СОДЕРЖАНИЕ

Том 61, номер 1, 2021 год

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

МАТЕМАТИЧЕСКАЯ ФИЗИКА

ИНФОРМАТИКА

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 3–19

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.6

EBR СХЕМЫ С КРИВОЛИНЕЙНЫМИ РЕКОНСТРУКЦИЯМИ ПЕРЕМЕННЫХ ВБЛИЗИ ОБТЕКАЕМЫХ ТЕЛ[1)]

© 2021 г. А. П. Дубень[1], Т. К. Козубская[1], П. В. Родионов[1,*], В. О. Цветкова[1]

[1] 125047 Москва, Миусская пл., 4, ИПМ им. М.В. Келдыша РАН, Россия

**e-mail: rodionov.cs@gmail.com*

Поступила в редакцию 13.05.2020 г.
Переработанный вариант 05.06.2020 г.
Принята к публикации 18.09.2020 г.

Работа посвящена развитию вершинно-центрированных EBR схем повышенной точности для расчета задач газовой динамики на неструктурированных сетках. Предлагается оснастить данные схемы возможностью использования квазиодномерного криволинейного шаблона для реконструкции переменных в областях структурированной или полуструктурированной анизотропной сетки вблизи обтекаемого тела. В двумерном случае использование криволинейных реконструкций приводит к естественной трансформации EBR схемы в структурированный конечно-объемный метод на структурированной сетке. В трехмерном случае для реализации криволинейных реконструкций переменных в призматических слоях разработан оригинальный алгоритм поиска точек шаблона и соответствующих метрических коэффициентов. Эффект от использования криволинейных реконструкций в EBR схемах при решении задач внешнего обтекания демонстрируется на известной тестовой задаче о течении вокруг аэродинамического профиля NACA0012, рассмотренной в двумерной и трехмерной постановках. Валидация нового алгоритма проводится путем сравнения с известными экспериментальными данными, а также результатами расчетов других авторов. Возможность использования криволинейных реконструкций в EBR схеме приводит к улучшению устойчивости метода и повышению точности численных результатов. Библ. 14. Фиг. 14. Табл. 4.

Ключевые слова: EBR схема, квазиодномерная реконструкция, криволинейный шаблон, полуструктурированная сетка, пограничный слой, задача внешнего обтекания.

DOI: 10.31857/S0044466920120030

ВВЕДЕНИЕ

Корректное моделирование турбулентных течений в задачах внешнего обтекания тел произвольной криволинейной конфигурации принципиальным образом зависит от точности воспроизведения формирующихся пограничных слоев. Необходимым условием для этого является достаточно подробное в нормальном направлении сеточное разрешение при сгущении сетки к поверхности обтекаемого тела. В продольном направлении значения переменных изменяются более плавно, а потому шаг сетки в этом направлении может быть существенно большим. Такой особенности приграничного турбулентного течения наилучшим образом отвечают слои анизотропной сетки, окружающие обтекаемое тело. При использовании неструктурированных сеток приграничные замкнутые слои или отдельные слоистые участки образуют структурированные или полуструктурированные подобласти. При этом на протяженных участках обтекаемого объекта течение направлено вдоль разделяющих слои криволинейных поверхностей (или линий в двумерном случае).

Учет в численном методе особенностей пристеночного течения и наличия структуры сетки в области, прилегающей к поверхности тела, может существенным образом повысить точность моделирования. Особенно это касается методов, использующих пространственные аппроксимации переменных на расширенных шаблонах.

[1)]Работа выполнена при финансовой поддержке РФФИ (код проекта 18-01-00445) и выполнена с использованием суперкомпьютеров ЦКП ИПМ им. М.В. Келдыша РАН и оборудования Центра коллективного пользования сверхвысокопроизводительными вычислительными ресурсами МГУ им. М.В. Ломоносова.

В работе рассматриваются алгоритмы, основанные на использовании вершинно-центрированных EBR (Edge-Based Reconstruction) схем для неструктурированных сеток [1]. Повышенная точность данных схем обеспечивается за счет реконструкций потоковых переменных на расширенных реберно-ориентированных шаблонах, а их экономность – за счет квазиодномерной природы такого подхода. EBR схемы, изначально разработанные для произвольных тетраэдральных сеток, допускают обобщение на гибридные неструктурированные сетки [2]. Однако на криволинейной структурированной сетке при большой степени анизотропии сеточных элементов использование EBR схем в их оригинальной формулировке, подразумевающей *прямолинейную* реконструкцию (т.е. реконструкцию на прямолинейных шаблонах), может приводить как к снижению точности расчета, так и к развитию неустойчивости. В настоящей работе исследуются причины таких негативных явлений. Для преодоления возникающих трудностей предлагается использование криволинейных шаблонов для реконструкции переменных (иначе говоря, *криволинейных* реконструкций), учитывающих структуру сетки в пристеночной области. Ранее аналогичный подход был реализован для метода коррекции потока [3], [4].

Целью настоящей работы является реализация техники криволинейных реконструкций для их использования в EBR схемах в двумерном случае, а главное, разработке нового эффективного алгоритма построения криволинейных реконструкций для полуструктурированных сеток в трехмерном случае. Эффект от использования криволинейных реконструкций демонстрируется на численном решении одной из классических валидационных задач об обтекании профиля NACA0012 [5], рассмотренной в двумерной и трехмерной постановках.

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Рассмотрим систему уравнений Навье–Стокса для сжимаемого газа

$$\frac{\partial \mathbf{Q}}{\partial t} + \nabla \cdot F(\mathbf{Q}) = \nabla \cdot F_v(\mathbf{Q}, \nabla \mathbf{Q}), \tag{1}$$

записанную относительно консервативных переменных, где

$$\mathbf{Q} = \begin{pmatrix} \rho \\ \rho \mathbf{u} \\ E \end{pmatrix}, \quad F = \begin{pmatrix} \rho \mathbf{u} \\ \rho \mathbf{u}\mathbf{u} + pI \\ (E + p)\mathbf{u} \end{pmatrix}, \quad F_v = \begin{pmatrix} 0 \\ \sigma \\ \sigma \mathbf{u} - \mathbf{q} \end{pmatrix}.$$

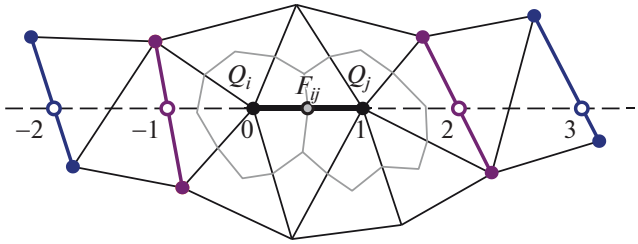
Здесь ρ – плотность, $\mathbf{u}$ – вектор скорости, p – давление, E – полная энергия, I – единичная матрица. Тензор напряжений и вектор теплового потока определяются как $\sigma = \mu(\nabla \mathbf{u} + (\nabla \mathbf{u})^{\mathrm{T}} - (2/3)(\nabla \cdot \mathbf{u})I)$ и $\boldsymbol{q} = -k\nabla T$ соответственно, где μ – коэффициент динамической вязкости, k – коэффициент теплопроводности, T – температура. Коэффициент динамической вязкости будем определять согласно закону Сазерленда.

В настоящей работе используется также система уравнений Навье–Стокса (1), осредненных по Рейнольдсу (RANS, Reynolds-Averaged Navier–Stokes). Входящий в систему RANS уравнений тензор рейнольдсовых напряжений замкнем с помощью линейной модели турбулентности Спаларта–Аллмараса (SA) [6], записанной относительно модифицированной турбулентной вязкости. При этом общий вид вязких потоков F_v останется неизменным с точностью до турбулентной вязкости, которая добавляется к динамической.

2. ЧИСЛЕННЫЙ МЕТОД НА ОСНОВЕ ОРИГИНАЛЬНОЙ EBR СХЕМЫ

Для численного решения системы уравнений (1) на произвольной вычислительной сетке построим схему с определением переменных в сеточных узлах. Далее такие схемы будем называть *вершинно-центрированными*. Вокруг каждого узла определим медианные ячейки, для которых, согласно конечно-объемному подходу, сформулируем разностные аналоги законов сохранения. Примеры медианных ячеек для двумерной сетки изображены на фиг. 1. Определив сеточную функцию $\mathbf{Q}_i$ как интегральное среднее функции $\mathbf{Q}$ по построенной вокруг узла i ячейке и использовав формулу Остроградского–Гаусса, перепишем систему (1) в векторно-матричном виде

$$V_i \frac{d\mathbf{Q}_i}{dt} + \sum_{j \in N_1(i)} \mathbf{F}_{ij} s_{ij} = F_{i,v},$$



Фиг. 1. Шаблон схемы EBR5 для ребра ij на неструктурированной треугольной сетке.

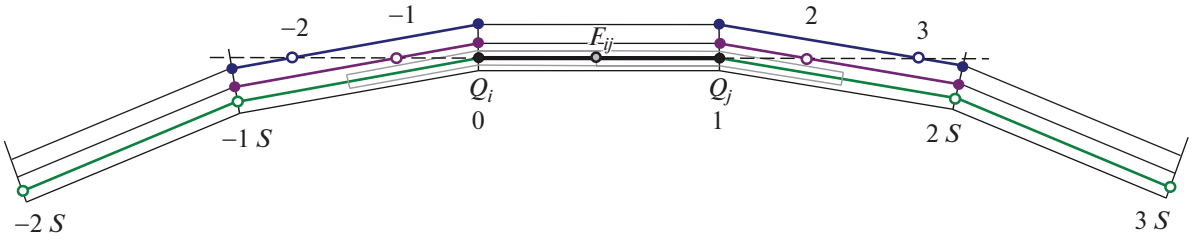
где V_i – объем ячейки, соответствующей узлу i, $\mathbf{F}_{ij}$ – интегральное среднее функции $F\mathbf{n}$ по разделяющей узлы i и j грани ячеек, площадь которой равна s_{ij}, $\mathbf{n}$ – вектор единичной нормали, $N_1(i)$ – множество соседей первого порядка узла i, $F_{i,v}$ – интеграл функции вязкого потока F_v по ячейке, отвечающей узлу i. Для вычисления конвективных потоков $\mathbf{F}_{ij}$ будем использовать метод Роу приближенного решения задачи о распаде разрыва:

$$\mathbf{F}_{ij} = \frac{1}{2}(\mathbf{F}_{ij}^R + \mathbf{F}_{ij}^L) - \frac{1}{2}|A_{ij}|(\mathbf{Q}_{ij}^R - \mathbf{Q}_{ij}^L).$$

Значения $\mathbf{Q}_{ij}^{L/R}$ слева и справа от интерфейса определим при помощи квазиодномерных реконструкций $R_{ij}^{L/R}\{\mathbf{Q}\}$, определенных на шаблонах, точки которых принадлежат прямой, содержащей ребро ij. Значения потоков $\mathbf{F}_{ij}^{L/R}$ будем полагать равными $F(R_{ij}^{L/R}\{\mathbf{Q}\})\mathbf{n}_{ij}$ либо $R_{ij}^{L/R}\{F\mathbf{n}_{ij}\}$ в зависимости от выбранного типа реконструкций [7]. Здесь $\mathbf{n}_{ij}$ – интегральное среднее вектора $\mathbf{n}$ по общей грани между ячейками, соответствующими узлам i и j, $|A_{ij}| = \left|\frac{dF\mathbf{n}}{d\mathbf{Q}}(\mathbf{Q}_{ij})\right| = S_{ij}|\Lambda_{ij}|S_{ij}^{-1}$, $\mathbf{Q}_{ij}$ – среднее по Роу, вычисленное по значениям $\mathbf{Q}_{ij}^{L/R}$, S_{ij} и Λ_{ij} – соответствующие $\mathbf{Q}_{ij}$ матрицы собственных векторов и собственных значений оператора $\frac{dF\mathbf{n}}{d\mathbf{Q}}$. Построенную таким образом схему, использующую реберно-ориентированные реконструкции переменных, можно в широком смысле отнести к классу EBR схем. В более узком смысле, согласно работе [1], в EBR схемах используются такие реконструкции, которые на трансляционно-инвариантных (переходящих в себя при переносе на вектор любого сеточного ребра) сетках приводят к трансформации данного метода в конечно-разностную схему высокого порядка. При этом схема данного семейства называется EBR n схемой, если в линейном случае ее порядок на трансляционно-инвариантных сетках равен n.

Опишем метод построения квазиодномерных реконструкций, используемых в оригинальной формулировке EBR схем [1], на примере схемы EBR5 в двумерной постановке (фиг. 1). Рассмотрим ребро ij, в середине которого необходимо реконструировать значение функции $\mathbf{Q}$, и для каждого из его узлов построим множества топологических соседей первого и второго порядка. Обозначим точку пересечения луча ji с множеством граней, все узлы которых являются соседями второго порядка узла i, индексом -2, а точку пересечения данного луча с множеством граней, все узлы которых являются соседями первого порядка узла i, индексом -1. В случае неединственности первой точки, индексом -2 обозначим точку, наиболее удаленную от узла i. Аналогично для луча ij и узла j получим точки с индексами 3 и 2 соответственно. Значения функции $\mathbf{Q}$ в точках $\{-2, -1, 2, 3\}$ определим при помощи линейной интерполяции по соответствующим пересекаемым лучом граням. Если присвоить узлу i индекс 0, а узлу j – индекс 1, то операторы реконструкции функции $\mathbf{Q}$ в терминах разделенных разностей

$$\Delta_m^L\{\mathbf{Q}\} = \frac{\mathbf{Q}_{m+1} - \mathbf{Q}_m}{|\mathbf{r}_{m+1} - \mathbf{r}_m|}, \quad \Delta_m^R\{\mathbf{Q}\} = \Delta_{-m}^L\{\mathbf{Q}\}$$



Фиг. 2. Шаблоны схем EBR5 и EBR5 IJK (EBR5 SS) для ребра ij на анизотропной структурированной сетке вблизи обтекаемого тела.

могут быть записаны как

$$R_{ij}^{L}\{\mathbf{Q}\} = \mathbf{Q}_i + \frac{|\mathbf{r}_i - \mathbf{r}_j|}{2}\sum_m \beta_m \Delta_m^L\{\mathbf{Q}\},$$
$$R_{ij}^{R}\{\mathbf{Q}\} = \mathbf{Q}_j - \frac{|\mathbf{r}_i - \mathbf{r}_j|}{2}\sum_m \beta_m \Delta_m^R\{\mathbf{Q}\}, \qquad (2)$$

где $\beta_{-2} = -1/15$, $\beta_{-1} = 11/30$, $\beta_0 = 4/5$, $\beta_1 = -1/10$ [1]. В схеме EBR3, для определения которой нужен более короткий шаблон, данные коэффициенты принимают значения $\beta_{-2} = 0$, $\beta_{-1} = 1/3$, $\beta_0 = 2/3$, $\beta_1 = 0$.

Как уже отмечено выше, для линеаризованных уравнений на трансляционно-инвариантных сетках теоретический порядок точности схем EBR5 и EBR3 равен пятому и третьему соответственно. В произвольном случае численный порядок точности схем EBR3 и EBR5 в зависимости от качества используемой неструктурированной сетки может варьироваться от 5/4 [8] до 3 [1].

Используемый в работе численный метод решения системы уравнений (1) имеет в основе своей схему EBR5 или EBR3 для аппроксимации конвективных потоков. Аппроксимация же вязких потоков проводится при помощи метода Галеркина с кусочно-линейными базисными функциями (с диагонализированной матрицей масс). Для интегрирования по времени применяется неявная схема первого порядка с линеаризацией системы сеточных уравнений по Ньютону. Решение системы линейных алгебраических уравнений в рамках одной итерации по методу Ньютона осуществляется с помощью метода бисопряженных градиентов с ILU0 предобуславливателем.

3. EBR СХЕМА С КРИВОЛИНЕЙНЫМИ РЕКОНСТРУКЦИЯМИ В СТРУКТУРИРОВАННЫХ ОБЛАСТЯХ ДВУМЕРНОЙ СЕТКИ ВБЛИЗИ ОБТЕКАЕМОГО ТЕЛА

Описанные в предыдущем разделе EBR схемы могут быть применены, вообще говоря, на произвольных структурированных и неструктурированных сетках. Однако, как уже было замечено в предыдущем разделе, точность этих схем напрямую связана с качеством сетки. Так, опыт показывает, что на криволинейных сетках с высокой степенью анизотропии ячеек, часто используемых в областях пограничного слоя (фиг. 2) в задачах аэродинамического обтекания, стандартная процедура реконструкции на *прямолинейном* шаблоне согласно оригинальной формулировке EBR схем может привести к заметной ошибке численного решения. Это вызвано следующими причинами: во-первых, прямолинейность шаблона реконструкции приводит к сильному перепаду длин в соседних шагах шаблона, что может являться причиной возникновения схемной неустойчивости; во-вторых, точки прямолинейного шаблона в силу, вообще говоря, криволинейной геометрии обтекаемого тела попадают в разные области пограничного слоя, что, ввиду высоких градиентов течения, приводит к увеличению вариации значений функций в данных точках и, как следствие, к увеличению погрешности аппроксимации.

В значительной мере преодолеть указанные трудности можно путем замены прямолинейной реконструкции в EBR схемах на *криволинейную* (т.е. реконструкцию, использующую криволинейный шаблон), которая естественным образом соответствует структуре сетки в пристеночной области и свойствам течения в пограничном слое.

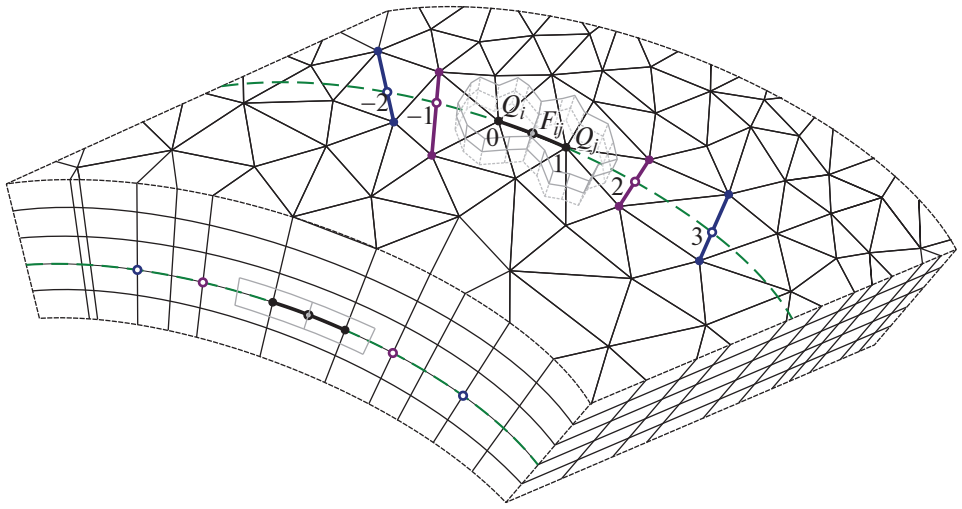
Сначала рассмотрим криволинейные реконструкции в EBR схемах и методы построения соответствующих им криволинейных шаблонов для двумерных задач аэродинамического обтекания, где в областях пограничных слоев, как правило, используются слои структурированных сеток, состоящих из трапециевидных элементов (фиг. 2). Учитывая структурированность данных сеток, в качестве точек шаблона криволинейной реконструкции будем выбирать не точки пересечения прямой ij с соответствующими изоповерхностями порядка соседства (точки $\{-2, -1, 2, 3\}$ на фиг. 2), а то или иное количество узлов справа и слева от узлов i и j, являющихся их структурными соседями (узлы $\{-2\ S, -1\ S, 2\ S, 3\ S\}$ на фиг. 2). Видно, что данный выбор обеспечивает примерно равный шаг между точками шаблона реконструкции, а также малую вариацию значений реконструируемой функции в пристеночной области. Выбор такого криволинейного шаблона вместо прямолинейного повышает устойчивость итоговой схемы и снижает ограничения на допустимую геометрию сетки при использовании тех же формул (2) для вычисления коэффициентов реконструкций. Стоит также отметить, что областях с резкими изменениями направлений сеточных линий, например вблизи острого угла геометрии, криволинейные реконструкции могут терять указанные свойства, и даже наоборот приводить к повышению ошибки и неустойчивости счета. Для устранения данного недостатка достаточно задать ограничение на максимальный угол между прямой ij и направлениями криволинейного шаблона реконструкции, при нарушении которого будет происходить переключение на оригинальную схему EBR.

Алгоритм перехода от криволинейных 5-точечных шаблонов реконструкций схемы EBR5 вблизи обтекаемого тела к стандартным прямолинейным реконструкциям при удалении от него удобно организовать поэтапно на основе анализа локальной структурированности сетки. Так, при отсутствии структурных соседей второго порядка, но при наличии соответствующих соседей первого порядка можно использовать укороченный криволинейный шаблон реконструкции, соответствующий схеме EBR3, а переходить к прямолинейным шаблонам реконструкции – только в случае отсутствия структурных соседей первого порядка. Отметим, что такой подход к построению шаблонов в структурированной подобласти сетки применим вдоль сеточных линий как в тангенциальном, так и нормальном направлении относительно обтекаемого тела. Описанный метод выбора шаблонов реконструкций уже использовался в работе [9]. В дальнейшем при описании численных результатов двумерные EBR схемы, явно использующие *ijk-топологию* при построении криволинейных шаблонов вблизи тела, будем обозначать “EBR IJK”.

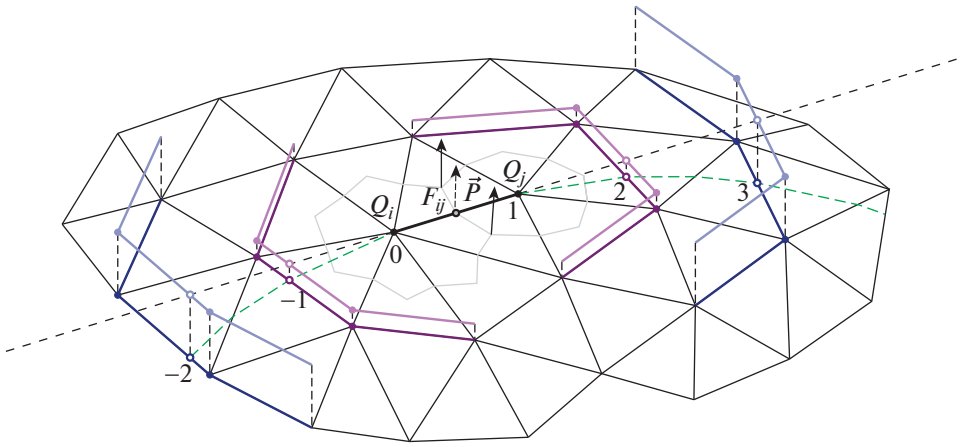
Сформулируем несколько иной вариант алгоритма построения шаблонов криволинейных реконструкций, преимущество которого заключается в простоте обобщения на трехмерный случай. Согласно этому алгоритму каждому узлу приграничной сетки сопоставляется уровень удаленности от поверхности обтекаемого тела, определяемый как минимальное число сеточных ребер, которыми этот узел может быть соединен с поверхностью. Для узлов на поверхности тела уровень удаленности полагается равным нулю. Если ребро ij лежит на изолинии уровня удаленности, то шаблон реконструкции будем составлять из узлов требуемого порядка соседства, имеющих тот же уровень удаленности. Формулы (2) для вычисления коэффициентов реконструкций, как и ранее, оставим неизменными. Нетрудно видеть, что построенные таким образом шаблоны реконструкции вдоль сеточных линий тангенциального направления будут совпадать с шаблонами реконструкции, построенными с использованием автоматического поиска структурированности сетки. Реконструкции же для ребер, расположенных в нормальном направлении по отношению к телу и вершины которых имеют разный уровень, будем проводить согласно оригинальной EBR схеме, т.е. при помощи прямолинейных шаблонов. Заметим, что в силу топологии приграничной структурированной сетки использование прямолинейных или криволинейных реконструкций в нормальном направлении не приводит к сколько-нибудь существенной разнице в результатах.

4. EBR СХЕМА С КРИВОЛИНЕЙНЫМИ РЕКОНСТРУКЦИЯМИ В ПРИЗМАТИЧЕСКИХ СЛОЯХ СЕТКИ ВБЛИЗИ ОБТЕКАЕМОГО ТЕЛА

В трехмерной постановке в областях пограничных слоев на практике, как правило, используются слоистые гексаэдральные структурированные или призматические полуструктурированные сетки. Последние называются *полуструктурированными* [10]–[13], поскольку обладают структурой в нормальном направлении (у каждого узла в данном направлении соседний узел однозначно определен), но не имеют ее в тангенциальных направлениях (фиг. 3). Если шаблоны реконструкции для ребер, располагающихся в нормальном направлении к телу, строятся в полной аналогии с двумерным случаем и их построение не представляет проблемы, то выбор шаб-



Фиг. 3. Шаблоны схемы EBR5 SS для ребра ij на призматической сетке вблизи обтекаемого тела.



Фиг. 4. Алгоритм нахождения точек криволинейного шаблона реконструкции переменных для схемы EBR5 SS.

лонов реконструкции на ребрах, лежащих на поверхностях раздела призматических слоев сетки, осложняется отсутствием сеточной структуры на этих поверхностях.

Для определения шаблонов криволинейных реконструкций в тангенциальном направлении построим следующий алгоритм. На стадии препроцессинга по аналогии с двумерным случаем определим уровни удаленности сеточных узлов от поверхности обтекаемого тела. Как и в двумерном случае, численный поток между узлами, имеющими разные уровни удаленности, будем определять согласно оригинальной EBR схеме, использующей прямолинейные реконструкции. Реконструкцию переменных на ребре ij, соединяющем соседние узлы i и j с одним уровнем удаленности, будем проводить по следующему алгоритму, иллюстрация к которому представлена на фиг. 4.

1. Для каждого из узлов ребра ij построим множества узлов, являющихся их соседями 1-го и 2-го порядка.

2. Из этих множеств исключим узлы, уровень удаленности которых отличается от уровня удаленности узлов i и j.

3. Каждому из построенных множеств узлов сопоставим множество сеточных ребер, оба узла которых принадлежат соответствующему множеству узлов.

4. Зададим проекционную плоскость, проходящую через ребро ij, посредством вектора $\vec{P}$, являющегося полусуммой нормалей к граням, инцидентным ребру ij и принадлежащим изопо-

верхности уровня удаленности узлов i и j. Если инцидентная грань только одна, вектор $\vec{P}$ положим равным нормали к ней, а если такие грани отсутствуют, реконструкцию вдоль ребра ij будем проводить согласно оригинальной схеме EBR с прямолинейными реконструкциями, пропуская последующие шаги данного алгоритма.

5. Спроектируем полученные в п. 3 множества ребер на проекционную плоскость, задаваемую вектором $\vec{P}$ и содержащую ребро ij, и построим на ней двумерный шаблон прямолинейной реконструкции в полном соответствии с оригинальной EBR схемой в двумерной постановке.

6. Криволинейную реконструкцию будем определять формулами (2) с метрическими коэффициентами, относящимися не к точкам шаблона прямолинейной реконструкции в проекционной плоскости, а к их прообразам, лежащим на изоповерхности соответствующего уровня удаленности и определяющим тем самым точки шаблона криволинейной реконструкции.

При использовании в качестве базовой схемы EBR5 в полуструктурированных областях сетки переключение между криволинейными и прямолинейными реконструкциями в неструктурированных зонах также предлагается выполнять поэтапно по аналогии с двумерным случаем. В дальнейшем при описании численных результатов трехмерные EBR схемы с криволинейными реконструкциями в полуструктурированных областях, построенными согласно вышеописанному алгоритму, будем обозначать “EBR SS” (*Semi-Structured*).

Заметим, что подобный подход к построению криволинейных реконструкций можно применять и в случае гексаэдральных слоев структурированных сеток, однако в этом случае проще использовать чисто структурированный подход, при котором шаблоны реконструкций определяются вдоль сеточных линий при заданной ijk-топологии.

5. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

5.1. Физическая постановка задачи

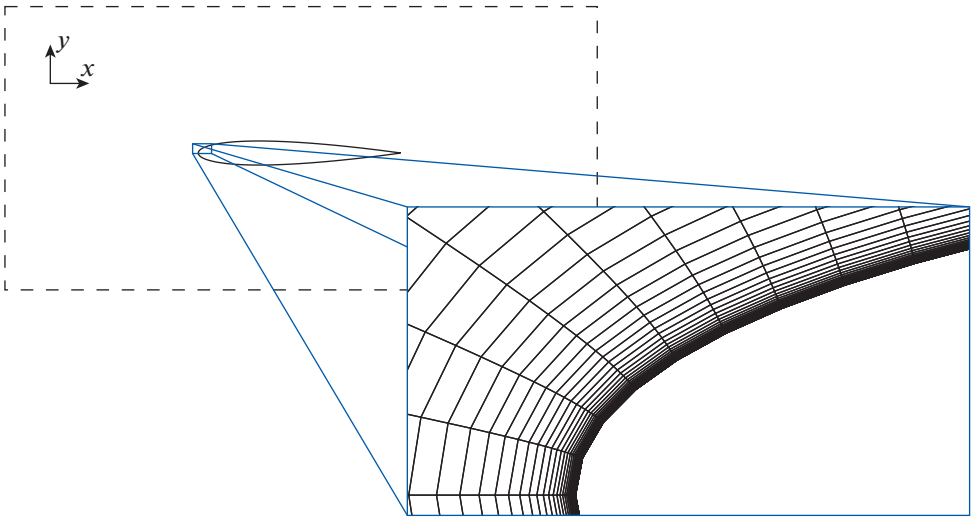
Для оценки модификаций схем EBR, описанных в предыдущем разделе, были проведены серии расчетов задачи об обтекании аэродинамического профиля NACA0012 [5]. Постановка задачи формулируется следующим образом. Профиль NACA0012 с хордой единичной длины помещается в однородный поток воздуха с числом Маха $M_\infty = 0.15$ и температурой $T_\infty = 300$ K. Число Рейнольдса относительно длины хорды c составляет $\mathrm{Re}_c = 6\times10^6$. В настоящей работе рассмотрены углы атаки 0°, 10° и 15°.

5.2. Вычислительная постановка задачи

При проведении расчетов используется фоновый поток, характеризующийся высоким уровнем турбулентности. Это обеспечивается за счет задания входного условия $\nu_t/\nu = 1$, где ν_t – коэффициент турбулентной вязкости, а ν – коэффициент молекулярной вязкости.

В двумерной постановке расчетная область представляет собой квадрат $-500 \le x/c, y/c \le 500$, в центре которого, в точке $(0,0)$, находится передняя кромка профиля. В трехмерной постановке расчетной областью является прямоугольный параллелепипед $-500 \le x/c,\ y/c \le 500$, $0 \le z/c \le 0.25$, передняя кромка профиля в данном случае принадлежит прямой $(0,0,z)$. На поверхности профиля задаются условия прилипания $\mathbf{u} = 0$, $\nu_t = 0$ и адиабатичности. На свободных границах $x/c = \pm500$ и $y/c = \pm500$ в случае входа держатся все параметры потока, кроме давления, которое экстраполируется, в случае выхода, наоборот, держится только давление, а основные параметры экстраполируются из внутренней области. В трехмерной постановке на границах $z/c = 0$ и $z/c = 0.25$ задаются условия периодичности.

Для проведения двумерных расчетов использовалась последовательность гибридных сеток, являющихся структурированными трапециевидными вблизи профиля и неструктурированными треугольными в остальной области (фиг. 5). Характерные параметры указанных сеток приведены в табл. 1, где N соответствует общему числу узлов сетки, а N_{surf} – числу узлов, принадлежащих границе обтекаемого профиля. Для проведения трехмерных расчетов использовалась аналогичная последовательность гибридных сеток, являющихся полуструктурированными призматическими вблизи профиля и неструктурированными тетраэдральными в остальной области. Сетки данной последовательности совпадали с соответствующими указанными ранее двумерными сетками на границах $z/c = 0$ и $z/c = 0.25$. Характерные параметры последовательности трехмерных



Фиг. 5. Конфигурация расчетных сеток.

сеток сведены в табл. 2, где при помощи $N_{\text{surf},\, z=0}$ обозначено количество точек на поверхности профиля в плоскости $z = 0$. Отметим, что названия “x1”, ..., “x8” введены исключительно для удобства дальнейших обозначений и не являются следствием последовательного разбиения.

Помимо указанных сеток, для валидационных расчетов в настоящей работе используется двумерная структурированная сетка 897 × 257, применяющаяся в [5] для получения эталонных численных результатов. Стоит отметить, что геометрия профиля в построенных сетках и сетке 897 × 257 отличались в пределах 1%.

Распределения значений безразмерной высоты первой пристеночной ячейки y^+, соответствующих обтеканию профиля под углами атаки 0°, 10° и 15°, для двумерных сеток изображены на фиг. 6. Аналогичные значения y^+ для трехмерных сеток практически не отличаются от приведенных.

Расчеты проводились до установления по абсолютной невязке (по полной энергии и турбулентной вязкости), а также до выхода значений коэффициентов подъемной силы и сопротивления формы к асимптотике.

5.3. Валидация EBR схем с криволинейными реконструкциями

Для тестирования схем EBR IJK и EBR SS использовались самые подробные двумерные и трехмерные сетки x8, а также эталонная сетка 897 × 257. Валидация данных схем и их программной реализации проводилась путем сопоставления значений безразмерных коэффициентов подъемной силы (C_l) и сопротивления формы (C_d) для различных углов атаки. Сравнение полученных результатов с данными из [5] представлено в табл. 3.

