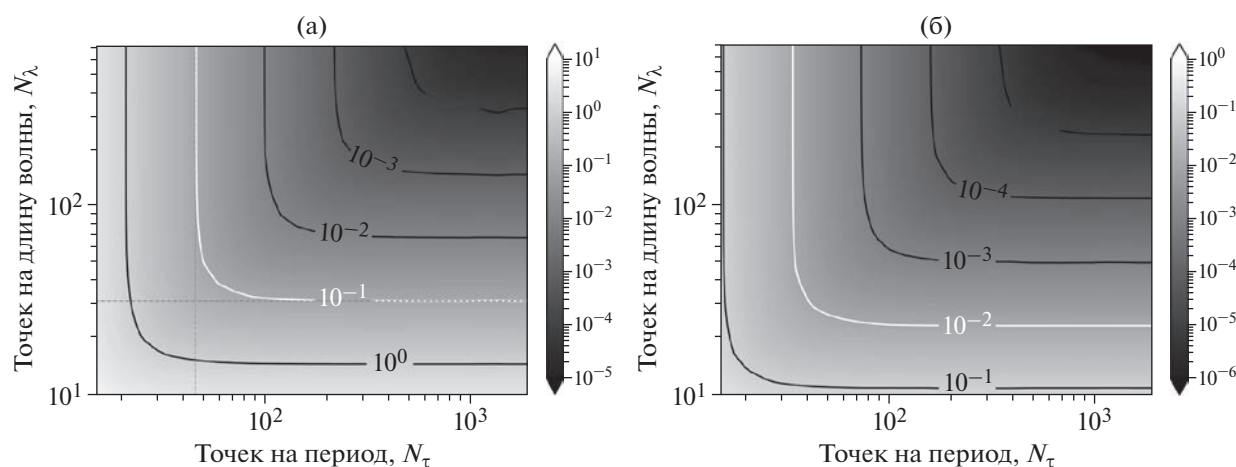
Расчетная сетка построена в прямоугольной области $x \in [0, 2] \times [0, 0.4] \ni y$. Для исследования пространственно-временной сходимости результатов рассматривается поле возмущения давления p' (или температуры T' в случае энтропийной волны) в момент времени $t_0 \approx 3$. Проводятся расчеты для всевозможных комбинаций $N_x \times N_t$ с различным количеством сеточных узлов вдоль $x - N_x \in [501, 595, 707, 841, 1001, 1190, 1415, 1501, 1683, 2001, 2380, 2501, 2830, 3001, 3366, 4003, 4501, 4760, 5661, 6001, 6732, 7501, 9001, 9520, 11322, 13464, 16012, 18001, 19041, 22644, 36001]$ – и различным суммарным количеством временных шагов – $N_t \in [938, 1115, 1326, 1577, 1875, 2230, 2652, 3153, 3750, 4460, 5303, 6307, 7500, 8919, 10607, 12613, 15000, 17838, 21213, 25227, 30000, 35676, 42426, 50454, 60000, 71352, 84853, 100908, 120000]$. Такой объем расчетных случаев представляет собой прямоугольную сетку из 31×29 расчетов. В координатах $\log_2 N_x, \log_2 N_t$ эта сетка близка к равномерной с шагом $1/4$. Количество узлов на длину волны составляет $N_\lambda = N_x \omega (x_{\max} - x_{\min}) / (2\pi c) = N_x \omega / (\pi c)$, где c – фазовая скорость возмущения ($1 - 1/M_\infty$ для медленной акустической волны или 1 для энтропийной волны). Количество временных шагов на период возмущения составляет $N_\tau = N_t t_0 \omega / (2\pi) = 1.5 N_t \omega / \pi$. Финальный момент времени t_0 может незначительно отличаться из-за равномерности шага по времени в течение всего расчета.

Декремент α_i определяется из результатов расчета путем аппроксимации локальных максимумов $|p'(x)|_{\max}$ с помощью экспоненциальной зависимости и метода наименьших квадратов. Точным аналитическим решением данной задачи является незатухающая монохроматическая волна, занимающая всю расчетную область:

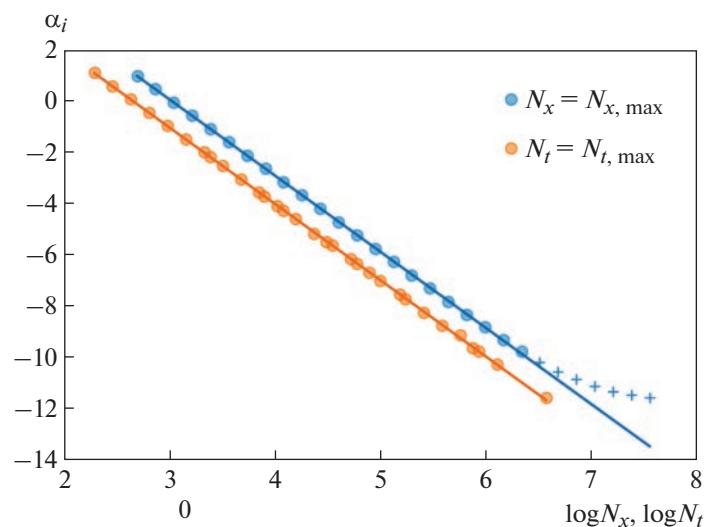
$$p'_{\text{теор}}(t, x) = \varepsilon \sin\left(\omega\left(\frac{x}{c} - t\right)\right).$$

Квинтэссенцией проведенного многопараметрического исследования является фиг. 1, на которой изображено поле логарифмического декремента акустической волны в базовом случае для всей сетки расчетов. Сходимость к точному решению $\alpha_{i, \text{теор}} = 0$ наблюдается при одновременном увеличении пространственного и временного разрешения. При увеличении только одного из них величина численной диссипации выходит на некоторый постоянный уровень и более не изменяется. Этот факт отражает теоретическую точность разностных схем, которая математически выражается в виде остаточного члена $o(\Delta x^m + \Delta t^p)$ при анализе порядка аппроксимации схемы с помощью формулы Тейлора вблизи некоторой точки (t_0, x_0) .

Таким образом, при $N_{\lambda, \max}$ можно оценить порядок сходимости численной диссипации возмущений по времени, а при $N_{\tau, \max}$ – по пространству. Результаты такого анализа представлены на фиг. 2: в обоих случаях достигается практически *третий* порядок сходимости (порядок по пространству 2.98; порядок по времени 2.97). При расчете последнего несколько точек, отмеченных крестиками, отклоняются от прямолинейной зависимости и исключены из рассмотрения.



Фиг. 1. Численная диссипация в зависимости от пространственно-временного разрешения. а — для $\alpha_{i,\text{числ}}$; б — для $\alpha_{i,\text{числ}}\lambda$.

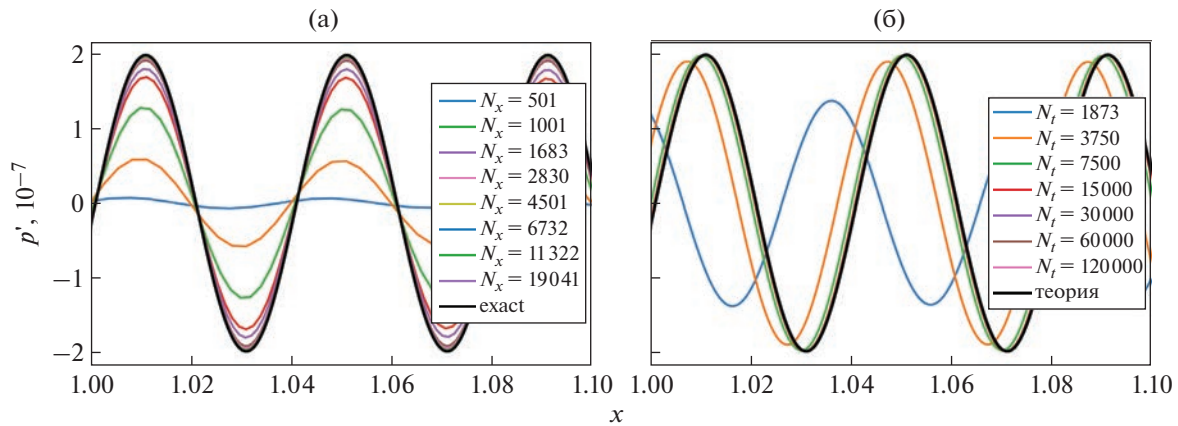


Фиг. 2. Сходимость $\alpha_{i,\text{числ}}$ к нулю по пространству и по времени. Крестиками отмечены точки, не участвующие в линейной регрессии.

Отклонение объясняется недостаточным пространственным разрешением N_λ : при $N_x = N_{x,\text{max}}$ с ростом временного разрешения N_t в некоторый момент пространственная составляющая ошибки начинает преобладать над временной составляющей. В результате суммарная ошибка выходит на плато, что проявляется в виде поведения, отмеченного крестиками на фиг. 2, и соответствует фиг. 1.

4.2. Сходимость по времени

Используемый численный метод [8] основан на аппроксимации временных и пространственных производных со вторым порядком точности, хотя для реконструкции газодинамических переменных на гранях ячейки используется схема третьего порядка. Тем не менее полученный третий порядок временной сходимости для декремента затухания $\alpha_{i,\text{числ}}$ не противоречит следствию



Фиг. 3. Распределение возмущения давления в медленной акустической волне при разных пространственных и временных разрешениях, $t = 3$; а – для $N_t = N_{t,\max}$; б – для $N_x = N_{x,\max}$.

теоремы Филиппова–Рябенского, в соответствии с которым порядок аппроксимации должен совпадать с порядком сходимости, если схема устойчива.

Рассмотрим одностороннюю аппроксимацию временной производной на временном шаге $n + 1$, представив числитель формулой Тейлора в окрестности $t^{[n+1]}$ с учетом обозначений $Q^{[n]} \equiv Q(t^{[n+1]} - \Delta t)$ и $Q^{[n-1]} \equiv Q(t^{[n+1]} - 2\Delta t)$, где Δt – шаг численного интегрирования по времени:

$$\left(\frac{\partial Q}{\partial t}\right)_h^{[n+1]} = \frac{3Q^{[n-1]} - 4Q^{[n]} + Q^{[n+1]}}{2\Delta t} = \frac{\partial Q^{[n+1]}}{\partial t} - \frac{\Delta t^2}{3} \frac{\partial^3 Q^{[n+1]}}{\partial t^3} + \frac{\Delta t^3}{4} \frac{\partial^4 Q^{[n+1]}}{\partial t^4} + o(\Delta t^3),$$

при $\Delta t \rightarrow 0$.

Очевидно, что в общем случае достигается второй порядок аппроксимации. Однако в рассматриваемом случае слабозатухающей монохроматической волны в однородном потоке все нечетные производные поля течения обращаются в ноль в экстремумах гармонического возмущения – вблизи огибающей возмущения порядок аппроксимации повышается до третьего. Поэтому численный декремент затухания, который получен из аппроксимации этой огибающей, сходится к теоретическому значению также с третьим порядком точности по времени ($p = 3$):

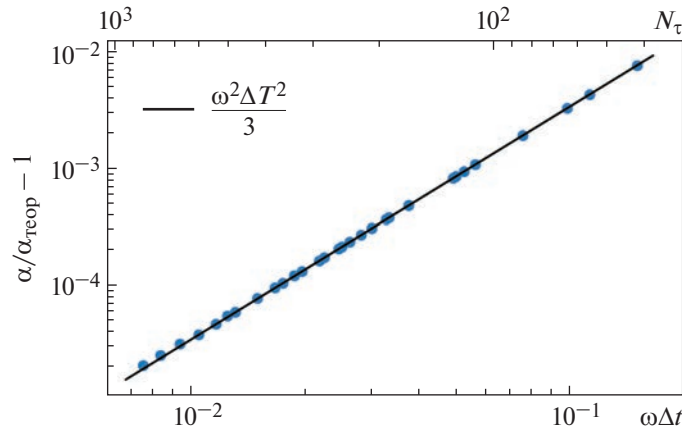
$$\max_x |\alpha_i - \alpha_{i,\text{теор}}| \equiv \|\alpha_i - \alpha_{i,\text{теор}}\|_\infty \leq C\Delta t^p.$$

Теперь рассмотрим сходимость по времени в классическом определении

$$\|q - q_{\text{теор}}\|_\infty \equiv \max_x |q - q_{\text{теор}}| \leq C\Delta t^p. \quad (3)$$

На фиг. 3 представлены примеры распределения возмущений в зависимости от пространственного и от временного сеточного разрешения. В первом случае фиксировано наиболее подробное временное разрешение $N_t = \text{const}$. Численное решение развивается практически синфазно с точным решением, а амплитуда стремится к амплитуде точного решения с ростом N_x . Во втором случае фиксировано наиболее подробное пространственное разрешение $N_x = \text{const}$. Поведение амплитуды волны не изменяется, однако при недостаточном временном разрешении проявляется численная дисперсия, которая приводит к появлению разности фаз между вычисленным и точным решениями в некоторый момент времени. Разность фаз стремится к нулю с ростом N_t .

Разность фаз обусловлена тем, что фазовая скорость акустической волны при численном моделировании отличается от истинной на величину $\frac{(\omega\Delta t)^2}{3} \sim N_t^{-2}$. Это следует из теоретического



Фиг. 4. Изменение волнового числа в численном решении при применении конечно-разностной схемы дифференцирования по времени.