Видно, что в численных данных, полученных известными программными комплексами, значения C_l отличаются друг от друга на 1%, а значения C_d – на 4%. При добавлении к ним данных тестирования схем EBR5 и EBR5 IJK на сетке 897 × 257 разница по C_l остается в пределах 1%, а отклонение по C_d увеличивается до 6%. При добавлении к исходным данным результатов тестирования схем EBR5, EBR5 IJK и EBR5 SS на двумерных и трехмерных сетках ×8 максимальное отличие по C_l увеличивается до 2%, а отклонение C_d составляет 5%. Отметим также, что в рамках

Таблица 1. Параметры двумерных сеток

2D сетка	x1	x2	x3	x4	x8	897 × 257
N	55 тыс.	51 тыс.	69 тыс.	83 тыс.	84 тыс.	231 тыс.
N_{surf}	102	162	246	442	930	513

Таблица 2. Параметры трехмерных сеток

3D сетка	x1	x2	x3	x4	x8
N	515 тыс.	801 тыс.	1.4 млн.	2.9 млн.	9.7 млн.
N_{surf}	3 тыс.	7 тыс.	15 тыс.	40 тыс.	176 тыс.
$N_{\text{surf},\, z=0}$	102	162	246	442	930

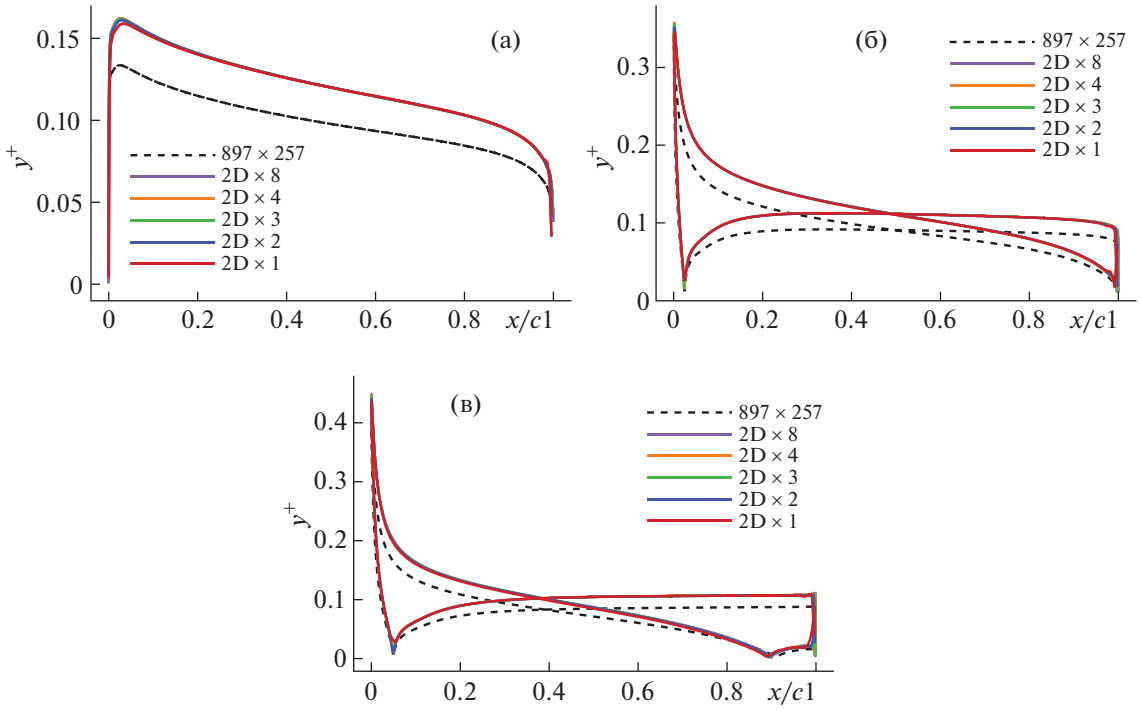
одной и той же сетки при использовании различных типов реконструкций разница по C_l сохраняется в пределах 0.1%, по C_d – в пределах 1%.

Сравнение численно полученных распределений коэффициентов давления (C_p) и трения (C_f) для различных углов атаки с данными [5] приведено на фиг. 7. Видно, что результаты расчетов соответствуют экспериментальным данным, а сами результаты расчетов по коэффициентам давления и трения практически совпадают друг с другом.

В итоге, несмотря на некоторое увеличение отклонений по интегральным характеристикам от соответствующих референсных значений и с учетом небольших различий в геометрии профиля между используемыми в расчетах сетками и эталонной сеткой 897 × 257, можно заключить, что полученные с помощью схем EBR IJK и EBR SS численные результаты подтверждают применимость данных схем для решения стационарных задач аэродинамического обтекания, а также корректность их программной реализации.

5.4. Сравнительный анализ результатов, полученных EBR схемами с прямолинейными и криволинейными реконструкциями

Рассмотрим теперь результаты расчетов, проведенных с использованием схем EBR с прямолинейными и криволинейными реконструкциями на последовательностях двумерных и трехмерных гибридных неструктурированных сеток x, x2, x3, x4 и x8. Как и ранее, на двумерных сетках будем использовать схему с криволинейными реконструкциями EBR IJK, а на трехмерных – схему с криволинейными реконструкциями EBR SS. Отметим, что применение прямолинейных



Фиг. 6. Безразмерная высота первой пристеночной ячейки y^+ для двумерных сеток при обтекании профиля под углами атаки 0° (а), 10° (б) и 15° (в).

Таблица 3. Результаты валидации схем EBR5 IJK и EBR5 SS по коэффициентам подъемной силы (C_l) и сопротивления формы (C_d)

Схема (сетка)	0°: C_l	10°: C_l	15°: C_l	0°: C_d	10°: C_d	15°: C_d
EBR5 (3D, x8)	~0	1.0862		0.00810	0.01234	
EBR5 SS (3D, x8)	~0	1.0865		0.00810	0.01233	
EBR5 (2D, x8)	~0	1.0875	1.5339	0.00811	0.01239	0.02179
EBR5 IJK (2D, x8)	~0	1.0871	1.5345	0.00812	0.01237	0.02166
EBR5 (897 × 257)	~0	1.0946	1.5437	0.00810	0.01264	0.02219
EBR5 IJK (897 × 257)	~0	1.0940	1.5436	0.00810	0.01259	0.02203
CFL3D (897 × 257)	~0	1.0909	1.5461	0.00819	0.01231	0.02124
FUN3D (897 × 257)	~0	1.0983	1.5547	0.00812	0.01242	0.02159
NTS (897 × 257)	~0	1.0891	1.5461	0.00813	0.01243	0.02105
JOE (897 × 257)	~0	1.0918	1.5490	0.00812	0.01245	0.02148
SUMB (897 × 257)	~0	1.0904	1.5446	0.00813	0.01233	0.02141
TURNS (897 × 257)	~0	1.1000	1.5642	0.00830	0.01230	0.02140
GGNS (897 × 257)	~0	1.0941	1.5576	0.00817	0.01225	0.02073

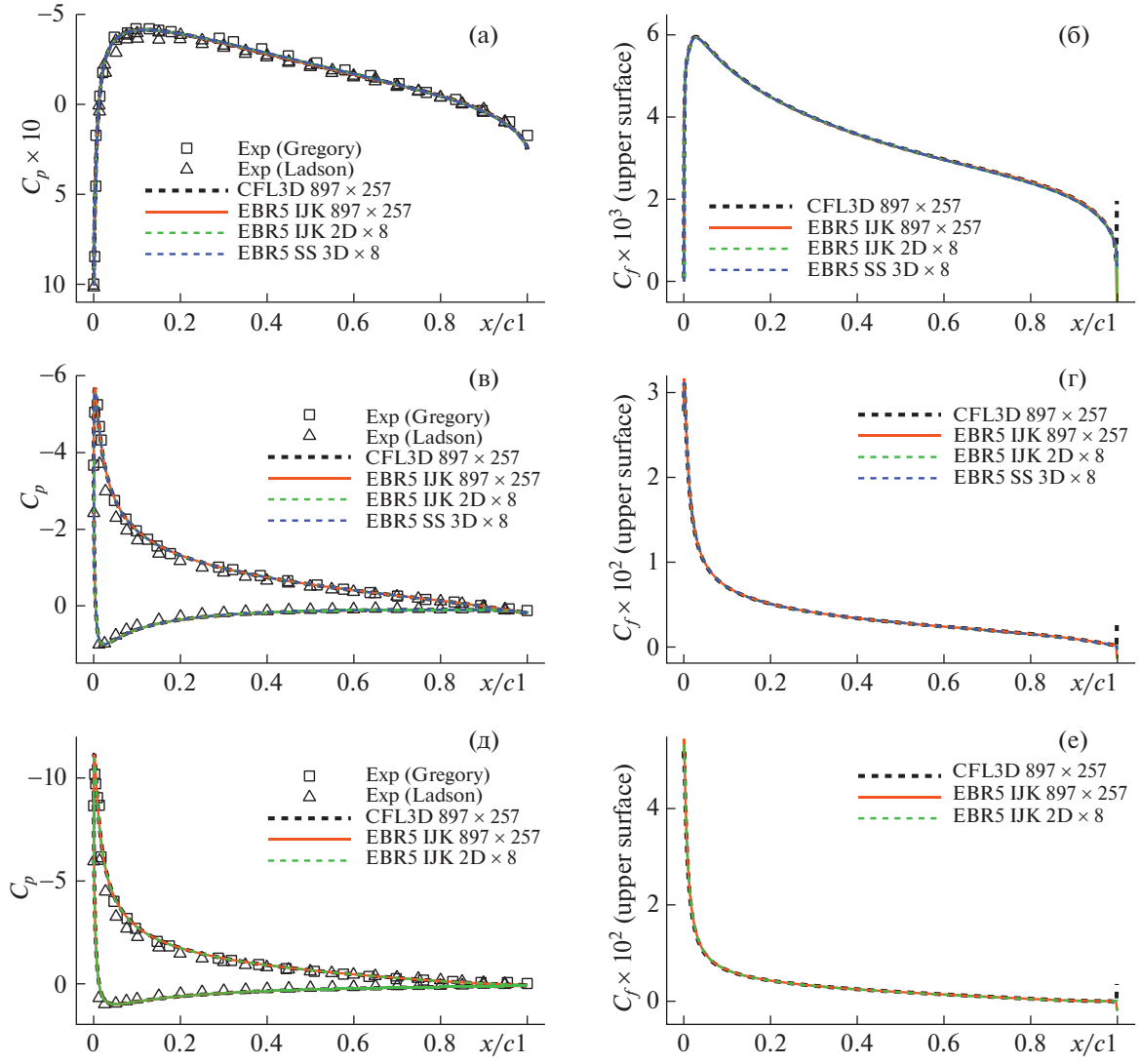
реконструкций при проведении расчетов на достаточно грубых сетках x1–x3, как правило, приводит к неустойчивости, возможные причины которой уже рассматривались. Для обеспечения устойчивости алгоритма в таких случаях использовалось ограничение максимального допустимого отношения длин шагов на шаблоне реконструкции и, при его несоблюдении, производилось переключение на схему EBR3 с изменением, при необходимости, коэффициента реконструкции β_{-1}: если в формулах (2) значение $|\mathbf{r}_0 - \mathbf{r}_{-1}| \times C_{\text{ratioLim}}$ было меньше, чем $|\mathbf{r}_i - \mathbf{r}_j|$, где C_{ratioLim} – глобально задаваемый ограничитель, значение β_{-1} умножалось на коэффициент $C_{\text{ratioLim}} \times |\mathbf{r}_0 - \mathbf{r}_{-1}| / |\mathbf{r}_i - \mathbf{r}_j|$. Для рассматриваемой задачи и сеток x1–x3 достаточным оказалось значение C_{ratioLim}, равное 20.

Начнем с анализа полученных интегральных характеристик для нулевого угла атаки (фиг. 8). Из приведенных результатов видно, что для всех схем на более грубых сетках ожидаемо получается менее точное решение, а на аналогичных двумерных и трехмерных сетках лучший результат получается на трехмерных.

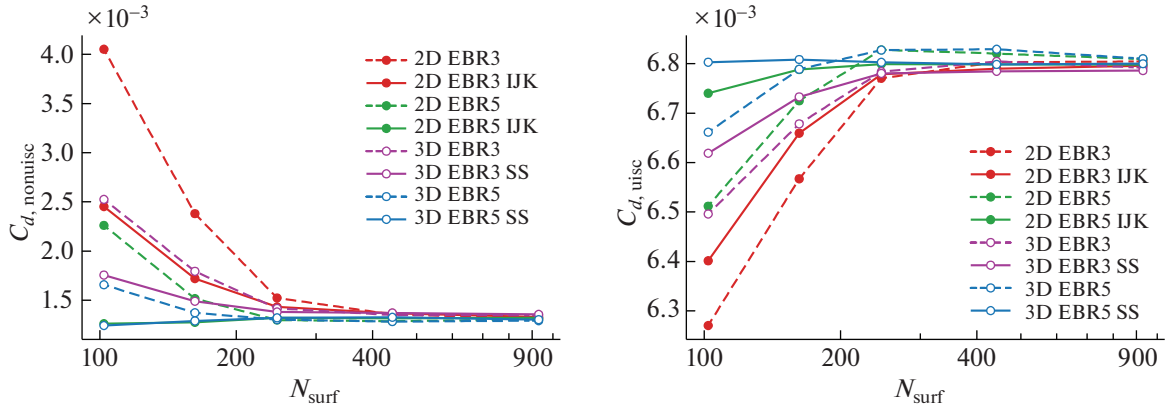
На грубых сетках x1–x3 схема EBR5 и ее версии с криволинейными реконструкциями ожидаемо демонстрируют более высокую точность в сравнении с соответствующими схемами, основанными на схеме EBR3. При этом использование криволинейных реконструкций приводит к существенному уменьшению ошибки таким образом, что значения исследуемых контрольных величин при использовании криволинейных реконструкций даже на достаточно грубых сетках x1 (как в двумерном, так и в трехмерном случаях) оказываются уже вполне близкими к истинным значениям, полученным на эталонных сетках. Также следует отметить, что при увеличении сеточного разрешения разница между результатами, полученными схемами с прямолинейными и криволинейным реконструкциями, уменьшается, что объясняется постепенным выпрямлением шаблонов криволинейных реконструкций в пограничном слое при сгущении сеток.

Приведенные выше выводы подтверждаются аналогичными результатами для углов атаки 10° и 15° (фиг. 9 и 10). Причем стоит заметить, что при ненулевом угле атаки на грубых сетках схемы EBR3, использующие укороченный криволинейный шаблон, показывает более высокую точность, чем схема EBR5, работающая на более протяженном, но прямолинейном шаблоне.

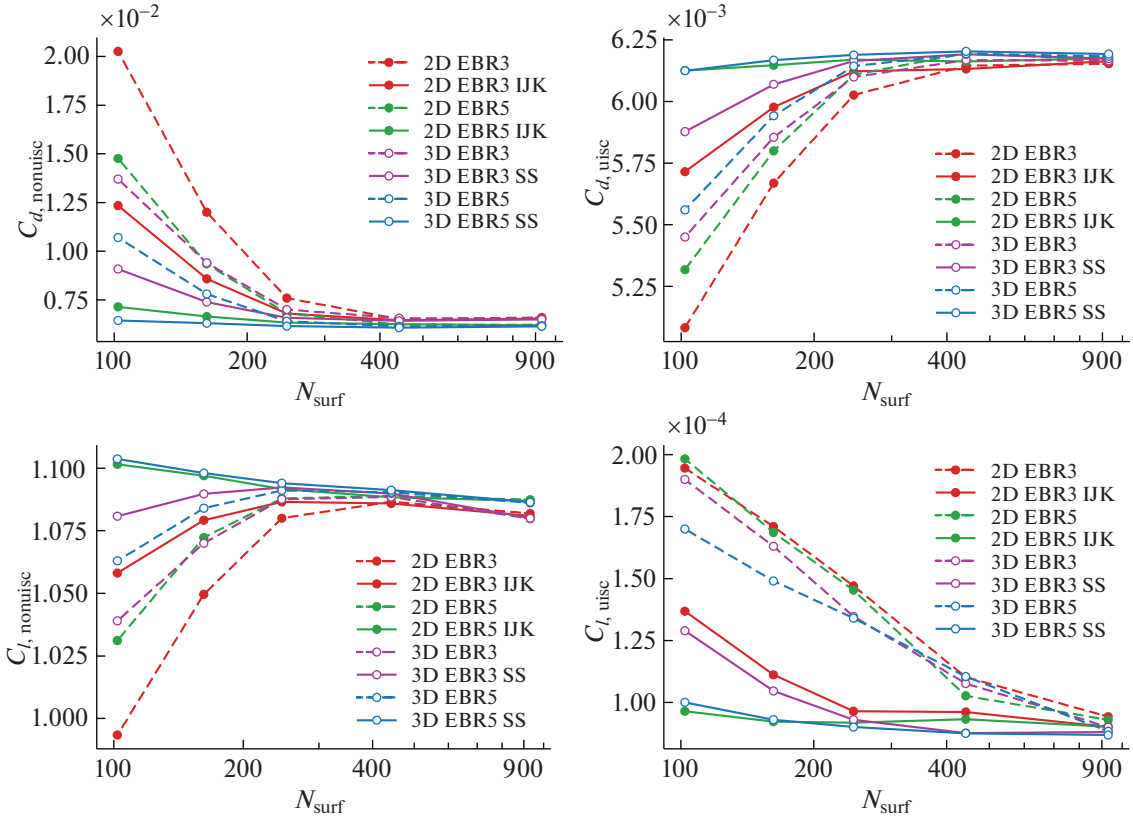
На фиг. 11 для иллюстрации работы рассматриваемых EBR схем на различных двумерных сетках приведены некоторые наиболее характерные распределения коэффициентов давления и трения по поверхности аэродинамического профиля. Из приведенных графиков видно, что для всех углов атаки наибольшее отклонение от эталонных численных результатов демонстрирует схема EBR3 на сетке x1, далее следуют схемы EBR5 на той же сетке, EBR3 на сетке x2, EBR5 на сетке x2 и EBR5 IJK на сетке x1. Описанная последовательность подтверждается данными, представленными на фиг. 8, 9 и 10. Еще раз отметим, что коэффициенты давления и трения, полученные



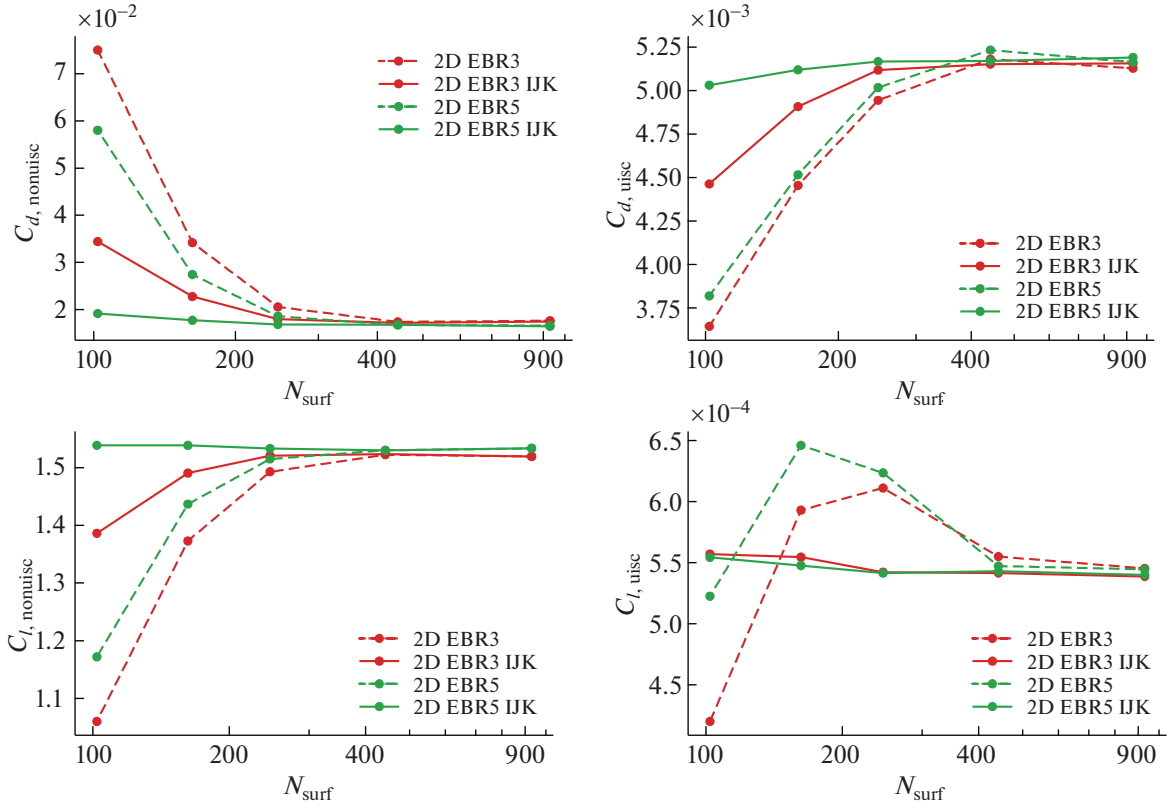
Фиг. 7. Результаты валидации схем EBR5 IJK и EBR5 SS по коэффициентам давления (C_p) и трения (C_f) для углов атаки 0° (а, б), 10° (в, г) и 15° (д, е).



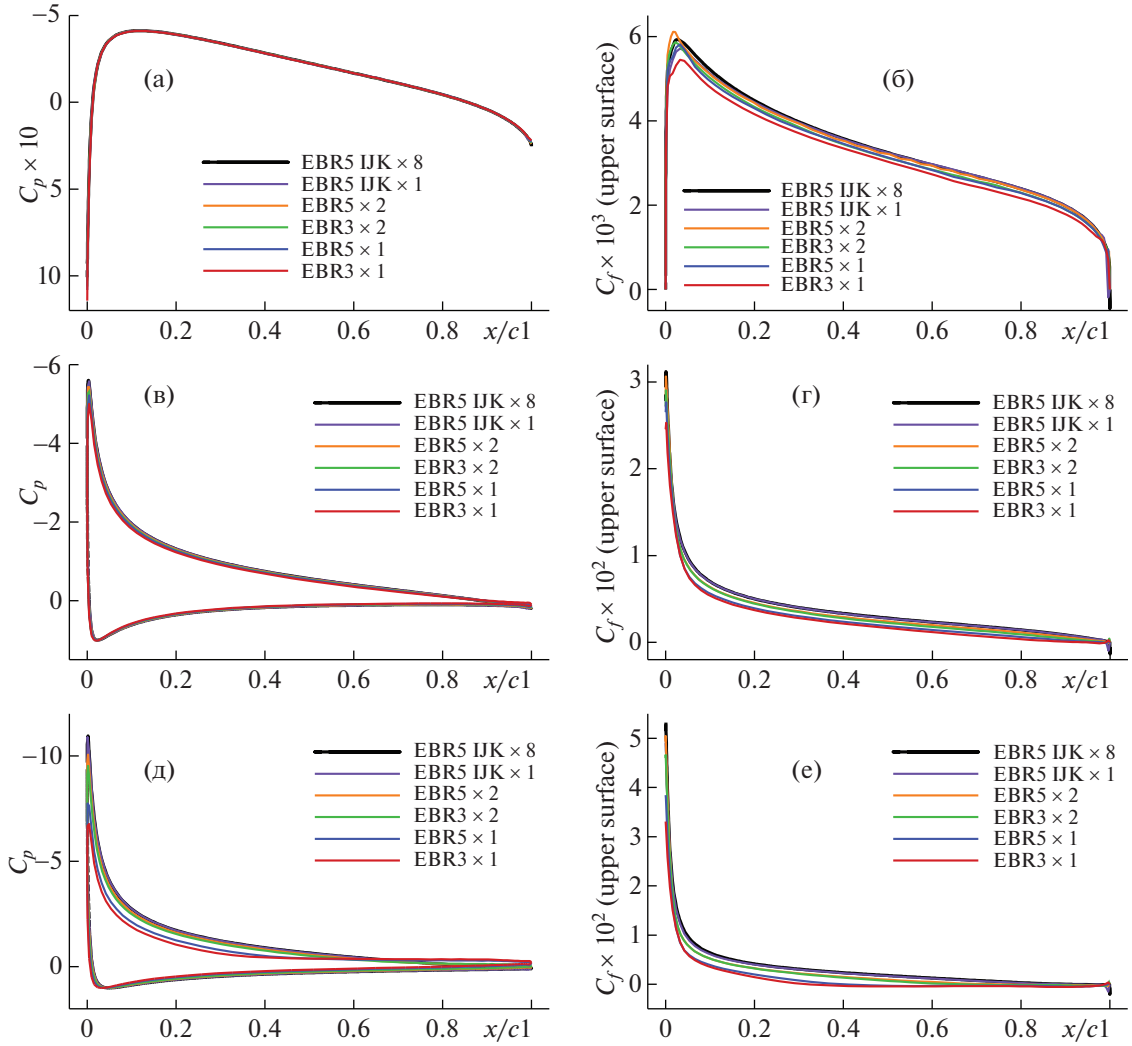
Фиг. 8. Невязкая ($C_{d,\text{nonvisc}}$) и вязкая ($C_{d,\text{visc}}$) компоненты коэффициента сопротивления формы, полученные с помощью различных схем на последовательностях двумерных и трехмерных сеток, для угла атаки 0°.



Фиг. 9. Компоненты коэффициентов сопротивления формы и подъемной силы, полученные с помощью различных схем на последовательностях двумерных и трехмерных сеток, для угла атаки 10°.



Фиг. 10. Компоненты коэффициентов сопротивления формы и подъемной силы, полученные с помощью различных схем на последовательности двумерных сеток, для угла атаки 15°.



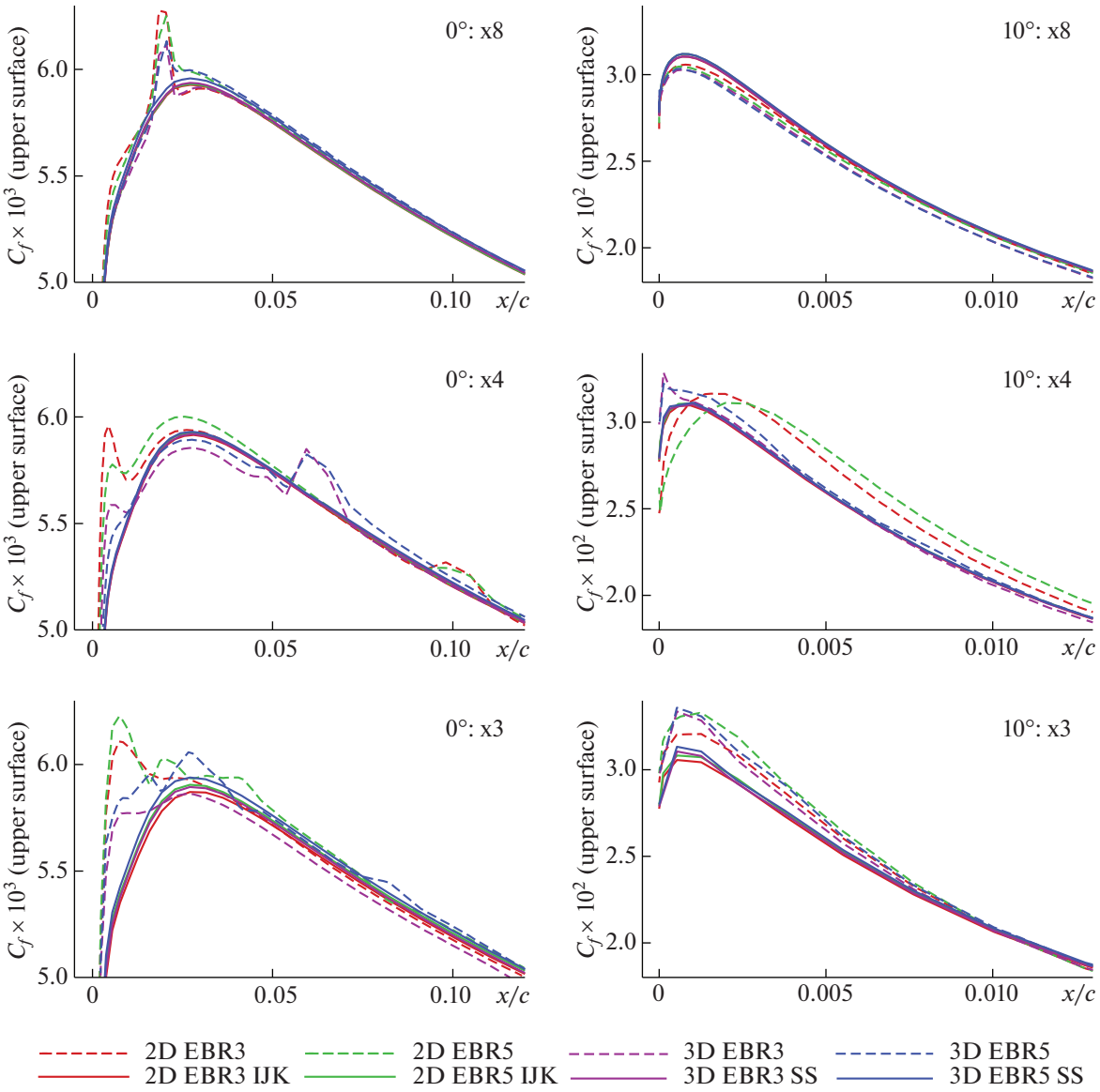
Фиг. 11. Наиболее характерные распределения коэффициентов давления (C_p) и трения (C_f), полученные в расчетах на последовательности двумерных сеток при углах атаки 0° (а, б), 10° (в, г) и 15° (д, е).

схемой EBR5 IJK с криволинейными реконструкциями, даже на самой грубой сетке x1 уже вполне хорошо согласуются с эталонными результатами, полученными на самой подробной сетке x8.

По приведенным на фиг. 11 распределениям коэффициента трения можно оценить вариации размеров зоны рециркуляции. Соответствующие координаты точки отрыва сведены в табл. 4.

Таблица 4. Координата x/c точки отрыва течения для некоторых расчетов, выполненных на последовательностях двумерных и трехмерных сеток

Схема (сетка)	10°	15°
EBR5 IJK (897 × 257)	–	0.91
EBR5 IJK (2D, x8)	–	0.90
EBR5 SS (3D, x8)	–	
EBR5 IJK (2D, x1)	–	0.89
EBR5 2D (2D, x2)	0.994	0.74
EBR3 2D (2D, x2)	0.987	0.66
EBR5 2D (2D, x1)	0.980	0.40
EBR3 2D (2D, x1)	0.935	0.32
EBR5 3D (3D, x1)	–	
EBR3 3D (3D, x1)	0.973	

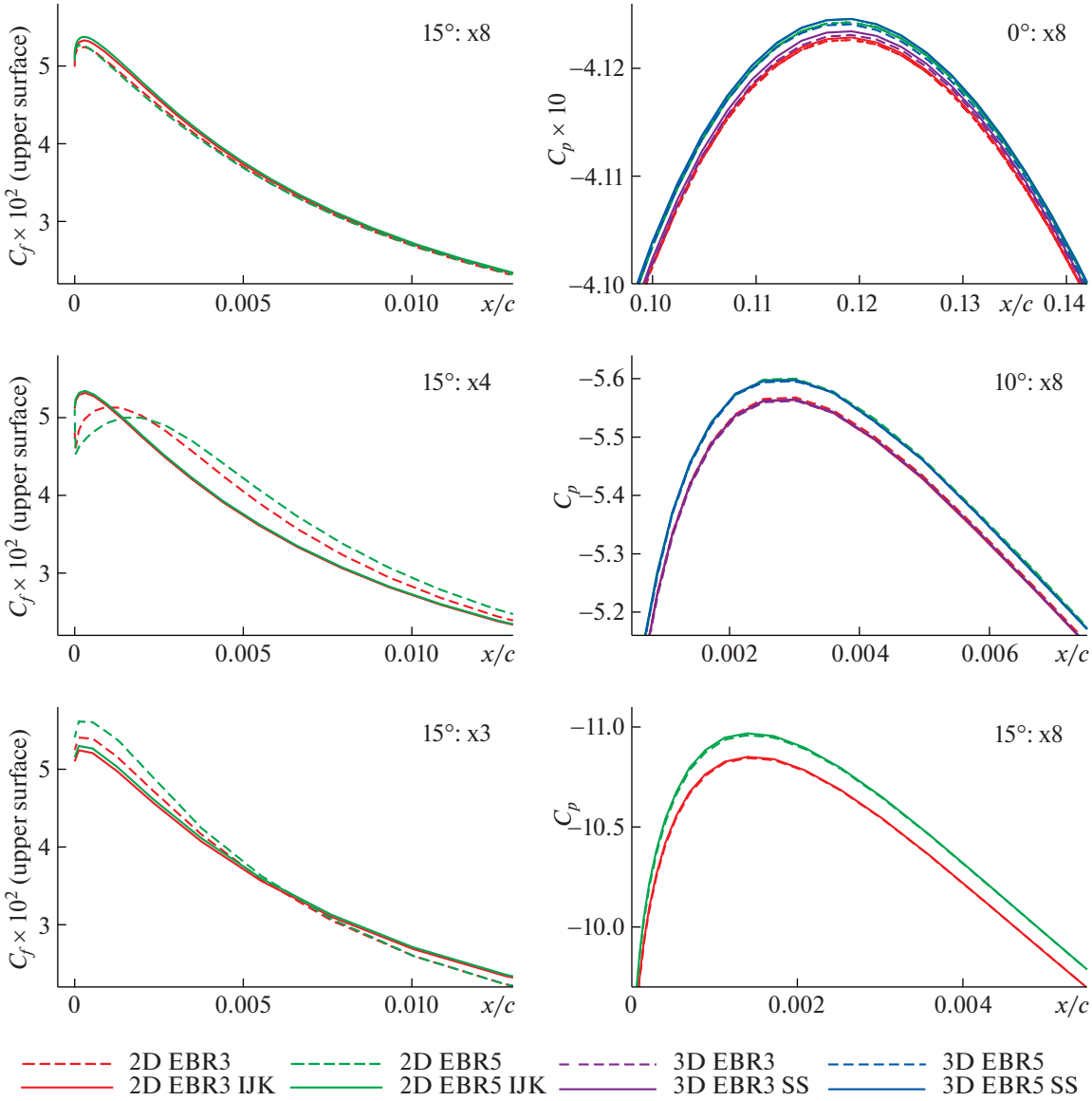


Фиг. 12. Распределения коэффициента трения в наиболее характерных областях, получаемые на подробных сетках при углах атаки 0° и 10°.

Из нее следует, что без использования криволинейных реконструкций в EBR схемах на грубых сетках полученный размер рециркуляционной зоны может существенно отличаться от соответствующего физически корректного результата. Особенно это заметно для угла атаки 15°. Также использование прямолинейных реконструкций на грубых сетках может приводить к ложному отрыву, что имеет место при угле атаки 10°.

Криволинейные реконструкции имеют смысл не только на грубых, но и на достаточно подробных сетках. Чтобы убедиться в этом, достаточно рассмотреть приведенные на фиг. 12 и 13 распределения коэффициента трения. Видно, что, во-первых, при использовании схем с прямолинейными реконструкциями как в двумерном, так и трехмерном случае коэффициент трения сходится к эталонному значению намного медленнее и менее регулярно в сравнении с аналогичными схемами с криволинейными реконструкциями, а во-вторых, что особенно заметно при угле атаки 0°, при общей сходимости могут сохраняться участки, в которых отличия от эталонных значений весьма существенны.

На фиг. 13 также приведены распределения коэффициента давления для наиболее характерных областей на самых подробных сетках x8 для различных углов атаки.

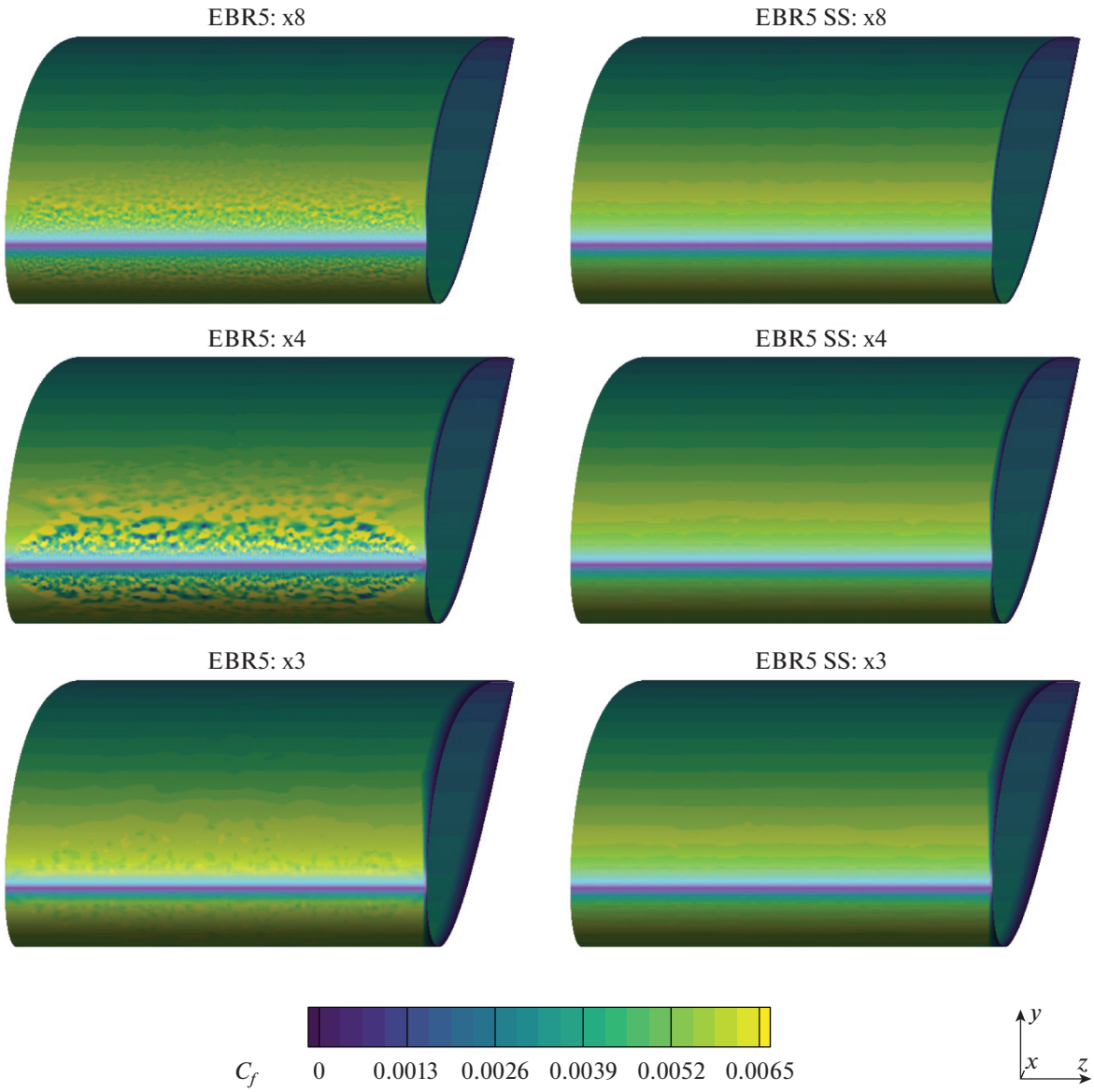


Фиг. 13. Распределения коэффициента трения в наиболее характерной области, получаемые на подробных сетках при угле атаки 15° (слева), и распределения коэффициента давления в наиболее характерных областях, получаемые для различных углов атаки на сетках x8 (справа).

Поверхностные распределения коэффициента трения для трехмерных расчетов обтекания с нулевым углом атаки по схемам EBR5 на подробных сетках изображены на фиг. 14. Из них видно, что использование прямолинейных реконструкций приводит к возникновению сеточных осцилляций, в то время как использование криволинейных реконструкций позволяет избежать данного эффекта. Аналогичная ситуация имеет место как для схем EBR3, так и для ненулевых углов атаки.

Из проведенного анализа можно заключить, что предлагаемые в работе обобщения EBR схем на случай криволинейных шаблонов реконструкции в областях структурированных или полуструктурированных сеток не только корректно работают, но и являются крайне необходимыми в расчетах на анизотропных сетках в областях пограничных слоев, которые используются при моделировании высокорейнольдсовых турбулентных течений.

Все описанные в настоящей работе расчеты выполнены с использованием программного комплекса NOISEtte [14].



Фиг. 14. Поверхностные распределения коэффициента трения, полученные с использованием схем EBR5 и EBR5 SS на трехмерных сетках x8, x4, x3 для угла атаки 0°.

ЗАКЛЮЧЕНИЕ

В работе предложено обобщение EBR схем на гибридных неструктурированных сетках за счет возможности проводить реберно-ориентированные реконструкции не вдоль прямой, содержащей данное ребро, а вдоль содержащей его кривой линии. Реализация такой возможности дает ряд важных преимуществ, существенных при моделировании течений вокруг тел произвольной формы.

Во-первых, реконструкции по криволинейному шаблону в структурированных и полуструктурированных областях пристеночной сетки позволяют получить более высокую точность численного решения, благодаря лучшей сонаправленности шаблона потоку.

Во-вторых, использование криволинейных реконструкций дает возможность избежать резкого перепада расстояний между узлами шаблона, который обычно возникает в оригинальных EBR схемах в результате пересечения прямой реконструкции слоистой сеточной структуры с анизотропными элементами. Сильная неравномерность шагов шаблона и большая вариация значений переменных в его узлах, возникающая из-за попадания точек прямолинейного шабло-

на в разные участки пограничного слоя, существенным образом снижают точность аппроксимации и устойчивость численного метода. Введение криволинейных шаблонов и криволинейных реконструкций на них решает указанные проблемы.

Предложенная реализация криволинейных реконструкций не увеличивает вычислительную стоимость численного алгоритма и не снижает его параллельную эффективность по сравнению с EBR схемой в ее оригинальной формулировке, использующей прямолинейные реконструкции.

На примере моделирования турбулентного течения около аэродинамического профиля NACA0012 показан выигрыш от применения криволинейных реконструкций, заключающийся в обеспечении устойчивого счета и получении более точного и приемлемого с инженерной точки зрения результата даже на достаточно грубых сетках.

Авторы благодарят П.А. Бахвалова за ценные замечания и полезные дискуссии.

СПИСОК ЛИТЕРАТУРЫ

1. *Abalakin I.V., Bakhvalov P.A., Kozubskaya T.K.* Edge-based reconstruction schemes for unstructured tetrahedral meshes // Int. J. Numer. Methods Fluids. 2016. V. 81. № 6. P. 331–356.
2. *Bakhvalov P.A., Kozubskaya T.K.* On Efficient Vertex-Centered Schemes on Hybrid Unstructured Meshes // AIAA Paper 2016–2966. 22nd AIAA/CEAS Aeroacoustics Conference. 2016.
3. *Katz A.J., Work D.* High-order flux correction/finite difference schemes for strand grids // J. Comput. Phys. 2015. V. 282. P. 360–380.
4. *Tong O., Katz A.J., Wissink A.M., Sitaraman J.* High-order methods for three-dimensional strand-cartesian grids // AIAA Paper 2015–0835. 53rd AIAA Aerospace Sciences Meeting. 2015.
5. NASA Langley Research Center. Turbulence Modeling Resource. 2DN00: 2D NACA 0012 Airfoil Validation Case URL: https://turbmodels.larc.nasa.gov/naca0012_val.html. SA Model Results URL: https://turbmodels.larc.nasa.gov/naca0012_val_sa.html.
6. *Spalart P.R., Allmaras S.R.* A One-Equation Turbulence Model for Aerodynamic Flows // AIAA Paper 92–0439. 30th Aerospace Science Meeting. 1992.
7. *Bakhvalov P.A., Kozubskaya T.K.* EBR-WENO scheme for solving gas dynamics problems with discontinuities on unstructured meshes // Comput. Fluids. 2017. V. 157. P. 312–324.
8. *Бахвалов П.А.* О порядке точности реберно-ориентированных схем на сетках специального вида // Препринты ИПМ им. М.В. Келдыша. 2017. № 79. 32 с.
9. *Duben A.P., Kozubskaya T.K.* On Scale-Resolving Simulation of Turbulent Flows Using Higher-Accuracy Quasi-1D Schemes on Unstructured Meshes // Progress in Hybrid RANS-LES Modelling. HRLM 2016. Notes on Numerical Fluid Mechanics and Multidisciplinary Design. V. 137. P. 169–178.
10. *Kallinderis Y., Ward S.* Prismatic grid generation for three-dimensional complex geometries // AIAA J. 1993. V. 31. № 10. P. 1850–1856.
11. *Connell S., Braaten M.* Semi-structured mesh generation for 3D Navier-Stokes calculations // AIAA Paper 92–0439. 12th Computational Fluid Dynamics Conference. 1995.
12. *Khawaja A., Kallinderis Y.* Hybrid grid generation for turbomachinery and aerospace applications // Int. J. Numer. Methods Eng. 2000. V. 49. № 1–2. P. 145–166.
13. *Athanasiadis A.N., Deconinck H.* A folding/unfolding algorithm for the construction of semi-structured layers in hybrid grid generation // Comput. Methods Appl. Mech. Eng. 2005. V. 194 № 48–49. P. 5051–5067.
14. *Gorobets A.* Parallel Algorithm of the NOISEtte Code for CFD and CAA Simulations // Lobachevskii J. Math. 2018. V. 39. № 4. P. 524–532.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 20–31

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 519.853.62

УСКОРЕННЫЙ МЕТААЛГОРИТМ ДЛЯ ЗАДАЧ ВЫПУКЛОЙ ОПТИМИЗАЦИИ[1)]

© 2021 г. А. В. Гасников[1,2], Д. М. Двинских[2,1,3], П. Е. Двуреченский[3,2], Д. И. Камзолов[1,*], В. В. Матюхин[1], Д. А. Пасечнюк[1], Н. К. Тупица[1], А. В. Чернов[1]

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, Московский физико-технический институт (национальный исследовательский университет), Россия

[2] 127051 Москва, Большой Каретный пер., 19, стр. 1, Институт проблем передачи информации им. А.А. Харкевича РАН, Россия

[3] Институт прикладного анализа и стохастики им. К. Вейерштрасса, Берлин, Германия

**e-mail: kamzolov.dmitry@phystech.edu*

Поступила в редакцию 18.04.2020 г.
Переработанный вариант 16.06.2020 г.
Принята к публикации 18.09.2020 г.

Предлагается оболочка, названная “ускоренный метаалгоритм”, которая позволяет единообразно получать ускоренные методы решения задач выпуклой безусловной минимизации в различных постановках на базе неускоренных вариантов. В качестве приложений приводятся квазиоптимальные алгоритмы для минимизации гладких функций с липшицевыми производными произвольного порядка, а также для решения гладких минимаксных задач. Предложенная оболочка является более общей, чем существующие, а также позволяет получать лучшие оценки скорости сходимости и практическую эффективность для ряда постановок задач. Библ. 26. Фиг. 2.

Ключевые слова: выпуклая оптимизация, проксимальный ускоренный метод, тензорные методы, неточный оракул, слайдинг, каталист.

DOI: 10.31857/S0044466921010051

1. ВВЕДЕНИЕ

В последние 15 лет в численных методах гладкой выпуклой оптимизации преобладают так называемые ускоренные методы. Прообразом таких методов является метод тяжелого шарика Б.Т. Поляка и моментный метод Ю.Е. Нестерова (см. [1], [2]). Оказалось, что для многих задач гладкой выпуклой оптимизации оптимальные методы (с точки зрения числа вычислений градиента функции; в общем случае, старших производных) могут быть найдены среди ускоренных методов (см. [1]–[3]). Появилось огромное число работ, в которых предлагаются различные варианты ускоренных методов для разных классов задач (см., например, обзор литературы в [1], [3]). Каждый раз процедура ускорения принимала свою причудливую форму. Естественно, возникло желание как-то унифицировать все это. В 2015 г. это было сделано для широкого класса (рандомизированных) градиентных методов с помощью проксимальной ускоренной оболочки, названной Каталист (см. [4]). (Здесь и далее в качестве названий подходов/алгоритмов иногда будут использоваться англицизмы. Дело в том, что дословный перевод исходно английских выражений на русский язык может только запутывать дело. Отметим также, что под “проксимальной оболочкой” здесь и далее имеется в виду просто проксимальный алгоритм. Слово “оболочка” подразумевает, что в проксимальном алгоритме на каждой итерации имеется своя внутренняя (вспомогательная) задача оптимизации, которую, как правило, нельзя решить аналитически. Ее нужно решать численно. Поэтому внешний проксимальный метод можно по-

[1)]Работа А.В. Гасникова выполнена при финансовой поддержке РФФИ (код проекта 18-31-20005 мол_а_вед в п. 2), работа Д.И. Камзолова выаолнена при финансовой поддержке РФФИ (код проекта 19-31-90170). Аспиранты в п. 3, работа П.Е. Двуреченского выаолнена при финансовой поддержке РФФИ (код проекта 18-29-03071 мк в п. 3). Работа Д.М. Двинских и В.В. Матюхина выполнена при финансовой поддержке Минобрнауки РФ (госзадание № 075-00337-20-03, номер проекта 0714-2020-0005).

нимать как “оболочку” для метода, использующегося для решения внутренней задачи.) С 2013 г. данные результаты стали активно переноситься на тензорные методы (использующие старшие производные) (см. [5]–[8]). В самое последнее время предпринимаются попытки унификации процедур ускорения для седловых задач и задач со структурой (композитных задач) (см. [9]–12]). Во всех этих направлениях по-прежнему использовалось значительное разнообразие ускоренных проксимальных оболочек (см. [1], [4]–[8], [10], [11], [13]– [16]). Метод из данной работы будет во многом базироваться на схеме из [14]. (Строго говоря, это даже не метод (алгоритм), а скорее оболочка (в смысле, определенном выше). В данной статье было выбрано название “ускоренный метаалгоритм”. Первое слово поясняет цель разрабатываемой оболочки – ускорение метода, использующегося в качестве базового (решающего внутреннюю задачу). Однако, в отличие от стандартной (ускоренной) оболочки, в предложенной в данной статье оболочке все же в ряде важных случаев вспомогательная задача решается аналитически и, стало быть, говорить об этой оболочке, как “оболочке”, а не как об обычном алгоритме, не совсем корректно. Поэтому было решено использовать более нейтральное в этом смысле слово – “метаалгоритм”.)

В данной работе показывается, что достаточно изучить всего одну ускоренную проксимальную оболочку, которая позволяет получать все известные нам ускоренные методы для задач гладкой выпуклой безусловной оптимизации. Причем в ряде случаев предложенный ускоренный метаалгоритм позволяет убирать логарифмические зазоры в оценках сложности (по сравнению с нижними оценками), имевшие место в предыдущих подходах.

2. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Рассмотрим следующую задачу (x_* – решение задачи):

$$\min_{x\in\mathbb{R}^d}\{F(x) \coloneqq f(x) + g(x)\}, \tag{1}$$

где f и g – выпуклые функции.

Везде в дальнейшем под $\|\cdot\|$ будем понимать обычную евклидову норму в пространстве $\mathbb{R}^d$,

$$D^k f(x)[h]^k = \sum_{i_1,\ldots,i_d\geq 0:\ \sum_{j=1}^d i_j = k} \frac{\partial^k f(x)}{\partial x_1^{i_1}\ldots\partial x_d^{i_d}} h_1^{i_1}\cdot\ldots\cdot h_d^{i_d},$$

$$\left\|D^k f(x)\right\| = \max_{\|h\|\leq 1}\left\|D^k f(x)[h]^k\right\|.$$

Будем считать, что f имеет липшицевы производные порядка p ($p\in\mathbb{N}$):

$$\left\|D^p f(x) - D^p f(y)\right\| \leq L_{p,f}\|x-y\|. \tag{2}$$

Здесь и далее (см., например, (7)) можно считать, что $x, y\in\mathbb{R}^d$ принадлежат евклидову шару с центром в точке x_* и радиусом $O\left(\left\|x_0 - x_*\right\|\right)$, где x_0 – точка старта (см. [6]).

Введем аппроксимацию рядом Тейлора функции f:

$$\Omega_p(f,x;y) = f(x) + \sum_{k=1}^p \frac{1}{k!} D^k f(x)[y-x]^k, \quad y\in\mathbb{R}^d.$$

Заметим, что из (2) следует (см. [17]), что

$$\left|f(y) - \Omega_p(f,x;y)\right| \leq \frac{L_{p,f}}{(p+1)!}\|y-x\|^{p+1}. \tag{3}$$

Доказательство следующей теоремы см. в Приложении 1 (литературный обзор см. в [11]).

Algorithm 1. Ускоренный Метаалгоритм (УМ) (УМ(x_0, f, g, p, H, k))

1: **Input:** $p\in\mathbb{N}$, $f\colon\mathbb{R}^d\to\mathbb{R}$, $g\colon\mathbb{R}^d\to\mathbb{R}$, $H>0$.

2: $A_0 = 0, y_0 = x_0$.

3: **for** $k = 0$ **to** $k = K-1$

4: Определить пару $\lambda_{k+1} > 0$ и $y_{k+1} \in \mathbb{R}^d$ из условий

$$\frac{1}{2} \le \lambda_{k+1} \frac{H\|y_{k+1} - \tilde{x}_k\|^{p-1}}{p!} \le \frac{p}{p+1},$$

где

$$y_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ \widetilde{\Omega}^k(y) := \Omega_p(f, \tilde{x}_k; y) + g(y) + \frac{H}{(p+1)!}\|y - \tilde{x}_k\|^{p+1} \right\}, \tag{4}$$

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1}A_k}}{2}, \quad A_{k+1} = A_k + a_{k+1},$$

$$\tilde{x}_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} x_k.$$

5: $x_{k+1} := x_k - a_{k+1}\nabla f(y_{k+1}) - a_{k+1}\nabla g(y_{k+1})$.

6: **end for**

7: **return** y_K

Теорема 1. *Пусть y_k – выход алгоритма* 1 *УМ*(x_0, f, g, p, H, k) *после k итераций при $p \ge 1$ и $H \ge (p+1)L_{p,f}$. Тогда*

$$F(y_k) - F(x_*) \le \frac{c_p H R^{p+1}}{k^{\frac{3p+1}{2}}}, \tag{5}$$

где $c_p = 2^{p-1}(p+1)^{\frac{3p+1}{2}}/p!$, $R = \|x_0 - x^*\|$.

Более того, при $p \ge 2$ для достижения точности ε: $F(y_k) - F(x_*) \le \varepsilon$ на каждой итерации УМ вспомогательную задачу (4) придется перерешивать для подбора пары (λ_{k+1}, y_{k+1}) не более чем $O(\ln(\varepsilon^{-1}))$ раз.

Заметим, что приведенная выше теорема будет справедлива и при условии $H \ge 2L_{p,f}$ (независимо от $p \in \mathbb{N}$). Это выводится из (3). Условие $H \ge (p+1)L_{p,f}$ было использовано, поскольку оно гарантирует выпуклость вспомогательной подзадачи (4) (см. [17]). При этом условии и $g \equiv 0$ для $p = 1, 2, 3$ существуют эффективные способы решения вспомогательной задачи (4) (см. [17]). Для $p = 1$ существует явная формула для решения (4), для $p = 2, 3$ сложность (4) такая же (с точностью до логарифмического по ε множителя), как у итерации метода Ньютона (см. [17]).

Отметим, что вспомогательную задачу (4) не обязательно решать точно: достаточно (см. [11], [18]) найти точку $\tilde{y}_{k+1}$, удовлетворяющую условию

$$\left\|\nabla\widetilde{\Omega}^k(\tilde{y}_{k+1})\right\| \le \frac{1}{4p(p+1)}\|\nabla F(\tilde{y}_{k+1})\|. \tag{6}$$

Такая модификация приведет лишь к появлению множителя $12/5$ в правой части (5).

Будем говорить, что функция F является r-равномерно выпуклой ($p+1 \ge r \ge 2$) с константой $\sigma_r > 0$, если

$$F(y) \ge F(x) + \langle \nabla F(x), y - x\rangle + \frac{\sigma_r}{r}\|y - x\|^r, \quad x, y \in \mathbb{R}^d. \tag{7}$$

В этом случае, используя [19]:

$$F(\tilde{y}_{k+1}) - F(x_*) \le \frac{r-1}{r}\left(\frac{1}{\sigma_r}\right)^{\frac{1}{r-1}}\|\nabla F(\tilde{y}_{k+1})\|^{\frac{r}{r-1}},$$

можно завязать критерий (6) на желаемую точностью (по функции) решения исходной задачи ε (см. [11]): $\left\|\nabla\widetilde{\Omega}^{k}(\tilde{y}_{k+1})\right\| = O((\varepsilon^{r-1}\sigma_r)^{1/r})$.

Более того, для $p = 1$ приведенные здесь выкладки можно уточнить, подчеркнув тем самым, что сложность решения вспомогательной задачи может даже не зависеть от ε. Оказывается (см. [16]), что условие

$$\left\|\tilde{y}_{k+1} - y^*_{k+1}\right\| \le \frac{H}{3H + 2L_1^g}\left\|\tilde{x}_k - y^*_{k+1}\right\|, \tag{8}$$

где y^*_{k+1} – точное решение задачи (4), а L_1^g – константа Липшица градиента ∇g, в теоретическом плане гарантирует то же, что и условие (6) при $p = 1$. А именно, теорема 1 останется верной с добавлением в правую часть (5) множителя 12/5.

Отметим, что оценка скорости сходимости (5) с точностью до числового множителя c_p не может быть улучшена для класса выпуклых задач (1) с липшицевой p-й производной и для широкого класса тензорных методов порядка p, описанном в [17]. При дополнительном предположении равномерной выпуклости F оптимальный метод можно построить на базе УМ с помощью процедуры рестартов (см. [11]) (см. алгоритм 2).

Algorithm 2. Рестартованный УМ(x_0, f, g, p, r, σ_r, H, k)

1: **Input:** r-равномерно выпуклая функция $F = f + g : \mathbb{R}^d \to \mathbb{R}$ с константой σ_r и УМ(x_0, f, g, p, H, K).

2: $z_0 = x_0$.

3: **for** $k = 0$ **to** K

4: $R_k = R_0 \cdot 2^{-k}$,

$$N_k = \max\left\{\left\lceil\left(\frac{rc_pH2^r}{\sigma_r}R_k^{p+1-r}\right)^{\frac{2}{3p+1}}\right\rceil, 1\right\}. \tag{9}$$

5: $z_{k+1} := y_{N_k}$, где y_{N_k} – выход УМ(z_k, f, g, p, H, N_k).

6: **end for**

7: **return** z_K

Теорема 2. *Пусть y_k – выход алгоритма* 2 *после k итераций. Тогда если $H \ge (p+1)L_{p,f}$, $\sigma_r > 0$, то общее число вычислений* (4) *для достижения $F(y_k) - F(x_*) \le \varepsilon$ будет*:

$$N = \tilde{O}\left(\left(\frac{HR^{p+1-r}}{\sigma_r}\right)^{\frac{2}{3p+1}}\right),$$

где $\tilde{O}(\,)$ – означает то же самое, что $O(\,)$ с точностью до множителя $\ln(\varepsilon^{-1})$.

Все, что было сказано после теоремы 1, можно отметить и в данном случае.

3. ПРИЛОЖЕНИЯ

3.1. Ускоренные методы композитной оптимизации

Если не думать о сложности решения подзадачи (4), например, считать, что g – какая-то простая функция и (4) решается по явным формулам (как, например, для задачи LASSO), то УМ описывает класс ускоренных методов (1-, 2-, 3-го, ... порядков) композитной оптимизации (см. [1], [2], [6]). При этом функция g не обязана быть гладкой. В общем случае в строчке 5 алгоритма 1 под $\nabla g(y_{k+1})$ следует понимать такой субградиент функции g в точке y_{k+1}, с которым суб-

градиент правой части (4) равен (близок) к нулю (немного переписав метод, от последнего ограничения можно отказаться). Отметим, что при $p = 1$ необходимость в поиске параметра λ_{k+1} исчезает, что делает метод заметно проще.

3.2. Ускоренные проксимальные методы. Каталист

Если считать $p = 1$, а $f \equiv 0$, $H > 0$, то получится ускоренный проксимальный метод. Отличительная особенность такого метода (см. также [16]) от других известных ускоренных проксимальных методов заключается в том, что не требуется очень точно решать вспомогательную задачу. Критерий (8) и сильная (2-равномерная) выпуклость вспомогательной подзадачи (4) указывают на то, что сложность решения (8) может не зависеть от желаемой точности решения исходной задачи ε. Таким образом, не теряется логарифмический множитель при использовании такой проксимальной оболочки для ускорения различных неускоренных процедур. Собственно, последнее направление получило название Каталист (см. [4]). До настоящего момента идея (Каталист) использования ускоренной проксимальной оболочки для ‘обертывания’ неускоренных методов, решающих вспомогательную задачу (4) на каждой итерации (при должном выборе параметра H), являлась наиболее общей идеей разработки ускоренных методов для разных задач. Мы получаем Каталист просто как частный случай УМ. Примеры использования Каталист будут приведены в п. 3.4.

3.3. Разделение оракульных сложностей

Если считать, что для g имеем $L_{p,g} < \infty$ (см. (2)) и на вспомогательную задачу (4) смотреть как на равномерно выпуклую достаточно гладкую задачу (с $f := g$, $g(x) := \Omega_p\left(f, \tilde{x}_k; x\right) + \frac{(p+1)L_{p,f}}{(p+1)!}\|x - \tilde{x}_k\|^{p+1}$), то для решения (4), в свою очередь, можно использовать Рестартованный УМ с $H \simeq (p+1)L_{p,g}$. В случае, когда $L_{p,f} \le L_{p,g}$ удается получить такие оценки сложности (см. [11], [3]) (см. теорему 1):

$$N_f = \tilde{O}\left(\left(\frac{L_{p,f}R^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}}\right) \text{ – число вызовов оракула для функции } f,$$

$$N_g = \tilde{O}\left(\left(\frac{L_{p,g}R^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}}\right) \text{ – число вызовов оракула для функции } g.$$

Вызов оракула подразумевает вычисление (старших) производных до порядка p включительно. Таким образом, число вызовов оракула для каждой из функции f, g является квазиоптимальным, т.е. оптимальным с точностью до логарифмического (от желаемой точности по функции) множителя. Аналогичные оценки можно получить и в r-равномерно ($r \ge 2$) выпуклом случае (см. п. 3.4).

Заметим, что при $p = 1$ внутреннюю задачу (4) не обязательно решать Рестартованным УМ. Можно использовать (ускоренные) покомпонентные и безградиентные методы, методы редукции дисперсии (см. [1], [3], [20]). Причем ускорение можно получить из базовых неускоренных вариантов этих методов с помощью УМ (см. п. 3.2). По сравнению с оболочкой, использованной в [10], УМ дает оценку сложности на логарифмический множитель лучше. Это следует из теоретического анализа и было подтверждено в экспериментах (см. [21]).

3.4. Ускоренные методы для седловых задач

Следуя, например, [9], [12], рассмотрим выпукло-вогнутую седловую задачу

$$\min_{x\in\mathbb{R}^{d_x}}\{F(x) := f(x) + \underbrace{\max_{y\in\mathbb{R}^{d_y}}\{G(x,y) - h(y)\}}_{g(x)=G(x,y^*(x))-h(y^*(x))}\}, \tag{10}$$

где $y^*(x) = \operatorname{argmax}_{y\in\mathbb{R}^{d_y}}\{G(x,y) - h(y)\}$. Будем считать, что ∇f, ∇G, ∇h являются соответственно L_f, L_G, L_h-липшицевыми. Также будем считать, что $f(x) + G(x,y)$ является μ_x-сильно (2-равномерно) выпуклой по x, а $G(x,y) - h(y)$ является μ_y-сильно (2-равномерно) вогнутой по y. Тогда $F(x)$ будет μ_x-сильно выпуклой, а ∇g будет $L_g = (L_G + 2L_G^2/\mu_y)$-липшицевым (см. [9], [12]).

Если считать, что доступен ∇g, то внешнюю задачу (10) можно решать ускоренным слайдингом (например, в варианте УМ с $p = 1$, см. п. 3.3) за $\tilde{O}(\sqrt{L_f/\mu_x})$ вычислений ∇f и $\tilde{O}(\sqrt{L_g/\mu_x})$ вычислений ∇g.

Чтобы приближенно посчитать $\nabla g(x) = \nabla_x G(x, y^*(x))$, надо решить (с достаточной точностью) вспомогательную задачу в (10), т.е. найти с нужной точностью $y^*(x)$. Это, в свою очередь, также можно сделать с помощью слайдинга (УМ с $p = 1$) за $\tilde{O}(\sqrt{L_h/\mu_y})$ вычислений ∇h и $\tilde{O}(\sqrt{L_G/\mu_y})$ вычислений $\nabla_y G$.

Резюмируя написанное, получаем, что исходную задачу (10) можно решить за $\tilde{O}(\sqrt{L_f/\mu_x})$ вычислений ∇f, $\tilde{O}(\sqrt{L_g/\mu_x}) \simeq \tilde{O}\left(\sqrt{L_G^2/(\mu_x\mu_y)}\right)$ вычислений $\nabla_x G$, $\tilde{O}\left(\sqrt{L_G^3/(\mu_x\mu_y^2)}\right)$ вычислений $\nabla_y G$, $\tilde{O}\left(\sqrt{L_h L_G^2/(\mu_x\mu_y^2)}\right)$ вычислений ∇h. Поменяв порядок взятия min и max аналогичным образом, можно прийти к оценкам $\tilde{O}(\sqrt{L_h/\mu_y})$ вычислений ∇h, $\tilde{O}\left(\sqrt{L_G^2/(\mu_x\mu_y)}\right)$ вычислений $\nabla_y G$, $\tilde{O}\left(\sqrt{L_G^3/(\mu_x^2\mu_y)}\right)$ вычислений $\nabla_x G$, $\tilde{O}\left(\sqrt{L_f L_G^2/(\mu_x^2\mu_y)}\right)$ вычислений ∇f.

Оценки, полученные на число вычислений $\nabla_x G$ и ∇f, в последнем случае не являются оптимальными (см. [12]). Чтобы улучшить данные оценки (сделать их оптимальными с точностью до логарифмических множителей (см. [12])), воспользуемся Каталистом (см. п. 3.2) (УМ, с $p = 1$, $H \gg \mu_x$, $f \equiv 0$, $g = F$, где F определяется (10)). Если параметр метода H, то число итераций метода будет $\tilde{O}(\sqrt{H/\mu_x})$ (см. теорему 2). На каждой итерации необходимо будет решать с должной точностью задачу вида (10), в которой $L_f := L_f + H$, $\mu_x := \mu_x + H \simeq H$. Таким образом, для решения внутренней седловой задачи потребуется $\tilde{O}(\sqrt{L_h/\mu_y})$ вычислений ∇h, $\tilde{O}\left(\sqrt{L_G^2/(H\mu_y)}\right)$ вычислений $\nabla_y G$, $\tilde{O}\left(\sqrt{L_G^3/(H^2\mu_y)}\right)$ вычислений $\nabla_x G$, $\tilde{O}\left(\sqrt{(L_f + H)L_G^2/(H^2\mu_y)}\right)$ вычислений ∇f. Считая для наглядности $L_f \geq L_G$, выберем $H = L_G$. Тогда итоговые оценки на число вычислений соответствующих градиентов будут такие: $\tilde{O}(\sqrt{L_h L_G/(\mu_x\mu_y)})$ вычислений ∇h, $\tilde{O}\left(\sqrt{L_G^2/(\mu_x\mu_y)}\right)$ вычислений $\nabla_y G$, $\tilde{O}\left(\sqrt{L_G^2/(\mu_x\mu_y)}\right)$ вычислений $\nabla_x G$, $\tilde{O}\left(\sqrt{L_f L_G/(\mu_x\mu_y)}\right)$ вычислений ∇f.

За счет использования УМ приведенная выше схема улучшает похожую схему рассуждений из [12] на логарифмический (по желаемой точности решения задачи) множитель, и обобщает ее на случай отличных от тождественного нуля функций f и h.

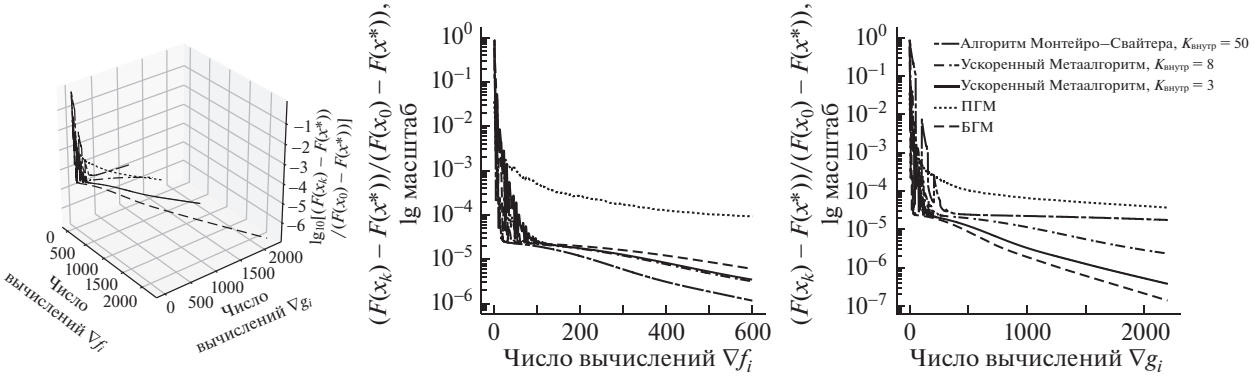
Приведенная здесь схема рассуждений наглядно демонстрирует, как из одной универсальной схемы ускорения удается получить (‘собрать как в конструкторе’) оптимальный метод с точностью до логарифмического (по желаемой точности множителя (при $f \equiv 0$ и $h \equiv 0$ – только при этих условиях известны нижние оценки (см. [12])).

3.5. Сравнение с алгоритмом Монтейро–Свайтера

Следуя [22], рассмотрим задачу оптимизации

$$\min_{x\in\mathbb{R}^n}\{F(x) := \underbrace{\log\left(\sum_{k=1}^{p}\exp(\langle A_k, x\rangle)\right)}_{=f(x)} + \underbrace{\frac{1}{2}\|Gx\|_2^2}_{=g(x)}\},$$

где $n = 500$, $p = 20\,000$, A – разреженная $p\times n$ матрица с коэффициентом разреженности 0.001 (под коэффициентом разреженности в данном случае понимается отношение числа ненулевых



Фиг. 1. Зависимость величины $(F(x_k) - F(x^*))/(F(x_0) - F(x^*))$ (в log масштабе) от числа вычислений компонент градиентов ∇f_i и ∇g_i. Двухмерные проекции.

элементов матрицы к общему числу ее элементов), чьи ненулевые элементы есть независимые одинаково распределенные случайные величины из равномерного распределения $\mathcal{U}(-1,1)$, а матрица G^2 получается из следующего выражения:

$$G^2 = \sum_{i=1}^{n} \lambda_i \tilde{e}_i^{\top} \tilde{e}_i,$$

где $\sum_{i=1}^{n} \lambda_i = 1$ и $[\tilde{e}_i]_j \sim \mathcal{U}(1,2)$ для каждой пары i, j.

Здесь f имеет липшицев градиент с константой Липшица:

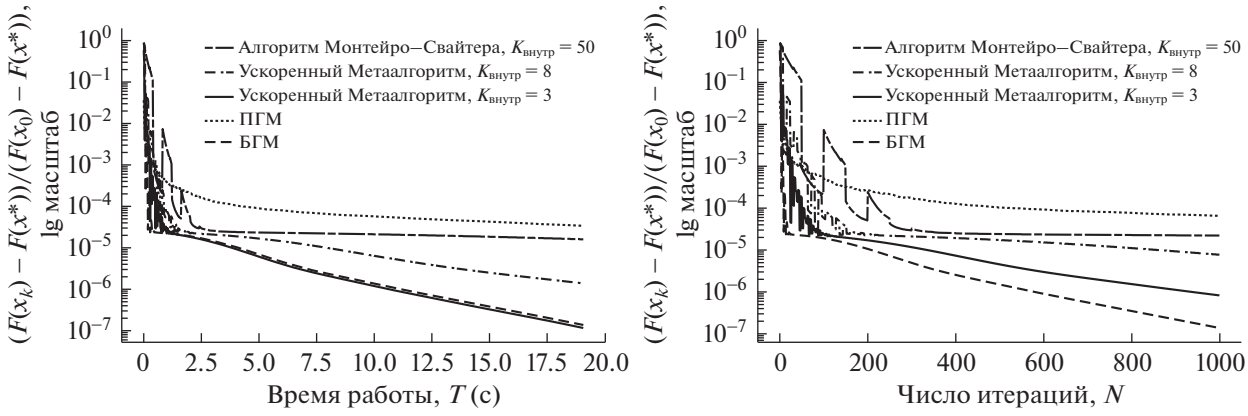
$$L_f = \max_{i=1,\ldots,n} \left\| A^{\langle k \rangle} \right\|_2^2,$$

где через $A^{\langle k \rangle}$ обозначен k-й столбец матрицы A.