анализа линеаризованных уравнений Эйлера с учетом аппроксимации временной производной и хорошо согласуется с проведенными методическими расчетами.

Применяя подход разд. 2 к уравнениям Эйлера и полагая процесс распространения малых возмущений адиабатическим, получаем систему дифференциальных уравнений в матричном виде:

$$\left[\begin{pmatrix} u & \rho \\ 1/\rho M_\infty^2 & u \end{pmatrix} \frac{\partial}{\partial x} + \mathbf{E} \frac{\partial}{\partial t} \right] \begin{pmatrix} \hat{\rho} \\ \hat{u} \end{pmatrix} e^{i\alpha x - i\omega t} = 0.$$

Действие операторов дифференцирования на экспоненциальную функцию сводится к умножению экспоненты на коэффициенты $i\alpha x$ и $-i\omega t$ соответственно. Исследование получившейся матрицы на собственные значения приводит к указанному ранее дисперсионному соотношению

$$\frac{\omega}{\alpha} = 1 \pm \frac{1}{M_\infty}.$$

Рассмотрим действие использованного оператора численного дифференцирования на экспоненциальную функцию:

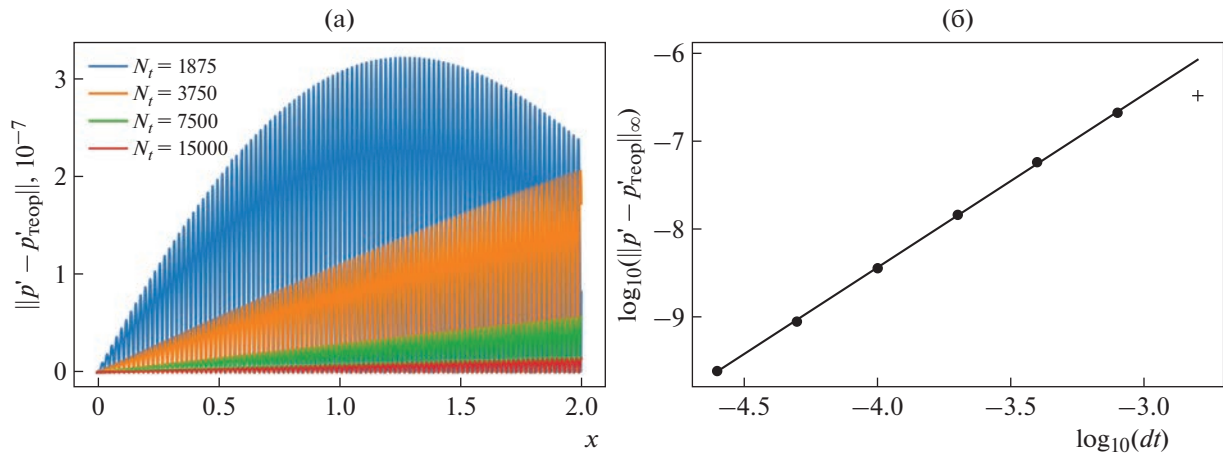
$$\left(\frac{\partial}{\partial x} \right)_h e^{-i\omega t} = \frac{e^{-i\omega t}}{2\Delta t} (3e^{-2i\omega\Delta t} - 4e^{-i\omega\Delta t} + 1) = -i\omega e^{-i\omega t} \left(1 + \frac{\omega^2 \Delta t^2}{3} + o(\omega^2 \Delta t^2) \right), \quad \omega^2 \Delta t^2 \rightarrow 0.$$

Подставляя такой оператор в матричное уравнение и пренебрегая остаточным членом $o(\omega^2 \Delta t^2)$, получим дисперсионное соотношение в виде

$$\frac{\omega}{\alpha} \left(1 + \frac{\omega^2 \Delta t^2}{3} \right) = 1 \pm \frac{1}{M_\infty}.$$

Поскольку частота волны, шаг интегрирования по времени и число Маха набегающего потока фиксированы при постановке задачи, появление нового множителя в дисперсионном соотношении с необходимостью проявляется в пропорциональном увеличении волнового числа. Действительно, изменение волнового числа в расчете отлично согласуется с теоретическим предсказанием, что проиллюстрировано на фиг. 4.

Из-за численной дисперсии абсолютная погрешность численного решения нарастает вниз по потоку и при разности фаз π достигает максимума (фиг. 5а) — истинный максимум гармонического возмущения становится минимумом при некотором x . Такая ситуация наблюдается только при достаточно грубом временном разрешении, которое далеко от практики. В остальных случаях наибольшая погрешность достигается вблизи правой границы расчетной области. Ис-



Фиг. 5. Распределение абсолютной погрешности численного решения $|p' - p'_{\text{теор}}|$ при $N_x = N_{x,\text{max}}$ и анализ временной сходимости.

пользуя классическое определение сходимости (3), можно вновь оценить порядок временной сходимости численного решения к точному. Он очень близок к двум (фиг. 5б). Следует отметить, что из линейной регрессии исключены результаты расчетов, в которых разность фаз превышала π (крестик на фиг. 5б).

4.3. Инвариантность численной диссипации

Аналогичные расчеты проведены с помощью того же численного метода (см. разд. 3) при числе $\text{Re}_{\infty,L} = 10^6$, для которого монохроматическая волна медленно затухает из-за вязкости и теплопроводности. Оказалось, что полученные выше результаты с высокой точностью повторяются для величины $\alpha_{i,\text{числ}}$ (2), где теоретическое значение вычисляется по формуле (1). Все иллюстрации, в частности, фиг. 1, визуально не изменяются и поэтому не приводятся повторно. Таким образом, численная диссипация является добавочным эффектом, который отделяется от физической картины явления.

Аналогичное методическое исследование проведено при различных числах Маха $M_{\infty} = 3, 4.5, 7.5, 9$ как для акустических, так и для энтропийных возмущений с различными частотами $\omega = 65.5, 131, 232$ и амплитудами $\varepsilon = 10^{-4}, 10^{-3}$. Установлено, что численное затухание возмущений на масштабе длины волны, $\alpha_{i,\text{числ}}\lambda$, зависит только от количества точек на длину волны N_{λ} и количества временных шагов на период возмущения N_{τ} . Используя эту инвариантность (независимость величины $\alpha_{i,\text{числ}}\lambda$ от физических параметров задачи), можно оценивать численную диссипацию, масштабируя фиг. 1 на случай произвольных параметров набегающего потока и элементарного возмущения.

5. ОБЩИЕ ЗАМЕЧАНИЯ

Корректное моделирование физического процесса затухания акустической волны требует чрезмерно подробного пространственно-временного разрешения возмущений. Однако в большинстве практически важных случаев не требуется точно воспроизводить вязкое затухание, допуская погрешность результатов нестационарного моделирования. Например, для 40 точек на длину волны и 200 точек на ее период относительная ошибка в расчете декремента α , оказывается большой из-за численной диссипации, но абсолютная величина декремента 0.04 остается достаточно малой, и амплитуда волны уменьшается лишь на 4% на единице расчетной области. Таким образом, численная диссипация акустической волны будет приводить к *ограниченной прогнозируемой погрешности* результатов численного моделирования.

Численный декремент $\alpha_{i, \text{числ}} = 0.1$ подразумевает, что амплитуда элементарного возмущения уменьшится примерно на 10% на единице длины расчетной области, или на 0.4% на масштабе длины волны базового возмущения. То есть, численная вязкость привносит в физическую задачу отрицательное интегральное усиление возмущений с фактором $N = -0.1$ ($q = q_0 e^{N(x)}$). Базовый расчетный случай является характерным для моделирования процессов ламинарно-турбулентного перехода в сверхзвуковых пограничных слоях в малозумных условиях. В таких задачах N -факторы достигают характерных значений более пяти на том же масштабе длины. Таким образом, численная диссипация приводит к ограниченной погрешности расчета эволюции возмущений в пределах нескольких процентов, что не может существенно повлиять на основные выводы, сделанные по результатам таких расчетов.

В полной пространственной постановке численная диссипация зависит от сеточного разрешения во всех трех направлениях. По предыдущему опыту расчетов, она проявляется в каждом направлении индивидуально, т.е. практически не зависит от сеточного разрешения в других направлениях. Поэтому анализ численной диссипации, проведенный выше на одномерной задаче, годится и в многомерном случае.

6. ЗАКЛЮЧЕНИЕ

Предложен вычислительный подход, который представляется универсальным способом оценки диссипативных свойств любого численного метода. Показано, что численная диссипация является добавочным эффектом, который отделяется от физической природы расчетной задачи.

Подход также пригоден для валидации численного метода: численное решение должно сходиться к теоретическому при уменьшении шага по пространству и по времени. Таким образом, результаты настоящей статьи можно рассматривать как успешную валидацию численного метода [8].

Вычислительный опыт авторов показывает, что сделанные выводы и полученные результаты количественно применимы в задачах развития возмущений в сверхзвуковых пограничных слоях — численная диссипация возмущений слабо зависит от особенностей течения и в большей мере определяется пространственно-временным разрешением конкретного возмущения. Тем не менее аккуратное исследование сходимости должно проводиться в каждом расчетном исследовании, а полученные выше базовые знания о схемной диссипации численного метода можно рассматривать как надежную отправную точку для прогнозирования эффекта численной диссипации и общей постановки расчетных задач.

Существенно нелинейный режим развития возмущений (например, в сверхзвуковом турбулентном пограничном слое) может приводить к появлению микроскачков, вблизи которых порядок аппроксимации численной схемы будет понижаться. В этом случае применимость предложенного подхода к исследованию диссипативных свойств численного метода требует дополнительной проверки.

СПИСОК ЛИТЕРАТУРЫ

1. *Петров И.Б., Лобанов А.И.* Лекции по вычислительной математике: учебное пособие М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2006. 523 с.
2. *Рождественский Б.Л., Яненко Н.Н.* Системы квазилинейных уравнений и их приложения к газовой динамике. (Изд. 2-е, перераб. и доп.) М.: Наука, 1978. 688 с.
3. *Аристова Е.Н., Астафуров Г.О.* Сравнение диссипативно-дисперсионных свойств компактных разностных схем для численного решения уравнения адвекции // Ж. вычисл. матем. и матем. физ. 2021. Т. 61. № 11. С. 1747–1758.
<https://doi.org/10.31857/S004446692111002>
4. *Шестаковская Е.С., Стариков Я.Е., Клиначева Н.Л.* Метод исследования диссипативных свойств разностных схем в эйлеровых координатах // Вестн. ЮУрГУ. Сер. Матем. моделирование и программирование. 2021. Т. 14. № 2. С. 108–116.

5. *Rogov B.V.* Dispersive and dissipative properties of the fully discrete bicomact schemes of the fourth order of spatial approximation for hyperbolic equations // *Applied Numerical Mathematics*. 2019. Vol. 139. P. 136–155.
6. *Головизнин В.М., Карабасов С.А., Козубская Т.К., Максимов Н.В.* Схема “Кабаре” для численного решения задач аэроакустики: обобщение на линеаризованные уравнения Эйлера в одномерном случае // *Ж. вычисл. матем. и матем. физ.* 2009. Т. 49. № 12. С. 2265–2280.
7. *Головизнин В.М., Соловьев А.В.* Дисперсионные и диссипативные характеристики разностных схем для уравнений в частных производных гиперболического типа. М.: МАКС Пресс, 2018, 198 с.; <http://lim.cs.msu.ru/index.php?id=9>
8. *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. 1064–1081 (*Comput. Math. Math. Phys.*, 56: 6 (2016), 1064).
9. *Ландау Л.Д., Лифшиц Е.М.* Гидродинамика. Теоретическая физика: Т.VI. (3-е изд., перераб.) М.: Наука, 1986. 736 с.

МАТЕМАТИЧЕСКАЯ
ФИЗИКА

УДК 519.63

THE EFFECT OF NONLOCAL SCALE VALUE AND PHASE LAGS
ON THERMOELASTIC WAVES IN A MULTILAYERED LEMV/CFRP
COMPOSITE CYLINDER¹⁾

© 2023 г. S. Mahesh^{1,*}, R. Selvamani^{2,**}, Farzad Ebrahimi³

¹ Department of Mathematics, V.S.B. Engineering College, Karur, India

² Department of Mathematics, Karunya Institute of Technology and Sciences, Coimbatore, India

³ Department of Mechanical Engineering, Imam Khomeini International University, Qazvin, Iran

*e-mail: maheshfuzzy1@gmail.com,

**e-mail: selvamani1729@gmail.com

Поступила в редакцию 07.10.2022 г.

Переработанный вариант 21.04.2023 г.

Принята к публикации 29.04.2023 г.

Влияние нелокальных масштабных значений и запаздываний по фазе на термоупругие волны в многослойном специальном композитном цилиндре из углепластика. Изучено влияние нелокального масштабного значения и двухфазных запаздываний на собственные колебания обобщенного линейного термоупругого многослойного композитного цилиндра из углепластика. Основные уравнения движения записаны вдоль продольной оси, и с использованием разделения переменных получена система дифференциальных уравнений. Краевым условием полагается отсутствие напряжений на внутренней и внешней границе, а также на поверхности раздела фаз. С помощью численных расчетов получены результаты для сдвига частоты, собственной частоты и термоупругого демпфирования. Результаты показывают, что нелокальный масштаб и параметры запаздывания по фазе существенно изменяют характер собственных колебаний.