На примере данной задачи сравним работу ускоренных методов, полученных с помощью алгоритма Монтейро–Свайтера (см. [7]) ($L = 20L_f$) и с помощью УМ ($H = L_f$) при использовании для решения вспомогательной задачи покомпонентного градиентного метода Нестерова (ПГМ) (см. [23]) ($\beta = 1/2$).

На фиг. 1 покомпонентный метод, ускоренный с помощью алгоритма Монтейро–Свайтера, сравнивается с методом, ускоренным оболочкой УМ с различным числом итераций метода для решений вспомогательной задачи ($K_{\text{внутр}} = k$ соответствует kn итерациям покомпонентного метода), и быстрым градиентным методом (БГМ). Представлен трехмерный график зависимости величины $(F(x_k) - F(x^*))/(F(x_0) - F(x^*))$ (в log масштабе, $F(x^*)$ выбирается равным значению F в точке, полученной после 25000 итераций БГМ) от числа вычислений компонент градиентов ∇f_i и ∇g_i, а также его двухмерные проекции. Так как некоторые методы требуют вычисления полного значения градиента (∇f или ∇g), обращения к оракулам в таком случае учитываются с весом $t_1/t_2 \approx 2.5$, где t_1 – среднее время вычисления полного градиента, t_2 – среднее время вычисления одной компоненты. Как можно видеть из графиков, число обращений к оракулу ∇f_i ускоренного с помощью оболочки УМ метода меньше, чем у быстрого градиентного метода. Кроме того, оболочка УМ позволяет значительно сократить число обращений к оракулу ∇g_i по сравнению с алгоритмом Монтейро–Свайтера.

На фиг. 2 сравнивается работа методов в зависимости от времени работы и числа итераций внутреннего метода. Как видно на фиг. 2, ускоренный с помощью оболочки УМ метод сходится по времени работы с большей скоростью, чем метод, ускоренный с помощью алгоритма Монтейро–Свайтера, а также с большей скоростью, чем быстрый градиентный метод.



Фиг. 2. Зависимость величины $(F(x_k) - F(x^*))/(F(x_0) - F(x^*))$ (в log масштабе) от времени работы и числа итераций внутреннего метода.

4. ВОЗМОЖНЫЕ ОБОБЩЕНИЯ

Приводимые выше конструкции существенным образом базируются на том, что рассматриваются задачи безусловной оптимизации и используется евклидова норма. На данный момент открытым остается вопрос о перенесении приведенных в статье результатов на задачи безусловной оптимизации с заменой евклидовой нормы на дивергенцию Брэгмана (см. [1], [5]). Тем более открытым остается вопрос об использовании других (более общих) моделей в построении мажоранты целевой функции (3) (см. [1]).

В [24] было подмечено (в том числе и в модельной общности), что для получения оптимальных версий ускоренных алгоритмов для задач стохастической оптимизации нужно уметь оценивать, как накапливается малый шум в градиенте в таких методах. Таким образом, строятся ускоренные стохастические градиентные методы на базе ускоренных не стохастических (детерминированных) и конструкции, названной минибатчингом (замена градиента в детерминированном методе его оценкой, построенной на базе стохастических градиентов). Насколько нам известно, для тензорных методов вопрос построения ускоренных методов для задач стохастической оптимизации остается открытым. В частности, не известен ответ на такой вопрос: верно ли, что для задач сильно выпуклой стохастической оптимизации требования к точности аппроксимации старших производных с помощью минибатчинга снижаются по мере роста порядка производных, как это имеет место в невыпуклом случае (см. [25])? Для ответа на этот вопрос для тензорных методов также как и для градиентных $p = 1$ может пригодиться анализ чувствительности исследуемых методов к неточности в вычислении производных. Некоторый задел в этом направлении уже имеется (см. [26]). В частности, при $p = 1$ УМ демонстрирует стандартное для ускоренных методов накопление неточностей в градиенте (см. [24]).

Из статьи может показаться, что для безусловных достаточно гладких задач выпуклой оптимизации предлагаемый в статье подход дает возможность всегда строить “оптимальные” методы. На самом деле это не совсем так. Во-первых, построение оптимальных методов даже на базе одного только УМ может быть совсем не простой задачей, как показывает пример из п. 3.4. Во-вторых, оговорка “с точностью до логарифмических множителей” весьма существенна. В частности, до сих пор остается открытым вопрос о том, устраним ли логарифмический мультипликативный зазор (по желаемой точности решения задачи по функции) между нижними оценками и тем, что дает УМ и другие ускоренные тензорные методы ($p \geq 2$) (см. теорему 1). В-третьих, упомянутые нижние оценки были получены для класса крыловских методов (для тензорных методов чуть иначе (см. [17])), однако, предлагаемая оболочка УМ в некоторых вариантах ее использования, в том числе в проксимальном варианте (Катализ) (см. п. 3.2), выводит из класса допустимых методов, для которого были полученые нижние оценки.

ПРИЛОЖЕНИЕ 1

В этом приложении представлено доказательство теоремы 1, основанное на доказательстве из [14], с учетом добавления композитной функции. Следующая теорема базируется на теореме 2.1 из [14].

Теорема 3. *Пусть $(y_k)_{k\geq 1}$ – это последовательность точек в $\mathbb{R}^d$, и $(\lambda_k)_{k\geq 1}$ – это последовательность в $\mathbb{R}_+$. Определим $(a_k)_{k\geq 1}$ такой, что $\lambda_k A_k = a_k^2$ и $A_k = \sum_{i=1}^k a_i$. Для любого $k \geq 0$ определим*

$$x_k = x_0 - \sum_{i=1}^k a_i(\nabla f(y_i) + g'(y_i)) \quad и \quad \tilde{x}_k := \frac{a_{k+1}}{A_{k+1}} x_k + \frac{A_k}{A_{k+1}} y_k.$$

Также предположим, что если для некоторого $\sigma \in [0,1]$ имеем

$$\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}\nabla f(y_{k+1}))\| \leq \sigma \|y_{k+1} - \tilde{x}_k\|, \tag{11}$$

тогда для любого $x \in \mathbb{R}^d$ верны неравенства

$$F(y_k) - F(x) \leq \frac{2\|x\|^2}{\left(\sum_{i=1}^k \sqrt{\lambda_i}\right)^2},$$

и

$$\sum_{i=1}^k \frac{A_i}{\lambda_i}\|y_i - \tilde{x}_{i-1}\|^2 \leq \frac{\|x^*\|^2}{1-\sigma^2}.$$

Для доказательства этой теоремы мы введем дополнительные леммы, основанные на леммах 2.2–2.5 и 3.1 из [14], леммы 2.6 и 3.3 могут использоваться без изменений.

Лемма 1. *Пусть $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2$, и по индукции определим $\psi_k(x) = \psi_{k-1}(x) + a_k\Omega_1(F, y_k, x)$, тогда $x_k = x_0 - \sum_{i=1}^k a_i(\nabla f(y_i) + g'(y_i))$ – это минимизатор функции ψ_k, и верно $\psi_k(x) \leq A_k F(x) + \frac{1}{2}\|x - x_0\|^2$, где $A_k = \sum_{i=1}^k a_i$.*

Лемма 2. *Пусть z_k такая, что*

$$\psi_k(x_k) - A_k F(z_k) \geq 0.$$

Тогда для любого x имеем

$$F(z_k) \leq F(x) + \frac{\|x - x_0\|^2}{2A_k}.$$

Доказательство. Из леммы 1 можно получить, что

$$A_k F(z_k) \leq \psi_k(x_k) \leq \psi_k(x) \leq A_k F(x) + \frac{1}{2}\|x - x_0\|^2.$$

Лемма 3. *Для любого x верно следующее неравенство*:

$$\psi_{k+1}(x) - A_{k+1}F(y_{k+1}) - (\psi_k(x_k) - A_k F(z_k)) \geq$$
$$\geq A_{k+1}(\nabla f(y_{k+1}) + g'(y_{k+1}))\left(\frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}z_k - y_{k+1}\right) + \frac{1}{2}\|x - x_k\|^2.$$

Доказательство. Во-первых, простыми вычислениями получим

$$\psi_k(x) = \psi_k(x_k) + \frac{1}{2}\|x - x_k\|^2,$$

и

$$\psi_{k+1}(x) = \psi_k(x_k) + \frac{1}{2}\|x - x_k\|^2 + a_{k+1}\Omega_1(f, y_{k+1}, x),$$

таким образом имеем

$$\psi_{k+1}(x) - \psi_k(x_k) = a_{k+1}\Omega_1(F, y_{k+1}, x) + \frac{1}{2}\|x - x_k\|^2. \tag{12}$$

Теперь мы хотим, чтобы $A_{k+1}F(z_{k+1}) - A_kF(z_k)$ было нижней оценкой неравенства (12), когда вычисляем $x = x_{k+1}$. Используя $\Omega_1(F, y_{k+1}, z_k) \le f(z_k)$, мы получаем

$$a_{k+1}\Omega_1(F, y_{k+1}, x) = A_{k+1}\Omega_1(F, y_{k+1}, x) - A_k\Omega_1(F, y_{k+1}, x) = A_{k+1}\Omega_1(F, y_{k+1}, x) - A_k\nabla F(y_{k+1})(x - z_k) -$$
$$- A_k\Omega_1(F, y_{k+1}, z_k) = A_{k+1}\Omega_1\left(F, y_{k+1}, x - \frac{A_k}{A_{k+1}}(x - z_k)\right) - A_k\Omega_1(F, y_{k+1}, z_k) \ge A_{k+1}F(y_{k+1}) - A_kF(z_k) +$$
$$+ A_{k+1}(\nabla f(y_{k+1}) + g'(y_{k+1}))\left(\frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}z_k - y_{k+1}\right),$$

что завершает доказательство.

Лемма 4. *Пусть* $\lambda_{k+1} := \dfrac{a_{k+1}^2}{A_{k+1}}$ *и* $\tilde{x}_k := \dfrac{a_{k+1}}{A_{k+1}}x_k + \dfrac{A_k}{A_{k+1}}y_k$, *тогда имеем*

$$\psi_{k+1}(x_{k+1}) - A_{k+1}F(y_{k+1}) - (\psi_k(x_k) - A_kF(y_k)) \ge$$
$$\ge \frac{A_{k+1}}{2\lambda_{k+1}}(\|y_{k+1} - \tilde{x}_k\|^2 - \|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}(\nabla f(y_{k+1})) + g'(y_{k+1}))\|^2).$$

А применив дополнительно неравенство (11), *получим*

$$\psi_k(x_k) - A_kF(y_k) \ge \frac{1 - \sigma^2}{2}\sum_{i=1}^{k}\frac{A_i}{\lambda_i}\|y_i - \tilde{x}_{i-1}\|^2.$$

Доказательство. Используем лемму 3 при $z_k = y_k$ и $x = x_{k+1}$ и получаем, что (при $\tilde{x} := \dfrac{a_{k+1}}{A_{k+1}}x + \dfrac{A_k}{A_{k+1}}y_k$)

$$(\nabla f(y_{k+1}) + g'(y_{k+1}))\left(\frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}y_k - y_{k+1}\right) + \frac{1}{2A_{k+1}}\|x - x_k\|^2 = (\nabla f(y_{k+1}) + g'(y_{k+1}))(\tilde{x} - y_{k+1}) +$$
$$+ \frac{1}{2A_{k+1}}\left\|\frac{A_{k+1}}{a_{k+1}}\left(\tilde{x} - \frac{A_k}{A_{k+1}}y_k\right) - x_k\right\|^2 = (\nabla f(y_{k+1}) + g'(y_{k+1}))(\tilde{x} - y_{k+1}) + \frac{A_{k+1}}{2a_{k+1}^2}\left\|\tilde{x} - \left(\frac{a_{k+1}}{A_k}x_k + \frac{A_k}{A_{k+1}}y_k\right)\right\|^2.$$

Откуда следует неравенство

$$\psi_{k+1}(x_{k+1}) - A_{k+1}F(y_{k+1}) - (\psi_k(x_k) - A_kF(y_k)) \ge$$
$$\ge A_{k+1}\min_{x\in\mathbb{R}^d}\left\{(\nabla f(y_{k+1}) + g'(y_{k+1}))(x - y_{k+1}) + \frac{1}{2\lambda_{k+1}}\|x - \tilde{x}_k\|^2\right\}.$$

Значение минимума можно легко посчитать.

Для первого выражения теоремы 3 достаточно объединить лемму 4 с леммой 2 и леммой 2.5 из [14]. Второе выражение в теореме 3 следует из леммы 4 и леммы 1.

Следующая лемма доказывает, что минимизация ряда Тэйлора порядка p для (4) может быть представлена как неявный градиентный шаг для некоторого большого размера шага.

Лемма 5. *Неравенство* (11) *верно при* $\sigma = 1/2$ *для* (4), *из этого следует, что*

$$\frac{1}{2} \le \lambda_{k+1}\frac{L_p\|y_{k+1} - \tilde{x}_k\|^{p-1}}{(p-1)!} \le \frac{p}{p+1}. \tag{13}$$

Доказательство. Из условия оптимальности имеем

$$\nabla_y f_p(y_{k+1}, \tilde{x}_k) + \frac{L_p(p+1)}{p!}(y_{k+1} - \tilde{x}_k)\|y_{k+1} - \tilde{x}_k\|^{p-1} + g'(y_{k+1}) = 0. \tag{14}$$

Откуда следует, что

$$y_{k+1} - (\tilde{x}_k - \lambda_{k+1}(\nabla f(y_{k+1}) + g'(y_{k+1}))) = \lambda_{k+1}(\nabla f(y_{k+1}) + g'(y_{k+1})) - \\ - \frac{p!}{L_p(p+1)\|y_{k+1} - \tilde{x}_k\|^{p-1}}(\nabla_y f_p(y_{k+1}, \tilde{x}_k) + g'(y_{k+1})).$$

Используя ряд Тэйлора, для градиента функции получаем

$$\|\nabla f(y) - \nabla_y f_p(y, x)\| \le \frac{L_p}{p!}\|y - x\|^p,$$

таким образом верно

$$\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}(\nabla f(y_{k+1}) + g'(y_{k+1})))\| \le \lambda_{k+1}\frac{L_p}{p!}\|y_{k+1} - \tilde{x}_k\|^p + \\ + \left|\lambda_{k+1} - \frac{p!}{L_p(p+1)\|y_{k+1} - \tilde{x}_k\|^{p-1}}\right|\|\nabla_y f_p(y_{k+1}, \tilde{x}_k) + g'(y_{k+1})\| \le \\ \le \|y_{k+1} - \tilde{x}_k\|\left(\lambda_{k+1}\frac{L_p}{p!}\|y_{k+1} - \tilde{x}_k\|^{p-1} + \left|\lambda_{k+1}\frac{L_p(p+1)\|y_{k+1} - \tilde{x}_k\|^{p-1}}{p!} - 1\right|\right) = \\ = \|y_{k+1} - \tilde{x}_k\|\left(\frac{\eta}{p} + \left|\eta\frac{p+1}{p} - 1\right|\right),$$

где мы используем (14) во втором неравенстве, и предполагаем $\eta := \lambda_{k+1}\dfrac{L_p\|y_{k+1} - \tilde{x}_k\|^{p-1}}{(p-1)!}$ в последнем равенстве. Итоговый результат получаем из предположения, что $1/2 \le \eta \le p/(p+1)$ в (13).

В заключение, если мы заменим $\|x^*\|$ на $\|x_0 - x^*\|$ в лемме 3.3 и используем лемму 3.4 из [14], то получим доказательство теоремы 1.

ПРИЛОЖЕНИЕ 2

Докажем теорему 2.

Доказательство. Так как функция F является r-равномерно выпуклой, то мы получаем

$$R_{k+1} = \|z_{k+1} - x_*\| \le \left(\frac{r(F(z_{k+1}) - F(x_*))}{\sigma_r}\right)^{1/r} \overset{(5)}{\le} \left(\frac{r\left(\dfrac{c_p L_p R_k^{p+1}}{N_k^{\frac{3p+1}{2}}}\right)}{\sigma_r}\right)^{1/r} = \left(\frac{rc_p L_p R_k^{p+1}}{\sigma_r N_k^{\frac{3p+1}{2}}}\right)^{1/r} \overset{(9)}{\le} \left(\frac{R_k^{p+1}}{2^r R_k^{p+1-r}}\right)^{1/r} = \frac{R_k}{2}.$$

Теперь вычислим общее число шагов метода 1:

$$\sum_{k=0}^{K} N_k \le \sum_{k=0}^{K}\left(\frac{rc_p L_p 2^r}{\sigma_r}R_k^{p+1-r}\right)^{\frac{2}{3p+1}} + K = \sum_{k=0}^{K}\left(\frac{rc_p L_p 2^r}{\sigma_r}(R_0 2^{-k})^{p+1-r}\right)^{\frac{2}{3p+1}} + K = \\ = \left(\frac{rc_p L_p 2^r R_0^{p+1-r}}{\sigma_r}\right)^{\frac{2}{3p+1}}\sum_{k=0}^{K} 2^{\frac{-2(p+1-r)k}{3p+1}} + K.$$

СПИСОК ЛИТЕРАТУРЫ

1. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. М.: МФТИ, 2018.
2. *Nesterov Yu.* Lectures on convex optimization. V. 137. Berlin, Germany: Springer, 2018.

3. *Lan G.* Lectures on optimization. Methods for Machine Learning // https://pwp.gatech.edu/guanghui-lan/publications/

4. *Lin H., Mairal J., Harchaoui Z.* Catalyst acceleration for first-order convex optimization: from theory to practice // J. Machine Learning Res. 2017. V. 18. No. 1. P. 7854–7907.

5. *Doikov N., Nesterov Yu.* Contracting proximal methods for smooth convex optimization // arXiv:1912.07972.

6. *Gasnikov A., Dvurechensky P., Gorbunov E., Vorontsova E., Selikhanovych D., Uribe C.A., Jiang B., Wang H., Zhang S., Bubeck S., Jiang Q.* Near Optimal Methods for Minimizing Convex Functions with Lipschitz p-th Derivatives // Proceed. Thirty-Second Conf. Learning Theory. 2019. P. 1392–1393.

7. *Monteiro R.D.C., Svaiter B.F.* An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods // SIAM J. Optimizat. 2013. V. 23. № 2. P. 1092–1125.

8. *Nesterov Yu.* Inexact Accelerated High-Order Proximal-Point Methods // CORE Discussion paper 2020/8.

9. *Alkousa M., Dvinskikh D., Stonyakin F., Gasnikov A.* Accelerated methods for composite non-bilinear saddle point problem // arXiv:1906.03620.

10. *Ivanova A., Gasnikov A., Dvurechensky P., Dvinskikh D., Tyurin A., Vorontsova E., Pasechnyuk D.* Oracle Complexity Separation in Convex Optimization // arXiv:2002.02706

11. *Kamzolov D., Gasnikov A., Dvurechensky P.* On the optimal combination of tensor optimization methods // arXiv:2002.01004

12. *Lin T., Jin C., Jordan M.* Near-optimal algorithms for minimax optimization // arXiv:2002.02417.

13. *Gasnikov A., Dvurechensky P., Gorbunov E., Vorontsova E., Selikhanovych D., Uribe C.A.* Optimal Tensor Methods in Smooth Convex and Uniformly ConvexOptimization // Proc. Thirty-Second Conf. Learning Theory. 2019. P. 1374–1391.

14. *Bubeck S., Jiang Q., Lee Y.T., Li Y., Sidford A.* Near-optimal method for highly smooth convex optimization // Proc. Thirty-Second Conf. Learning Theory. 2019. P. 492–507.

15. *Jiang B., Wang H., Zhang S.* An optimal high-order tensor method for convex optimization // Proc. Thirty-Second Conf. Learning Theory. 2019. P. 1799–1801.

16. *Ivanova A., Grishchenko D., Gasnikov A., Shulgin E.* Adaptive Catalyst for smooth convex optimization // arXiv:1911.11271

17. *Nesterov Yu.* Implementable tensor methods in unconstrained convex optimization // Math. Program. 2019. P. 1–27.

18. *Kamzolov D., Gasnikov A.* Near-Optimal Hyperfast Second-Order Method for convex optimization and its Sliding // arXiv:2002.09050

19. *Grapiglia G.N., Nesterov Yu.* On inexact solution of auxiliary problems in tensor methods for convex optimization // arXiv:1907.13023

20. *Dvurechensky P., Gasnikov A., Tiurin A.* Randomized Similar Triangles Method: A unifying framework for accelerated randomized optimization methods (Coordinate Descent, Directional Search, Derivative-Free Method) // arXiv:1707.08486

21. Ссылка: исходный код экспериментов на GitHub https://github.com/dmivilensky/composite-accelerated-method

22. *Spokoiny V., Panov M.* Accuracy of Gaussian approximation in nonparametric Bernstein–von Mises Theorem // arXiv preprint arXiv:1910.06028. 2019.

23. *Nesterov Yu., Stich S.U.* Efficiency of the accelerated coordinate descent method on structured optimization problems // SIAM Journal on Optimization. 2017. Т. 27. №. 1. С. 110–123.

24. *Dvinskikh D., Tyurin A., Gasnikov A., Omelchenko S.* Accelerated and nonaccelerated stochastic gradient descent with model conception // arXiv:2001.03443

25. *Lucchi A., Kohler J.* A Stochastic Tensor Method for Non-convex Optimization // arXiv:1911.10367

26. *Baes M.* Estimate sequence methods: extensions and approximations // Inst.r Operat. Res. ETH, Zürich, Switzerland, 2009.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 32–46

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 519.615

ОБ ОПТИМАЛЬНОМ ВЫБОРЕ ПАРАМЕТРОВ В ДВУХТОЧЕЧНЫХ ИТЕРАЦИОННЫХ МЕТОДАХ РЕШЕНИЯ НЕЛИНЕЙНЫХ УРАВНЕНИЙ[1)]

© 2021 г. Т. Жанлав[1,*], Х. Отгондорж[1,2,]**

[1] *13330 Улан-батор, Институт математики и информационных технологий, Монгольская Академия Наук, Монголия*

[2] *14191 Улан-батор, Факультет Прикладных Наук, Монгольский Государственный Университет Науки и Технологии, Монголия*

**e-mail: tzhanlav@yahoo.com*

***e-mail: otgondorj@gmail.com*

Поступила в редакцию 05.11.2019 г.
Переработанный вариант 07.07.2020 г.
Принята к публикации 18.09.2020 г.

Разрабатывается новый оптимальный двухпараметрический класс итерационных методов без производных с применением к итерациям типа Хансена–Патрика. Посредством самоускоряющихся параметров мы также получаем новые более высокого порядка методы с памятью. Впервые мы находим точные аналитические формулы для оптимального значения параметров. Увеличение порядка сходимости с 4 до 7 достигается без каких-либо дополнительных вычислений. Таким образом, предлагаемые методы с памятью обладают очень высокой вычислительной эффективностью. Численные примеры и сравнения с некоторыми существующими методами включены для подтверждения теоретических результатов и высокой вычислительной эффективности. Библ. 14. Табл. 6.

Ключевые слова: нелинейные уравнения, двухточечные итерации, методы с памятью, оптимальные методы.

DOI: 10.31857/S0044466920120182

1. ВВЕДЕНИЕ

В численном анализе и инженерных приложениях часто требуется решить нелинейное уравнение вида $f(x) = 0$, где $f : D \subset R \to R$ – скалярная функция, определенная на открытом интервале D. Основными методами решения такого уравнения являются метод Ньютона, заданный (см. [1] и так далее) $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$, $n \geq 0$, и метод Стеффенсена [13], заданный

$$x_{n+1} = x_n - \frac{f(x_n)^2}{f(x_n + f(x_n)) - f(x_n)}, \quad n \geq 0.$$

В последние годы было предложено много итерационных методов более высокого порядка [1]–[6], в которых была введена концепция увеличения порядка сходимости. Достоинство таких методов заключается в том, что они быстро сходятся к требуемым решениям. Однако с увеличением порядка итеративного метода увеличивается количество функциональных вычислений на каждом шаге. Недавно исследователи предложили несколько двухпараметрических простых двухшаговых методов с памятью и без памяти [2], [8], [13], [14]. Авторы этих работ использовали символьные вычисления для получения порядка сходимости и уравнения ошибки. Такая процедура существенно облегчает громоздкие выкладки. Так, в полученном уравнении ошибки присутствуют параметры итерации. Удачные выборы этих параметров позволяют не только повышать порядок сходимости, но и построить новые итерационные методы с памятью. Основная

[1)]Работа выполнена при частичной финансовой поддержке фонда науки и технологии Монголии в рамках гранта SST_18 /2018.

цель данной работы – найти оптимальный выбор параметров τ_n и γ, λ в двухточечных итерационных методах. Мы получили аналитические формулы для γ, λ без использования метода символьных вычислений.

В разд. 2 мы получаем оптимальные двухточечные итерации Хансена–Патрика без производных. В разд. 3 мы предлагаем семейство двух параметрических оптимальных итераций и доказываем теорему о локальной сходимости. В разд. 4 мы предлагаем новые двухточечные итерации как с памятью, так и без памяти. В разд. 5 мы представляем результаты численных экспериментов, которые подтверждают теоретический вывод о порядке сходимости и сравнение с другими известными методами того же порядка сходимости.

2. ОПТИМАЛЬНЫЕ ДВУХТОЧЕЧНЫЕ ИТЕРАЦИИ

Рассмотрим двухточечные итерации вида

$$y_n = x_n - \frac{f(x_n)}{f'(x_n)}, \quad x_{n+1} = x_n - \tau_n \frac{f(x_n)}{f'(x_n)}, \tag{2.1}$$

где τ_n– параметр итерации. Известно, что оптимальный выбор параметра позволяет расширить область сходимости и увеличить скорость сходимости итераций (2.1). Достаточным условием сходимости четвертого порядка [3] является

$$\tau_n = 1 + \theta_n + 2\theta_n^2 + O(f(x_n)^3), \tag{2.2}$$

где

$$\theta_n = \frac{f(y_n)}{f(x_n)}. \tag{2.3}$$

Условие (2.2) часто используется не только для проверки порядка сходимости итераций (2.1), но также для получения новых оптимальных методов. Для ясности мы напомним некоторые определения, нужные в дальнейшем. Многоточечные методы с порядком сходимости 2^{n-1} называем оптимальным [10], где n – количество вычислений функции на каждом шаге итерации. Еще одним из важных характеристик итерационных методов является их индекс эффективности $EI = \rho^{1/n}$, где ρ – порядок сходимости. В качестве примера рассмотрим известное кубически сходящееся семейство итераций Лагерра (или итерации Хансена–Патрика) (2.1) с параметром τ_n, заданным в виде

$$\tau_n = \frac{\alpha + 1}{\alpha + \operatorname{sign}(\alpha)\sqrt{1 - (\alpha + 1)\dfrac{f''(x_n)f(x_n)}{f'(x_n)^2}}}, \quad \alpha \neq -1. \tag{2.4}$$

Используя разложение Тейлора функции $f(y_n)$ в точке x_n, легко показать, что

$$\theta_n = \frac{f''(x_n)f(x_n)}{2f'(x_n)^2} + O(f(x_n)^2). \tag{2.5}$$

Тогда (2.4) приводит к

$$\tau_n = \frac{\alpha + 1}{\alpha \pm \sqrt{1 - 2(\alpha + 1)\theta_n + O(f(x_n)^2)}}, \quad \alpha \neq -1. \tag{2.6}$$

Пренебрегая малым членом $O(f(x_n)^2)$ в (2.6), получаем

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}}, \quad \alpha \neq -1. \tag{2.7}$$

Кансел и др. в [2] рассматривали итерации (2.1) с параметром, определенным (2.7). Используя известное отношение

$$(1 - x)^\alpha = 1 - \alpha x + \frac{\alpha(\alpha - 1)}{2}x^2 - \frac{\alpha(\alpha - 1)(\alpha - 2)}{6}x^3 + \ldots, \quad |x| < 1, \tag{2.8}$$

легко показать, что (2.7) имеет следующую асимптотику:

$$\tau_n = 1 + \theta_n + \frac{\alpha + 3}{2}\theta_n^2 + O(\theta_n^3). \tag{2.9}$$

Сравнение (2.9) с (2.2) показывает, что итерации (2.1) с параметром τ_n, заданным (2.7), не являются оптимальными. Так как здесь требуются три вычисления функции $f(x_n)$, $f(y_n)$ и $f'(x_n)$ на каждом шаге итерации и порядок сходимости равен 3. Исключение составляет только $\alpha = 1$, т.е.

$$\tau_n = \frac{2}{1 + \sqrt{1 - 4\theta_n}}, \tag{2.10}$$

который удовлетворяет условию (2.2). Следует отметить, что с помощью ускоряющей процедуры для τ_n в [4] была получена итерация четвертого порядка (2.1) с τ_n, заданным (2.10)). По этой причине значение, определенное (2.10), называется оптимальным. Отметим также, что в [6] была сделана попытка поиска оптимального параметра α семейства Лагерра с точки зрения сходимости. Как правило, итерация (2.1) с параметром τ_n, заданным (2.7), имеет только третий порядок сходимости. Используя условие (2.2), можно найти оптимальную модификацию семейства Хансена–Патрика четвертого порядка. С этой целью мы ищем τ_n в виде

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} H(\theta_n), \quad \alpha \neq -1, \tag{2.11}$$

где H – вещественная функция, удовлетворяющая условиям

$$H(0) = 1, \quad H'(0) = a, \quad H''(0) = 2b. \tag{2.12}$$

Находим a и b в (2.12) такими, что (2.11) удовлетворяет условию (2.2). Используя разложение Тейлора функций $H(\theta_n)$ и (2.9) в (2.11), получаем

$$\tau_n = 1 + (a + 1)\theta_n + \left(a + b + \frac{\alpha + 3}{2}\right)\theta_n^2 + \ldots. \tag{2.13}$$

Сравнение (2.13) с (2.2) дает

$$a = 0, \quad b = \frac{1 - \alpha}{2}.$$

Таким образом, получаем оптимальный вариант семейства Хансена–Патрика (2.1) с параметром, определяемым в виде

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}}\left(1 + \frac{1 - \alpha}{2}\theta_n^2\right), \quad \alpha \neq -1. \tag{2.14}$$

Когда $\alpha = 1$, (2.14) приводит к (2.10). Таким образом, показываем, что можно перейти от любых итераций третьего порядка (2.1) к оптимальным двухточечным итерациям, используя условие (2.2). Аналогично, легко показать, что итерации Хансена–Патрика имеют оптимальный четвертый порядок сходимости, если

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} + \frac{1 - \alpha}{2}\theta_n^2. \tag{2.15}$$

Обратим внимание, что в [1] авторы предложили новую оптимальную модификацию четвертого порядка семейства Хансена–Патрика (2.1) с параметром, определяемым формулой

$$\tau_n = \frac{\alpha + 1}{\alpha + \sqrt{\dfrac{1 - (\alpha + 3)\theta_n - (\alpha^2 - 1)\theta_n^2}{1 + (\alpha - 1)\theta_n}}}, \quad \alpha \neq -1. \tag{2.16}$$

Несмотря на то что они оптимальны с индексом эффективности $EI = 4^{1/3} \approx 1.587$, итерации (2.1) требуют вычисление производной первого порядка на каждом шаге итерации и поэтому не могут применяться к уравнениям с негладкими функциями. В [5] было дано правило для перехода ите-

раций (2.1) в их оптимальный вариант без производных и наоборот. Согласно этому правилу легко получить вариант (2.1) без производной с помощью (2.16). Это имеет форму

$$\begin{gathered} y_n = x_n - \frac{f(x_n)}{\phi_n}, \\ x_{n+1} = x_n - \tau_n \frac{f(x_n)}{\phi_n} \quad \text{или} \quad \left(x_{n+1} = y_n - \overline{\tau}_n \frac{f(y_n)}{\phi_n} \right), \end{gathered} \tag{2.17}$$

где

$$w_n = x_n + \gamma f(x_n), \quad \phi_n = f[x_n, w_n] = \frac{f(w_n) - f(x_n)}{w_n - x_n},$$

а также

$$\tau_n = 1 + \overline{\tau}_n \theta_n, \tag{2.18}$$

$$\overline{\tau}_n = \frac{1}{\theta_n} \left(\frac{\alpha + 1}{\alpha + \sqrt{\dfrac{1 - (\alpha + 3)\theta_n - (\alpha^2 - 1)\theta_n^2}{1 + (\alpha - 1)\theta_n}}} - 1 \right) + (\hat{d}_n - 2)\theta_n, \tag{2.19}$$

$$\hat{d}_n = \frac{2 + \gamma\phi_n}{1 + \gamma\phi_n}. \tag{2.20}$$

Аналогичным образом, используя достаточное условие сходимости четвертого порядка [7]

$$\overline{\tau}_n = 1 + \hat{d}_n \theta_n + O(f(x_n)^2), \tag{2.21}$$

для (2.17) можно легко построить вариант итераций без производных (2.1), (2.14). Его можно записать как (2.17) с параметром

$$\overline{\tau}_n = \frac{1}{\theta_n} \left(\frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha + 1)\theta_n}} \left(1 + \left(\hat{d}_n - 2 - \frac{\alpha - 1}{2} \right) \theta_n^2 \right) - 1 \right), \quad \alpha \neq -1. \tag{2.22}$$

Таким образом, имеем семейства итераций типа Хансена–Патрика без производных (2.17) с параметром, заданным двумя вариантами (2.19) и (2.22).