Ключевые слова: термоупругость, собственные колебания, композитный материал.

DOI: 10.31857/S0044466923090120, **EDN:** OYEOBT

¹⁾ Полный текст статьи печатается в английской версии журнала.

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.63

ВЛИЯНИЕ ДЕМПФИРОВАНИЯ ПОДАТЛИВОЙ ПОВЕРХНОСТИ НА НЕВЯЗКУЮ НЕУСТОЙЧИВОСТЬ НЕСЖИМАЕМОГО ПОГРАНИЧНОГО СЛОЯ НАД НЕЙ

© 2023 г. И. В. Савенков^{1,*}

¹ 119991 Москва, ул. Вавилова, 40, ВЦ ФИЦ ИУ РАН, Россия

*e-mail: isavenkov@mail.ru

Поступила в редакцию 02.11.2022 г.
Переработанный вариант 14.05.2023 г.
Принята к публикации 29.05.2023 г.

В рамках асимптотической теории свободного взаимодействия изучена неустойчивость несжимаемого пограничного слоя над податливой пластиной по отношению к невязким возмущениям в пределе высоких чисел Рейнольдса. Показано, что неустойчивые невязкие возмущения могут существовать только при учете инерционности и/или демпфирования пластины. Выявлена двоякая роль демпфирования: при одних условиях она приводит к подавлению неустойчивости, при других — к ее возникновению. Библ. 18. Фиг. 6.

Ключевые слова: несжимаемый пограничный слой, неустойчивость, волны Толлмина—Шлихтинга, податливая поверхность, инерционность, демпфирование, изгибная жесткость, продольное натяжение, асимптотические разложения, теория свободного взаимодействия.

DOI: 10.31857/S0044466923090144, EDN: SGZLHK

ВВЕДЕНИЕ

Податливые поверхности стали объектом довольно пристального внимания с ранних экспериментов (см. [1]), в которых было показано, что значительное затягивание процесса ламинарно-турбулентного перехода может быть достигнуто путем замены жесткой поверхности на деформируемую. Последовавшие исследования в рамках линейной теории устойчивости (см., например, [2–5]) показали, что инкременты нарастания неустойчивых волн в пограничном слое на податливой поверхности действительно могут быть меньше, чем на жесткой поверхности.

Несмотря на прогресс, достигнутый к настоящему времени (см. [6–9] и ссылки в них), остается немало вопросов из-за сложности уравнений, описывающих взаимодействие течения с податливой поверхностью. Первая сложность вызвана тем, что характерные числа Рейнольдса, при которых происходит ламинарно-турбулентный переход, довольно высоки, так что в уравнениях движения появляется малый параметр при старшей производной. Вторая сложность связана с большим числом параметров, характеризующим свойства податливой поверхности, к числу которых относятся ее инерционность, демпфирование, упругость, продольное натяжение и изгибная жесткость, так что трудно выявить характерные закономерности при таком обилии параметров.

Первую сложность можно преодолеть, используя концепцию свободно взаимодействующего пограничного слоя (см. [10–12]), справедливую при высоких числах Рейнольдса. Чтобы обойти вторую сложность, можно зафиксировать все параметры податливой поверхности (или пренебречь некоторыми из них), кроме одного, и смотреть, как течение реагирует на изменение этого параметра. В рамках такого подхода в ряде работ и изучалась неустойчивость несжимаемого пограничного слоя на базе концепции свободного взаимодействия (см. [13–17]). Настоящая работа является дальнейшим развитием [17].

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим двумерное обтекание податливой пластины равномерным потоком несжимаемой жидкости плотности ρ_∞^* и вязкости μ_∞^* со скоростью U_∞^* . Введем ортогональную систему ко-

ординат, поместив ее начало на некотором расстоянии L^* от передней кромки и направив ось x^* вниз по потоку. Введем малый параметр $\varepsilon = R^{-1/8}$, считая число Рейнольдса $R = \rho_\infty^* U_\infty^* L^* / \mu_\infty^*$ очень большим. Здесь и далее звездочками вверху обозначаются размерные величины.

Пусть на пластину локально воздействует внешнее давление $\Delta p^* \sim \delta^2 \rho_\infty^* U_\infty^{*2}$ на участке $\Delta x^* \sim \varepsilon^4 / \delta L^*$ с характерным временем $\Delta t^* \sim \varepsilon^4 / \delta^2 L^* / U_\infty^*$, вызывая локальные деформации стенки высотой $\Delta y^* \sim \varepsilon^4 \delta L^*$, где малый параметр δ таков, что $\varepsilon \ll \delta \ll 1$. Тогда возмущенное движение можно описать с помощью концепции свободного взаимодействия (см. [10–12]), развитой применительно к локально-невязким возмущениям (см., например, [18]). Согласно этой теории, вся область течения разбивается на следующие четыре характерные области.

1.1. Основная толщина пограничного слоя

В основной толще пограничного слоя, т.е. на расстояниях $y^* = R^{-1/2} L^* Y_m$ от стенки ($Y_m = O(1)$), течение в пограничном слое будет лишь слабовозмущенным:

$$\frac{u^*}{U_\infty^*} = U_0 + \delta u_{1m} + \dots, \quad \frac{v^*}{U_\infty^*} = \delta^2 v_{1m} + \dots, \quad \frac{p^* - p_\infty^*}{\rho_\infty^* U_\infty^{*2}} = \delta^2 p_{1m} + \dots, \quad (1.1)$$

где $U_0(Y_m)$ — профиль продольной компоненты скорости основного течения и функции с индексами $1m$ зависят от безразмерных переменных

$$t = \frac{\delta^2 \varepsilon^{-4} U_\infty^{*2} t^*}{L^*}, \quad x = \delta \varepsilon^{-4} \frac{x^* - L^*}{L^*}, \quad Y_m = \frac{\varepsilon^{-4} y^*}{L^*}.$$

Время t и продольная координата x будут “сквозными” для всех областей взаимодействия, тогда как поперечная координата меняется при переходе между ними.

После подстановки (1.1) в уравнения Навье—Стокса и тривиального интегрирования имеем

$$u_{1m} = A_1(t, x) \frac{dU_0}{dY_m}, \quad v_{1m} = -\frac{\partial A_1}{\partial x} U_0(Y_m), \quad p_{1m} = p_{1m}(t, x), \quad (1.2)$$

где функции $A_1(t, x)$ и $p_{1m}(t, x)$ пока остаются произвольными. Заметим, что функция $A_1(t, x)$ имеет физический смысл мгновенного смещения линий тока в основной толще пограничного слоя.

1.2. Нелинейный слой

В силу предельного поведения, $U_0 = \lambda_1 Y_m + \dots$ ($\lambda_1 = \text{const}$) при $Y_m \rightarrow 0$, разложения (1.1) нарушаются вблизи стенки: при $Y_m = O(\delta)$ первые два слагаемых в разложении для u^* становятся величинами одного порядка. Поэтому в нижнем подслое с $Y_a = \delta^{-1} Y_m = \delta^{-1} \varepsilon^{-4} y^* / L^* \sim 1$ возмущения становятся нелинейными, а ряды (1.1) трансформируются в

$$\frac{u^*}{U_\infty^*} = \delta u_{1a} + \dots, \quad \frac{v^*}{U_\infty^*} = \delta^3 v_{1a} + \dots, \quad \frac{p^* - p_\infty^*}{\rho_\infty^* U_\infty^{*2}} = \delta^2 p_{1a} + \dots, \quad (1.3)$$

где все функции с индексом $1a$ зависят от аргументов t, x, Y_a . Подстановка (1.3) в уравнения Навье—Стокса ведет к системе

$$\frac{\partial u_{1a}}{\partial x} + \frac{\partial v_{1a}}{\partial Y_a} = 0, \quad \frac{\partial p_{1a}}{\partial Y_a} = 0, \quad \frac{\partial u_{1a}}{\partial t} + u_{1a} \frac{\partial u_{1a}}{\partial x} + v_{1a} \frac{\partial u_{1a}}{\partial Y_a} = -\frac{\partial p_{1a}}{\partial x}, \quad (1.4)$$

в которой, в отличие от классической системы Прандтля, отсутствует член $\partial^2 u_{1a} / \partial Y_a^2$.

Сращивание разложений (1.1) и (1.3) с учетом решения (1.2) ведет к предельному условию

$$u_{1a} - \lambda_1 Y_a \rightarrow \lambda_1 A_1(t, x) \quad \text{при} \quad Y_a \rightarrow \infty. \quad (1.5)$$

Если предположить, что пограничный слой был изначально невозмущенным, то можно выписать точное решение системы (1.4), удовлетворяющее граничному условию (1.5):

$$u_{1a} = \lambda_1 Y_a + \lambda_1 A_1(t, x), \quad v_{1a} = -\frac{\partial A_1}{\partial t} - \lambda_1 Y_a \frac{\partial A_1}{\partial x} - \lambda_1 A_1 \frac{\partial A_1}{\partial x} - \frac{1}{\lambda_1} \frac{\partial p_{1a}}{\partial x}. \quad (1.6)$$

Учитывая условие непротекания на стенке

$$v^* = \frac{\partial F^*}{\partial t^*} + u^* \frac{\partial F^*}{\partial x^*} \quad \text{при} \quad Y^* = F^*(t^*, x^*) = \varepsilon^4 \delta L^* F_1(t, x),$$

преобразующееся в безразмерных переменных в

$$v_{1a} = \frac{\partial F_1}{\partial t} + u_{1a} \frac{\partial F_1}{\partial x} \quad \text{при} \quad Y_a = F_1(t, x),$$

из (1.6) получаем одно уравнение движения:

$$\frac{\partial(A_1 + F_1)}{\partial t} + \lambda_1(A_1 + F_1) \frac{\partial(A_1 + F_1)}{\partial x} = -\frac{1}{\lambda_1} \frac{\partial p_{1a}}{\partial x}, \quad (1.7)$$

связывающее три неизвестные функции: $A_1(t, x)$, $p_{1a}(t, x)$ и $F_1(t, x)$.

1.3. Вязкий подслой

Нетрудно видеть, что решение (1.6) не удовлетворяет условиям прилипания на стенке (должно быть $u_{1a} = 0$ при $Y_a = 0$), из-за чего требуется ввести вязкий подслой со своими асимптотическими разложениями следующего вида:

$$\frac{u^*}{U_\infty^*} = \delta u_{1l} + \dots, \quad \frac{v^*}{U_\infty^*} = \delta^3 \left(\frac{\partial F_1}{\partial t} + u_{1l} \frac{\partial F_1}{\partial x} \right) + \delta \varepsilon^2 v_{1l} + \dots, \quad \frac{p^* - p_\infty^*}{\rho_\infty^* U_\infty^{*2}} = \delta^2 p_{1l} + \dots,$$

где функции с нижним индексом $1l$ зависят от переменных t, x и $Y_l = \delta \varepsilon^{-6} (y^* - F^*)/L^*$. Подстановка этого разложения в уравнения Навье–Стокса ведет к классической системе Прандтля

$$\begin{aligned} \frac{\partial u_{1l}}{\partial x} + \frac{\partial v_{1l}}{\partial Y_l} &= 0, \quad \frac{\partial p_{1l}}{\partial Y_l} = 0, \\ \frac{\partial u_{1l}}{\partial t} + u_{1l} \frac{\partial u_{1l}}{\partial x} + v_{1l} \frac{\partial u_{1l}}{\partial Y_l} &= -\frac{\partial p_{1l}}{\partial x} + \frac{\partial^2 u_{1l}}{\partial Y_l^2}. \end{aligned} \quad (1.8)$$

Сращивание разложений (1.3) и (1.8) дает

$$p_{1l}(t, x) = p_{1a}(t, x) \quad \text{и} \quad u_{1l}(t, x, Y_l) \rightarrow A_1 \quad \text{при} \quad Y_l \rightarrow \infty,$$

откуда видно, что в данном случае вязкий подслой играет пассивную роль: решение (1.8) с соответствующими граничными условиями ($u_{1l}(t, x, 0) = 0$, $v_{1l}(t, x, 0) = 0$) определяется после нахождения функций $A_1(t, x)$ и $p_{1a}(t, x)$. Мы оставляем в стороне вопросы существования и единственности полной нелинейной задачи, формулирующейся на базе уравнений (1.8), поскольку в дальнейшем будем изучать только линеаризованные уравнения.

1.4. Внешняя область

В этой области течения, примыкающей сверху к пограничному слою, введем следующие асимптотические разложения:

$$\frac{u^*}{U_\infty^*} = 1 + \delta^2 u_{1u} + \dots, \quad \frac{v^*}{U_\infty^*} = \delta^2 v_{1u} + \dots, \quad \frac{p^* - p_\infty^*}{\rho_\infty^* U_\infty^{*2}} = \delta^2 p_{1u} + \dots, \quad (1.9)$$

с функциями, зависящими от t, x и $Y_u = \delta \varepsilon^{-4} y^*/L^*$.

Подстановка разложений (1.9) в уравнения Навье–Стокса ведет к уравнению Лапласа для давления

$$\partial^2 p_{lu} / \partial x^2 + \partial^2 p_{lu} / \partial Y_u^2 = 0 \quad (1.10)$$

со следующими зависимостями:

$$u_{lu} = -p_{lu}, \quad \partial v_{lu} / \partial x = -\partial p_{lu} / \partial Y_u. \quad (1.11)$$

Сращивание разложений (1.9) и (1.1) с учетом (1.10) и (1.11) дает так называемое условие свободного взаимодействия, связывающее давление $p_{la}(t, x)$ с функцией $A_1(t, x)$:

$$p_{la}(t, x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\partial A_1(t, \xi) / \partial \xi}{x - \xi} d\xi. \quad (1.12)$$

1.5. Податливая поверхность

Будем считать, что податливая поверхность описывается моделью Крамера (см. [1]), тогда уравнение движения такой поверхности принимает следующий вид:

$$p^* - p_0^* = -K^* F^* + T^* \frac{\partial^2 F^*}{\partial x^{*2}} - B^* \frac{\partial^4 F^*}{\partial x^{*4}} - \rho_s^* h_s^* \frac{\partial^2 F^*}{\partial t^{*2}} - d^* \frac{\partial F^*}{\partial t^*}, \quad (1.13)$$

где K^* , T^* , B^* , d^* , ρ_s^* и h_s^* — коэффициент упругости, продольное натяжение, изгибная жесткость, коэффициент демпфирования, плотность и толщина обтекаемой пластины соответственно, а p_0^* — внешнее давление под пластиной. С учетом этого уравнение движения пластины в безразмерных переменных принимает вид

$$\gamma_1 \frac{\partial^2 F_1}{\partial t^2} + d_1 \frac{\partial F_1}{\partial t} - \alpha_1 \frac{\partial^2 F_1}{\partial x^2} + \beta_1 \frac{\partial^4 F_1}{\partial x^4} + K_1 F_1 = -p_{la}(t, x) + p_0(t, x), \quad (1.14)$$

где безразмерные коэффициенты определяются следующим образом:

$$\gamma = \delta^3 \varepsilon^{-2} \frac{\rho_s^* h_s^*}{\rho_\infty^* L^*}, \quad d_1 = \delta^{-1} \varepsilon^2 \frac{d^*}{\rho_\infty^* U_\infty^{*2}}, \quad \alpha_1 = \delta \varepsilon^{-4} \frac{T^*}{\rho_\infty^* U_\infty^{*2} L^*},$$

$$\beta_1 = \delta^3 \varepsilon^{-12} \frac{B}{\rho_\infty^* U_\infty^{*2} L^{*3}}, \quad K_1 = \delta^{-1} \varepsilon^4 \frac{K^* L^*}{\rho_\infty^* U_\infty^{*2}}.$$

В дальнейшем будем считать их величинами порядка единицы.

Таким образом, мы получили три уравнения: (1.7), (1.12) и (1.14) для трех неизвестных функций F_1 , A_1 и p_{la} . Сделаем еще аффинное преобразование

$$t = \lambda_1^{-3/2} T, \quad x = \lambda_1^{-5/4} X, \quad A_1 = \lambda_1^{-3/4} A, \quad F_1 = \lambda_1^{-3/4} F, \quad p_{la} = \lambda_1^{1/2} P,$$

позволяющее избавиться от константы λ_1 , после чего эти уравнения принимают следующий вид:

$$\gamma \frac{\partial^2 F}{\partial T^2} - \alpha \frac{\partial^2 F}{\partial X^2} + \beta \frac{\partial^4 F}{\partial X^4} + KF = -(P(T, X) - P_0(T, X)),$$

$$\frac{\partial(A + F)}{\partial T} + (A + F) \frac{\partial(A + F)}{\partial X} = -\frac{\partial P}{\partial X}, \quad P(T, X) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\partial A(T, \xi) / \partial \xi}{X - \xi} d\xi, \quad (1.15)$$

где параметры $d = \lambda_1^{-3/2} d_1$, $\alpha = \lambda_1^{-1/2} \alpha_1$, $\beta = \lambda_1^2 \beta_1$ и $K = \lambda_1^{-3} K_1$.

2. ДИСПЕРСИОННОЕ СООТНОШЕНИЕ

Изучим собственные малые колебания течения, линеаризовав систему (1.15) по малому амплитудному параметру $\delta_1 \rightarrow 0$ и положив внешнее давление $P_0 = 0$:

$$(F, P, A) = \delta_1 (F', P', A'). \quad (2.1)$$

Выделив в явном виде гармоническую зависимость

$$(F', P', A') = (\bar{F}, \bar{P}, \bar{A}) \exp(\omega t + ikx),$$

получаем следующее уравнение относительно ω :

$$(\omega + k|k|)(\gamma\omega^2 + d\omega + \Lambda) - |k|\omega = 0,$$

где $\Lambda = K + \alpha k^2 + \beta k^4$ зависит только от волнового числа k и является параметром, если рассматривать k в качестве параметра. Решения этого уравнения обладают свойством $\omega(-k) = c.c.\omega(k)$ (где *c.c.* означает операцию комплексного сопряжения), в силу чего достаточно ограничиться рассмотрением только диапазона $k \geq 0$. Полагая далее $k \geq 0$ и переходя для удобства к переменной $\omega_0 = i\omega$, получаем кубическое уравнение относительно ω_0 :

$$(\omega_0 - k^2)(\gamma\omega_0^2 + id\omega_0 - \Lambda) + k\omega_0 = 0. \quad (2.2)$$

Нетрудно видеть, что при $d \neq 0$ корни уравнения (2.2) не могут быть вещественными, откуда следует, что при учете демпфирования в рассматриваемой системе не может быть нейтральных колебаний. Значит, каждый из трех корней уравнения (2.2) может быть только целиком (при всех k) устойчивым или целиком неустойчивым.

2.1. Безынерционная пластина

Начнем исследование неустойчивости с более простого случая, когда $\gamma = 0$, что соответствует безынерционной пластине. В этом случае уравнение (2.2) вырождается в квадратное уравнение

$$(\omega_0 - k^2)(id\omega_0 - \Lambda) + k\omega_0 = 0, \quad (2.3)$$

в котором без ограничения общности можно считать $d = 1$, поскольку от параметра d можно избавиться путем аффинного преобразования $\omega_0 \rightarrow d^{-2}\omega_0$, $k \rightarrow d^{-1}k$, $\Lambda \rightarrow d^{-1}\Lambda$, $K \rightarrow d^{-1}K$, $\alpha \rightarrow d\alpha$, $\beta \rightarrow d^3\beta$. Считая это преобразование выполненным, имеем два корня (2.3):

$$\omega_{0,\pm} = \frac{ik - i\Lambda + k^2 \pm i\sqrt{(ik^2 - k + \Lambda)^2 - 4i\Lambda k^2}}{2}, \quad (2.4)$$

где $\Lambda = K + \alpha k^2 + \beta k^4$.

Нетрудно получить следующие асимптотики корней (2.4):

– для $\beta \neq 0$:

$$\omega_{0,+} = k^2 + \frac{1}{\beta}k^{-1} + \frac{i-\alpha}{\beta^2}k^{-3} + \dots, \quad \omega_{0,-} = -i\beta k^4 + \dots \quad \text{при} \quad k \rightarrow \infty; \quad (2.5)$$

– для $\beta = 0$:

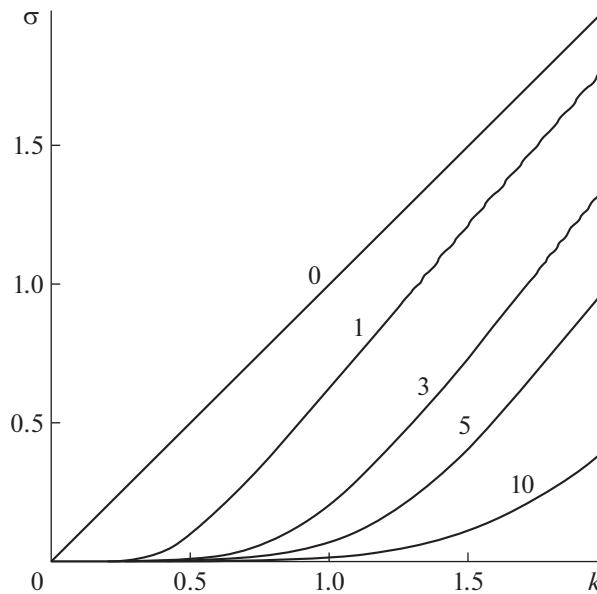
$$\omega_{0,+} = k^2 + \frac{\alpha+i}{\alpha^2+1}k + \dots, \quad \omega_{0,-} = -i\alpha k^2 + \dots \quad \text{при} \quad k \rightarrow \infty. \quad (2.6)$$

Несмотря на разный характер поведения решения при $k \rightarrow \infty$, эти асимптотики объединяет одно: инкремент нарастания “плюсового” корня

$$(\sigma = \text{Im } \omega_{0,+} \sim \frac{1}{\beta^2}k^{-3} \text{ или } \sigma = \text{Im } \omega_{0,+} \sim \frac{1}{\alpha^2+1}k) \text{ положителен в пределе больших } k, \text{ а значит,}$$

этот корень неустойчив при всех k (как было отмечено ранее, все корни или целиком неустойчивы, или целиком устойчивы, поскольку нейтральных решений не существует). По этой же причине “минусовой” корень всегда устойчив (ведь $\text{Im } \omega_{0,-} \sim -\beta k^4 < 0$ или $\text{Im } \omega_{0,-} \sim -\alpha k^2$ в пределе больших k).

Таким образом, можно сразу сделать вывод общего характера: наличие демпфирования безынерционной пластины всегда ведет к появлению неустойчивости, причем единственный неустойчивый корень неустойчив во всем диапазоне волновых чисел k .



Фиг. 1.

Далее, асимптотика (2.6) говорит о том, что инкремент нарастания $\sigma \sim \frac{1}{\alpha^2 + 1}k$ может стать сколь угодно большим с увеличением k , если пренебречь изгибной жесткостью пластины ($\beta = 0$). Сказанное иллюстрируют зависимости инкремента нарастания σ от k при $\alpha = \beta = 0$ для ряда значений K , приведенные на фиг. 1. Видно, что σ довольно быстро выходит на свою линейную асимптотику (2.6). Учет продольного натяжения ($\alpha \neq 0$) может только уменьшить скорость этого роста, но не подавить линейную зависимость (см. фиг. 2 для ряда значений α при $K = 1, \beta = 0$). И только наличие изгибной жесткости радикально меняет ситуацию, уводя σ на алгебраическое затухание $\sim \frac{1}{\beta^2}k^{-3}$ в пределе больших k , согласно асимптотике (2.5). Насколько сильно подавляется неустойчивость уже при малых β , можно проследить по фиг. 3 для ряда значений β (при $K = 1, \alpha = 0$).

2.2. Общий случай

Рассмотрим общий случай инерционной пластины, когда $\gamma \neq 0$. Путем аффинного преобразования

$$\begin{aligned} \omega_0 &\rightarrow \gamma^{-2/3}\omega_0, & k &\rightarrow \gamma^{-1/3}k, & d &\rightarrow \gamma^{1/3}d, & \Lambda &\rightarrow \gamma^{-1/3}\Lambda, \\ K &\rightarrow \gamma^{-1/3}K, & \alpha &\rightarrow \gamma^{1/3}\alpha, & \beta &\rightarrow \gamma\beta \end{aligned} \quad (2.7)$$

можно избавиться от коэффициента γ в уравнении (2.2), которое тогда принимает вид

$$(\omega_0 - k^2)(\omega_0^2 + id\omega_0 - \Lambda) + k\omega_0 = 0. \quad (2.8)$$

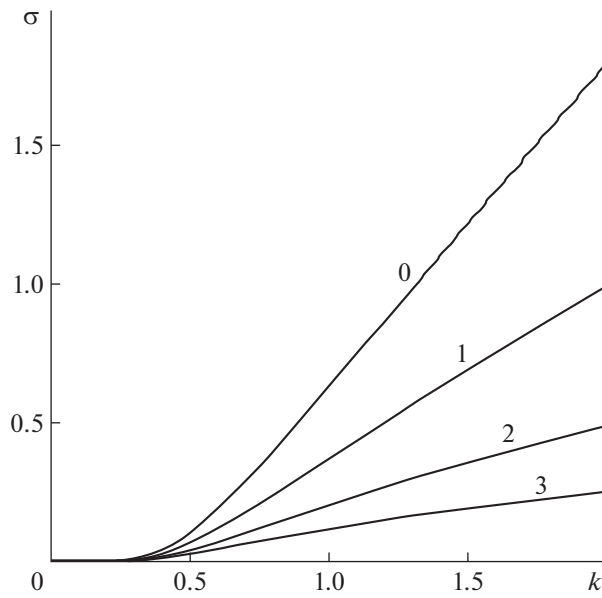
В дальнейшем будем считать такое преобразование выполненным и исследовать решения дисперсионного соотношения (2.8), представляющего собой кубическое уравнение.

Начнем изучение (2.8) с предельного случая волновых чисел k , стремящихся к бесконечности. В таком случае нетрудно получить следующие асимптотики дисперсионного соотношения (2.8).