Замечание 1. В общем, мы можем рассмотреть следующую весовую функцию:

$$W(\theta_n, \alpha, m) = \frac{\alpha + 1}{\alpha + \sqrt[m]{1 - m(\alpha + 1)\theta_n}} = 1 + \theta_n + \left(1 - \frac{1 - m}{2}(\alpha + 1) \right) \theta_n^2 + \ldots \tag{2.23}$$

в итерации (2.1). Мы называем $W(\theta_n, \alpha, m)$ обобщенной весовой функцией Хансена–Патрика. Функция $W(\theta_n, \alpha, 2)$ приводит к (2.7). Легко показать, что итерационные методы (2.1) имеют оптимальный четвертый порядок сходимости, когда τ_n удовлетворяет одному из следующих условий:

$$\tau_n = W(\theta_n, \alpha, m) + \left(1 + \frac{1 - m}{2}(\alpha + 1) \right) \theta_n^2, \quad \alpha \neq -1, \tag{2.24}$$

и

$$\tau_n = W(\theta_n, \alpha, m) H(\theta_n), \tag{2.25}$$

где H – вещественная функция, удовлетворяющая следующим условиям:

$$H(0) = 1, \quad H'(0) = 0, \quad H''(0) = 2\left(1 + \frac{1 - m}{2}(\alpha + 1) \right). \tag{2.26}$$

Что касается H, можно выбрать, например, следующие функции:

$$H_1 = 1 + \left(1 + \frac{1-m}{2}(\alpha+1)\right)\theta_n^2,$$
$$H_2 = \frac{1}{1 - \left(1 + \frac{1-m}{2}(\alpha+1)\right)\theta_n^2},$$
$$H_3 = \sqrt{1 + (2 + (1-m)(\alpha+1))\theta_n^2}.$$

3. СЕМЕЙСТВО ДВУХПАРАМЕТРИЧЕСКИХ ОПТИМАЛЬНЫХ ИТЕРАЦИЙ, НЕ СОДЕРЖАЩИХ ПРОИЗВОДНЫЕ

Теперь рассмотрим двухпараметрические итерации без производных

$$y_n = x_n - \frac{f(x_n)}{\phi_n + \lambda f(w_n)}, \quad \lambda \in R,$$
$$x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n + \lambda f(w_n)}, \tag{3.1}$$

где $w_n = x_n + \gamma f(x_n)$, $\gamma \in R$, $\phi_n = f[x_n, w_n] = \dfrac{f(w_n) - f(x_n)}{\gamma f_n}$. Наша цель – найти $\bar{\tau}_n$ в (3.1) так, чтобы итерации (3.1) имели оптимальную сходимость четвертого порядка. Для этого сначала используем разложение Тейлора функции $f(w_n) = f(x_n)(1 + \gamma\phi_n)$ в точке x_n. Тогда мы получим

$$\phi_n = f'(x_n)\left(1 + \frac{a_n}{2} f'(x_n)\gamma\right) + O(f_n^2), \quad (f_n) = f(x_n), \tag{3.2}$$

где

$$a_n = \frac{f''(x_n) f(x_n)}{f'(x_n)^2}. \tag{3.3}$$

Пусть $\eta_n = \dfrac{f'(x_n)}{\phi_n}$. Тогда, используя (3.2), получаем

$$\eta_n = \frac{1}{1 + \frac{a_n}{2} f'(x_n)\gamma + O(f_n^2)} = 1 - \frac{a_n}{2} f'(x_n)\gamma + O(f_n^2). \tag{3.4}$$

Разложение Тейлора $f(y_n)$ в точке x_n дает

$$f(y_n) = f(x_n)\left(1 - \eta_n\left(1 - \frac{\lambda f(w_n)}{\phi_n}\right)\right) + O(f_n^2) = f(x_n)\left(1 - \left(1 - \frac{a_n}{2} f'(x_n)\gamma\right)\left(1 - \frac{\lambda f(w_n)}{\phi_n}\right)\right) + O(f_n^2). \tag{3.5}$$

Согласно (3.3), имеем $f(y_n) = O(f(x_n)^2)$. Аналогично, из второго шага в (3.1) получаем

$$f(x_{n+1}) = \left(1 - \bar{\tau}_n \frac{f'(y_n)}{\phi_n + \lambda f(w_n)}\right) f(y_n) + O(f(y_n)^2). \tag{3.6}$$

Из (3.5) и (3.6) ясно, что получим

$$f(x_{n+1}) = O(f_n^4), \tag{3.7}$$

если выбирать $\bar{\tau}_n$ так, чтобы

$$1 - \bar{\tau}_n \frac{f'(y_n)}{\phi_n + \lambda f(w_n)} = O(f_n^2)$$

или

$$\bar{\tau}_n = \frac{\phi_n + \lambda f(w_n)}{f'(y_n)} + O(f_n^2). \tag{3.8}$$

Разложение Тейлора $f'(y_n)$ в точке x_n дает

$$f'(y_n) = f'(x_n)\left(1 - \frac{f_n'' f_n}{f_n'\phi_n\left(1 + \frac{\lambda f(w_n)}{\phi_n}\right)}\right) + O(f_n^2), \quad f_n' = f'(x_n).$$

Используя (3.3) и (3.4) в последнем соотношении, мы получаем

$$f'(y_n) = f_n'(1 - a_n) + O(f_n^2). \tag{3.9}$$

Подставляя (3.2) и (3.9) в (3.8), получаем

$$\begin{aligned}\bar{\tau}_n &= \frac{1 + \frac{a_n}{2} f_n'\gamma + \lambda\frac{(1+\gamma\phi_n)f_n}{f_n'} + O(f_n^2)}{1 - a_n + O(f_n^2)} = \left(1 + \frac{a_n}{2} f_n'\gamma + \frac{\lambda(1+\gamma\phi_n)f_n}{f_n'}\right)(1 + a_n) + O(f_n^2) = \\ &= 1 + \left(1 + \frac{f_n'\gamma}{2}\right)a_n + \frac{\lambda(1+\gamma\phi_n)f_n}{f_n'} + O(f_n^2).\end{aligned} \tag{3.10}$$

Согласно (3.3) и (3.4) имеем $f_n' = \phi_n + O(f_n)$. Следовательно, можно заменить f_n' через φ_n в (3.10) без потери точности. В результате имеем

$$\bar{\tau}_n = 1 + \frac{2 + \gamma\phi_n}{2} a_n + \frac{\lambda(1+\gamma\phi_n)f_n}{\phi_n} + O(f_n^2). \tag{3.11}$$

Далее, используя разложение Тейлора $f(y_n)$ и (3.4), легко получить

$$\begin{aligned}\theta_n &= \frac{a_n}{2}(1 + \gamma f_n') + \frac{\lambda(1+\gamma\phi_n)}{\phi_n} f(x_n) + O(f_n^2) = \frac{a_n}{2}(1 + \gamma\phi_n) + \frac{\lambda(1+\gamma\phi_n)}{\phi_n} f(x_n) + O(f_n^2) = \\ &= (1 + \gamma\varphi_n)\left(\frac{a_n}{2} + \lambda\frac{f(x_n)}{\phi_n}\right) + O(f_n^2).\end{aligned} \tag{3.12}$$

Отсюда мы находим

$$\frac{a_n}{2} = \frac{\theta_n}{1 + \gamma\phi_n} - \frac{\lambda f(x_n)}{\phi_n} + O(f_n^2). \tag{3.13}$$

Подставляя (3.13) в (3.11), получаем

$$\bar{\tau}_n = 1 + \hat{d}_n\theta_n - \frac{\lambda f(x_n)}{\phi_n} + O(f_n^2). \tag{3.14}$$

Таким образом, мы можем сформулировать полученные результаты.

Теорема 1. *Предположим, что функция $f : D \subset R \to R$ достаточно дифференцируема и имеет простой ноль $x^* \in D$. Пусть начальное приближение x_0 достаточно близко к x^*, а параметр $\bar{\tau}_n$ удовлетворяет условию* (3.14). *Тогда итерационные методы* (3.1) *имеют оптимальную сходимость четвертого порядка.*

Кансал и соавт. в [2] предложили новое трехпараметрическое оптимальное семейство итераций типа Хансена–Патрика без производных

$$\begin{aligned}y_n &= x_n - \frac{f(x_n)}{\phi_n + \lambda f(w_n)}, \\ x_{n+1} &= y_n - \bar{\tau}_n\frac{f(y_n)}{f[y, w_n] + \lambda f(w_n)},\end{aligned} \tag{3.15}$$

где

$$\bar{\tau}_n = \frac{1}{\theta_n}\left(\frac{\alpha + 1}{\alpha + \sqrt{1 - 2(\alpha+1)\theta_n}} - 1\right)H(\theta_n), \quad \alpha \neq -1. \tag{3.16}$$

Здесь H – вещественная весовая функция, удовлетворяющая условию

$$H(0) = 1, \quad H'(0) = -\frac{\alpha+1}{2}, \quad |H''(0)| < \infty. \tag{3.17}$$

Обратим внимание, что итерации (3.15) имеют различие в знаменателе во втором этапе по сравнению с (3.1). Используя легко проверяемое соотношение

$$f[y, w_n] + \lambda f(w_n) = (\phi_n + \lambda f(w_n))\left(1 - \frac{\phi_n\theta_n - \lambda f(w_n)}{(1+\gamma\phi_n)(\phi + \lambda f(w_n))}\right) + O(f_n^2), \tag{3.18}$$

второй шаг в (3.15) можно переписать в виде

$$x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n + \lambda f(w_n)},$$

где

$$\bar{\tau}_n = \left(1 + \frac{\phi_n\theta_n - \lambda f(w_n)}{(1+\gamma\phi_n)(\phi + \lambda f(w_n))}\right)\frac{1}{\theta_n}\left(\frac{\alpha+1}{\alpha + \sqrt{1 - 2(\alpha+1)\theta_n}} - 1\right)H(\theta_n), \quad \alpha \neq -1. \tag{3.19}$$

Нетрудно показать, что $\bar{\tau}_n$, заданный по формуле (3.19), удовлетворяет условию (3.14). То есть, доказываем, что итерации (3.15)–(3.17) имеют сходимость четвертого порядка без использования символических вычислений, используемых в [2]. Двухпараметрическая итерация (3.1) с $\bar{\tau}_n$, заданным (3.19), представляет новый вариант семейства итераций типа Хансена–Патрика без производной. Аналогично, используя формулу (3.18), легко показать, что двухпараметрические методы четвертого порядка без производных, приведенные в [8], [10], [13], удовлетворяют условию (3.14).

Пусть $\gamma = 0$ в (3.1). Тогда (3.1) приводит к итерациям с одним параметром

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n) + \lambda f(x_n)}, \\ x_{n+1} &= y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n) + \lambda f(x_n)}. \end{aligned} \tag{3.20}$$

По теореме 1 итерации (3.20) имеют оптимальную сходимость четвертого порядка, когда $\bar{\tau}_n$ определяется в виде

$$\bar{\tau}_n = 1 + 2\theta_n - \frac{\lambda f(x_n)}{f'(x_n)} + O(f_n^2). \tag{3.21}$$

Итерации (3.20) требуют трех функциональных вычислений: $f(x_n)$, $f(y_n)$ и $f'(x_n)$. Индекс эффективности итераций составляет $EI = \sqrt[3]{4} \approx 1.587$. Теперь попробуем найти оптимальное значение свободного параметра λ. Для этого сначала используем разложение Тейлора $f(y_n)$ в точке x_n и соотношение (2.8). В результате имеем

$$f(y_n) = f(x_n)^2\left(\frac{\lambda}{f_n'} + \frac{f_n''}{2f_n'^2} - \frac{\lambda^2 f_n}{f_n'^2} - \frac{\lambda f_n f_n''}{f_n'^3} - \frac{f_n''' f_n}{6f_n'^3}\right) + O(f_n^4). \tag{3.22}$$

Это означает, что $f(y_n) = O(f_n^2)$ для любого λ. Если выбираем

$$\lambda = \lambda_n = -\frac{f_n''}{2f_n'} \tag{3.23}$$

в (3.22), то получаем $f(y_n) = O(f_n^3)$. Назовем значение λ_n, заданное по формуле (3.23), оптимальным в том смысле, что оно увеличивает порядок сходимости последовательности y_n с двух до трех. При выборе (3.23) выражение (3.22) можно записать в виде

$$f(y_n) = f(x_n)\left(\left(\frac{a_n}{2}\right)^2 - \frac{f_n''' f_n^2}{6f_n'^3}\right) + O(f_n^4). \tag{3.24}$$

Отсюда следует, что

$$\theta_n = \left(\frac{a_n}{2}\right)^2 - \frac{f_n''' f_n^2}{6 f_n'^3} + O(f_n^3),$$

что подразумевает

$$\frac{f_n''' f_n^2}{f_n'^3} = \frac{3}{2} a_n^2 - 6\theta_n + O(f_n^3). \tag{3.25}$$

Теперь рассмотрим разложение Тейлора $f(x_{n+1})$ в точке y_n:

$$f(x_{n+1}) = f(y_n)\left(1 - \frac{f'(y_n)}{f'(x_n)}\overline{\tau}_n\left(1 - \frac{\lambda f_n}{f_n'}\right)\right) + O(f(y_n)^2). \tag{3.26}$$

Выше было показано, что $f(y_n) = O(f_n^3)$ при выборе (3.23). Следовательно, из (3.26) ясно, что

$$f(x_{n+1}) = O(f(x_n)^6), \tag{3.27}$$

если мы выберем $\overline{\tau}_n$ такой, что

$$1 - \frac{f'(y_n)}{f'(x_n)}\overline{\tau}_n\left(1 - \frac{\lambda f_n}{f_n'}\right) = O(f_n^3)$$

или

$$\overline{\tau}_n = \frac{1}{1 - \dfrac{\lambda f_n}{f_n'}}\frac{f'(x_n)}{f'(y_n)} + O(f_n^3). \tag{3.28}$$

Используя разложение Тейлора $f'(y_n)$ в точке x_n, легко показать, что

$$f'(y_n) = f'(x_n)\left(1 - a_n - \frac{a_n^2}{2} + \frac{f_n''' f_n^2}{2 f_n'^3}\right) + O(f_n^3).$$

Учитывая (3.25), получаем

$$\frac{f'(x_n)}{f'(y_n)} = \frac{1}{1 - a_n + a_n^2/4 - 3\theta_n + O(f_n^3)} = 1 + a_n + \frac{3a_n^2}{4} + 3\theta_n + O(f_n^3). \tag{3.29}$$

Подставив (3.23) и (3.29) в (3.28), получим

$$\overline{\tau}_n = 1 + \frac{a_n}{2} + \frac{a_n^2}{4} + 3\theta_n + O(f_n^3). \tag{3.30}$$

Это означает, что соотношение (3.27) выполняется при выборах (3.30) и (3.23). Таким образом, верна

Теорема 2. *Предположим, что функция $f : D \subset R \to R$ достаточно дифференцируема и имеет простой ноль $x^* \in D$. Пусть начальное приближение x_0 достаточно близко к x^*, а параметры λ_n и $\overline{\tau}_n$ удовлетворяют условиям* (3.23) *и* (3.30). *Тогда итерационные методы* (3.20) *имеют сходимость шестого порядка.*

На основе оптимального выбора параметров λ_n и $\overline{\tau}_n$ можно построить новые сходящиеся итерации шестого порядка с памятью:

x_0, λ_0 заданные. Тогда,

$$\lambda_n = -\frac{\Delta_n}{2 f_n'}, \quad n = 1, 2, \ldots,$$

$$\overline{\tau}_n = 1 - \frac{\lambda_n f_n}{f_n'} + \left(\frac{\lambda_n f_n}{f_n'}\right)^2 + 3\theta_n, \tag{3.31}$$

$$y_n = x_n - \frac{f(x_n)}{f'(x_n) + \lambda_n f(x_n)},$$

$$x_{n+1} = y_n - \overline{\tau}_n \frac{f(y_n)}{f'(x_n) + \lambda_n f(x_n)}, \quad n = 0, 1, \ldots,$$

где

$$\Delta_n = \frac{f(x_n + \gamma f(x_n)) - 2f(x_n) + f(x_n - \gamma f(x_n))}{(\gamma f(x_n))^2}, \quad \gamma \in R\backslash\{0\}.$$

Видно, что $f''(x_n) = \Delta_n + O(f_n^2)$.

Замечание 2. Уравнение ошибки, полученное с помощью символьных вычислений, играет важную роль в создании новых методов (без производных) с памятью [1], [2], [8]–[13]. Например, с выбором

$$\lambda = -c_2 = -\frac{f''(x^*)}{2f'(x^*)}, \quad \lim_{n\to\infty} \lambda_n = \lambda,$$

порядок сходимости методов увеличивается, в то время как на каждом шаге итерации мы имеем точную аналитическую формулу (3.23).

Аналогично, легко показать, что

$$f(x_{n+1}) = O(f(x_n)^5), \quad \text{если} \quad \overline{\tau}_n = 1 + \frac{a_n}{2} + O(f_n^2). \tag{3.32}$$

Следует отметить, что подобные с (3.20) методы были рассмотрены Вангом и соавт. в [8], где рассматривается

$$y_n = x_n - \frac{f(x_n)}{f'(x_n) + \lambda f(x_n)},$$

$$x_{n+1} = y_n - \frac{f(y_n)}{f'(x_n) + \gamma f(x_n)} G(\theta_n), \quad \lambda, \gamma \in R, \tag{3.33}$$

и показано, что (3.33) имеет сходимость четвертого порядка, когда

$$\gamma = 2\lambda, \quad G(0) = 1, \quad G'(0) = 2, \quad |G''(0)| < \infty. \tag{3.34}$$

В [8] использовался самоускоряющийся параметр

$$\lambda_n = -\frac{H_m''(x_n)}{2H_m'(x_n)}, \quad m = 2, 3, 4, \tag{3.35}$$

в (3.33) и доказано, что порядок сходимости R – итерационных методов (3.33) с параметром (3.35) с памятью составляет не менее $(5 + \sqrt{17})/2 \approx 4.5616$, $(5 + \sqrt{21})/2 \approx 4.7913$ и 5 соответственно. Здесь $H_m(x_n)$ – интерполяционный полином Эрмита со степенью $m = 2, 3, 4$, удовлетворяющий условию $H_m'(x_n) = f'(x_n)$. Итерации (3.33) и (3.34) можно переписать как (3.20) с $\overline{\tau}_n$ вида

$$\overline{\tau}_n = \left(1 - \frac{\lambda f_n}{f_n' + \lambda f_n} + \cdots\right)(1 + 2\theta_n + \cdots) = 1 + \frac{a_n}{2} + O(f_n^2),$$

т.е. $\overline{\tau}_n$ удовлетворяет условию (3.32). Если мы выберем λ_n как в (3.31), тогда получим итерации с памятью:

$$x_0, \lambda_0 \quad \text{заданные.} \quad \text{Тогда,}$$

$$\lambda_n = -\frac{\Delta_n}{2f_n'}, \quad a_n = -\frac{2\lambda_n f_n}{f_n'}, \quad \overline{\tau}_n = 1 + \frac{a_n}{2},$$

$$y_n = x_n - \frac{f(x_n)}{f'(x_n) + \lambda_n f(x_n)}, \tag{3.36}$$

$$x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{f'(x_n) + \lambda_n f(x_n)},$$

имеющие пятый порядок сходимости.

Замечание 3. Как уже упоминалось выше, при рассмотрении итераций типа Хансена–Патрика $\bar{\tau}_n$ определяется как (3.19).

4. НОВЫЕ ИТЕРАЦИОННЫЕ МЕТОДЫ С ПАМЯТЬЮ

Теперь мы приступим к созданию новых итерационных методов с памятью из (3.1), используя два самоускоряющихся параметра γ и λ. Легко показать, что

$$w_n = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{4.1}$$

и

$$f(w_n) = \frac{f_n'' f_n^2}{2f_n'^2} + O(f_n^3), \tag{4.2}$$

при выборе

$$\gamma = \gamma_n = -\frac{1}{f_n'}. \tag{4.3}$$

Пусть $f(x_n) \in C^4(I)$. Используя разложение Тейлора для $f(w_n)$ и (4.3), получаем

$$\phi_n = f_n'\left(1 - \frac{a_n}{2} + \frac{f_n''' f_n^2}{6f_n'^3}\right) + O(f_n^3).$$

Следовательно,

$$\eta_n = \frac{f_n'}{\phi_n} = 1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6f_n'^3} + O(f_n^3). \tag{4.4}$$

Разложение Тейлора $f(y_n)$ в точке x_n дает

$$f(y_n) = f(x_n)\left(1 - \frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}} + \frac{a_n}{2}\left(\frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}}\right)^2 - \frac{f_n''' f_n^2}{6f_n'^3}\left(\frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}}\right)^3\right) + O(f_n^4). \tag{4.5}$$

В силу (4.2) и (4.4) имеем

$$\begin{aligned}\frac{\eta_n}{1 + \frac{\lambda f(w_n)}{\phi_n}} &= \left(1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6f_n'^3} + \cdots\right)\left(1 - \frac{\lambda f(w_n)}{\phi_n} + \cdots\right) = \\ &= \left(1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6f_n'^3} - \frac{\lambda f(w_n)}{\phi_n} + O(f_n^3)\right).\end{aligned} \tag{4.6}$$

Используя (4.6) в (4.5), получаем

$$\begin{aligned}f(y_n) &= f(x_n)\left(-\frac{a_n}{2} - \frac{a_n^2}{4} + \frac{f_n''' f_n^2}{6f_n^3} + \frac{\lambda f(w_n)}{\phi_n} + \frac{a_n}{2}(1 + a_n) - \frac{f_n''' f_n^2}{6f_n'^3}\right) + O(f_n^4) = \\ &= f(x_n)\left(\frac{a_n^2}{4} + \frac{\lambda f(w_n)}{\phi_n}\right) + O(f_n^4).\end{aligned} \tag{4.7}$$

Из (4.7) ясно, что

$$f(y_n) = O(f_n^4), \tag{4.8}$$

если

$$\frac{a_n^2}{4} + \frac{\lambda f(w_n)}{\phi_n} = 0, \tag{4.9}$$

или

$$\lambda_n = -\frac{a_n^2 \phi_n}{4 f(w_n)}. \tag{4.10}$$

Используя (4.2) и (4.4) в (4.10), получаем

$$\lambda_n = -\frac{f_n''}{2f_n'}, \tag{4.11}$$

т.е. соотношение (4.8) выполняется при выборе (4.11). Далее из (3.6) и (4.8) ясно, что

$$f(x_{n+1}) = O(f_n^7), \tag{4.12}$$

если

$$\bar{\tau}_n = -\frac{\phi_n + \lambda f(w_n)}{f'(y_n)} + O(f_n^3) \tag{4.13}$$

или

$$\bar{\tau}_n = \frac{\phi_n}{f'(y_n)}\left(1 - \frac{a_n^2}{4}\right) + O(f_n^3). \tag{4.14}$$

Разложение Тейлора $f'(y_n)$ в точке x_n дает

$$f'(y_n) = f'(x_n)\left(1 - a_n\left(\frac{\eta_n}{1 - \frac{a_n^2}{4}}\right) + \frac{f_n''' f_n^2}{2f_n'^3}\left(\frac{\eta_n}{1 - \frac{a_n^2}{4}}\right)^2\right) + O(f_n^3). \tag{4.15}$$

Поскольку

$$\frac{\eta_n}{1 - \frac{a_n^2}{4}} = \left(1 + \frac{a_n}{2} + \frac{a_n^2}{4} - \frac{f_n''' f_n^2}{6f_n'^3} + \ldots\right)\left(1 + \frac{a_n^2}{4} + \ldots\right) = 1 + \frac{a_n}{2} + \frac{a_n^2}{2} - \frac{f_n''' f_n^2}{6f_n'^3} + O(f_n^3), \tag{4.16}$$

из (4.15) получаем

$$f'(y_n) = f'(x_n)\left(1 - a_n - \frac{a_n^2}{2} + \frac{f_n''' f_n^2}{2f_n'^3}\right) + O(f_n^3). \tag{4.17}$$

Используя последнее выражение и (4.4) в (4.14), получаем

$$\bar{\tau}_n = 1 - \frac{a_n}{2} + \frac{3}{4}a_n^2 + 2(1 + \gamma_n \phi_n) + O(f_n^3), \tag{4.18}$$

где использована формула

$$1 + \gamma_n \phi_n = \frac{a_n}{2} - \frac{f_n''' f_n^2}{6f_n'^3} + O(f_n^3). \tag{4.19}$$

Таблица 1. $f_1 = e^{x^3-x} - \cos(x^2-1) + x^3 + 1$, $x_0 = -1.5$, $x^* = -1$ [14]

Методы	n	$\bar{\tau}_n$	$\lvert x_n - x^*\rvert$	ρ
(3.1) ($\lambda = -0.1$, $\gamma = -0.01$)	4	(3.14)	0.1014e−217	4.00
(3.1) ($\alpha = 1$, $\gamma = -0.01$)	4	(3.19)	0.1544e−224	4.00
(3.20) ($\lambda = -0.1$)	4	(3.21)	0.6919e−229	4.00
Dzunic [13] ($p = -0.1$, $\gamma = -0.01$, $g(\theta_n) = 1 + \theta_n$)	4		0.4682e−222	4.00
Wang-Zhang [8] ($t = 8$, $\lambda = -0.1$, $G(\theta_n) = 1 + 2 * \theta_n + t * \theta_n^2$)	4		0.1974e−192	4.00
Kung-Traub [11]	4		0.9297e−173	4.00
Chebyshev-Halley [11]	4		0.5980e−175	4.00

Таким образом, можно сформулировать полученные результаты в виде теоремы.

Теорема 3. *Предположим, что функция $f : D \subset R \to R$ достаточно дифференцируема и имеет простой ноль $x^* \in D$. Пусть начальное приближение x_0 достаточно близко к x^*, а параметры γ и λ в* (3.1) *выбираются как*

$$\gamma = \gamma_n = -\frac{1}{f'(x_n)}, \quad \lambda = \lambda_n = -\frac{f''(x_n)}{2f'(x_n)},$$

и $\bar{\tau}_n$ определяется формулой (3.14) (*либо* (4.18)). *Тогда итерационные методы* (3.1) *имеют порядок сходимости семь.*

Таким образом, оптимальный выбор параметров позволяет увеличить порядок сходимости с 4 до 7. Однако значения $f'(x_n)$ и $f''(x_n)$ на практике недоступны, и такое ускорение сходимости не может быть реализовано. Но мы бы приблизили параметры γ_n и λ_n. Они могут быть вычислены с использованием информации, доступной из текущей и предыдущей итераций. На основе вариантов (4.3) и (4.11) можно построить двухточечные итерации (без производных) с памятью и имеющие седьмой порядок сходимости:

$$\begin{gathered} x_0, \lambda_0, \gamma_0 \quad \text{заданные.} \quad \text{Тогда} \quad w_0 = x_0 + \gamma_0 f(x_0), \\ \gamma_n = -\frac{1}{N_3'(x_n)}, \quad w_n = x_n + \gamma_n f(x_n), \quad \lambda_n = -\frac{N_4''(x_n)}{2N_4'(x_n)}, \quad n = 1, 2, \ldots, \\ y_n = x_n - \frac{f(x_n)}{\phi_n + \lambda_n f(w_n)}, \\ x_{n+1} = y_n - \bar{\tau}_n \frac{f(y_n)}{\phi_n + \lambda_n f(w_n)}, \quad n = 0, 1, \ldots, \end{gathered} \tag{4.20}$$

где $\bar{\tau}_n$ удовлетворяет условию (3.14). Здесь $N_3(t, x_n, y_{n-1}, x_{n-1}, w_{n-1})$ и $N_4(t, w_n, x_n, w_{n-1}, y_{n-1}, x_{n-1})$ – интерполяционные полиномы Ньютона третьей и четвертой степени, построенные через доступные узловые точки $(x_n, x_{n-1}, y_{n-1}, w_{n-1})$ и $(x_n, w_n, x_{n-1}, y_{n-1}, w_{n-1})$ соответственно. Заметим, что в [13] получено вышеуказанное свойство при выборе

$$\lambda_n = -\frac{N_4''(w_n)}{2N_4'(w_n)}.$$

5. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

Чтобы продемонстрировать поведение сходимости и эффективность методов (3.1), (3.20), (3.36), (4.20), мы рассмотрим несколько числовых примеров и сделаем сравнения с существующими методами того же порядка. Расчеты были выполнены в Maple 18 с использованием ариф-

Таблица 2. $f_1 = e^{x^3-x} - \cos(x^2-1) + x^3 + 1$, $x_0 = -1.5$, $x^* = -1$ [14]

Методы	n	$\overline{\tau}_n$	$\lvert x_n - x^*\rvert$	ρ
(3.20) ($\lambda_n = -f_n''/2f_n'$)	3	(3.21)	0.1735e–56	5.00
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	3	(3.32)	0.7578e–99	5.00
(3.36) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	3		0.4079e–85	5.02
(3.33)–(3.35) [8] ($\lambda_n = -H_4''/2f_n'$, $\lambda_0 = -0.1$)	3		0.2404e–89	5.09
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	3	(3.30)	0.6559e–176	6.00
(3.31) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	3		0.4538e–125	6.00
(3.1) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	3	(4.18)	0.3128e–93	7.00
(4.20) ($\lambda_n = -N_4''(x_n)/2N_4'(x_n)$, $\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	3	(3.21)	0.4294e–162	7.06
Dzunic [13] ($p_0 = -0.1$, $\gamma_0 = -0.01$), $g(\theta_n) = 1 + \theta_n$	3		0.1404e–157	7.06
Cordero [14] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	3		0.1114e–157	7.06
Kansal [2] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$), $\alpha = \beta = 1/2$	3		0.1095e–100	7.08

Таблица 3. $f_2 = e^{x^3-3x}\sin x + \log(x^2+1)$, $x_0 = 1$, $x^* = 0$ [12]

Методы	n	$\overline{\tau}_n$	$\lvert x_n - x^*\rvert$	ρ
(3.1) ($\lambda = -0.1$, $\gamma = -0.01$)	4	(3.14)	0.1469e–82	4.00
(3.1) ($\alpha = 1$, $\gamma = -0.01$)	4	(3.19)	0.6589e–68	3.99
(3.20) ($\lambda = -0.1$)	4	(3.21)	0.3650e–83	4.00
Dzunic [13] ($p = -0.1$, $\gamma = -0.01$, $g(\theta_n) = 1 + \theta_n$)	4		0.7008e–66	4.00
Wang-Zhang [8] ($t = 8$, $\lambda = -0.1$, $G(\theta_n) = 1 + 2 * \theta_n + t * \theta_n^2$)	5		0.7447e–204	4.00
Kung-Traub [11]	4		0.1469e–82	4.00
Chebyshev-Halley [11]	4		0.1975e–88	4.00

Таблица 4. $f_2 = e^{x^3-3x}\sin x + \log(x^2+1)$, $x_0 = 1$, $x^* = 0$ [12]

Методы	n	$\overline{\tau}_n$	$\lvert x_n - x^*\rvert$	ρ
(3.20) ($\lambda_n = -f_n''/2f_n'$)	4	(3.21)	0.2170e–217	5.00
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(3.32)	0.3916e–220	5.00
(3.36) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	4		0.2326e–136	5.00
(3.33)–(3.35) [8] ($\lambda_n = -H_4''/2f_n'$, $\lambda_0 = -0.1$)	4		0.2069e–122	5.00
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(3.30)	0.3111e–233	6.00
(3.31) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	4		0.2699e–291	6.00
(3.1) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(4.18)	0.1560e–119	7.00
(4.20) ($\lambda_n = -N_4''(x_n)/2N_4'(x_n)$, $\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	4	(3.21)	0.3134e–416	7.00
Dzunic [13] ($p_0 = -0.1$, $\gamma_0 = -0.01$), $g(\theta_n) = 1 + \theta_n$	4		0.3892e–330	6.99
Cordero [14] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	4		0.5524e–284	6.99
Kansal [2] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$), $\alpha = \beta = 1/2$	4		0.8391e–293	6.98

метики с кратной точностью и с 1000 цифрами. Для численных расчетов мы использовали следующие функции [12], [13] и [14]:

$$f_1 = e^{x^3-x} - \cos(x^2-1) + x^3 + 1, \quad x^* = -1,$$
$$f_2 = e^{x^3-3x}\sin x + \log(x^2+1), \quad x^* = 0,$$
$$f_3 = (x^6 + x^{-6} + 4)(x-1)\sin x^2, \quad x^* = 1,$$

Таблица 5. $f_3 = (x^6 + x^{-6} + 4)(x-1)\sin x^2$, $x_0 = 0.8$, $x^* = 1$ [13]

Методы	n	$\overline{\tau}_n$	$\lvert x_n - x^*\rvert$	ρ
(3.1) ($\lambda = -0.1$, $\gamma = -0.01$)	4	(3.14)	0.3589e–140	4.00
(3.1) ($\alpha = 1$, $\gamma = -0.01$)	4	(3.19)	0.9036e–111	4.00
(3.20) ($\lambda = -0.1$)	4	(3.21)	0.1007e–138	4.00
Dzunic [13] ($p = -0.1$, $\gamma = -0.01$, $g(\theta_n) = 1 + \theta_n$)	4		0.4671e–130	4.00
Wang-Zhang [8] ($t = 8$, $\lambda = -0.1$, $G(\theta_n) = 1 + 2 * \theta_n + t * \theta_n^2$)	4		0.1552e–96	4.00
Kung-Traub [11]	4		0.2972e–132	4.00
Chebyshev-Halley [11]	4		0.2847e–118	4.00

Таблица 6. $f_3 = (x^6 + x^{-6} + 4)(x-1)\sin x^2$, $x_0 = 0.8$, $x^* = 1$ [13]

Методы	n	$\overline{\tau}_n$	$\lvert x_n - x^*\rvert$	ρ
(3.20) ($\lambda_n = -f_n''/2f_n'$)	3	(3.21)	0.1344e–53	4.99
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(3.32)	0.2239e–250	5.00
(3.36) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	4		0.5113e–183	5.00
(3.33)–(3.35) [8] ($\lambda_n = -H_4''/2f_n'$, $\lambda_0 = -0.1$)	4		0.1142e–213	5.00
(3.20) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(3.30)	0.2116e–259	6.00
(3.31) ($\lambda_n = -\Delta_n/2f_n'$, $\lambda_0 = -0.1$)	3		0.1080e–60	5.96
(3.1) ($\lambda_n = -f_n''/2f_n'$, $\lambda_0 = -0.1$)	4	(4.18)	0.1802e–315	7.00
(4.20) ($\lambda_n = -N_4''(x_n)/2N_4'(x_n)$, $\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	3	(3.21)	0.6532e–107	7.03
Dzunic [13] ($p_0 = -0.1$, $\gamma_0 = -0.01$), $g(\theta_n) = 1 + \theta_n$	3		0.1364e–87	7.05
Cordero [14] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$)	3		0.8409e–80	7.04
Kansal [2] ($\lambda_0 = -0.1$, $\gamma_0 = -0.01$), $\alpha = \beta = 1/2$	3		0.9084e–72	7.02

и критерий остановки $\lvert x_n - x^*\rvert < 10^{-60}$. Результаты расчетов приведены в табл. 1–6, где указаны числа итераций (n), абсолютные погрешности $\lvert x_n - x^*\rvert$ и вычислительный порядок сходимости (ρ), заданный по формуле

$$\rho \approx \frac{\ln(\lvert x_n - x^*\rvert / \lvert x_{n-1} - x^*\rvert)}{\ln(\lvert x_{n-1} - x^*\rvert / \lvert x_{n-2} - x^*\rvert)}.$$

Из табл. 1–6 видно, что результаты расчетов полностью подтверждают теоретический порядок сходимости, полученный в предыдущих разделах.

6. ВЫВОДЫ

Мы предлагаем новый класс оптимальных двухточечных итерационных методов, не содержащих производные, которые включают в себя два свободных параметра. Впервые мы нашли точные аналитические формулы для оптимальных значений этих параметров, что позволяет повышать порядок сходимости. На этой основе мы предлагаем новые итерационные методы с высоким порядком сходимости как с памятью, так и без памяти.

СПИСОК ЛИТЕРАТУРЫ

1. *Kansal M., Kanwar V., Bhatia S.* New modifications of Hansen–Patrick's family with optimal fourth and eighth orders of convergence // Appl. Math.Comput. 2015. V. 269. P. 507–519.
2. *Kansal M., Kanwar V., Bhatia S.* Efficient derivative-free variants of Hansen-Patrick's family with memory for solving nonlinear equations // Numer. Algor. 2016. V. 73. P. 1017–1036.