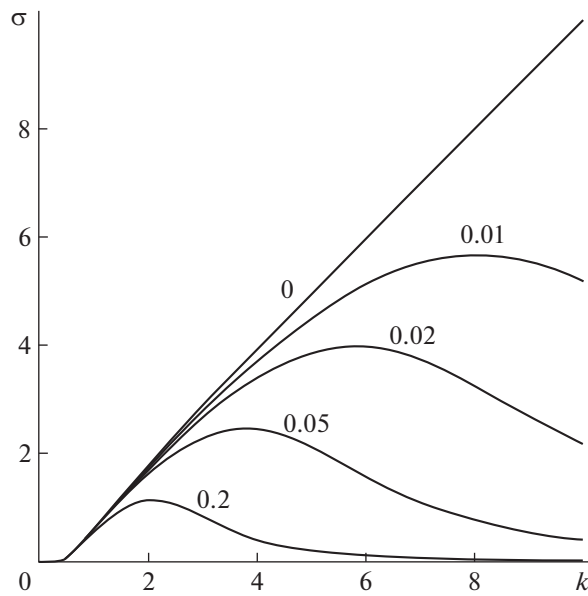
Для первого корня:

– при $\beta \neq 1$:

$$\omega_{0,1} = k^2 - \frac{1}{1-\beta}k + \frac{id - \alpha}{(1-\beta)^2}k^{-3} + \dots \quad \text{при } k \rightarrow \infty; \quad (2.9)$$



Фиг. 2.



Фиг. 3.

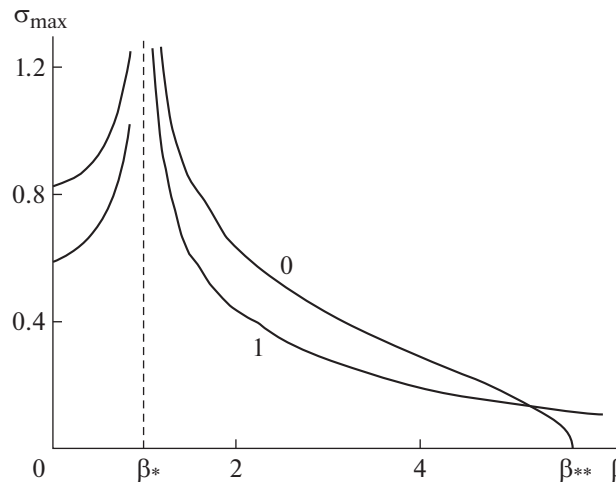
– при $\beta = 1$:

$$\omega_{0,1} = k^2 + i \frac{1}{\sqrt{2}} k^{1/2} + \frac{\alpha - id}{2} + \dots \quad \text{при } k \rightarrow \infty. \quad (2.10)$$

Для второго и третьего корней:

– при $\beta \neq 0$:

$$\omega_{0,2,3} = \pm \beta^{1/2} k^2 \pm \frac{\alpha}{2\beta^{1/2}} - \frac{1}{2} id + \dots \quad \text{при } k \rightarrow \infty; \quad (2.11)$$



Фиг. 4.

– при $\beta = 0$:

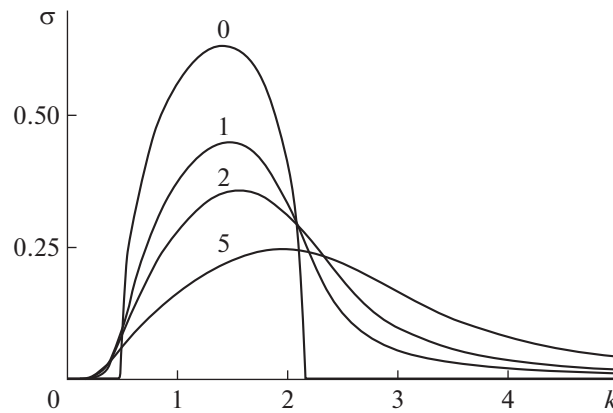
$$\omega_{0,2,3} = \pm \alpha^{1/2} k - \frac{1}{2} id + \dots \quad \text{при } k \rightarrow \infty. \quad (2.12)$$

Как видно из (2.11) и (2.12), инкремент нарастания $\sigma = \text{Im } \omega_{0,2,3} \sim -1/2d < 0$ в пределе $k \rightarrow \infty$, значит, и второй, и третий корень устойчивы при больших k и будут оставаться таковыми при всех k (как мы ранее заметили, корни могут быть только целиком устойчивыми или целиком неустойчивыми).

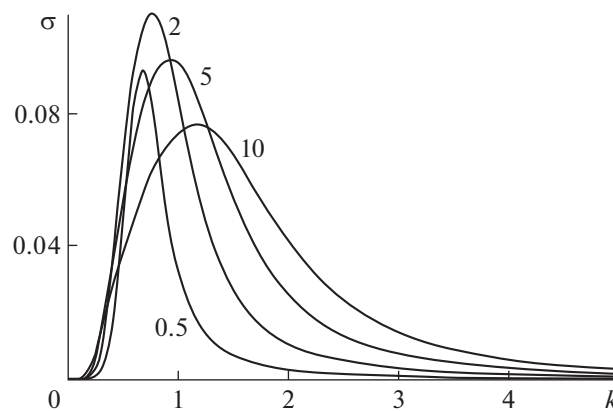
Первый же корень, напротив, всегда неустойчив (при всех $k \neq 0$, K , α , β и $d \neq 0$), причем согласно (2.9) он затухает алгебраически: $\sigma = \text{Im } \omega_{0,1} \sim \frac{d}{(1-\beta)^2} k^{-3}$ в пределе $k \rightarrow \infty$ при $\beta \neq 1$. Но самое интересное с точки зрения неустойчивости происходит при $\beta = 1$, когда этот корень согласно (2.10) неограниченно растет: $\sigma = \text{Im } \omega_{0,1} \sim \frac{1}{\sqrt{2}} k^{1/2}$ в пределе $k \rightarrow \infty$. Собственно, этот эффект был обнаружен ранее в работе [17], в которой не учитывалось демпфирование пластины. Учет демпфирования в настоящей работе показывает, что хотя демпфирование и несколько подавляет этот рост (на постоянную величину $-1/2d$), но оно не в силах справиться с ним. Сказанное иллюстрирует фиг. 4, на которой представлены зависимости $\sigma_{\max}(\beta) = \max \sigma(-\infty < k < \infty)$ при $K = 1$ и $\alpha = 0$ для значений $d = 0, 1$ (отмечены соответствующими цифрами рядом).

Отметим еще одну характерную особенность, прослеживающуюся по фиг. 4. Как было выяснено в [17], в случае бездемпферной пластины ($d = 0$) существует пороговое значение $\beta = \beta_{**}$ (зависящее от K и α), при превышении которого неустойчивость пропадает ($\beta_{**} = 5.67$ для $K = 1$ и $\alpha = 0$ на фиг. 4). При учете демпфирования ($d \neq 0$) такое пороговое значение β_{**} не может существовать, поскольку тогда первый корень неустойчив всегда; и на фиг. 4 видно, что $\sigma_{\max}(\beta; d = 1)$ плавно затухает при $\beta \rightarrow \infty$, никак не “ощущая” переход через пороговое значение $\beta_{**} = 5.67$, при котором $\sigma_{\max}(\beta; d = 0)$ резко падает до нуля. Таким образом, хотя демпфирование при умеренных β подавляет неустойчивость, при достижении порогового значения β_{**} и выше оно оказывает противоположный эффект, приводя к появлению неустойчивости (хотя и слабой).

Характер влияния демпфирования на неустойчивость при увеличении d можно проследить по типичным фиг. 5 и фиг. 6, на которых отображена зависимость инкремента нарастания $\sigma(k)$ при $K = 1$ и $\alpha = 0$ для ряда значений d (отмечены соответствующими цифрами рядом). Как показывают расчеты, в “допороговом” случае (см. фиг. 5 для $\beta = 2 < \beta_{**} = 5.67$) максимальный инкремент нарастания $\sigma_{\max}$ монотонно снижается с ростом d (т.е., чем больше демпфирование, тем больше оно подавляет неустойчивость), тогда как в “запороговом” случае (см. фиг. 6 для



Фиг. 5.



Фиг. 6.

$\beta = 6 > \beta_{**} = 5.67$) максимальный инкремент нарастания $\sigma_{\max}$ сначала растет, а затем падает (т.е. демпфирование сначала приводит к появлению и росту неустойчивости, а затем начинает давить ее).

3. ВЫВОДЫ

1. Итак, в рамках теории свободного взаимодействия показано, что в пределе высоких чисел Рейнольдса невязкая неустойчивость может существовать только за счет инерционности или демпфирования пластины (если инерционность и демпфирование пренебрежимо малы, то течение будет устойчивым по отношению к невязким возмущениям).

2. Если при наличии демпфирования инерционность и изгибная жесткость пренебрежимо малы, то это приводит к катастрофической потере неустойчивости: инкремент нарастания возмущений неограниченно растет в коротковолновом пределе ($k \rightarrow \infty$).

3. При наличии инерционности существует критическое значение изгибной жесткости $\beta = \beta_* = 1$, при которой инкремент нарастания возмущений неограниченно растет в коротковолновом пределе. Демпфирование не может справиться с этой “катастрофой” (только уменьшает инкремент нарастания на конечную величину).

4. При наличии демпфирования всегда (при любых $k \neq 0$, K , α и β) существует ровно одна неустойчивая мода собственных колебаний течения, в том числе и при превышении порогового значения $\beta = \beta_{**}$ (зависящего от K и α), когда неустойчивость пропадает, если пренебрегать демпфированием.

СПИСОК ЛИТЕРАТУРЫ

1. *Kramer M.O.* Boundary-layer stabilization by distributed damping // *J. Aeronaut. Sci.* 1957. V. 24. P. 458–460.
2. *Benjamin T.B.* Effects of a flexible boundary on hydrodynamic stability // *J. Fluid Mech.* 1960. № 9. P. 513–532.
3. *Benjamin T.B.* The threefold classification of unstable disturbances in flexible surfaces bounding inviscid flows // *J. Fluid Mech.* 1963. № 16. P. 436–450.
4. *Landahl M.T.* On the stability of a laminar incompressible boundary layer over a flexible surface // *J. Fluid Mech.* 1962. № 13. P. 609–632.
5. *Carpenter P.W., Garrad A.D.* The hydrodynamic stability of flow over Kramer-type compliant surfaces. Part 1. Tollmien-Schlichting instabilities // *J. Fluid Mech.* 1985. V. 155. P. 465–510.
6. *Riley J.J., Gad-el-Hak, Metcalfe R.W.* Compliant coatings // *Ann. Rev. Fluid Mech.* 1988. № 20. P. 393–420.
7. Flow Past Highly Compliant Boundaries and in Collapsible Tubes // *Proc. of the IUTAM Symp., University of Warwick, UK, 26–30 March 2001.* Kluwer Acad. Publ., 2003.
8. *Gad-el-Hak M.* Compliant coatings for drag reduction // *Progress in Aerospace Sciences.* 2002. V. 38. Iss. 1. January 2002. P. 77–99.
9. *Carpenter P.W.* Recent progress in the use of compliant walls for laminar flow control // *Progress in Industrial Mathematics at ECMI 2006. Math. in Industry.* 2008. № 12. P. 178.
10. *Нейланд В.Я.* К теории отрыва ламинарного пограничного слоя в сверхзвуковом потоке // *Изв. АН СССР. Сер. механ. жидкости и газа.* 1969. № 4. С. 53–58.
11. *Stewartson K., Williams P.G.* Self-induced separation // *Proc. Roy. Soc. A.* 1969. V. 312. № 1509. P. 181–206.
12. *Messiter A.F.* Boundary-layer flow near the trailing edge of a flat plate // *SIAM J. Appl. Math.* 1970. V. 18. № 1. P. 241–257.
13. *Савенков И.В.* Подавление роста нелинейных волновых пакетов упругостью обтекаемой поверхности // *Ж. вычисл. матем. и матем. физ.* 1995. Т. 35. № 1. С. 95–103.
14. *Савенков И.В.* Об абсолютной неустойчивости несжимаемого пограничного слоя на податливой поверхности // *Ж. вычисл. матем. и матем. физ.* 2018. Т. 58. № 2. С. 281–290.
15. *Савенков И.В.* Влияние инерционности податливой поверхности на вязкую неустойчивость несжимаемого пограничного слоя // *Ж. вычисл. матем. и матем. физ.* 2019. Т. 59. № 4. С. 707–715.
16. *Walker J.D.A., Fletcher A., Ruban A.I.* Instabilities of a flexible surface in supersonic flow // *Q. Jl Mech. Appl. Math.* 2006. V. 59. № 2. P. 253–276.
17. *Савенков И.В.* Невязкая неустойчивость несжимаемого пограничного слоя на податливой поверхности // *Ж. вычисл. матем. и матем. физ.* 2020. Т. 60. № 7. С. 1268–1280.
18. *Жук В.И.* Волны Толлмина–Шлихтинга и солитоны. М.: Наука, 2001. 167 с.

УДК 519.62

АГРЕГИРОВАНИЕ И ДЕКОМПОЗИЦИЯ СИСТЕМ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ И СИСТЕМ УПРАВЛЕНИЯ С РАСПРЕДЕЛЕННЫМИ ПАРАМЕТРАМИ

© 2023 г. В. И. Елкин^{1,*}¹ 119333 Москва, ул. Вавилова, 40, ФИЦ ИУ РАН, Россия

*e-mail: elk_v@mail.ru

Поступила в редакцию 09.09.2022 г.
Переработанный вариант 08.04.2023 г.
Принята к публикации 29.05.2023 г.

Рассматривается вопрос об агрегировании (укрупненном, упрощенном представлении) систем дифференциальных уравнений с частными производными и систем управления с распределенными параметрами. Получены условия декомпозиции на основе агрегирования. Библ. 14.

Ключевые слова: агрегирование, декомпозиция, факторизация, управляемые системы.

DOI: 10.31857/S0044466923090089, **EDN:** DNGTRN

ВВЕДЕНИЕ

Под агрегированием сложных математических моделей, например, моделей большой размерности, обычно понимается сопоставление исходной модели некоторой упрощенной модели (агрегированной модели), оперирующей укрупненными характеристиками, обычно имеющей меньшую размерность, и функционирующей в определенном смысле в силу функционирования исходной модели. Информация об агрегированной модели может быть использована для изучения исходной модели. В частности, агрегированию может сопутствовать декомпозиция модели на совокупность более простых моделей, одна из которых как раз является агрегированной моделью.

Хорошо известным примером агрегирования является случай управляемых динамических систем (см. [1, 2])

$$\dot{y}^i = g^i(t, y, u), \quad i = 1, \dots, n, \quad (t, y) \in M \subset R^{n+1}, \quad u \in U \subset R^r. \quad (1)$$

Здесь рассматриваются произвольные системы дифференциальных уравнений с частными производными первого порядка

$$\Lambda_v(t, y, p) = 0, \quad v = 1, \dots, l,$$

где символ p обозначает частные производные y по t . Как известно, такую систему путем введения дополнительных переменных можно привести к специальному виду в параметрической форме, разрешенной относительно всех производных (см. [3]),

$$\partial_k y^i = g_k^i(t, y, u), \quad \text{где} \quad \partial_k = \partial / \partial t_k. \quad (2)$$

Для таких систем открывается возможность применения методов агрегирования и декомпозиции теории динамических систем с управлением, используя некоторую аналогию данных объектов.

Здесь следует отметить родственное направление исследований факторизации (декомпозиции) дифференциальных уравнений с позиций теории групп. Это направление известно под на-

званием групповой анализ дифференциальных уравнений — раздел математики, изучающий свойства симметрий дифференциальных уравнений относительно различных преобразований зависимых и независимых переменных; включает в себя методы и прикладные аспекты дифференциальной геометрии, теории групп и алгебр Ли, вариационного исчисления и является, в свою очередь, эффективным инструментом исследования в теории ОДУ, ДУЧП и математической физике. Считается, что Л.В. Овсянников положил начало этому научному направлению (см. [4]), хотя по существу он возродил исследования Софуса Ли XIX в. Л.В. Овсянников ввел понятие инвариантного решения, которое сопутствует некоторой факторизации (агрегированию) систем уравнений, что позволяет снижать размерность задач, в частности, уравнений газовой динамики. Известными учениками Л.В. Овсянникова являются Н.Х. Ибрагимов (см. [5]) и Ю.Н. Павловский, который, кстати, первоначально использовал именно методы группового анализа для изучения агрегирования управляемых динамических систем (см. [1, 2]).

Кроме того, существует школа А.В. Виноградова, который также под влиянием идей Софуса Ли начал систематическое исследование оснований геометрической теории дифференциальных уравнений в частных производных (см. [6, 7]).

Подчеркнем, что применение группового анализа для дифференциальных уравнений основано на использовании симметрий, нахождение которых сводится к решению также дифференциальных (определяющих) уравнений, которые могут быть весьма сложными. Между тем применение подхода настоящей работы сводится к исследованию на совместность стандартного типа дифференциальных уравнений с одинаковой главной частью (см. [8, 9]).

1. АГРЕГИРОВАНИЕ (ФАКТОРИЗАЦИЯ) ДИНАМИЧЕСКИХ СИСТЕМ С УПРАВЛЕНИЕМ

Рассматриваются динамические системы с управлением вида (1). Предполагается, что функции g^i , $\partial f^i / \partial y^j$, $\partial g^i / \partial u^\alpha$ являются гладкими. Обычно y называют фазовыми переменными (состояниями), u — управлениями (внешними воздействиями). Множество M , называемое фазовым пространством, — область, U — область. Управления могут быть кусочно-непрерывными функциями $u(t)$, $t \in [t_0, t_1]$. В этом случае они называются допустимыми. Решением или фазовой траекторией системы (1) называется непрерывная кусочно-гладкая функция $y(t)$, $t \in [t_0, t_1]$, для которой существует такое допустимое управление $u(t)$, $t \in [t_0, t_1]$, что функции $y(t)$, $u(t)$ удовлетворяют соотношениям (1).

Под декомпозицией управляемой системы (1) в этом разделе понимается приведение этой системы с помощью замены фазовых переменных $y \mapsto z = \varphi(t, y)$ к виду

$$\dot{z}^k = h^k(t, z^1, \dots, z^m, u), \quad k = 1, \dots, m, \quad (3)$$

$$\dot{z}^l = h^l(t, z^1, \dots, z^n, u), \quad l = m + 1, \dots, n. \quad (4)$$

Представление (3), (4) характеризуется тем, что уравнения (3) образуют замкнутую систему. Поэтому любое решение $z(t)$ системы (3), (4) может быть получено следующим образом. Сначала нужно найти решение $z^1(t), \dots, z^m(t)$ системы (3) (соответствующее некоторому допустимому управлению $u(t)$), а затем, после подстановки $z^1(t), \dots, z^m(t)$ в систему (4), которая становится замкнутой системой, найти оставшиеся функции $z^{m+1}(t), \dots, z^n(t)$. Таким образом, декомпозиция (3), (4) позволяет свести процесс нахождения решения системы к нахождению решений для двух систем, фазовые пространства которых имеют размерности, меньшие, чем n .

Если существует такое представление, то говорят также, что система (1) допускает агрегирование или факторизацию порядка $n - m$. При этом среди функций

$$z^i = \varphi^i(t, y), \quad i = 1, \dots, n, \quad (5)$$

осуществляющих соответствующую замену фазовых координат $y \mapsto z = \varphi(t, y)$, первые m функций называются агрегатами, а система (3) — агрегированной системой. Функционирование агрегированной системы происходит в силу функционирования исходной системы в следующем

смысле: если $y(t)$ — решение системы (1), то $z^k(t) = \varphi^k(y(t))$, $k = 1, \dots, m$, — решение системы (3). Заметим также, что с точки зрения теории категорий систему (3) можно трактовать как фактор-объект (фактор-систему) объекта (1) в категории систем (1) с морфизмами, являющимися отображениями фазовых пространств и переводящими решения в решения.

Для изучения возможностей такой декомпозиции в теории динамических систем применяется так называемый метод полных семейств векторных полей (см. [1, 2, 8]). Согласно этому методу, отыскиваются условия, которым должны удовлетворять полные семейства векторных полей, для которых агрегаты (5) являются интегралами. Этот метод будем применять и для декомпозиции дифференциальных уравнений с частными производными в следующем разделе, где он будет подробно описан.

2. ДЕКОМПОЗИЦИЯ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Рассматривается система дифференциальных уравнений с частными производными первого порядка

$$\Delta_v(t, y, p) = 0, \quad v = 1, \dots, l, \quad (6)$$

содержащая $t = (t^1, \dots, t^m)$, $y = (y^1, \dots, y^n)$ и производные от функций y^i первого порядка, причем используется обозначение $p = (\dots, \partial_k y^i, \dots)$, где $\partial_k y^i = \partial y^i / \partial x^k$. В дальнейшем предполагаем для простоты, что функции Δ_v , а также все встречающиеся функции (в том числе решения) гладкие. Под гладкостью понимаем бесконечную дифференцируемость. Считаем, что система (6) максимального ранга и матрица Якоби размера $(l \times mn)$ производных от функций Δ_v по p имеет ранг l всюду, где $\Delta(t, y, p) = 0$. При этом условии соотношения (6) задают гладкое многообразие $\mathcal{S}_\Delta$ в пространстве переменных (t, y, p) или, иначе говоря, пространстве 1-струй функций $y(t)$. Относительно системы (6) будем считать также, что она локально разрешима. Согласно [10], система локально разрешима в точке $(x_0, y_0, p_0) \in \mathcal{S}_\Delta$, если существует гладкое решение $y = f(t)$, определенное для t из некоторой окрестности точки t_0 , имеющее предписанные начальные условия $y_0 = f(t_0)$, $p_0 = (\partial f / \partial t)_{t=t_0}$. Система называется локально разрешимой, если она локально разрешима в каждой точке многообразия $\mathcal{S}_\Delta$.

Уравнения (6) задают многообразие $\mathcal{S}_\Delta$ в неявном виде. Это многообразие можно (в силу невырожденности) представить и в параметрическом виде (в некоторой локальной карте) путем разрешения (6) относительно максимального числа производных, причем если остальные (параметрические) производные обозначить через u как параметрические переменные типа $\partial_\delta y^\beta = u^\gamma$, то получим локальное представление $\mathcal{S}_\Delta$ в параметрической форме, разрешенной относительно всех производных (см. [3]):

$$\partial_k y^i = g_k^i(t, y, u), \quad (7)$$

$$t \in I \subset R^m, \quad y \in L \subset R^n, \quad u \in U \subset R^s, \quad (8)$$

где I, L, U — некоторые области. Такое представление произвольной системы дифференциальных уравнений с частными производными известно как специальный вид (см. [3]). В (7) остальные (параметрические) производные обозначены через u как параметрические переменные.

Пример 1. Рассмотрим систему

$$\frac{\partial y^1}{\partial x_1} = \frac{\partial y^2}{\partial x_1} x_2, \quad \frac{\partial y^2}{\partial x_2} = y^1 y^2, \quad (9)$$

$$\frac{\partial y^1}{\partial x_2} = y^2(1 + x_2 y^1). \quad (10)$$

Полагая $\partial y^2 / \partial x_1 = u$, запишем данную систему в виде (7)

$$\frac{\partial y^1}{\partial x_1} = ux_2, \quad \frac{\partial y^2}{\partial x_1} = u, \quad (11)$$

$$\frac{\partial y^1}{\partial x_2} = y^2(1 + x_2 y^1), \quad \frac{\partial y^2}{\partial x_2} = y^1 y^2. \quad (12)$$

Далее будем рассматривать системы уравнений в специальном виде и в карте (8). Условие локальной разрешимости в данном случае означает, что для любой точки $t_0 \in I$, $y_0 \in M$, $u_0 \in U$ существует окрестность $V \subset L$ точки t_0 , в которой определено такое гладкое решение $y(t)$, что $y(t_0) = y_0$, $u(t_0) = u_0$.

Сравнивая (7) и (1), можно заметить некоторую аналогию этих систем. Действительно, каждое решение $y(t)$ управляемой системы (1) получается после подстановки в правую часть управлений $u(t)$ из некоторого класса допустимых функций. С другой стороны, решения $y(t)$ системы (7) соответствуют некоторому выбору параметрической функции $u(t)$ также из некоторого класса. Однако есть существенное различие в классах допустимых “управлений”. Для управляемых систем классы допустимых управлений достаточно известны и широки: от класса кусочно-непрерывных функций до класса измеримых функций. Для систем (7) заранее задать класс допустимых “управлений” затруднительно, ибо далеко не каждый выбор функций $u(t)$ приводит к решению $y(t)$. Препятствием является, в частности, возможность несовместности полученной системы после подстановки $u(t)$. Тем не менее идеология теории управлений может быть полезна.

Далее рассматриваются дифференциальные уравнения с частными производными специального вида, причем для простоты изложения и не ограничивая общности, “автономные”, т.е. с правыми частями, не зависящими от аргументов t :

$$\partial_k y^i = g_k^i(y, u). \quad (13)$$

Речь пойдет о декомпозиции системы (13) с помощью замены зависимых переменных y , точнее о возможности преобразования системы с помощью замены переменных

$$z^i = \varphi^i(y), \quad i = 1, \dots, n, \quad (14)$$

к виду

$$\partial_k z^l = h_k^l(z^1, \dots, z^m, u), \quad l = 1, \dots, m, \quad (15)$$

$$\partial_k z^i = h_k^i(z^1, \dots, z^n, u), \quad i = m+1, \dots, n. \quad (16)$$

Если такое представление возможно, то будем говорить также, как в разд. 1, что система (13) допускает агрегирование по зависимым переменным порядка $n - m$, причем первые m функций в замене переменных (14)

$$z^i = \varphi^i(y), \quad i = 1, \dots, m, \quad (17)$$

называются агрегатами. Далее рассматривается применение метода полных семейств для исследования возможности декомпозиции (15), (16) для систем (13). Этот метод впервые применен Ю.Н. Павловским для агрегирования динамических систем (см. [1, 2]). Полным семейством векторных полей или операторов называется семейство полей

$$Z_a = b_a^i(y) \frac{\partial}{\partial y^i}, \quad i = 1, \dots, p \quad (18)$$

(по повторяющемуся индексу здесь и далее производится суммирование), если:

1)

$$\text{rank} \|b_a^i(y)\| = p,$$

т.е. векторы $Z_a(y)$, $a = 1, \dots, p$, линейно независимы в каждой точке $y \in M$ или, как еще говорят, линейно несвязаны в области M ;

2)

$$[Z_a, Z_c] = h_{ac}^d(y)Z_d, \quad a, c, d = 1, \dots, p, \quad (19)$$

т.е. все коммутаторы $[Z_a, Z_c]$ семейства выражаются линейно с переменными коэффициентами $h_{ac}^d(y)$ через остальные поля семейства.