3. *Zhanlav T., Ulziibayar V., Chuluunbaatar O.* Necessary and sufficient conditions for the convergence of two- and three-point Newton-type iterations // Comput. Math. Math. Phys. 2017. V. 57. P. 1090–1100.
4. *Zhanlav T., Chuluunbaatar O., Ulziibayar V.* Accelerating the convergence of Newton-type iterations // J. Numer. Anal. Approx. Theory. 2017. V. 46. P. 162–180.
5. *Zhanlav T., Mijiddorj R., Otgondorj Kh.* Constructive theory of designing optimal eighth-order derivative-free methods for solving nonlinear equations // AM. J. Comput. Math. 2020. V. 10. P. 100–117.
6. *Petković L.D., Petković M.S., Neta B.* On optimal parameter of Laguerre's family of zero-finding methods // Inter. Journal of Comput. Math. 2018. V. 95. 692–707.
7. *Zhanlav T., Chuluunbaatar O., Otgondorj Kh.* A derivative-free families of optimal two-and three-point iterative methods for solving nonlinear equations // Comput. Math. Math. Phys. 2019. V. 59. P. 920–936.
8. *Wang X., Zhang T.* A new family of Newton-type iterative methods with and without memory for solving nonlinear equations // Calcolo 2014. V. 51. P. 1–15.
9. *Wang X.* A new accelerating technique applied to a variant of Cordero–Torregrosa method // J. Comput. Appl. Math. 2018. V. 330. P. 695–709.
10. *Petković M.S., Ilic S., Dzunić J.* Derivative-free two-point methods with and without memory for solving nonlinear equations // Appl. Math. Comput. 2010. V. 217. P. 1887–1895.
11. *Argyros I.K., Kansal M., Kanwar V., Bajaj S.* Higher-order derivative-free families of Chebyshev-Halley type methods with or without memory for solving nonlinear equations // Appl. Math. Comput. 2017. V. 315. P. 224–245.
12. *Lotfi T., Soleymani F., Ghorbanzadeh M., Assari P.* On the construction of some tri-parametric iterative methods with memory // Numer. Algor. 2015. V. 70. P. 835–845.
13. *Dzunić J.* On efficient two-parameter methods for solving nonlinear equations // Numer. Algor. 2013. V. 63. P. 549–569.
14. *Cordero A., Lotfi T., Bakhtiari P., Torregrosa J.R.* An efficient two-parametric family with memory for nonlinear equations // Numer. Algor. 2015. V. 68. P. 323–335.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 47–56

ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

УДК 519.642

ВЫБОР ПАРАМЕТРА РЕГУЛЯРИЗАЦИИ НА ОСНОВЕ РЕКОНСТРУКЦИИ РЕГУЛЯРИЗИРОВАННОГО РЕШЕНИЯ В ЗАДАЧЕ АДАПТИВНОЙ КОРРЕКЦИИ СИГНАЛОВ

© 2021 г. М. Л. Маслаков

199178 Санкт-Петербург, 11-я линия В.О., 66, АО "Российский институт мощного радиостроения", Россия

e-mail: maslakovml@gmail.com

Поступила в редакцию 23.01.2020 г.
Переработанный вариант 10.07.2020 г.
Принята к публикации 18.09.2020 г.

Рассматривается адаптивная коррекция сигналов как решение обратной некорректной задачи. Данная задача сводится к интегральному уравнению типа свертки, а для его решения используется метод регуляризации. Для выбора параметра регуляризации предлагается осуществить реконструкцию регуляризированного решения. Представлены результаты численных экспериментов. Библ. 18. Фиг. 3.

Ключевые слова: некорректная задача, интегральное уравнение типа свертки, регуляризация, параметр регуляризации.

DOI: 10.31857/S0044466921010063

1. ВВЕДЕНИЕ

В статье рассматривается задача адаптивной коррекции частотно ограниченных информационных сигналов, передаваемых через нестационарные замирающие каналы связи (см. [1]). Данная задача сводится к решению интегрального уравнения типа свертки I рода (см. [2]), которое запишем в операторном виде

$$HS = U_\delta, \tag{1.1}$$

где $H \in R^{k\times m}$, $k \geq m$, – матрица коэффициентов импульсной характеристики канала, $S \in R^m$ – вектор отсчетов передаваемого сигнала, $U_\delta \in R^k$ – вектор отсчетов принятого сигнала.

Вектор U_δ представляет собой результат измерений на фоне белого гауссовского шума, т.е.

$$U_\delta = \bar{U} + \xi, \tag{1.2}$$

где $\bar{U}$ – точные значения вектора отсчетов принятого сигнала, ξ – аддитивный белый гауссовский шум с нулевым средним и дисперсией σ_ξ^2.

Матрица H состоит из коэффициентов импульсной характеристики канала $h(t)$ (см. [3]), причем

$$h(t) \equiv 0, \quad t \leq 0, \tag{1.3}$$

$$\int_{-\infty}^{\infty} |h(t)|\, dt < \infty. \tag{1.4}$$

Матрица H представляет собой циркулянтную матрицу $k \times m$ вида

$$H = \begin{bmatrix} h(t_1) & 0 & \cdots & 0 \\ h(t_2) & h(t_1) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h(t_k) & & \ddots & h(t_1) \\ 0 & \ddots & & h(t_2) \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & h(t_k) \end{bmatrix}, \tag{1.5}$$

где $h(t_k)$ – отсчеты импульсной характеристики $h(t)$.

Отметим, что на практике коэффициенты матрицы H получают из решения уравнения (1.1) путем передачи тестового сигнала и получении отклика на него (см. [4], [5]). Таким образом, H в общем случае является регуляризированным решением, а значит, $H \equiv H_\alpha$. В рамках рассматриваемой задачи полагаем

$$\max\left|\bar{H} - H\right| \le \delta_H, \tag{1.6}$$

где $\bar{H}$ – точные значения коэффициентов импульсной характеристики канала.

Точные оценки σ_ξ^2 и δ_H отсутствуют.

Для выбора параметра регуляризации применяют различные эвристические методы, описание и сравнительный анализ некоторых из них приводится, например, в [6]–[13]. Применение конкретного метода определяется особенностью постановки решаемой задачи.

В данной работе автором предлагается метод выбора параметра регуляризации путем реконструкции получаемого решения. Для этого предполагается осуществить поэлементную оценку регуляризированного решения и формирование опорных функций.

Статья организована следующим образом. В разд. 2 даны определение реконструированного регуляризированного решения и его свойства. В разд. 3 приведен метод выбора параметра регуляризации на основе реконструированного регуляризированного решения. При этом рассмотрены случаи точно известной и приближенной матрицы H. Результаты численного эксперимента представлены в разд. 4. Выводы по работе сформулированы в разд. 5.

2. РЕКОНСТРУКЦИЯ РЕГУЛЯРИЗИРОВАННОГО РЕШЕНИЯ

2.1. Получение реконструированного регуляризированного решения

Вектор передаваемого информационного сигнала S представляет собой отсчеты фазоманипулированного одночастотного сигнала вида (см. [14], [15])

$$s(t) = A\sum_{n=1}^{N} \cos\left(\omega_0 t + \varphi(n)\right) p(t - (n-1)T_{\text{sym}}), \quad t \in [0; NT_{\text{sym}}], \tag{2.1}$$

где N – количество передаваемых символов, A – амплитуда передаваемого сигнала, $\omega_0 = 2\pi f_0$ – несущая частота, $\varphi(n)$, $n = 1, 2, \ldots, N$ – фазы передаваемых символов, T_{sym} – длительность символа, $p(t)$ – импульсная функция вида

$$p(t) = \begin{cases} 1, & t \in [0; NT_{\text{sym}}], \\ 0, & t \notin [0; NT_{\text{sym}}]. \end{cases}$$

Введем оператор Y, осуществляющий операцию модуляции, т.е. формирующий сигнал (2.1) в соответствии с входным вектором $B = \{b(n)\}, n = 1, 2, \ldots, N$, состоящим из последовательности передаваемых информационных бит, т.е.

$$S = YB. \tag{2.2}$$

Последовательность фаз $\varphi(n)$, $n = 1, 2, \ldots, N$, однозначно соответствует вектору B. При этом для различной позиционности фазовой модуляции (ФМ) определенному набору бит соответствуют строго определенные значения фаз (см. [15]):

$$b = \{0;1\} \Leftrightarrow \varphi = \{0;\pi\} - \text{для двухпозиционной ФМ (BPSK)};$$

$$b = \{00;01;11;10\} \Leftrightarrow \varphi = \left\{0;\frac{\pi}{2};\pi;\frac{3\pi}{2} \equiv -\frac{\pi}{2}\right\} - \text{для четырехпозиционной ФМ (QPSK)};$$

и т.д. с учетом кода Грея (см. [16]).

Также, для удобства, будем полагать, что

$$T_{\text{sym}} F_s \in \mathbb{N}, \tag{2.3}$$

где F_s – частота дискретизации.

Решение уравнения (1.1) осуществляется методом регуляризации (в работе использован метод регуляризации Тихонова из [12]). Соответствующее численное решение этого уравнения обозначим вектором $S(\alpha)$, зависящим от параметра регуляризации α.

Далее введем оператор Y^-, осуществляющий оптимальный когерентный прием, т.е. демодуляцию (см. [15], [16]). Данный оператор введен лишь для удобства последующей записи математических выражений. Описание (реализация) Y^- приведено в [5]. Тогда, осуществив демодуляцию регуляризированного решения уравнения (1.1), определяемого вектором $S(\alpha)$, получим последовательность бит $B(\alpha) = \{b(n,\alpha)\}$, $n = 1, 2, \ldots, N$, которая также будет зависеть от параметра α, т.е.

$$B(\alpha) = Y^- S(\alpha). \tag{2.4}$$

Тогда решение

$$S_Y(\alpha) = Y(Y^- S(\alpha)) \tag{2.5}$$

будем называть *реконструированным регуляризированным решением* уравнения (1.1).

2.2. Свойства реконструированного регуляризированного решения

Рассмотрим квадратичные нормы исходного вектора S, а также регуляризированных решений уравнения (1.1) – $S(\alpha)$ и $S_Y(\alpha)$:

$$E = \|S\|^2, \tag{2.6}$$

$$E(\alpha) = \|S(\alpha)\|^2, \tag{2.7}$$

$$E_Y(\alpha) = \|S_Y(\alpha)\|^2. \tag{2.8}$$

При этом для (2.7) и (2.8) имеют место следующие предельные соотношения:

$$\lim_{\alpha \to +\infty} E(\alpha) = 0 \quad (\text{см., например, [11], [12]}), \tag{2.9}$$

$$\lim_{\alpha \to +\infty} E_Y(\alpha) = E. \tag{2.10}$$

Последнее следует из выражения (2.1) с учетом допущения (2.3). Вообще говоря, для любого вектора B имеем

$$\|YB\|^2 = E. \tag{2.11}$$

Иными словами

$$E_Y(\alpha) \equiv \text{const} = E. \tag{2.12}$$

Отметим, что в случае сигналов BPSK соотношение (2.11) справедливо и без допущения (2.3).

Оптимальное значение параметра регуляризации, согласно [5], соответствует минимуму функционала, представляющего собой количество битовых (символьных) ошибок

$$q(\alpha) = \sum_{n=1}^{N} (b(n) \oplus b(n,\alpha)), \tag{2.13}$$

где $\oplus$ – знак сложения по модулю два.

При этом отметим, что информационная последовательность бит $b(n)$, $n = 1, 2, \ldots, N$, на практике неизвестна.

Очевидно, что в случае, когда существует такое $\alpha^* > 0$, при котором достигается минимум функционала (2.13)

$$q(\alpha^*) = 0, \tag{2.14}$$

имеет место

$$S_Y(\alpha^*) \equiv S. \tag{2.15}$$

Однако из [5] следует, что (2.14) не всегда выполняется, т.е. для $\forall \alpha > 0$ может быть, что

$$\min_{\alpha} q(\alpha) > 0. \tag{2.16}$$

В этом случае (2.15) выполняется только лишь на некоторых отрезках (интервалах) из $[0; NT_{\text{sym}}]$, т.е.

$$s_Y(t,\alpha) \equiv s(t), \quad [t_1;t_2] \cup [t_3;t_4] \cup [t_5;t_6] \ldots \in [0; NT_{\text{sym}}]. \tag{2.17}$$

Здесь $s_Y(t,\alpha)$ – фазоманипулированный сигнал, аналогичный (2.1), значения фаз которого соответствуют вектору $B(\alpha)$. Вектор $S_Y(\alpha)$ определяется отсчетами сигнала $s_Y(t,\alpha)$. Отметим, что интервалы $[t_k;t_{k+1}]$ кратны длительности символа T_{sym}.

3. ВЫБОР ПАРАМЕТРА РЕГУЛЯРИЗАЦИИ

3.1. Случай точно известных коэффициентов матрицы H

Пусть матрица коэффициентов импульсной характеристики канала известна точно, т.е. $\delta_H = 0$.

Обозначим приближенное представление реконструированного решения в форме

$$U_Y(\alpha) = HS_Y(\alpha), \tag{3.1}$$

при этом

$$\|U_Y(\alpha)\|^2 = E_{U_Y}(\alpha) > 0 \quad (\forall \alpha > 0). \tag{3.2}$$

Отметим, что в общем случае $E_{U_Y}(\alpha) \neq \text{const}$, в отличие от (2.12).

Рассмотрим соотношение

$$\begin{aligned} r(\alpha) &= \|U_Y(\alpha) - U_\delta\|^2 = \|U_Y(\alpha) - \bar{U} + \bar{U} - U_\delta\|^2 = \\ &= \|U_Y(\alpha) - \bar{U}\|^2 + \|\bar{U} - U_\delta\|^2 + 2\|U_Y(\alpha) - \bar{U}\|\|\bar{U} - U_\delta\|. \end{aligned} \tag{3.3}$$

С учетом (1.2) данное выражение примет вид

$$r(\alpha) = \|U_Y(\alpha) - \bar{U}\|^2 + \|\xi\|^2 + 2\|U_Y(\alpha) - \bar{U}\|\|\xi\|. \tag{3.4}$$

Допустим, что (2.14) и, следовательно, (2.15) выполняется. В этом случае $U_Y(\alpha^*) \equiv \bar{U}$ и тогда выражение (3.5) преобразуется к виду

$$r(\alpha^*) = \|\xi\|^2 = \sigma_\xi^2. \tag{3.5}$$

Допустим, что выполняется условие (2.16) и, следовательно, имеет место (2.17). Значит, существуют такие интервалы $[t_1';t_2'] \cup [t_3';t_4'] \cup [t_5';t_6']$, на которых выполняется

$$U_Y(\alpha) \equiv \bar{U}. \tag{3.6}$$

Отметим, что данные интервалы, вообще говоря, не равны интервалам из (2.17).

Обозначим объединение всех интервалов, на которых выполняется (3.6), как T_0, а где не выполняется – T_e. При этом

$$T_0 \cup T_e \equiv [0;k]. \tag{3.7}$$

Тогда $\forall\alpha > 0$ имеют место следующие утверждения:

$$\left\|U_Y(\alpha) - \bar{U}\right\|^2_{T_0} = 0, \tag{3.8}$$

$$0 < \left\|U_Y(\alpha) - \bar{U}\right\|^2_{T_e} \le \left\|2\bar{U}\right\|^2_{T_e}. \tag{3.9}$$

Утверждение (3.8) очевидно и следует из (3.6). Утверждение (3.9) следует из того, что в худшем случае

$$U_Y(\alpha) = -\bar{U}. \tag{3.10}$$

В результате выражение (3.4) преобразуется к виду

$$\begin{aligned} r(\alpha) &= \left\|U_Y(\alpha) - \bar{U}\right\|^2_{T_0\cup T_e} + \left\|\xi\right\|^2_{T_0\cup T_e} + 2\left\|U_Y(\alpha) - \bar{U}\right\|_{T_0\cup T_e}\left\|\xi\right\|_{T_0\cup T_e} = \\ &= \left\|U_Y(\alpha) - \bar{U}\right\|^2_{T_e} + \left\|\xi\right\|^2 + 2\left\|U_Y(\alpha) - \bar{U}\right\|_{T_e}\left\|\xi\right\| = \left\|U_Y(\alpha) - \bar{U}\right\|^2_{T_e} + \sigma^2_\xi + 2\left\|U_Y(\alpha) - \bar{U}\right\|_{T_e}\sigma_\xi. \end{aligned} \tag{3.11}$$

Пусть для некоторого $\alpha_1 > 0$ $T_e^1 = [t_1';t_2']$, а для $\alpha_2 > 0$ $T_e^2 = [t_1';t_3']$. При этом $t_2' < t_3'$. Очевидно, что в этом случае

$$\left\|U_Y(\alpha_1) - \bar{U}\right\|^2_{T_e^1} < \left\|U_Y(\alpha_2) - \bar{U}\right\|^2_{T_e^2} \tag{3.12}$$

и, следовательно,

$$r(\alpha_1) < r(\alpha_2), \tag{3.13}$$

а значит, $\min r(\alpha)$ соответствует лучшему приближению $U_Y(\alpha)$ к $\bar{U}$ и в качестве оптимального значения параметра регуляризации можно взять

$$\alpha_{\text{opt}} = \arg\left(\min_\alpha r(\alpha)\right), \tag{3.14}$$

что также обеспечивает минимум функционала (2.13).

3.2. Случай приближенной матрицы H

Пусть матрица коэффициентов импульсной характеристики канала является регуляризированным приближенным решением H_α таким, что (1.6), $\delta_H > 0$ (δ_H – неизвестно). Как показано в [4], оптимальное значение параметра регуляризации для матрицы H_α (т.е. обеспечивающее лучшее приближение H_α к $\bar{H}$) равно оптимальному значению параметра регуляризации для решения исходного уравнения (1.1). При этом

$$\lim_{\alpha\to+\infty}\left\|H_\alpha\right\| = 0, \tag{3.15}$$

а значит,

$$\lim_{\alpha\to+\infty}\left\|U_Y(\alpha)\right\| = 0. \tag{3.16}$$

Допустим выполнение (2.14) и (2.15). Тогда

$$r(\alpha^*) = \left\|U_Y(\alpha^*) - U_\delta\right\|^2 = \left\|H_{\alpha^*}S_Y(\alpha^*) - U_\delta\right\|^2 = \left\|H_{\alpha^*}\bar{S} - U_\delta\right\|^2. \tag{3.17}$$

В этом случае (3.17) представляет собой невязку для регуляризированных коэффициентов матрицы H_α, свойства которой приведены в [11], [12]. Отличительной особенностью выражения (3.17) является то, что коэффициенты матрицы H_α получены при решении уравнения (1.1) в условиях другой реализации шумовой составляющей в U_δ.

Выражение (3.17) можно также представить в виде

$$r(\alpha^*) = \left\|H_{\alpha^*}\bar{S} - \bar{U} + \bar{U} - U_\delta\right\|^2 = \left\|H_{\alpha^*}\bar{S} - \bar{U}\right\|^2 + \left\|\bar{U} - U_\delta\right\|^2 - 2\left\|H_{\alpha^*}\bar{S} - \bar{U}\right\|\left\|\bar{U} - U_\delta\right\| = \\ = \left\|\vartheta(\alpha^*)\right\|^2 + \left\|\xi\right\|^2 - 2\left\|\vartheta(\alpha^*)\right\|\left\|\xi\right\|, \tag{3.18}$$

где $\vartheta(\alpha^*)$ – представляет собой “отбеленный” шум.

Причем $\vartheta(\alpha)$ и ξ – независимы. Рассмотрим подробнее $\left\|\vartheta(\alpha^*)\right\|^2$:

$$\left\|\vartheta(\alpha^*)\right\|^2 = \left\|H_{\alpha^*}\bar{S} - \bar{U}\right\|^2 = \left\|H_{\alpha^*}\bar{S} - \bar{H}\bar{S}\right\|^2 \le E\left\|H_{\alpha^*} - \bar{H}\right\|^2. \tag{3.19}$$

Лучшее приближение H_{α^*} обеспечит минимум $\left\|\vartheta(\alpha^*)\right\|$. При этом в [5] доказано, что условие (2.14) выполняется лишь в ограниченной области возможных значений $\{\alpha^*\}$, любое из которых будет являться оптимальным с точки зрения минимума функционала (2.13).

Пусть выполняется условие (2.16) и, следовательно, имеет место (2.17). При этом существуют интервалы T_0, на которых имеет место (3.19), т.е.

$$\left\|\vartheta(\alpha)\right\|_{T_0}^2 = \left\|H_\alpha S_Y(\alpha) - \bar{U}\right\|_{T_0}^2 = \left\|H_\alpha\bar{S} - \bar{U}\right\|_{T_0}^2, \tag{3.20}$$

а также интервалы T_e, на которых

$$\left\|\vartheta(\alpha)\right\|_{T_e}^2 = \left\|H_\alpha S_Y(\alpha) - \bar{U}\right\|_{T_e}^2. \tag{3.21}$$

С учетом (3.16) имеют место следующие предельные соотношения:

$$\lim_{\alpha\to+\infty}\left\|\vartheta(\alpha)\right\|_{T_0}^2 = \left\|\bar{U}\right\|_{T_0}^2, \tag{3.22}$$

$$\lim_{\alpha\to+\infty}\left\|\vartheta(\alpha)\right\|_{T_e}^2 = \left\|\bar{U}\right\|_{T_e}^2, \tag{3.23}$$

$$\lim_{\alpha\to+\infty}\left\|\vartheta(\alpha)\right\|^2 = \left\|\bar{U}\right\|^2 = E_U. \tag{3.24}$$

Свойства невязки $r(\alpha)$ приведены в [12], в частности,

$$\lim_{\alpha\to+\infty} r(\alpha) = \left\|U_\delta\right\|^2 = E_{U_\delta}, \tag{3.25}$$

$$\lim_{\alpha\to0+} r(\alpha) = \mu^2, \tag{3.26}$$

где μ^2 – мера несовместности.

Кроме того, в рамках рассматриваемой задачи имеет место очевидное неравенство

$$\lim_{\alpha\to0+}\left\|r(\alpha)\right\|_{T_0}^2 < \lim_{\alpha\to0+}\left\|r(\alpha)\right\|_{T_e}^2. \tag{3.27}$$

Таким образом, с учетом неравенства (3.12), из выражения $\left\|\bar{U} - U_\delta\right\|^2 = \left\|\xi\right\|^2 = \text{const}$ следует, что минимуму $\left\|\vartheta(\alpha)\right\|^2 = \left\|H_\alpha S_Y(\alpha) - \bar{U}\right\|^2$ соответствует минимум невязки $r(\alpha)$. А выбор α_{opt} из (3.14) обеспечивает минимум функционала (2.13).

3.3. Некоторые дополнительные замечания

Рассмотрим выбор параметра регуляризации по невязке для решения уравнения (1.1) в форме $S(\alpha)$ (см. [17]). С учетом принятых обозначений невязка есть

$$\hat{r}(\alpha) = \left\|HS(\alpha) - U_\delta\right\|^2. \tag{3.28}$$

Различные модификации данного подхода рассмотрены, например, в [13], как для случая $\delta_H = 0$, так и для случая $\delta_H \ne 0$ (но не в случае $H \equiv H_\alpha$).

Рассмотрим подробнее этот случай ($H \equiv H_\alpha$), имеющий место в рассматриваемой задаче коррекции сигналов. Тогда вместо (3.26) выражение для невязки примет вид

$$\tilde{r}(\alpha) = \left\|H_\alpha S(\alpha) - U_\delta\right\|^2. \tag{3.29}$$

В основном свойства невязки $\tilde{r}(\alpha)$ и ее предельные значения не отличаются от случаев $r(\alpha)$ и $\hat{r}(\alpha)$. Тем не менее имеются определенные различия.

1. $\exists\alpha_0 : \forall\alpha > \alpha_0$ выполняется неравенство

$$\|H_\alpha S(\alpha)\|^2 < \|H_\alpha S_Y(\alpha)\|^2. \tag{3.30}$$

Доказательство следует из того, что $\|H_\alpha\|^2$ и $\|S(\alpha)\|^2$ – монотонно не возрастающие функции (см. [12]). Однако для $\|S_Y(\alpha)\|^2$ при $\forall\alpha$ выполняется (2.12).

Тогда имеем

$$\|H_\alpha S(\alpha)\|^2 \le \|H_\alpha\|^2 \|S(\alpha)\|^2, \tag{3.31}$$

$$\|H_\alpha S_Y(\alpha)\|^2 \le \|H_\alpha\|^2 E. \tag{3.32}$$

С учетом (2.9) приходим к выводу, что найдется такое $\alpha_0 : \forall\alpha > \alpha_0$ имеет место неравенство

$$\|S(\alpha)\|^2 < E, \tag{3.33}$$

а значит, неравенство (3.30) также выполняется.

Аналогично можно доказать выполнение неравенства

$$\|\bar{H}S(\alpha)\|^2 < \|\bar{H}S_Y(\alpha)\|^2. \tag{3.34}$$

При этом отметим, что α_0 может быть ≥ 0.

2. Функция $\|H_\alpha S(\alpha)\|^2$ монотонно не возрастающая $\forall\alpha > 0$, однако функция $\|H_\alpha S_Y(\alpha)\|^2$ не является монотонной на интервале $[0;\alpha']$. Данное свойство менее очевидно и следует из того, что в общем случае для различных векторов B_1 и B_2: $B_1 \ne B_2$ имеет место

$$\|\bar{H}YB_1\|^2 \ne \|\bar{H}YB_2\|^2. \tag{3.35}$$

Следовательно, $\exists\alpha_1, \alpha_2 \in [0;\alpha'] : \alpha_1 < \alpha_2$, при которых

$$\|\bar{H}S_Y(\alpha_1)\|^2 < \|\bar{H}S_Y(\alpha_2)\|^2, \tag{3.36}$$

и, как следствие, имеем

$$\|H_{\alpha_1}S_Y(\alpha_1)\|^2 < \|H_{\alpha_2}S_Y(\alpha_2)\|^2. \tag{3.37}$$

Это приводит к тому, что зависимость невязки от параметра регуляризации может иметь несколько локальных минимумов.

4. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

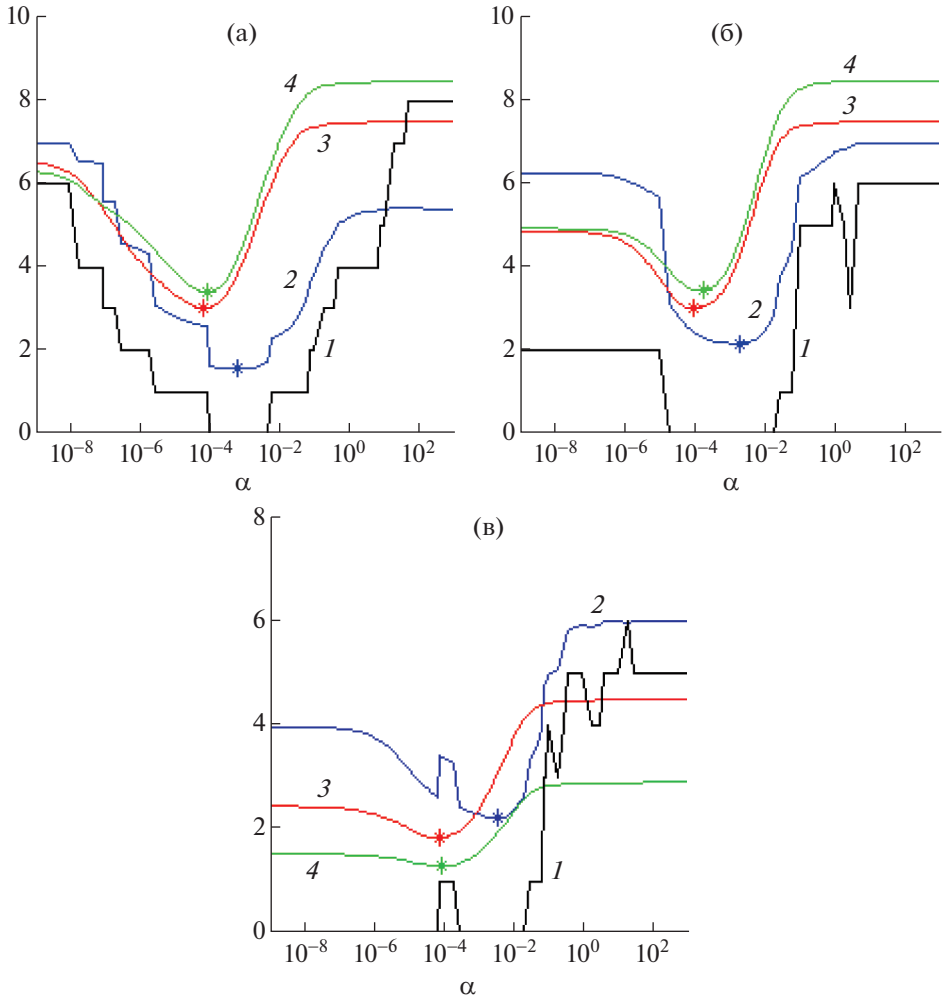
В данном разделе представлены результаты численного моделирования, демонстрирующие характерные свойства невязки при использовании реконструированного регуляризированного решения, а также эффективность предложенного метода выбора параметра регуляризации применительно к задаче адаптивной коррекции сигналов.

В качестве модели канала (функции $h(t)$) использована модель Ваттерсона (см. [18]), применяемая при моделировании коротковолновых каналов связи. При моделировании выбраны следующие параметры модели канала: 2 луча, интервал между лучами 2 мс, замирания каждого луча по закону Рэлея, используемая полоса частот 0/3–3/4 кГц. Параметры передаваемого сигнала: несущая частота сигнала $\omega_0 = 2\pi f_0$ при $f_0 = 1.8$ Гц; длительность символа $T_{\text{sym}} = 0.625$ мс. Частота дискретизации 16 кГц.

На фиг. 1 представлены характерные зависимости функционала $q(\alpha)$, невязок $r(\alpha)$, $\tilde{r}(\alpha)$, а также $\bar{r}(\alpha)$, определяемой из выражения

$$\bar{r}(\alpha) = \|\bar{H}S(\alpha) - U_\delta\|^2. \tag{4.1}$$

Отметим, что представленные на фиг. 1 зависимости $r(\alpha)$, $\tilde{r}(\alpha)$ и $\bar{r}(\alpha)$ для наглядности были нормированы.



Фиг. 1. Характерный вид зависимостей, полученных при ОСШ 5 дБ: $q(\alpha)$ (черная *1*); $r(\alpha)$ (синяя *2*); $\tilde{r}(\alpha)$ (красная *3*); $\bar{r}(\alpha)$ (зеленая *4*).

Значения минимумов для зависимостей на фиг. 1а, б отмечены знаком *.

Дополнительно для демонстрации свойств 1, 2 на фиг. 2 приведены соответствующие зависимости $\|H_\alpha S_Y(\alpha)\|^2$, $\|H_\alpha S(\alpha)\|^2$ и $\|\bar{H} S(\alpha)\|^2$.

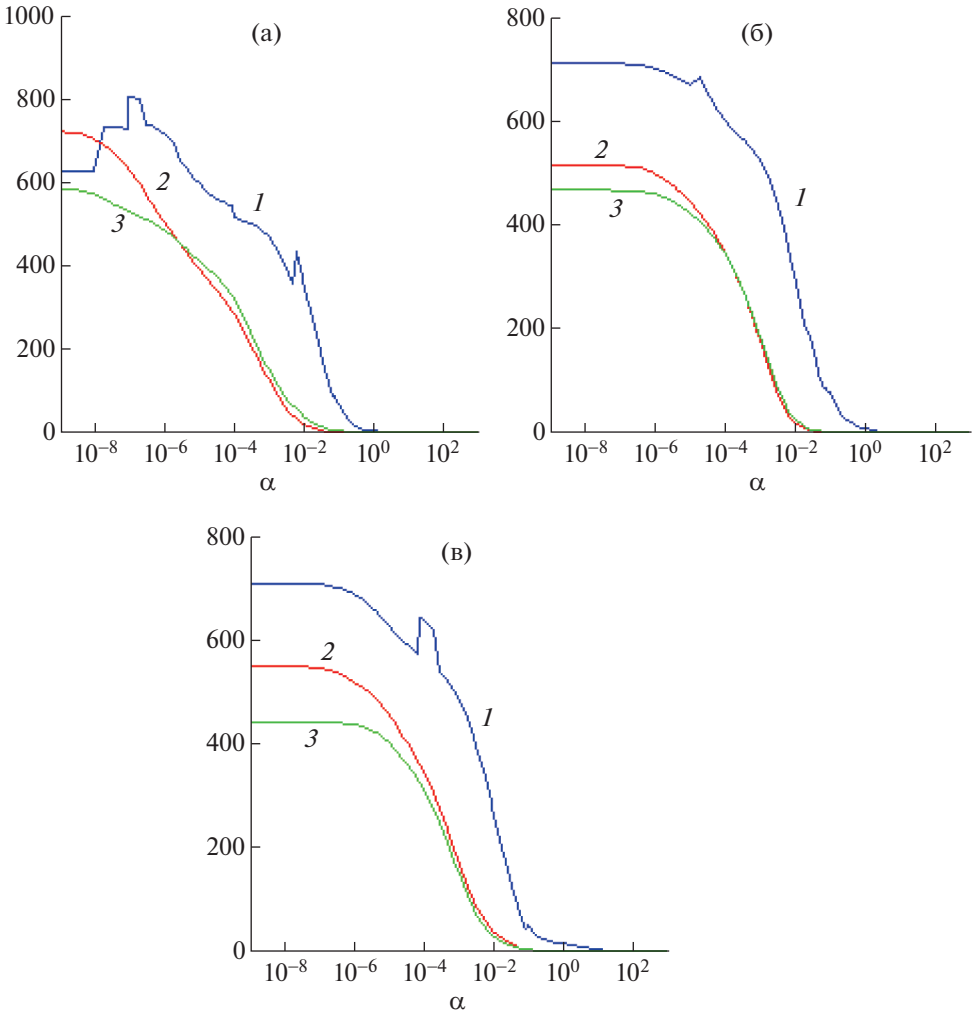
Для демонстрации эффективности применения предложенного метода выбора параметра регуляризации проведен эксперимент и получены зависимости вероятности ошибки на бит от ОСШ. Вероятность ошибки на бит определяется из выражения

$$P = \frac{1}{LN}\sum_{l=1}^{L} q(l, \hat{\alpha}_l), \tag{4.2}$$

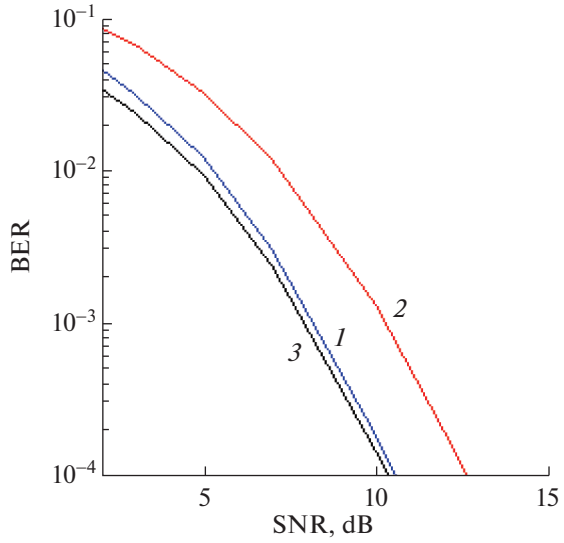
где L – объем выборки, $q(l, \hat{\alpha}_l)$ – число ошибок в l-м эксперименте, $\hat{\alpha}_l$ – выбранное значение параметра регуляризации.

При моделировании были заданы параметры $L = 20\,000$, $N = 15$. Обозначим вероятности ошибки на бит следующим образом: P – при выборе параметра регуляризации предлагаемым методом; $\tilde{P}$ – при выборе параметра регуляризации по невязке, определяемой из выражения (3.29).

Нижняя граница на фиг. 3 получена экспериментально при условии, что последовательность $b_m(n)$, $n = 1, 2, \ldots, N$, известна точно в каждом из M опытов.



Фиг. 2. Зависимости $\|H_\alpha S_Y(\alpha)\|^2$ (синяя *1*), $\|H_\alpha S(\alpha)\|^2$ (красная *2*), $\|\bar{H}S(\alpha)\|^2$ (зеленая *3*).



Фиг. 3. Зависимости вероятности ошибки на бит от ОСШ (SNR): P (синяя *1*), $\tilde{P}$ (красная *2*), нижняя граница (черная *3*).

Эффективность предложенного метода обусловлена следующим. Представим регуляризированное решение уравнения (1.1) в форме

$$S(\alpha) = \bar{S} + \eta(\alpha), \tag{4.3}$$

где $\eta(\alpha)$ назовем погрешностью (или ошибкой) вычисления, при этом

$$\|\eta(\alpha)\|^2 > 0. \tag{4.4}$$

Аналогично, для реконструированного регуляризированного решения

$$S_Y(\alpha) = \bar{S} + \eta_Y(\alpha). \tag{4.5}$$

При этом из свойства (2.17) следует, что существуют такие интервалы $\hat{T}_0$, на которых

$$\|\eta_Y(\alpha)\|^2_{\hat{T}_0} \equiv 0, \tag{4.6}$$

а значение нормы ошибки вычисления $\|\eta_Y(\alpha)\|^2_{\hat{T}e}$ тем меньше, чем меньше суммарная длительность интервала $|\hat{T}_e|$, что и соответствует минимуму функционала (2.13). При этом в ряде случаев может быть, что $|\hat{T}_e| = 0$, и соответственно $\|\eta_Y(\alpha)\|^2 \equiv 0$ на всей длительности сигнала, т.е. $\hat{T}_0 = [0; NT_{\text{sym}}]$.

5. ЗАКЛЮЧЕНИЕ

Вычислительные эксперименты подтвердили применимость и эффективность предложенного метода выбора параметра регуляризации на основе реконструкции регуляризированного решения в задаче адаптивной коррекции сигналов. Данный метод не требует знания или получения оценок о погрешности правой части уравнения (1.1) и оценки ошибки вычисления матрицы H_α. Однако требуется априорное знание определенных характеристик искомого решения S для получения оценок элементов функции $s(t)$, необходимых для формирования опорных функций.

СПИСОК ЛИТЕРАТУРЫ

1. *Eleftheriou E., Falconer D.* Adaptive equalization techniques for HF channels // IEEE J.l on Selected Areas in Communicat. 1987. V. 5. I. 2. P. 238–247.
2. *Santamarina J.C., Fratta D.* Discrete Signals and Inverse Problems. John Wiley & Sons, Ltd, 2005.
3. *Haykin S.* Adaptive Filter Theory. 5-th ed. Boston: Pearson, 2014.
4. *Маслаков М.Л.* Применение двухпараметрических стабилизирующих функций при решении интегрального уравнения типа свертки методом регуляризации // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 4. С. 541–549.
5. *Маслаков М.Л.* Выбор параметра регуляризации в задачах адаптивной фильтрации // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 6. С. 951–560.
6. *Bauer F., Lukas M.A.* Comparing parameter choice methods for regularization of ill-posed problem // Math. Comput. Simul. 2011. V. 81. I. 9. P. 1795–1841.
7. *Hansen P.C.* Rank-Deficient and Discrete Ill-Posed Problems, SIAM, Philadelphia, 1998.
8. *Lu S., Pereverzev S.V.* Regularization Theory for Ill-posed Problems. De Gruyter, Berlin, 2013.
9. *Hochstenbach M.E., Reichel L., Rodriguez G.* Regularization parameter determination for discrete ill-posed problems // J. Comput. Appl. Math. 2015. V. 273. P. 132–149.
10. *Hamarik U., Palm R., Raus T.* A family of rules for parameter choice in Tikhonov regularization of ill-posed problems with inexact noise level // J. Comput. Appl. Math. 2012. V. 236. P. 2146–2157.
11. *Гончарский А.В., Леонов А.С., Ягола А.Г.* Обобщенный принцип невязки // Ж. вычисл. матем. и матем. физ. 1973. Т. 13. № 2. С. 294–302.
12. *Тихонов А.Н., Гончарсикй А.В., Степанов В.В., Ягола А.Г.* Численные методы решения некорректных задач. М.: Наука, 1990.
13. *Сизиков В.С.* О способах невязки при решении некорректных задач // Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 9. С. 1294–1312.
14. *Johnson E.E., Koski E., Furman W.N., Jorgenson M., Nieto J.* Third-Generation and Wideband HF Radio Communications. Artech House, Inc, Boston, 2013.
15. *Xiong F.* Digital Modulation Techniques, Second Edition. Artech House, Inc, Boston, 2006.
16. *Proakis J.G., Salehi M.* Digital Communications, Fifth Edition. New York: McGraw-Hill, 2008.
17. *Морозов В.А.* О принципе невязки при решении операторных уравнений методом регуляризации // Ж. вычисл. матем. и матем. физ. 1968. Т. 8. № 2. С. 295–309.
18. *Watteson C.C., Juroshek J.R., Bensema W.D.* Experimental Confirmation of an HF Channel Model // IEEE Transact. Communicat. Technology. 1970. V. COM-18. № 6. P. 792–803.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 57–84

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.928.4

АСИМПТОТИКА КОНТРАСТНОЙ СТРУКТУРЫ ТИПА СТУПЕНЬКИ В СТАЦИОНАРНОЙ ЧАСТИЧНО ДИССИПАТИВНОЙ СИСТЕМЕ УРАВНЕНИЙ[1)]

© 2021 г. В. Ф. Бутузов

119992 Москва, Ленинские горы, МГУ, физ. ф-т, Россия

e-mail: butuzov@phys.msu.ru

Поступила в редакцию 18.06.2020 г.
Переработанный вариант 18.06.2020 г.
Принята к публикации 18.09.2020 г.

Рассматривается краевая задача для системы двух обыкновенных дифференциальных уравнений, одно из которых второго, а другое – первого порядка, с малым параметром при производных в каждом уравнении. Установлены условия, при которых существует решение этой задачи, обладающее внутренним переходным слоем в окрестности некоторой точки, где происходит быстрый переход решения из малой окрестности одного решения соответствующей вырожденной системы в малую окрестность другого решения вырожденной системы. Решение такого типа называется контрастной структурой типа ступеньки (КСТС). Построено и обосновано асимптотическое приближение КСТС по малому параметру. Оно имеет определенные отличия от КСТС в других сингулярно возмущенных задачах. Это касается, прежде всего, структуры асимптотики решения в переходном слое. Обоснование построенной асимптотики проводится с помощью асимптотического метода дифференциальных неравенств, применение которого в рассмотренной задаче также имеет свои качественные особенности. Библ. 10.

Ключевые слова: сингулярно возмущенная стационарная частично диссипативная система уравнений, контрастная структура типа ступеньки, асимптотический метод дифференциальных неравенств.

DOI: 10.31857/S0044466920120029

1. ВВЕДЕНИЕ. ПОСТАНОВКА ЗАДАЧИ

1.1. Рассмотрим систему уравнений

$$\begin{aligned} \varepsilon^2\left(\frac{d^2u}{dx^2} - w(x)\frac{du}{dx}\right) &= F(u,v,x,\varepsilon), \\ \varepsilon^2\frac{dv}{dx} &= f(u,v,x,\varepsilon), \quad x \in (0;1), \end{aligned} \tag{1}$$

с краевыми условиями

$$u(0,\varepsilon) = u^0, \quad v(0;\varepsilon) = v^0, \quad u(1,\varepsilon) = u^1. \tag{2}$$

Здесь u и v – искомые скалярные функции, $\varepsilon > 0$ – малый параметр, w, F и f – заданные достаточно гладкие функции соответственно на отрезке $0 \le x \le 1$ и в области

$$D = \{(u,v,x,\varepsilon) : u \in I_u,\ v \in I_v,\ x \in [0;1],\ \varepsilon \in [0,\varepsilon_0]\}, \tag{3}$$

где I_u, I_v – некоторые интервалы, $\varepsilon_0 > 0$.

Система вида (1) относится к классу так называемых частично диссипативных систем, поскольку член со второй производной (диффузионный член) содержится только в одном уравнении. Такие системы возникают, в частности, в стационарных задачах химической кинетики в

[1)]Работа выполнена при финансовой поддержке РФФИ, грант № 18-01-00424.

случае быстрых реакций. В этом случае u и v – концентрации реагирующих веществ, ε^{-2} – так называемая константа скорости быстрой реакции (большая величина).

При $\varepsilon = 0$ из (1) получаем вырожденную систему

$$F(u, v, x, 0) = 0, \quad f(u, v, x, 0) = 0. \tag{4}$$

Цель работы – доказать, что при определенных условиях существует решение задачи (1), (2) с переходным слоем в окрестности некоторой внутренней точки x_* отрезка $0 \le x \le 1$ (точки перехода), где решение задачи совершает резкий переход из малой окрестности одного решения вырожденной системы (4) в малую окрестность другого решения системы (4) (образуется "ступенька"). Такое решение называется *контрастной структурой типа ступеньки* (КСТС).

Наряду с доказательством существования КСТС будет построено ее асимптотическое приближение по малому параметру ε.

Отметим, что в [1] для системы (1) с краевыми условиями (2) построена и обоснована асимптотика пограничного решения, т.е. такого решения, которое при $\varepsilon \to 0$ стремится на всем интервале $0 < x < 1$ к одному и тому же решению вырожденной системы (4) и отлично от него только в малых окрестностях точек $x = 0$ и $x = 1$ (пограничных слоях). Результаты работы [1] будут использоваться в данной работе как при построении асимптотики КСТС, так и при ее обосновании, поскольку искомая КСТС будет получена в результате объединения двух пограничных решений системы (1), построенных раздельно на отрезках $[0, x_*]$ и $[x_*, 1]$.

Опишем кратко структуру работы.

В п. 1.2 представлены условия, которые будут обеспечивать существования искомой КСТС в задаче (1), (2). В разд. 2 при этих условиях построена формальная асимптотика КСТС, причем построение ведется раздельно на отрезках $[0, x_*]$ и $[x_*, 1]$, где x_* – искомая точка перехода, а затем в результате сшивания в точке x_* формальных асимптотик, построенных на этих двух отрезках, получено представление x_* в виде асимптотического ряда по степеням ε. В разд. 3 и 4 рассмотрены две вспомогательные краевые задачи для системы (1) соответственно на отрезках $[0, x_\delta]$ и $[x_\delta, 1]$, где точка x_δ выбирается с использованием ряда для x_*, полученного в разд. 2. Доказано существование решений этих задач, обладающих построенной в разд. 2 асимптотикой. В разд. 5 показано, что точку x_δ можно выбрать так, что функции $u(x, \varepsilon)$ и $v(x, \varepsilon)$, составленные из решений двух вспомогательных задач, образуют искомую КСТС. В разд. 6 содержатся некоторые замечания в отношении рассмотренной задачи, а также других возможных задач о контрастных структурах в частично диссипативных системах уравнений.

Отметим, что контрастные структуры в различных сингулярно возмущенных задачах исследовались во многих работах, например, [2]–[8]. Асимптотика КСТС в данной работе имеет свои качественные особенности, относящиеся, прежде всего, к переходному слою.

1.2. Условия

Сформулируем условия, при которых будет доказано существование КСТС в задаче (1), (2) (для достаточно малых ε) и построено асимптотическое приближение КСТС.

В п. 1.1 говорилось о достаточной гладкости заданных функций w, F, f. Как обычно, требуемый порядок гладкости зависит от порядка асимптотики, которую хотят построить. Поскольку речь пойдет об асимптотике произвольного порядка, будем считать эти функции бесконечно дифференцируемыми.

Условие A1. $w(x) \in C^\infty[0; 1]$, $F \in C^\infty(D)$, $f \in C^\infty(D)$,

где область D определена в (3), и пусть $u^0 \in I_u$, $v^0 \in I_v$, $u^1 \in I_u$, где I_u и I_v – интервалы, фигурирующие в определении области D.

Следующее условие относится к вырожденной системе (4).

Условие A2. Уравнение

$$f(u, v, x, 0) = 0$$

имеет бесконечно дифференцируемый простой (т.е. однократный) корень

$$v = \varphi(u, x) \in I_v \quad \text{при} \quad u \in I_u, \quad x \in [0; 1],$$

а уравнение

$$g(u,x) := F(u,\varphi(u,x),x,0) = 0 \tag{5}$$

имеет ровно три бесконечно дифференцируемых простых корня

$$u = \psi_i(x), \quad x \in [0;1], \quad i = 1,2,3,$$

причем

$$\psi_1(x) < \psi_2(x) < \psi_3(x), \quad \psi_i(x) \in I_u \quad \text{при} \quad x \in [0;1]. \tag{6}$$

Следующее условие относится к уравнению относительно x_0:

$$I(x_0) := \int_{\psi_1(x_0)}^{\psi_3(x_0)} g(u,x_0)du = 0. \tag{7}$$

Условие А3. Уравнение (7) имеет корень $x_0 = \bar{x}_0 \in (0;1)$, и

$$I'(\bar{x}_0) \neq 0. \tag{8}$$

Забегая вперед, отметим, что искомая точка перехода x_* будет иметь представление

$$x_* = \bar{x}_0 + O(\varepsilon).$$

Остальные условия связаны с производными функций g, f, F. Чтобы сформулировать эти условия, определим несколько кривых на плоскости переменных (u,x) и в пространстве переменных (u,v,x).

Кривые на плоскости (u,x):

$$l_1 = \{(u,x) : u \in [u^0, \psi_1(0)],\ x = 0\},$$
$$l_2 = \{(u,x) : u = \psi_1(x),\ x \in [0,\ \bar{x}_0]\},$$
$$l_3 = \{(u,x) : u \in [\psi_1(\bar{x}_0), \psi_2(\bar{x}_0)],\ x = \bar{x}_0\},$$
$$l_4 = \{(u,x) : u \in [\psi_2(\bar{x}_0), \psi_3(\bar{x}_0)],\ x = \bar{x}_0\},$$
$$l_5 = \{(u,x) : u = \psi_3(x),\ x \in [\bar{x}_0,\ 1]\},$$
$$l_6 = \{(u,x) : u \in [\psi_3(1),\ u^1],\ x = 1\}, \quad l = \bigcup_{i=1}^{6} l_i.$$

Отметим, что кривые l_i $(i = 1, 2, \ldots, 6)$ являются гладкими, а кривая l – непрерывная кривая, составленная из шести гладких звеньев $l_1, \ldots, l_6$.

Кривые в пространстве (u,v,x):

$$L_0 = \{(u,v,x) : u = u^0,\ v \in [v^0, \varphi(u^0,0)],\ x = 0\},$$
$$L_i\{(u,v,x) : v = \varphi(u,x), (u,x) \in l_i\}, \quad i = 1,2,\ldots,6, \tag{9}$$
$$L^{(-)} = \bigcup_{i=0}^{3} L_i, \quad L^{(+)} = \bigcup_{i=4}^{6} L_i, \quad L = L^{(-)} \cup L^{(+)}.$$

Отметим также, что некоторые из введенных кривых могут вырождаться в точку. Например, если $v^0 = \varphi(u^0,0)$, то отрезок L_0 вырождается в точку $(u^0,v^0,0)$, которая является одним из концов кривой L_1. Для определенности будем считать, что L_0 – невырожденный отрезок.

Сформулируем теперь остальные условия.

Условие А4. $\dfrac{\partial g}{\partial u}(u,x) > 0$ в точках кривых $l_1 \cup l_2$ и $l_5 \cup l_6$.

Условие А5. $\dfrac{\partial f}{\partial v}(u,v,x,0) < 0$ в точках кривой $\bigcup_{i=1}^{6} L_i$, и $f(u,v,x,0) \neq 0$ на отрезке L_0, за исключением его конца $(u^0,\varphi(u^0,0),0)$.

Условие А6. $\dfrac{\partial F}{\partial v}(u,v,x,0) < 0$ в точках кривой L.

Условие А7. $\dfrac{\partial f}{\partial u}(u,v,x,0) > 0$ в точках кривой L.

Условие А8. $R^{(-)}(u,\bar{x}_0) := \bar{F}_u^{(-)}(\bar{x}_0) + \bar{F}_v^{(-)}(\bar{x}_0)\varphi_u(u,\bar{x}_0) > 0$ при $\psi_1(\bar{x}_0) \le u \le \psi_2(\bar{x}_0)$, т.е. в точках кривой l_3;

$R^{(+)}(u,\bar{x}_0) := \bar{F}_u^{(+)}(\bar{x}_0) + \bar{F}_v^{(+)}(\bar{x}_0)\varphi_u(u,\bar{x}_0) > 0$ при $\psi_2(\bar{x}_0) \le u \le \psi_3(\bar{x}_0)$, т.е. в точках кривой l_4; здесь

$$\begin{aligned} \bar{F}_u^{(-)}(x) &= \frac{\partial F}{\partial u}\big(\psi_1(x), \varphi(\psi_1(x),x), x, 0\big), \\ \bar{F}_v^{(-)}(x) &= \frac{\partial F}{\partial v}\big(\psi_1(x), \varphi(\psi_1(x),x), x, 0\big), \end{aligned} \tag{10}$$

$$\begin{aligned} \bar{F}_u^{(+)}(x) &= \frac{\partial F}{\partial u}\big(\psi_3(x), \varphi(\psi_3(x),x), x, 0\big), \\ \bar{F}_v^{(+)}(x) &= \frac{\partial F}{\partial v}\big(\psi_3(x), \varphi(\psi_3(x),x), x, 0\big), \quad \varphi_u(u,x) = \frac{\partial \varphi}{\partial u}(u,x). \end{aligned} \tag{11}$$

Приведем простой пример функций F и f, удовлетворяющих условиям А1–А8:

$$F(u,v,x,\varepsilon) = g(u,x) + u - v + \varepsilon F_1(u,v,x,\varepsilon),$$
$$f(u,v,x,\varepsilon) = u - v + \varepsilon f_1(u,v,x,\varepsilon),$$

где

$$g(u,x) = \big(u - \psi_1(x)\big)\big(u - \psi_2(x)\big)\big(u - \psi_3(x)\big),$$

причем выполнены неравенства (6) и условие А3, а граничные значения u^0 и u^1 достаточно близки соответственно к $\psi_1(0)$ и $\psi_3(1)$.

Заметим, что если в определениях кривых l и L заменить $\bar{x}_0$ на x_*, то неравенства в условиях А4–А8 останутся верными для всех значений x_* из некоторой достаточно малой и независящей от ε окрестности точки $\bar{x}_0$. Будем этим пользоваться при построении формальной асимптотики КСТС в разд. 2.

2. ПОСТРОЕНИЕ ФОРМАЛЬНОЙ АСИМПТОТИКИ КСТС

2.1. Вид асимптотики

Возьмем произвольное значение x_* из указанной в конце п. 1.2 достаточно малой окрестности точки $\bar{x}_0$ и будем строить формальную асимптотику КСТС в задаче (1), (2) в виде

$$U(x,\varepsilon) = \begin{cases} U^{(-)}(x,\varepsilon), & x \in [0, x_*], \\ U^{(+)}(x,\varepsilon), & x \in [x_*, 1], \end{cases} \quad V(x,\varepsilon) = \begin{cases} V^{(-)}(x,\varepsilon), & x \in [0, x_*], \\ V^{(+)}(x,\varepsilon), & x \in [x_*, 1], \end{cases}$$

где

$$U^{(-)}(x,\varepsilon) = \bar{u}^{(-)}(x,\varepsilon) + \Pi^{(-)}u(\xi,\varepsilon) + P^{(-)}u(\zeta,\varepsilon) + Q^{(-)}u(\sigma,\varepsilon), \tag{12}$$

$$V^{(-)}(x,\varepsilon) = \bar{v}^{(-)}(x,\varepsilon) + \Pi^{(-)}v(\xi,\varepsilon) + P^{(-)}v(\zeta,\varepsilon) + Q^{(-)}v(\sigma,\varepsilon), \tag{13}$$

$$U^{(+)}(x,\varepsilon) = \bar{u}^{(+)}(x,\varepsilon) + Q^{(+)}u(\sigma,\varepsilon) + \Pi^{(+)}u(\tilde{\xi},\varepsilon), \tag{14}$$

$$V^{(+)}(x,\varepsilon) = \bar{v}(x,\varepsilon) + Q^{(+)}v(\sigma,\varepsilon) + \Pi^{(+)}v(\tilde{\xi},\varepsilon), \tag{15}$$

$\bar{u}^{(\pm)}$, $\bar{v}^{(\pm)}$ – регулярные части асимптотики; $\Pi^{(-)}u$, $\Pi^{(-)}v$ и $P^{(-)}u$, $P^{(-)}v$ – пограничные части, описывающие пограничное поведение решения в окрестности точки $x = 0$, $\xi = x/\varepsilon$ и $\zeta = x/\varepsilon^2$ – пограничные переменные; $Q^{(-)}u$, $Q^{(-)}v$ и $Q^{(+)}u$, $Q^{(+)}v$ – внутренние части асимптотики, опи-

сывающие поведение решения в окрестности точки перехода x_* (во внутреннем переходном слое) слева и справа от точки x_*, $\sigma = (x - x_*)/\varepsilon$ – внутрислойная переменная; $\Pi^{(+)}u$, $\Pi^{(+)}v$ – пограничные части асимптотики в окрестности точки $x = 1$, $\xi = (x-1)/\varepsilon$ – пограничная переменная.

Точку x_* определим условием

$$U^{(-)}(x_*, \varepsilon) = U^{(+)}(x_*, \varepsilon) = \psi_2(x_*). \tag{16}$$

Все слагаемые в правых частях (12)–(15) будут построены в виде рядов по целым степеням ε с помощью известного алгоритма А.Б. Васильевой (см. [9]). При этом будут использоваться краевые условия, вытекающие из (2) и (16):

$$U^{(-)}(0, \varepsilon) = u^0, \quad V^{(-)}(0, \varepsilon) = v^0, \quad U^{(-)}(x_*, \varepsilon) = \psi_2(x_*), \tag{17}$$

$$U^{(+)}(x_*, \varepsilon) = \psi_2(x_*), \quad U^{(+)}(1, \varepsilon) = u^1. \tag{18}$$

2.2. Построение асимптотики на отрезке $[0, x_*]$

На отрезке $[0, x_*]$ асимптотика $U^{(-)}(x, \varepsilon)$, $V^{(-)}(x, \varepsilon)$ вида (12), (13) является асимптотикой пограничного типа. Построение такой асимптотики подробно описано в [1], поэтому ограничимся здесь более кратким описанием.

2.2.1. Регулярные части асимптотики. Построим их в виде

$$\overline{u}^{(-)}(x, \varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i \overline{u}_i^{(-)}(x), \quad \overline{v}^{(-)}(x, \varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i \overline{v}_i^{(-)}(x). \tag{19}$$

Стандартным способом, т.е. подставив ряды (19) в систему (1) вместо u и v, разложив правые части уравнений в ряды по степеням ε и приравняв коэффициенты при одинаковых степенях ε в левой и правой частях каждого уравнения, получим последовательно для $i = 0, 1, 2, \ldots$ системы уравнений относительно $u_i^{(-)}(x)$, $\overline{v}_i^{(-)}(x)$. Для $\overline{u}_0^{(-)}(x)$, $\overline{v}_0^{(-)}(x)$ получается вырожденная система (4):

$$F(\overline{u}_0^{(-)}, \overline{v}_0^{(-)}, x, 0) = 0, \quad f(\overline{u}_0^{(-)}, \overline{v}_0^{(-)}, x, 0) = 0.$$

В качестве ее решения возьмем (см. условие А2)

$$\overline{u}_0^{(-)}(x) = \psi_1(x), \quad \overline{v}_0^{(-)}(x) = \varphi\big(\psi_1(x), x\big), \quad x \in [0; x_*].$$

Для $\overline{u}_i^{(-)}(x)$, $\overline{v}_I^{(-)}(x)$ при $i \geq 1$ получается система линейных уравнений

$$\begin{aligned} \overline{F}_u^{(-)}(x)\overline{u}_i^{(-)} + \overline{F}_v^{(-)}(x)\overline{v}_i^{(-)} &= F_i^{(-)}(x), \\ \overline{f}_u^{(-)}(x)\overline{u}_i^{(-)} + \overline{f}_v^{(-)}(x)\overline{v}_i^{(-)} &= f_i^{(-)}(x), \end{aligned} \tag{20}$$

где $\overline{F}_u^{(-)}(x)$ и $\overline{F}_v^{(-)}(x)$ определены в (10), $\overline{f}_u^{(-)}(x)$ и $\overline{f}_v^{(-)}(x)$ имеют аналогичные выражения, а функции $F_i(x)$ и $f_i(x)$ выражаются рекуррентно через $\overline{u}_j^{(-)}(x)$, $\overline{v}_j^{(-)}(x)$ с номерами $j < i$.

Определитель $\Delta^{(-)}(x)$ системы (20) запишем в виде

$$\Delta^{(-)}(x) = \overline{F}_u^{(-)}(x)\overline{f}_v^{(-)}(x) - \overline{F}_v^{(-)}(x)\overline{f}_u^{(-)}(x) = \overline{f}_v^{(-)}(x)\overline{g}_u^{(-)}(x),$$

где

$$\overline{g}_u^{(-)}(x) := \frac{\partial g}{\partial u}(\psi_1(x), x).$$

Так как $\overline{f}_v^{(-)}(x) = \dfrac{\partial f}{\partial v}(u, v, x, 0)$ при $(u, v, x) \in L_2$, т.е. производная $\overline{f}_v^{(-)}(x)$ вычисляется в точках кривой L_2, то в силу условия А5 справедливо неравенство

$$\overline{f}_v^{(-)}(x) < 0, \quad x \in [0; x_*].$$

Аналогично, $\overline{g}_u^{(-)}(x) = \dfrac{\partial g}{\partial u}(u,x)$ при $(u,x) \in l_2$, поэтому в силу условия A4

$$\overline{g}_u^{(-)}(x) > 0, \quad x \in [0; x_*].$$

Следовательно, $\Delta^{(-)}(x) < 0$, $x \in [0, x_*]$, и, значит, система (20) имеет единственное решение.

Таким образом, ряды (19) построены.

2.2.2. Погранслойные части асимптотики $\Pi^{(-)}\mathbf{u}$, $\Pi^{(-)}\mathbf{v}$ и $\mathbf{P}^{(-)}\mathbf{u}$, $\mathbf{P}^{(-)}\mathbf{v}$. Построим их в виде

$$\begin{gathered}\Pi^{(-)}u(\xi,\varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i\Pi_i^{(-)}u(\xi),\\ \Pi^{(-)}v(\xi,\varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i\Pi_i^{(-)}v(\xi), \quad \xi = x/\varepsilon \geq 0;\end{gathered} \tag{21}$$

$$\begin{gathered}P^{(-)}u(\zeta,\varepsilon) = \varepsilon^2\sum_{i=0}^{\infty}\varepsilon^i P_i^{(-)}u(\zeta),\\ P^{(-)}v(\zeta,\varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i P_i^{(-)}v(\zeta), \quad \zeta = x/\varepsilon^2 \geq 0.\end{gathered} \tag{22}$$

Стандартным способом (см. [9]) для $\Pi^{(-)}u$, $\Pi^{(-)}v$ получается система уравнений

$$\begin{gathered}\frac{d^2\Pi^{(-)}u}{d\xi^2} - \varepsilon w(\varepsilon\xi)\frac{d\Pi^{(-)}u}{d\xi} = \Pi^{(-)}F := F(\overline{u}^{(-)}(\varepsilon\xi,\varepsilon) + \Pi^{(-)}u, \overline{v}^{(-)}(\varepsilon\xi,\varepsilon) +\\ + \Pi^{(-)}v, \varepsilon\xi, \varepsilon) - F(\overline{u}^{(-)}(\varepsilon\xi,\varepsilon), \overline{v}^{(-)}(\varepsilon\xi,\varepsilon), \varepsilon\xi, \varepsilon),\\ \varepsilon\frac{d\Pi^{(-)}v}{d\xi} = \Pi^{(-)}f, \quad \xi \geq 0,\end{gathered} \tag{23}$$

где $\Pi^{(-)}f$ имеет выражение, аналогичное $\Pi^{(-)}F$, а для $P^{(-)}u$, $P^{(-)}v$ получается система уравнений

$$\begin{gathered}\frac{1}{\varepsilon^2}\frac{d^2P^{(-)}u}{d\zeta^2} - w(\varepsilon^2\zeta)\frac{dP^{(-)}u}{d\zeta} = P^{(-)}F := F(\overline{u}^{(-)}(\varepsilon^2\zeta,\varepsilon) + \Pi^{(-)}u(\varepsilon\zeta,\varepsilon) +\\ + P^{(-)}u, \overline{v}^{(-)}(\varepsilon^2\zeta,\varepsilon) + \Pi^{(-)}v(\varepsilon\zeta,\varepsilon) + P^{(-)}v, \varepsilon^2\zeta, \varepsilon) - F(\overline{u}^{(-)}(\varepsilon^2\zeta,\varepsilon) +\\ + \Pi^{(-)}u(\varepsilon\zeta,\varepsilon), \overline{v}^{(-)}(\varepsilon^2\zeta,\varepsilon) + \Pi^{(-)}v(\varepsilon\zeta,\varepsilon), \varepsilon^2\zeta, \varepsilon),\\ \frac{dP^{(-)}v}{d\zeta} = P^{(-)}f, \quad \zeta \geq 0,\end{gathered} \tag{24}$$

где $P^{(-)}f$ имеет выражение, аналогичное $P^{(-)}F$.

Из (23) будем извлекать последовательно для $i = 0, 1, 2, \ldots$ уравнения относительно $\Pi_i^{(-)}u$, $\Pi_i^{(-)}v$, а из (24) – уравнения относительно $P_i^{(-)}u$, $P_i^{(-)}v$. Для каждого i эти функции будут определяться в таком порядке:

$$\Pi_i^{(-)}u \to \Pi_i^{(-)}v \to P_i^{(-)}v \to P_i^{(-)}u.$$

Для $\Pi_0^{(-)}u$, $\Pi_0^{(-)}v$ из (23) следует система уравнений

$$\begin{gathered}\frac{d^2\Pi_0^{(-)}u}{d\xi^2} = F(\overline{u}_0^{(-)}(0) + \Pi_0^{(-)}u, \overline{v}_0^{(-)}(0) + \Pi_0^{(-)}v, 0, 0),\\ 0 = f(\overline{u}_0^{(-)}(0) + \Pi_0^{(-)}u, \overline{v}_0^{(-)}(0) + \Pi_0^{(-)}v, 0, 0), \quad \xi \geq 0.\end{gathered} \tag{25}$$

Из второго уравнения, используя условие A2, получаем

$$\overline{v}_0^{(-)}(0) + \Pi_0^{(-)}v = \varphi(\overline{u}_0^{(-)}(0) + \Pi_0^{(-)}u, 0). \tag{26}$$

Подставляя в первое уравнение, приходим к уравнению для $\Pi_0^{(-)}u$:

$$\frac{d^2\Pi_0^{(-)}u}{d\xi^2} = g(\bar{u}_0^{(-)}(0) + \Pi_0 u, 0), \quad \xi \geq 0. \tag{27}$$

К этому уравнению нужно добавить граничные условия.

Чтобы получить граничное условие при $\xi = 0$, подставим выражение (12) для $U^{(-)}(x,\varepsilon)$ в граничное условие $U^{(-)}(0,\varepsilon) = u^0$ (см. (17)) с учетом того, что все члены ряда $Q^{(-)}u(\xi,\varepsilon)$ равны нулю при $x = 0$ (см. замечание 1 в конце пп. 2.2.3). Получим равенство

$$\sum_{i=0}^{\infty} \varepsilon^i (\bar{u}_i^{(-)}(0) + \Pi_i^{(-)}u(0) + \varepsilon^2 P_i^{(-)}u(0)) = u^0. \tag{28}$$

Отсюда имеем $\bar{u}_0^{(-)}(0) + \Pi_0^{(-)}u(0) = u^0$, и, следовательно, граничное условие для $\Pi_0^{(-)}u(\xi)$ при $\xi = 0$ имеет вид

$$\Pi_0^{(-)}u(0) = u^0 - \bar{u}_0(0) = u^0 - \psi_1(0). \tag{29}$$

В качестве второго граничного условия для $\Pi_0^{(-)}u(\xi)$ и также для остальных функций $\Pi_i^{(-)}u(\xi)$ возьмем стандартное для пограничных функций условие на бесконечности

$$\Pi_i^{(-)}u(\infty) = 0, \quad i = 0, 1, 2, \ldots . \tag{30}$$

Заметим, что $g(\bar{u}_0(0),0) = g(\psi_1(0),0) = 0$ в силу условия A2, поэтому, если $u^0 = \psi_1(0)$, то $\Pi_0^{(-)}u(\xi) = 0$ при $\xi \geq 0$.

Если же $u^0 \neq \psi_1(0)$, то воспользуемся тем, что в силу условия A4 производная $\dfrac{\partial g}{\partial u}(u,x) > 0$ на кривой l_1, т.е. при $\{u \in [u^0, \psi_1,(0)],\ x = 0\}$, и, следовательно, $g(\bar{u}_0(0) + \Pi_0^{(-)}u, 0) \neq 0$ при $\Pi_0^{(-)}u \in [u^0 - \psi_1(0), 0]$. Поэтому задача для $\Pi_0^{(-)}u$ сводится стандартным образом к уравнению первого порядка

$$\frac{d\Pi_0^{(-)}u}{d\xi} = \pm\left[2\int_0^{\Pi_0^{(-)}u} g(\psi_1(0) + s, 0)ds\right]^{1/2}, \quad \xi \geq 0, \tag{31}$$

с начальным условием (29), причем в правой части (31) берется знак плюс, если $u^0 < \psi_1(0)$, и знак минус, если $u^0 > \psi_1(0)$. Уравнение (31) интегрируется в квадратурах, функция $\Pi_0^{(-)}u(\xi)$ является монотонной функцией при $\xi \geq 0$ и имеет экспоненциальную оценку

$$\left|\Pi_0^{(-)}u(\xi)\right| \leq c\exp(-\kappa\xi), \quad \xi \geq 0. \tag{32}$$

Такого же вида оценка верна для производной $\dfrac{d\Pi_0^{(-)}u}{d\xi}(\xi)$ и функции $\Pi_0^{(-)}v(\xi)$, которая определяется теперь из (26).

Здесь и в дальнейшем буквами c и κ (иногда через c_1, κ_1, …) обозначаются не зависящие от ε подходящие положительные числа, вообще говоря, различные в разных оценках.