Гладкая функция $\Phi(y)$, $y \in M$, называется интегралом поля Z_a или корнем оператора Z_a , если

$$Z_a(\Phi(y)) = b_a^i(y) \frac{\partial \Phi}{\partial y^i} = 0.$$

Теорема 1 (см. [10]). Полное семейство (18), $p < n$, имеет в окрестности каждой точки $m = n - p$ функционально независимых интегралов (полный набор)

$$\Phi^k(y), \quad \text{rank} \left\| \frac{\partial \Phi^k}{\partial y^i} \right\|_{i=1, \dots, n}^{k=1, \dots, n-m} = n - m, \quad (20)$$

причем любой интеграл $\Phi(y)$ функционально выражается через полный набор

$$\Phi(y) = F(\Phi^1(y), \dots, \Phi^m(y)).$$

Теорема 2 (см. [10]). Пусть в области M заданы независимые функции (20). Тогда в окрестности каждой точки M найдется полное семейство (18), $p - n - m$, для которого функции (20) составляют полный набор интегралов.

Доказательство. Рассмотрим систему однородных линейных алгебраических уравнений

$$\xi^i \frac{\partial \Phi^k}{\partial y^i} = 0, \quad k = 1, \dots, m, \quad (21)$$

относительно компонент векторного поля $\xi = \xi^i \partial / \partial y^i$. Для любой точки y найдется окрестность, в которой некоторый минор m -го порядка матрицы системы уравнений (21) отличен от нуля. Фундаментальная система решений в этой окрестности определяет линейно несвязанное семейство векторных полей (18), для которого функции (20) являются интегралами. Любое поле $\xi = \xi^i \partial / \partial y^i$, удовлетворяющее (21), должно линейно выражаться через поля (18). В частности, так как коммутаторы $[\xi_a, \xi_b]$ в силу свойства

$$[\xi, \eta](\Phi) = \xi(\eta(\Phi)) - \eta(\xi(\Phi)) \quad (22)$$

удовлетворяют (21), то справедливо свойство 2 семейства (18). Следовательно, построенное семейство (18) является полным. Функции (20) составляют максимальный набор независимых интегралов семейства (18), ибо в противном случае получилось бы противоречие с теоремой 1.

Переходим к условиям существования декомпозиции (15), (16). Введем в рассмотрение операторы полного дифференцирования системы (13) по переменным t^l (здесь отсутствует явное дифференцирование $\partial / \partial t^l$ по t^l в силу вышеупомянутой “автономности”):

$$X_l = g_l^i(y, u) \partial / \partial y^i, \quad l = 1, \dots, m. \quad (23)$$

Замена зависимых переменных (14) в системе (13) осуществляется следующим образом: нужно подействовать операторами (23) на функции (14)

$$X_l(\Phi^k(y)) = g_l^i(y, u) \frac{\partial \Phi^k}{\partial y^i} \quad (24)$$

и выразить функционально полученные функции через (14)

$$g_l^i(y, u) \frac{\partial \Phi^k}{\partial y^i} = h^k(\Phi^1(y), \dots, \Phi^n(y)) u. \quad (25)$$

Функции

$$h^k(z^1, \dots, z^n, u) \quad (26)$$

являются новыми правыми частями системы, при этом декомпозиция (15), (16) возникает, когда первые m функций (25) функционально выражаются только через агрегаты (17), т.е.

$$g_l^i(y, u) \frac{\partial \varphi^k}{\partial y^i} = h^k(\varphi^1(y), \dots, \varphi^m(y), u). \quad (27)$$

Непосредственное применение этого условия не дает конструктивного метода нахождения агрегатов и проверки возможности декомпозиции. Поэтому применяется опосредственный подход, когда сначала с помощью решения некоторых систем дифференциальных уравнений отыскиваются полные семейства векторных полей, для которых агрегаты являются интегралами, а затем находятся агрегаты, т.е. интегралы этих семейств. При этом совместность упомянутых систем дифференциальных уравнений определяет возможность декомпозиции.

Теорема 3. Система (13) тогда и только тогда локально приводится к виду (15), (16) с помощью замены переменных (14), когда существует такое полное семейство (18), $p = n - m < n$, что

$$[X_l, Z_a] = \mu_{la}^b(y, u) Z_b, \quad l = 1, \dots, r, \quad a, b = 1, \dots, p, \quad (28)$$

причем агрегаты (17) составляют полный набор интегралов семейства (18).

Доказательство. Необходимость. Для агрегатов (17) по теореме 2 существует полное семейство (18), интегралами которого являются агрегаты. Подействуем коммутаторами $[X_l, Z_a]$ на эти функции, используя свойство коммутатора (22). Получим

$$[X_l, Z_a](\varphi^i(y)) = X_l(Z_a(\varphi^i(y))) - Z_a(X_l(\varphi^i(y))) = 0. \quad (29)$$

Здесь использовано то обстоятельство, что функции $X_l(\varphi^i(y))$, согласно (27), функционально выражаются через полный набор интегралов семейства (18) и поэтому также являются интегралами этого семейства. В доказательстве теоремы 2 отмечено, что условие (29) влечет (28).

Достаточность. Для семейства (18) возьмем полный набор интегралов $\varphi^i(y)$, $i = 1, \dots, m = n - p$, и рассмотрим опять выражения (29). Из (29) вытекает, что функции $X_l(\varphi^i(y))$ являются интегралами семейства (18) и по теореме 1 функционально выражаются через полный набор интегралов и, следовательно, $\varphi^i(y)$ являются агрегатами для системы (13).

Выражения (29) можно трактовать как систему дифференциальных уравнений

$$X_l(b_a^i) = Z_a(g^i) + \mu_{la}^b(y, u) b_c^i, \quad i = 1, \dots, m, \quad a, c = 1, \dots, p, \quad u \in U, \quad (30)$$

относительно неизвестных компонент семейства (18). В (30) выражения $\xi_u(b_a^i)$, $Z_a(f^i)$ представляют собой действия полей ξ_u , Z_a на функции $b_a^i(y)$, $f^i(y, u)$ как операторов. После решения этой системы нужно найти полный набор интегралов $\varphi^i(y)$, $i = 1, \dots, m = n - p$, полного семейства (18), что, согласно [11], сводится к решению некоторых систем обыкновенных дифференциальных уравнений. Для нахождения правых частей фактор-системы следует функции $f^i(y, u) \frac{\partial \varphi^k}{\partial y^i}$

выразить функционально через функции $\varphi^i(y)$, $i = 1, \dots, m = n - p$, т.е. представить их в виде

$$g^i(y, u) \frac{\partial \varphi^k}{\partial y^i} = h^k(\varphi^1(y), \dots, \varphi^m(y), u).$$

Построенные функции h^k и определяют искомую фактор-систему

$$\dot{z}^k = h^k(z, u), \quad k = 1, \dots, m.$$

Система (30) содержит, помимо $b_a^i(y)$, еще неизвестные параметрические переменные $\mu_{ia}^b(y, u)$. Последнее обстоятельство усложняет дело. Кроме того, нужно добавить условие полноты $[Z_b, Z_a] = \kappa_{ba}^c(y)Z_c$ семейства (18), которое тоже представляет собой систему дифференциальных уравнений, причем с неизвестными параметрическими переменными $\kappa_{ba}^c(y)$. Рассмотрим метод исключения параметрических переменных; осуществляется с помощью использования якобиевых семейств (для динамических систем эта процедура предложена автором в [8]). Сначала докажем одно утверждение, которое связывает полные семейства и их интегралы. Для того чтобы исключить функции μ_{ia}^c из (30), следует задаться некоторым свойством, которым должны обладать агрегаты (17). Это свойство заключается в определенном расположении базисного минора в якобиевой матрице $\|\partial\phi^k/\partial y^i\|$.

Теорема 4. Пусть (18) — полное семейство векторных полей, (20) — полный набор интегралов этого семейства, а I — некоторое подмножество из p элементов множества индексов $\{1, \dots, n\}$, $\bar{I} = \{1, \dots, n\} \setminus I$. Тогда следующие условия эквивалентны:

$$(a) \quad |\xi_a^i|_{a=1, \dots, p}^{i \in I} \neq 0,$$

$$(б) \quad \left| \frac{\partial \phi^k}{\partial y^i} \right|_{i \in \bar{I}}^{k=1, \dots, m} \neq 0.$$

Доказательство. Пусть выполняется условие (а). Рассмотрим семейство полей η_b , $b \in I$, которое получается из исходного семейства линейным невырожденным преобразованием: $\eta_b = \mu_b^a(y)\xi_a$, $b \in I$. Здесь функции μ_b^a составляют матрицу, обратную к матрице $\|\xi_a^i\|_{a=1, \dots, p}^{i \in I}$. Построенное семейство полей будет иметь следующий вид:

$$\eta_b = \frac{\partial}{\partial y^b} + \eta_b^c(y) \frac{\partial}{\partial y^c}, \quad b \in I, \quad c \in \bar{I}, \quad (31)$$

где η_b^c — некоторые функции. Ясно, что функции ϕ^k являются интегралами и полей η_b , т.е.

$$\frac{\partial \phi^k}{\partial y^b} = -\eta_b^c \frac{\partial \phi^k}{\partial y^c}.$$

Отсюда вытекает, что выполняется условие (б).

Пусть теперь выполняется условие (б). Рассмотрим систему линейных однородных уравнений относительно неизвестного векторного поля $\xi = \xi^i(y)\partial/\partial y^i$:

$$\xi^i \frac{\partial \phi^k}{\partial y^i} = 0, \quad k = 1, \dots, m.$$

Очевидно, что поля ξ_a , $a = 1, \dots, p$, образуют фундаментальную систему решений этой системы уравнений. С другой стороны, нетрудно видеть, что существует фундаментальная система решений вида (31). Два семейства (18) и (31) должны быть связаны линейным невырожденным преобразованием. Отсюда вытекает справедливость условия (а).

Для того чтобы исключить функции h_{ia}^c из (30), следует задаться некоторым свойством, которым должны обладать агрегаты (17). Это свойство заключается в определенном расположении базисного минора в якобиевой матрице $\|\partial\phi^k/\partial y^i\|$. Предположим, что для искомым агрегатов (17)

$$\left| \frac{\partial \phi^k}{\partial y^i} \right|_{i=p+1, \dots, n}^{k=1, \dots, m} \neq 0, \quad p = n - m. \quad (32)$$

Справедлива

Теорема 5. Для того чтобы система (1) допускала (локальную) факторизацию порядка $p = n - m$, характеризующуюся свойством (32), необходимо и достаточно, чтобы была совместна система дифференциальных уравнений относительно неизвестных функций $b_a^l(y)$:

$$X_l(b_a^l) = Z_a(f^l) - b_c^l Z_a(f^c), \quad (33)$$

$$Z_a(b_c^l) = Z_c(b_a^l), \quad (34)$$

где

$$Z_a = \frac{\partial}{\partial y^a} + b_a^l(y) \frac{\partial}{\partial y^l}, \quad (35)$$

$$u \in U, \quad a, c = 1, \dots, p, \quad l = p+1, \dots, n.$$

Доказательство. Необходимость. Если система (1) допускает факторизацию со свойством (32), то, согласно доказательству теоремы 4, существует семейство вида (35). Семейство (35) должно удовлетворять соотношениям (28). Запишем эти соотношения покомпонентно:

$$\xi_a(\delta_a^i) - Z_a(f^i) = h_a^c \delta_c^i, \quad (36)$$

$$\xi_\beta(b_a^l) - Z_a(f^l) = h_a^c b_c^l, \quad (37)$$

$$i, a, c = 1, \dots, p, \quad l = p+1, \dots, n, \quad \delta_a^i - \text{символ Кронекера.}$$

Из (36) имеем $h_a^i = -Z_a(f^i)$. Подставляя h_a^i в (37), получим (33). Равенства (34) представляют собой условие полноты семейства (35). Действительно, согласно доказательству теоремы 1, семейство вида (35) является полным тогда и только тогда, когда оно якобиево, т.е. $[Z_a, Z_c] = 0$, $a, c = 1, \dots, p$, что равносильно равенствам (34).

Достаточность. Если совместна система (33), (34), то семейство полей (35) является полным. Далее, полагая $h_a^i = -Z_a(f^i)$, из (33) получим (36), (37), откуда вытекает (28). Следовательно, система (1) допускает факторизацию порядка p . Свойство (32) вытекает из вида семейства (35) и теоремы 4. Теорема 5 доказана.

Уравнения (33) представляют собой систему уравнений относительно b_a^l без параметрических переменных.

При ином расположении базисного минора в матрице $\|\partial \varphi^k / \partial y^i\|$ условия факторизации (33), (34) преобразуются очевидным образом. Итак, задача решения системы дифференциальных уравнений (30) с параметрическими переменными сведена к задаче решения нескольких систем дифференциальных уравнений вида (33) без параметрических переменных. Решение системы (33), т.е. набор компонент семейства полей (35), следует подставить для проверки в уравнения (34), представляющие собой условия полноты семейства (35). Лишь те решения системы (33), которые одновременно являются и решениями системы (34), определяют семейство, порождающее факторизацию порядка p .

Система такого вида является частным случаем так называемых уравнений в частных производных с одинаковой главной частью

$$a_k^j(x) \frac{\partial y^i}{\partial x^j} = b_k^i(x, y), \quad (38)$$

$$i = 1, \dots, n, \quad j = 1, \dots, m, \quad k = 1, \dots, r.$$

Функции a_k^j, b_k^i — гладкие в области $V \times U$, где $V \subset R^m, U \subset R^n$.

Замечание 1. Системы (38) для $r = 1$ рассматривались в [12]. Другой частный случай систем вида

$$\frac{\partial y^i}{\partial x^j} = b_j^i(x, y), \quad i = 1, \dots, n, \quad j = 1, \dots, m, \quad (39)$$

также хорошо известен (см., например, [10]). Для проверки совместности следует рассмотреть равенство смешанных производных:

$$\frac{\partial^2 y^i}{\partial x^j \partial x^s} = \frac{\partial^2 y^i}{\partial x^s \partial x^j}, \quad (40)$$

в силу (39). Если эти уравнения тождественно удовлетворяются, то система (38) вполне интегрируема. В противном случае получаем систему конечных уравнений, устанавливающих условия на y , как функций от x . Если каждое из этих уравнений продифференцировать по x^i и заменить производные $\partial y^i / \partial x^i$ их значениями из (39), то полученные уравнения либо будут следствиями полученных уравнений, либо составят новую систему. Продолжая этот процесс, получим ряд систем, которые должны быть совместны, если уравнения (39) имеют решения. В противном случае решений у (39) нет. Данные рассуждения и выводы можно применять также и тогда, когда заданы какие-либо дополнительные функциональные соотношения между переменными y и x :

$$\eta^\alpha(y, x) = 0, \quad \alpha = 1, \dots, q. \quad (41)$$

В [8, 9] автором получен алгоритм проверки совместности и нахождения решений систем уравнений систем (38), точнее в [7] были исследованы частные случаи систем типа (151), которые возникают при рассмотрении агрегирования управляемых динамических систем, а в [8] уже рассмотрен общий случай.

Наконец, рассмотрим один весьма общий случай систем, когда можно проверить совместность системы (33), (34), не решая предварительно систему (33). Но сначала следует рассмотреть важный частный случай декомпозиции, когда правые части агрегированной системы равны нулю. Он связан с понятием первого интеграла систем дифференциальных уравнений.

2.1. Первые интегралы

Также как и для обыкновенных дифференциальных уравнений введем понятие первого интеграла для систем (13). Гладкую функцию $\Phi(y)$ назовем первым интегралом системы (13), если на любом решении $y(t)$ системы (7) она принимает постоянное значение

$$\Phi(y(t)) = \text{const.}$$

Замечание 2. Введенное определение первого интеграла, по существу, совпадает с определением для обыкновенных дифференциальных уравнений. (Для неавтономных систем в качестве первых интегралов можно рассматривать и более общие функции $\Phi(t, y)$.) Однако оно практически не используется в литературе для уравнений с частными производными. Для управляемых динамических систем это понятие появилось в связи с важным практическим вопросом — управляемости. Причина, видимо, в редкости случаев существования и отсутствии алгоритмов нахождения. Более актуальным считается более общая проблема существования законов сохранения (см., например, [5, 6, 10]). Тем не менее, если первые интегралы существуют и найдены, то это может существенно упростить исследование систем уравнений (снизить порядок, см. далее пример). Аналогичное определение применяется и для управляемых динамических систем. Там это понятие появилось в связи с важным практическим вопросом — управляемости (см. [13]).

В силу гладкости рассматриваемых функций, условием того, что функция $\Phi(y)$ есть первый интеграл, является тождественное равенство нулю производных по всем независимым переменным от этой функции в силу каждого решения системы (7):

$$h_k^i(y(t), u(t)) \frac{\partial \Phi(y(t))}{\partial y^i} \equiv 0, \quad t \in I, \quad k = 1, \dots, m. \quad (42)$$

(Напоминаем, что по повторяющемуся индексу производится суммирование.) Для неавтономных систем следует добавить и явное дифференцирование по t^k для исследуемых функций $\Phi(t, y)$.

Введем ассоциированные семейства векторных полей (операторов) C_k , $k = 1, \dots, m$, в области M (для неавтономных — в области $I \times M$). Каждое семейство C_k состоит из операторов полного дифференцирования по переменной t^k

$$h_k^i(y, u) \frac{\partial}{\partial y^i} = 0, \quad k = 1, \dots, m, \quad (43)$$

с фиксированным $k = 1, \dots, m$, где u пробегает все множество постоянных значений из U . Объединение этих семейств обозначим через C_0 (для управляемых динамических систем имеется только одно семейство).

Теорема 6. *Функция $\Phi(t, y)$ является первым интегралом тогда и только тогда, когда*

$$h_k^i(y, u) \frac{\partial \Phi(y)}{\partial y^i} \equiv 0, \quad t \in I, \quad k = 1, \dots, m, \quad (t, y) \in M, \quad u \in U. \quad (44)$$

Доказательство. Достаточность очевидна, ибо выполнение (44) для всех $y \in M$, $u \in U$ влечет (42). Необходимость вытекает из условия локальной разрешимости. Действительно, возьмем первый интеграл $\Phi(y)$, точки $y_0 \in M$, $u_0 \in U$ и соответствующее гладкое решение $y(t)$ такое, что $y(t_0) = y_0$, $u(t_0) = u_0$. Из (42) для этого решения следует (44) при $y_0 \in M$, $u_0 \in U$.

Из закона преобразования дифференциальных уравнений при замене зависимых переменных следует, что если в качестве первых использовать набор независимых первых интегралов, то в результате получится система

$$\partial_k z^i = 0, \quad i = 1, \dots, q, \quad \partial_k z^j = h_k^j(t, x, u), \quad j = q + 1, \dots, n. \quad (45)$$

Из работы [13], посвященной управляемым динамическим системам, следует алгоритм проверки существования и нахождения первых интегралов. Нужно выделить в каждом ассоциированном семействе (43) максимальное число линейно несвязанных векторных полей, объединить эти поля и пополнить в области изменения переменных $I \times M$. Если число полей s в полученном полном семействе меньше числа переменных $n + m$, то нетривиальные первые интегралы существуют, причем число функционально независимых интегралов равно $n + m - s$. Они находятся с помощью решения соответствующей полной системы дифференциальных уравнений.

Пример 2. Рассмотрим систему (11), (12) из примера 1. Семейство C_0 состоит из двух подсемейств:

$$X_{1u} = \frac{\partial}{\partial x^1} + ux^2 \frac{\partial}{\partial y^1} + u \frac{\partial}{\partial y^2},$$

$$X_{2u} = \frac{\partial}{\partial x^2} + y^2(1 + x^2 y^1) \frac{\partial}{\partial y^1} + y^1 y^2 \frac{\partial}{\partial y^2}.$$

Легко видеть, что поля $Y_1 = X_{10}$, $Y_2 = X_{11}$, $Y_3 = X_{2u}$ образуют линейно несвязанное подсемейство D семейства C_0 . Вычисляя коммутаторы, получим

$$[Y_1, Y_2] = 0, \quad [Y_1, Y_3] = 0, \quad [Y_2, Y_3] = (x^2 y^2 + y^1) Y_2 - (x^2 + y_1) Y_1.$$

Таким образом, поля $Y_1 = X_{10}$, $Y_2 = X_{11}$, $Y_3 = X_{2u}$ составляют полное семейство, т.е. процесс пополнения заканчивается на первом шаге. Полная система дифференциальных уравнений имеет решение $\Phi(x, y) = y^2 x^2 - y^1$. После замены переменных $z^1 y^2 x^1 - y^1$, $z^2 = y^2$ исходная система уравнений приводится к виду

$$\frac{\partial z^1}{\partial x^1} = \frac{\partial z^1}{\partial x^2} = 0, \quad (46)$$

$$\frac{\partial z^2}{\partial x^2} = z^2(z^2 x^2 - z^1). \quad (47)$$

2.2. Системы в общем положении

Это такие системы, у которых отсутствуют нетривиальные первые интегралы. Термин системы, находящейся в общем положении, был введен для динамических систем с управлением (см. [14])

$$\dot{y}^i = f^i(y, u), \quad i = 1, \dots, n, \quad y \in M \subset R^n, \quad u \in U \subset R^r. \quad (48)$$

В определенном смысле почти все системы вида (48) находятся в общем положении (см. [14]). Аналогичное свойство выполняется и для систем дифференциальных уравнений с частными производными и систем управления с распределенными параметрами в силу одинаковости применяемых конструкций.

В предыдущем разделе описан алгоритм нахождения первых интегралов для системы (13). Он заключается в процессе пополнения. Операторы полного дифференцирования (23) после подстановки в них всевозможных постоянных $u \in U$ дают ассоциированное семейство векторных полей в M , компоненты которых зависят от y . Далее нужно выделить в ассоциированном семействе максимальное число линейно несвязанных векторных полей, объединить эти поля и пополнить в области изменения переменных M . Процесс пополнения заключается в постепенном добавлении коммутаторов до тех пор, пока все новые коммутаторы не будут линейно выражаться через уже полученные. Если число полей s в полученном полном семействе меньше числа переменных n , то нетривиальные первые интегралы существуют, причем число функционально независимых интегралов равно $n - s$. Таким образом, нетривиальные первые интегралы отсутствуют, если $s = n$, т.е. полученное полное семейство состоит из n линейно несвязанных полей. Следовательно, матрица из компонент этого семейства является квадратной $n \times n$ невырожденной матрицей. Теперь заметим, что ассоциированные поля удовлетворяют соотношениям (28). Прямым вычислением подтверждается, что коммутаторы этих полей также удовлетворяют (28), а следовательно, и поля построенного полного семейства. Они также удовлетворяют и соотношениям (33), ибо эти соотношения являются следствиями соотношений (28). Для построенного полного семейства соотношения (33) характеризуются следующим обстоятельством: левую часть можно трактовать как умножение невырожденной матрицы на производные от b_a^j . В результате эти производные можно выразить через остальные члены уравнений, и уравнения преобразуются в систему типа (39). После подстановки производных в (34) эти дифференциальные связи заменяются на конечные связи.

Согласно замечанию 1, в этом случае при проведении алгоритма исследования на совместность системы (33) дополнительные дифференциальные связи типа уравнений (34) можно заменить на конечные связи и рассматривать их в рамках проведения данного алгоритма.

ЗАКЛЮЧЕНИЕ

Рассмотрена возможность применения методов теории управления в теории уравнений с частными производными. При этом использована специальная форма представления любой системы уравнений с частными производными. Эта форма дает возможность трактовать параметрические переменные как некоторые управления, порождающие все решения. Такая трактовка позволяет в определенном смысле использовать идеологию теории управления, но, главное, позволяет ввести аналогичные математические объекты, которые показали свою эффективность для решения задач управления. Здесь показана эффективность этих методов для исследования проблемы декомпозиции систем уравнений с частными производными. Следующая проблема — исследование симметрий — для следующих работ.

СПИСОК ЛИТЕРАТУРЫ

1. Павловский Ю.Н. Групповые свойства управляемых систем и фазовые организационные структуры. // Ж. вычисл. матем. и матем. физ. 1974. Т. 14. № 4. С. 862–872.
2. Павловский Ю.Н. Групповые свойства управляемых систем и фазовые организационные структуры. // Ж. вычисл. матем. и матем. физ. 1974. Т. 14. № 5. С. 1093–1103.
3. Рашевский П.К. Геометрическая теория уравнений с частными производными. М.—Л.: Гостехиздат, 1947. 354 с.
4. Овсянников Л.В. Групповой анализ дифференциальных уравнений. М.: Наука, 1978, 389 с.

5. *Ибрагимов Н.Х., Авдонина Е.Д.* Нелинейная самосопряженность, законы сохранения и построение решений уравнений в частных производных с помощью законов сохранения // *Успехи матем. наук.* 2013. Т. 68. Вып. 5(413). С. 111–146.
6. *Vinogradov A.M.* Symmetries and conservation laws of partial differential equations: basic notions and results // *Special Issue of Acta Applicandae Mathematicae.* V. 15. Iss. 1/2, *Acta Applicandae Mathematicae.* 1989. V. 15. P. 3–21.
7. *Прохорова М.Ф.* Факторизация уравнения реакции-диффузии, волнового уравнения и других // *Тр. ИММ УрО РАН.* 2013. Т. 19. № 4. С. 203–213.
8. *Елкин В.И.* Об условиях агрегирования управляемых динамических систем // *Ж. вычисл. матем. и матем. физ.* 1978. Т. 18. № 4. С. 928–934.
9. *Елкин В.И.* Общее решение систем уравнений в частных производных с одинаковой главной частью // *Дифференц. ур-ния.* 1985. Т. 21. № 8. С. 1389–1398.
10. *Олвер П.* Приложения групп Ли к дифференциальным уравнениям. М.: Мир, 1989. 639 с.
11. *Эйзенхарт Л.П.* Непрерывные группы преобразований. М.: Изд-во иностр. лит., 1947. 359 с.
12. *Курант Р.* Уравнения с частными производными. М.: Мир, 1964, 830 с.
13. *Яковенко Г.Н.* Теория управления регулярными системами. М.: БИНОМ. Лаборатория знаний, 2008. 264 с.
14. *Лобри К.* Динамические полисистемы и теория управления. В кн.: *Математические методы в теории систем.* М.: Мир, 1979. С. 134–173.