Для $P_0^{(-)}u$, $P_0^{(-)}v$ из (24) получаем систему уравнений

$$\frac{d^2P_0^{(-)}u}{d\zeta^2} = F(u^0, \varphi(u^0,0) + P_0^{(-)}v, 0, 0) - F(u^0, \varphi(u^0,0), 0, 0), \tag{33}$$

$$\frac{dP_0^{(-)}v}{d\zeta} = f(u^0, \varphi(u^0,0) + P_0^{(-)}v, 0, 0), \quad \zeta \geq 0. \tag{34}$$

Зададим для $P_0^{(-)}u(\zeta)$ граничное условие на бесконечности

$$P_0^{(-)}u(\infty) = 0, \tag{35}$$

а для $P_0^{(-)}v(\zeta)$ – начальное условие при $\zeta = 0$. Чтобы его получить, подставим выражение (13) для $V^{(-)}(x,\varepsilon)$ в граничное условие $V^{(-)}(0,\varepsilon) = v^0$ (см. (17)), учитывая, что все члены ряда $Q^{(-)}v(\xi,\varepsilon)$ равны нулю при $x = 0$ (см. замечание 1 в конце пп. 2.2.3). Получим равенство

$$\sum_{i=0}^{\infty} \varepsilon^i (\overline{v}_i^{(-)}(0) + \Pi_i^{(-)}v(0) + P_i^{(-)}v(0)) = v^0. \tag{36}$$

Отсюда имеем

$$P_0^{(-)}v(0) = v^0 - (\overline{v}_0^{(-)}(0) + \Pi_0^{(-)}v(0)) = v^0 - \varphi(u^0,0) =: P_0^{(-)}. \tag{37}$$

Заметим, что $f(u^0,\varphi(u^0,0),0,0) = 0$ в силу условия A2, и, значит, $P_0^{(-)}v = 0$ является точкой покоя уравнения (34) асимптотически устойчивой в силу неравенства $\dfrac{\partial f}{\partial v}(u^0,\varphi(u^0,0),0,0) < 0$ (см. условие A5). Если $v^0 = \varphi(u^0,0)$, то $P_0^{(-)} = 0$, и тогда $P_0^{(-)}v(\zeta) = 0$ при $\zeta \geq 0$. Если же $P_0^{(-)} \neq 0$, то $f(u^0,\varphi(u^0,0) + s,0,0) \neq 0$ при $s \in (0, P_0^{(-)}]$ в силу условия A5, поэтому решение задачи (34), (37) является монотонной функцией и имеет экспоненциальную оценку

$$\left|P_0^{(-)}v(\zeta)\right| \leq c\exp(-\kappa\zeta), \quad \zeta \geq 0. \tag{38}$$

Так как функция $P_0^{(-)}v(\zeta)$ найдена, то правая часть уравнения (33) является теперь известной функцией, имеющей такую же экспоненциальную оценку, как (38). Обозначив эту функцию $\chi_0^{(-)}(\zeta)$, запишем решение уравнения (33) с граничным условием (35) в виде

$$P_0^{(-)}u(\zeta) = \int_{\infty}^{\zeta} ds \int_{\infty}^{s} \chi_0^{(-)}(t)dt. \tag{39}$$

Отсюда следует, что $P_0^{(-)}u(\zeta)$ и ее производная $\dfrac{dP_0^{(-)}u}{d\zeta}(\zeta)$ имеют оценки вида (38).

Таким образом, главные члены погранслойных рядов (21) и (22) определены и имеют оценки вида (32) и (38).

При $i \geq 1$ для $\Pi_i^{(-)}u$, $\Pi_i^{(-)}v$ из (23) получается система уравнений

$$\begin{gathered} \frac{d^2\Pi_i^{(-)}u}{d\xi^2} = F_u^{(-)}(\xi)\Pi_i^{(-)}u + F_v^{(-)}(\xi)\Pi_i^{(-)}v + r_i^{(-)}(\xi), \\ f_u^{(-)}(\xi)\Pi_i^{(-)}u + f_v^{(-)}(\xi)\Pi_i^{(-)}v + \varrho_i^{(-)}(\xi) = 0, \end{gathered} \tag{40}$$

где

$$F_u^{(-)}(\xi) := \frac{\partial F}{\partial u}(\overline{u}_0^{(-)}(0) + \Pi_0^{(-)}u(\xi), \overline{v}_0^{(-)}(0) + \Pi_0^{(-)}v(\xi), 0, 0), \tag{41}$$

и такой же смысл имеют обозначения $F_v^{(-)}(\xi)$, $f_u^{(-)}(\xi)$, $f_v^{(-)}(\xi)$, а $r_i^{(-)}(\xi)$ и $\varrho_i^{(-)}(\xi)$ – известные на i-м шаге функции, рекуррентно выражающиеся через уже найденные функции $\Pi_j^{(-)}u(\xi)$, $\Pi_j^{(-)}v(\xi)$ с номерами $j < i$ и имеющие экспоненциальные оценки вида (32), если такие же оценки имеют функции $\Pi_j^{(-)}u$, $\dfrac{d\Pi_j^{(-)}u}{d\xi}$, $\Pi_j^{(-)}v$ с номерами $j < i$.

Так как

$$f_v^{(-)}(\xi) := \frac{\partial f}{\partial v}(\bar{u}_0(0) + \Pi_0^{(-)}u(\xi), \varphi(\bar{u}_0^{(-)}(0) + \Pi_0^{(-)}u(\xi), 0), 0, 0) = \\ = \frac{\partial f}{\partial v}(u, \varphi(u, 0), 0, 0) \quad \text{при} \quad u = \bar{u}_0(0) + \Pi_0^{(-)}u(\xi), \tag{42}$$

и так как

$$u = (\bar{u}_0^{(-)}(0) + \Pi_0^{(-)}u(\xi)) \in [u^0, \psi_1(0)) \quad \text{при} \quad \xi \geq 0$$

(в силу монотонности $\Pi_0^{(-)}u(\xi)$ при $\xi \geq 0$), то значения производной $f_v^{(-)}(\xi)$ при $\xi \geq 0$ совпадают со значениями $\frac{\partial f}{\partial v}(u, v, x, 0)$ на кривой L_1. Поэтому в силу условия A5

$$f_v^{(-)}(\xi) \leq -\kappa < 0 \quad \text{при} \quad \xi \geq 0. \tag{43}$$

Это дает возможность выразить $\Pi_i^{(-)}v$ через $\Pi_i^{(-)}u$ из второго уравнения системы (40):

$$\Pi_i^{(-)}v = \varphi_u^{(-)}(\xi)\Pi_i^{(-)}u - (f_v^{(-)}(\xi))^{-1}\varrho_i^{(-)}\xi), \tag{44}$$

где

$$\varphi_u^{(-)}(\xi) := -(f_v^{(-)}(\xi))^{-1}f_u^{(-)}(\xi) = \frac{\partial \varphi}{\partial u}(\bar{u}_0^{(-)}(0) + \Pi_0^{(-)}u(\xi), 0). \tag{45}$$

Подставляя выражение (44) в первое уравнение системы (40), приходим к уравнению для $\Pi_i^{(-)}u(\xi)$:

$$\frac{d^2\Pi_i^{(-)}u}{d\xi^2} = g_u^{(-)}(\xi)\Pi_i^{(-)}u + \pi_i^{(-)}(\xi), \quad \xi \geq 0, \tag{46}$$

где

$$g_u^{(-)}(\xi) := \frac{\partial g}{\partial u}(u_0^{(-)}(0) + \Pi_0^{(-)}u(\xi), 0),$$

$\pi_i^{(-)}(\xi)$ – известная функция, имеющая оценку вида (32).

Из (28) и (30) получаем граничные условия для $\Pi_i^{(-)}u(\xi)$:

$$\Pi_i^{(-)}u(0) = -\bar{u}_i(0) - P_{i-2}^{(-)}u(0) =: \Pi_i^0, \quad \Pi_i^{(-)}u(\infty) = 0, \tag{47}$$

где $P_{i-2}^{(-)}u(0)$ – известное на i-м шаге число, в частности, $P_{-1}^{(-)}u(0)$ считаем равным нулю. Решение задачи (46), (47) запишем в виде

$$\Pi_i^{(-)}u(\xi) = \Phi(\xi)\Phi^{-1}(0)\Pi_i^0 + \Phi(\xi)\int_0^\xi \Phi^{-2}(s)\int_\infty^s \Phi(t)\pi_i^{(-)}(t)dtds, \tag{48}$$

где $\Phi(\xi) = \frac{d\Pi_0^{(-)}u}{d\xi}(\xi)$. Используя оценки вида (32) для $\Phi(\xi)$ и $\pi_i^{(-)}(\xi)$, из (48) получаем экспоненциальную оценку для $\Pi_i^{(-)}u(\xi)$:

$$\left|\Pi_i^{(-)}u(\xi)\right| \leq c\exp(-\kappa\xi), \quad \xi \geq 0. \tag{49}$$

Такую же оценку имеют производная $\frac{d\Pi_i^{(-)}u}{d\xi}(\xi)$ и функция $\Pi_i^{(-)}v(\xi)$, которая определяется теперь равенством (44).

Перейдем к функциям $P_i^{(-)}u(\zeta)$, $P_i^{(-)}v(\zeta)$ при $i \geq 1$. Для них из (24) получается система уравнений

$$\frac{d^2P_i^{(-)}u}{d\zeta^2} = \hat{F}_v^{(-)}(\zeta)P_i^{(-)}v(\zeta) + \chi_i^{(-)}(\zeta), \tag{50}$$

$$\frac{dP_i^{(-)}v}{d\zeta} = \hat{f}v^{(-)}(\zeta)P_i^{(-)}v + p_i^{(-)}(\zeta), \quad \zeta \geq 0, \tag{51}$$

где

$$\begin{aligned} \hat{F}_v^{(-)}(\zeta) &:= \frac{\partial F}{\partial v}(u^0, \varphi(u^0, 0) + P_0^{(-)}v(\zeta), 0, 0), \\ \hat{f}_v^{(-)}(\zeta) &:= \frac{\partial f}{\partial v}(u^0, \varphi(u^0, 0) + P_0^{(-)}v(\zeta), 0, 0), \end{aligned} \tag{52}$$

а $\chi_i^{(-)}(\zeta)$ и $p_i^{(-)}(\zeta)$ – известные на i-м шаге функции, рекуррентно выражающиеся через $P_j^{(-)}u(\zeta)$, $P_j^{(-)}v(\zeta)$ с номерами $j < i$ и имеющие экспоненциальные оценки вида (38), если такие же оценки имеют функции $P_j^{(-)}u$, $\dfrac{dP_j^{(-)}u}{d\zeta}$, $P_j^{(-)}v$ с номерами $j < i$.

Зададим для $P_i^{(-)}u(\zeta)$ граничное условие, аналогичное (35):

$$P_i^{(-)}u(\infty) = 0, \tag{53}$$

а для $P_i^{(-)}v(\zeta)$ из (36) получаем начальное условие

$$P_i^{(-)}v(0) = -\overline{v}_i^{(-)}(0) - \Pi_i^{(-)}v(0) =: P_i^{(-)}. \tag{54}$$

Решение задачи (51), (54) имеет вид

$$P_i^{(-)}v(\zeta) = K^{(-)}(\zeta, 0)P_i^{(-)} + \int_0^\zeta K^{(-)}(\zeta, s)p_i^{(-)}(s)ds, \tag{55}$$

где

$$K^{(-)}(\zeta, s) = \exp\left(\int_s^\zeta \hat{f}_v^{(-)}(t)dt\right).$$

Так как (см. (52) и (42))

$$\hat{f}_v^{(-)}(\zeta) = \frac{\partial f}{\partial v}(u^0, \varphi(u^0, 0), 0, 0) + O(P_0^{(-)}v(\zeta)) = f_v^{(-)}(\xi)\Big|_{\xi=0} + O(P_0^{(-)}v(\zeta)),$$

то (см. (43) и (38))

$$\hat{f}_v^{(-)}(\zeta) \leq -\varkappa + c\exp(-\varkappa_1\zeta), \quad \zeta \geq 0.$$

Поэтому

$$K^{(-)}(\zeta, s) \leq c_1\exp\big(-\varkappa(\zeta - s)\big), \quad 0 \leq s \leq \zeta.$$

В силу этой оценки и оценки вида (38) для $p_i^{(-)}(\zeta)$ из (55) получается экспоненциальная оценка для $P_i^{(-)}v(\zeta)$:

$$\left|P_i^{(-)}v(\zeta)\right| \leq c\exp(-\varkappa\zeta), \quad \zeta \geq 0. \tag{56}$$

Поскольку функция $P_i^{(-)}v(\zeta)$ найдена, то правая часть уравнения (50) является теперь известной функцией, имеющей оценку вида (56).

Решение задачи (50), (53) имеет вид, аналогичный (39):

$$P_i^{(-)}u(\zeta) = \int_{\infty}^{\zeta} ds \int_{\infty}^{s} (\hat{F}_v^{(-)}(t)P_i^{(-)}v(t) + \chi_i^{(-)}(t))dt,$$

откуда следует, что $P_i^{(-)}u(\zeta)$ и ее производная $\dfrac{dP_i^{(-)}u}{d\zeta}(\zeta)$ имеют оценки вида (56).

Итак, погранслойные ряды (21) и (22) построены, причем пограничные функции $\Pi_i^{(-)}u$, $\Pi_i^{(-)}v$ и $P_i^{(-)}u$, $P_i^{(-)}v$ имеют экспоненциальные оценки вида (49) и (56).

2.2.3. Внутрислойные части асимптотики $Q^{(-)}u$, $Q^{(-)}v$. Такое название мы дали рядам $Q^{(-)}u(\sigma,\varepsilon)$ и $Q^{(-)}v(\sigma,\varepsilon)$, имея в виду, что они будут описывать быстрое изменение решения исходной задачи (1), (2) в переходном слое слева от точки x_*. Эти ряды построим в виде

$$\begin{aligned} Q^{(-)}u(\sigma,\varepsilon) &= \sum_{i=0}^{\infty} \varepsilon^i Q_i^{(-)}u(\sigma), \\ Q^{(-)}v(\sigma,\varepsilon) &= \sum_{i=0}^{\infty} \varepsilon^i Q_i^{(-)}v(\sigma), \quad \sigma = (x - x_*)/\varepsilon \le 0. \end{aligned} \tag{57}$$

Для $Q^{(-)}u$, $Q^{(-)}v$ стандартным способом получается система уравнений

$$\begin{aligned} \frac{d^2Q^{(-)}u}{d\sigma^2} - \varepsilon w(x_* + \varepsilon\sigma)\frac{dQ^{(-)}u}{d\sigma} &= Q^{(-)}F := F(\bar{u}^{(-)}(x_* + \varepsilon\sigma,\varepsilon) + Q^{(-)}u, \bar{v}^{(-)}(x_* + \varepsilon\sigma,\varepsilon) + \\ &+ Q^{(-)}v, x_* + \varepsilon\sigma,\varepsilon) - F(\bar{u}^{(-)}(x_* + \varepsilon\sigma,\varepsilon), \bar{v}^{(-)}(x_* + \varepsilon\sigma,\varepsilon), x_* + \varepsilon\sigma,\varepsilon), \\ \varepsilon\frac{dQ^{(-)}v}{d\sigma} &= Q^{(-)}f, \quad \sigma \le 0, \end{aligned} \tag{58}$$

где $Q^{(-)}f$ имеет выражение, аналогичное $Q^{(-)}F$.

Из системы (58) для $Q_0^{(-)}u$, $Q_0^{(-)}v$ следует система уравнений, аналогичная (25):

$$\begin{aligned} \frac{d^2Q_0^{(-)}u}{d\sigma^2} &= F(\bar{u}_0^{(-)}(x_*) + Q_0^{(-)}u, \bar{v}_0^{(-)}(x_*) + Q_0^{(-)}v, x_*, 0), \\ 0 &= f(\bar{u}_0^{(-)}(x_*) + Q_0^{(-)}u, \bar{v}_0^{(-)}(x_*) + Q_0^{(-)}v, x_*, 0), \quad \sigma \le 0. \end{aligned}$$

Из второго уравнения, используя условие A2, получаем

$$\bar{v}_0^{(-)}(x_*) + Q_0^{(-)}v = \varphi(\bar{u}_0^{(-)}(x_*) + Q_0^{(-)}u, x_*). \tag{59}$$

Подставляя в первое уравнение, приходим к уравнению для $Q_0^{(-)}u$ такого же типа, как уравнение (27) для $\Pi_0^{(-)}u$:

$$\frac{d^2Q_0^{(-)}u}{d\sigma^2} = g(\bar{u}_0^{(-)}(x_*) + Q_0^{(-)}u, x_*), \quad \sigma \le 0. \tag{60}$$

Чтобы получить для $Q_0^{(-)}u(\sigma)$ граничное условие при $\sigma = 0$, подставим выражение (12) для $U^{(-)}(x,\varepsilon)$ в граничное условие $U^{(-)}(x_*,\varepsilon) = \psi_2(x_*)$ (см. (17)) с учетом того, что все члены рядов $\Pi^{(-)}u$ и $P^{(-)}u$ равны нулю при $x = x_*$ (см. замечание 1 в конце этого подпункта). Получим равенство

$$\sum_{i=0}^{\infty} \varepsilon^i \bar{u}_i^{(-)}(x_*) + \sum_{i=0}^{\infty} \varepsilon^i Q_i^{(-)}u(0) = \psi_2(x_*),$$

откуда имеем $\bar{u}_0^{(-)}(x_*) + Q_0^{(-)}u(0) = \psi_2(x_*)$, и, следовательно, граничное условие для $Q_0^{(-)}u(\sigma)$ при $\sigma = 0$ имеет вид

$$Q_0^{(-)}u(0) = \psi_2(x_*) - \bar{u}_0^{(-)}(x_*) = \psi_2(x_*) - \psi_1(x_*) > 0. \tag{61}$$

Второе граничное условие для $Q_0^{(-)}u(\sigma)$ – стандартное условие на бесконечности

$$Q_0^{(-)}u(-\infty) = 0. \tag{62}$$

Задача (60)–(62) для $Q_0^{(-)}u(\sigma)$ сводится стандартным образом к уравнению первого порядка

$$\frac{dQ_0^{(-)}u}{d\sigma} = \left[2\int_0^{Q_0^{(-)}u} g(\psi_1(x_*) + s, x_*)ds\right]^{1/2}, \quad \sigma \le 0, \tag{63}$$

с начальным условием (61). Отметим, что функция $g(\psi_1(x_*) + s, x_*)$ равна нулю при $s = 0$ и также при $s = \psi_2(x_*) - \psi_1(x_*)$ и не равна нулю при $0 < s < \psi_2(x_*) - \psi_1(x_*)$ в силу условия A2, а так как ее производная $\dfrac{\partial g}{\partial u} > 0$ при $s = 0$ в силу условия A4, то

$$g(\psi_1(x_*) + s, x_*) > 0 \quad \text{при} \quad 0 < s < \psi_2(x_*) - \psi_1(x_*).$$

Уравнение (63) интегрируется в квадратурах, его решение с начальным условием (61) является положительной возрастающей функцией на полупрямой $\sigma \le 0$ и имеет экспоненциальную оценку

$$\left|Q_0^{(-)}u(\sigma)\right| \le c\exp(\kappa\sigma), \quad \sigma \le 0. \tag{64}$$

Такого же типа оценка верна для $\dfrac{dQ_0^{(-)}u}{d\sigma}(\sigma)$.

Из (63) при $\sigma = 0$ получаем

$$\frac{dQ_0^{(-)}u}{d\sigma}(0) = \left[2\int_0^{\psi_2(x_*)-\psi_1(x_*)} g(\psi_1(x_*) + s, x_*)ds\right]^{1/2} = \left[2\int_{\psi_1(x_*)}^{\psi_2(x_*)} g(u, x_*)du\right]^{1/2}. \tag{65}$$

Эта формула будет использована в п. 2.4.

Зная $Q_0^{(-)}u(\sigma)$, из (59) находим функцию $Q_0^{(-)}v(\sigma)$, которая также имеет оценку вида (64).

Функции $Q_i^{(-)}u(\sigma)$ и $Q_i^{(-)}v(\sigma)$ при $i \ge 1$ определяются аналогично тому, как в пп. 2.2.2 были определены функции $\Pi_i^{(-)}u(\xi)$ и $\Pi_i^{(-)}v(\xi)$, и имеют оценки вида (64).

Итак, ряды (57) построены, и тем самым завершено построение формальной асимптотики на отрезке $[0, x_*]$.

Замечание 1. При построении рядов (21), (22) и (57) говорилось о том, что все функции $Q_i^{(-)}u$ и $Q_i^{(-)}v$ равны нулю при $x = 0$, а функции $\Pi_i^{(-)}u$, $\Pi_i^{(-)}v$ и $P_i^{(-)}u$, $P_i^{(-)}v$ равны нулю при $x = x_*$. Это достигается применением стандартной процедуры умножения этих функций на срезающие функции (см. [1]), что не влияет на построенные асимптотические разложения. За подправленными пограничными функциями сохраняем старые обозначения. Будем считать, что

$$\begin{gathered} Q_i^{(-)}u = Q_i^{(-)}v = 0 \quad \text{при} \quad x \in [0; x_*/2], \\ \Pi_i^{(-)}u = \Pi_i^{(-)}v = P_i^{(-)}u = P_i^{(-)}v = 0 \quad \text{при} \quad x \in [x_*/2; x_*]. \end{gathered} \tag{66}$$

Замечание 2. Обозначим через $U_k^{(-)}(x, \varepsilon)$ и $V_k^{(-)}(x, \varepsilon)$ частичные суммы k-го порядка построенных рядов (12) и (13):

$$U_k^{(-)}(x, \varepsilon) = \sum_{i=0}^{k} \varepsilon^i(\bar{u}_i^{(-)}(x) + \Pi_i^{(-)}u(\xi) + \varepsilon^2 P_i^{(-)}u(\zeta) + Q_i^{(-)}u(\sigma)), \tag{67}$$

$$V_k^{(-)}(x,\varepsilon) = \sum_{i=0}^{k} \varepsilon^i (\bar{v}_i^{(-)}(x) + \Pi_i^{(-)} v(\xi) + P_i^{(-)} v(\zeta) + Q_i^{(-)} v(\sigma)). \tag{68}$$

Из самого способа построения рядов (12) и (13) следует, что для $U_k^{(-)}$, $V_k^{(-)}$ при $k = 0, 1, 2, \ldots$ справедливы равенства

$$L_\varepsilon(U_k^{(-)}, V_k^{(-)}) := \varepsilon^2 \left(\frac{d^2 U_k^{(-)}}{dx^2} - w(x) \frac{dU_k^{(-)}}{dx} \right) - F(U_k^{(-)}, V_k^{(-)}, x, \varepsilon) = O(\varepsilon^{k+1}), \quad x \in (0, x_*),$$

$$M_\varepsilon(V_k^{(-)}, U_k^{(-)}) := \varepsilon^2 \frac{dV_k^{(-)}}{dx} - f(U_k^{(-)}, V_k^{(-)}, x, \varepsilon) = O(\varepsilon^{k+1}), \quad x \in (0, x_*), \tag{69}$$

$$U_k^{(-)}(0,\varepsilon) = u^0 + O(\varepsilon^{k+1}), \quad V_k^{(-)}(0,\varepsilon) = v^0, \quad U_k^{(-)}(x_*,\varepsilon) = \psi_2(x_*).$$

2.3. Построение асимптотики на отрезке $[x_*, 1]$

Заметим, прежде всего, что вид асимптотики $U^{(+)}(x,\varepsilon)$, $V^{(+)}(x,\varepsilon)$ на отрезке $[x_*, 1]$ (см. (14), (15)) существенно отличается от вида $U^{(-)}(x,\varepsilon)$, $V^{(-)}(x,\varepsilon)$ (см. (12), (13)). Отличие состоит в том, что $U^{(+)}$ и $V^{(+)}$ не содержат P-функций. Это соответствует тому, что краевые условия (18) не содержат условия для $V^{(+)}$ (в отличие от (17)). На первый взгляд может показаться, что для $V^{(+)}(x,\varepsilon)$ следует задать в точке x_* краевое условие $V^{(+)}(x_*,\varepsilon) = V^{(-)}(x_*,\varepsilon)$, чтобы обеспечить непрерывное сшивание асимптотик $V^{(+)}$ и $V^{(-)}$ в точке x_*. Однако, как будет показано ниже, непрерывное и, более того, сколь угодно гладкое сшивание $V^{(+)}$ и $V^{(-)}$ в точке x_* будет достигнуто за счет выбора точки x_*.

Регулярные части асимптотики $\bar{u}^{(+)}(x,\varepsilon)$ и $\bar{v}^{(+)}(x,\varepsilon)$ строятся в виде, аналогичном (19):

$$\bar{u}^{(+)}(x,\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i \bar{u}_i^{(+)}(x), \quad \bar{v}^{(+)}(x,\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i \bar{v}_i^{(+)}(x).$$

Главные члены $\bar{u}_0^{(+)}(x)$ и $\bar{v}_0^{(+)}(x)$ этих рядов являются решением вырожденной системы (4), связанным с корнем $u = \psi_3(x)$ уравнения (5):

$$\bar{u}_0^{(+)}(x) = \psi_3(x), \quad \bar{v}_0^{(+)}(x) = \varphi(\psi_3(x), x).$$

Функции $\bar{u}_i^{(+)}(x)$, $\bar{v}_i^{(+)}(x)$ при $i \geq 1$ определяются из линейных систем вида (20) (с заменой индекса $(-)$ на $(+)$), определитель которых $\Delta^{(+)}(x) = f_v^{(+)}(x) \bar{g}_u^{(+)}(x) < 0$ в силу неравенств из условий A5 и A4, относящихся к кривым L_5 и l_5.

Внутрислойные части асимптотики $Q^{(+)}u(\sigma,\varepsilon)$, $Q^{(+)}v(\sigma,\varepsilon)$ построим в виде

$$Q^{(+)}u(\sigma,\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i Q_i^{(+)}u(\sigma), \quad Q^{(+)}v(\sigma,\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i Q_i^{(+)}v(\sigma), \quad \sigma = (x - x_*)/\varepsilon \geq 0.$$

Для $Q^{(+)}u$, $Q^{(+)}v$ стандартным способом получается система такого же типа, как (58), откуда для $Q_0^{(+)}u$, $Q_0^{(+)}v$ имеем систему уравнений

$$\frac{d^2 Q_0^{(+)}u}{d\sigma^2} = F(\bar{u}_0^{(+)}(x_*) + Q_0^{(+)}u, \bar{v}_0(x_*) + Q_0^{(+)}v, x_*, 0),$$

$$0 = f(\bar{u}_0^{(+)}(x_*) + Q_0^{(+)}u, \bar{v}_0^{(+)}(x_*) + Q_0^{(+)}v, x_*, 0), \quad \sigma \geq 0.$$

Из второго уравнения, используя условие A2, получаем

$$\bar{v}_0^{(+)}(x_*) + Q_0^{(+)}v = \varphi(\bar{u}_0^{(+)}(x_*) + Q_0^{(+)}u, x_*). \tag{70}$$

Подставляя в первое уравнение, приходим к уравнению для $Q_0^{(+)}u$, аналогичному уравнению (60) для $Q_0^{(-)}u$:

$$\frac{d^2Q_0^{(+)}u}{d\sigma^2} = g(\bar{u}_0^{(+)}(x_*) + Q_0^{(+)}u, x_*), \quad \sigma \geq 0. \tag{71}$$

Также стандартным способом добавляем граничные условия

$$Q_0^{(+)}u(0) = \psi_2(x_*) - u_0^{(+)}(x_*) = \psi_2(x_*) - \psi_3(x_*) < 0, \tag{72}$$

$$Q_0^{(+)}u(\infty) = 0 \tag{73}$$

и сводим задачу (71)–(73) к уравнению первого порядка

$$\frac{dQ_0^{(+)}u}{d\sigma} = \left[2\int_0^{Q_0^{(+)}u} g(\psi_3(x_*) + s, x_*)ds\right]^{1/2}, \quad \sigma \geq 0, \tag{74}$$

с начальным условием (72).

Уравнение (74) интегрируется в квадратурах, его решение с начальным условием (72) является отрицательной возрастающей функцией и имеет экспоненциальную оценку

$$\left|Q_0^{(+)}u(\sigma)\right| \leq c\exp(-\kappa\sigma), \quad \sigma \geq 0. \tag{75}$$

После этого функция $Q_0^{(+)}v(\sigma)$ находится из (70) и также имеет оценку вида (75). Такую же оценку имеет производная $\dfrac{dQ_0^{(+)}u}{d\sigma}(\sigma)$.

Из (74) при $\sigma = 0$ получаем

$$\frac{dQ_0^{(+)}u}{d\sigma}(0) = \left[2\int_{\psi_3(x_*)}^{\psi_2(x_*)} g(u, x_*)du\right]^{1/2}. \tag{76}$$

Эта формула будет использована в п. 2.4.

Функции $Q_i^{(+)}u(\sigma)$ и $Q_i^{(+)}v(\sigma)$ при $i \geq 1$ определяются аналогично тому, как в пп. 2.2.2 были определены функции $\Pi_i^{(-)}u(\xi)$ и $\Pi_i^{(-)}v(\xi)$, и имеют оценки вида (75).

Итак, внутрислойные ряды $Q^{(+)}u(\sigma,\varepsilon)$ и $Q^{(+)}v(\sigma,\varepsilon)$ построены.

Погранслойные части асимптотики $\Pi^{(+)}u(\tilde{\xi},\varepsilon)$, $\Pi^{(+)}v(\tilde{\xi},\varepsilon)$ строятся в виде рядов

$$\Pi^{(+)}u(\tilde{\xi},\varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i\Pi_i^{(+)}u(\tilde{\xi}),$$

$$\Pi^{(+)}v(\tilde{\xi},\varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i\Pi_i^{(+)}v(\tilde{\xi}), \quad \tilde{\xi} = (x-1)/\varepsilon \leq 0,$$

аналогично построению рядов $\Pi^{(-)}u(\xi,\varepsilon)$, $\Pi^{(-)}v(\xi,\varepsilon)$.

Стандартным способом для $\Pi_0^{(+)}u$, $\Pi_0^{(+)}v$ получается система уравнений, аналогичная (25):

$$\frac{d^2\Pi_0^{(+)}u}{d\tilde{\xi}^2} = F(\bar{u}_0^{(+)}(1) + \Pi_0^{(+)}u, \bar{v}_0^{(+)}(1) + \Pi_0^{(+)}v, 1, 0),$$

$$0 = f(\bar{u}_0^{(+)}(1) + \Pi_0^{(+)}u, v_0^{(+)}(1) + \Pi_0^{(+)}v, 1, 0), \quad \tilde{\xi} \leq 0.$$

Из второго уравнения имеем

$$\bar{v}_0^{(+)}(1) + \Pi_0^{(+)}v = \varphi(\bar{u}_0^{(+)}(1) + \Pi_0^{(+)}u, 1). \tag{77}$$

Подставляя в первое уравнение, приходим к уравнению для $\Pi_0^{(+)}u$, аналогичному (27):

$$\frac{d^2\Pi_0^{(+)}u}{d\tilde{\xi}^2} = g(\bar{u}_0^{(+)}(1) + \Pi_0^{(+)}u, 1), \quad \tilde{\xi} \le 0. \tag{78}$$

Также стандартным образом получаем граничные условия

$$\Pi_0^{(+)}u(0) = u^1 - \bar{u}_0^{(+)}(1) = u^1 - \psi_3(1), \quad \Pi_0^{(+)}u(-\infty) = 0. \tag{79}$$

Если $u^1 = \psi_3(1)$, то $\Pi_0^{(+)}u(\tilde{\xi}) = 0$ при $\tilde{\xi} \le 0$, а если $u^1 \ne \psi_3(1)$, то задача (78), (79) сводится к уравнению первого порядка

$$\frac{d\Pi_0^{(+)}u}{d\tilde{\xi}} = \pm\left[2\int_0^{\Pi_0^{(+)}u} g(\psi_3(1) + s, 1)ds\right]^{1/2}, \quad \tilde{\xi} \le 0, \tag{80}$$

с начальным условием при $\tilde{\xi} = 0$ из (79), причем в правой части (80) берется знак плюс, если $u^1 > \psi_3(1)$, и знак минус, если $u^1 < \psi_3(1)$.

Уравнение (80) интегрируется в квадратурах, функция $\Pi_0^{(+)}u(\tilde{\xi})$ является монотонной при $\tilde{\xi} \le 0$ и имеет экспоненциальную оценку

$$\left|\Pi_0^{(+)}u(\tilde{\xi})\right| \le c\exp(\varkappa\tilde{\xi}), \quad \tilde{\xi} \le 0. \tag{81}$$

Такую же оценку имеют производная $\frac{d\Pi_0^{(+)}u}{d\tilde{\xi}}(\tilde{\xi})$ и функция $\Pi_0^{(+)}v(\tilde{\xi})$, которая определяется теперь из (77).

Функции $\Pi_i^{(+)}u(\tilde{\xi})$, $\Pi_i^{(+)}v(\tilde{\xi})$ при $i \ge 1$ определяются аналогично тому, как в пп. 2.2.2 были определены функции $\Pi_i^{(-)}u(\xi)$, $\Pi_i^{(-)}v(\xi)$ и имеют оценки вида (81).

Таким образом, завершено построение формальной асимптотики на отрезке $[x_*, 1]$.

Замечание 3. Как и при построении асимптотики на отрезке $[0, x_*]$, считаем, что все функции $Q_i^{(+)}u$, $Q_i^{(+)}v$ и $\Pi_i^{(+)}u$, $\Pi_i^{(+)}v$ умножены на соответствующие срезающие функции.

Замечание 4. Обозначим через $U_k^{(+)}(x,\varepsilon)$, $V_k^{(+)}(x,\varepsilon)$ частичные суммы k-го порядка построенных рядов (14) и (15):

$$\begin{aligned} U_k^{(+)}(x,\varepsilon) &= \sum_{i=0}^{k}\varepsilon^i(\bar{u}_i^{(+)}(x) + Q_i^{(+)}u(\sigma) + \Pi_i^{(+)}u(\tilde{\xi})), \\ V_k^{(+)}(x,\varepsilon) &= \sum_{i-0}^{k}\varepsilon^i(\bar{v}_i^{(+)}(x) + Q_i^{(+)}v(\sigma) + \Pi_i^{(+)}v(\tilde{\xi})). \end{aligned} \tag{82}$$

Из самого способа построения рядов (14) и (15) следует, что $U_k^{(+)}$ и $V_k^{(+)}$ удовлетворяют для любого $k = 0, 1, 2, \ldots$ равенствам (операторы L_ε и M_ε определены в (69))

$$L_\varepsilon(U_k^{(+)}, V_k^{(+)}) = O(\varepsilon^{k+1}), \quad x \in (x_*, 1), \tag{83}$$

$$M_\varepsilon(V_k^{(+)}, U_k^{(+)}) = O(\varepsilon^{k+1}), \quad x \in (x_*, 1), \tag{84}$$

$$U_k^{(+)}(x_*, \varepsilon) = \psi_2(x_*), \quad U_k^{(+)}(x_*, \varepsilon) = u^1. \tag{85}$$

2.4. Сшивание формальных асимптотик в точке x_*

Построенные формальные ряды $U^{(-)}(x,\varepsilon)$ и $U^{(+)}(x,\varepsilon)$ удовлетворяют равенству

$$U^{(-)}(x_*, \varepsilon) = U^{(+)}(x_*, \varepsilon), \tag{86}$$

так как обе части этого формального равенства равны $\psi_2(x_*)$ (см. (17) и (18)). Аналогичное равенство для $V^{(-)}(x,\varepsilon)$ и $V^{(+)}(x,\varepsilon)$ не имеет места при произвольном x_*. Оказывается, однако, что формальное равенство

$$V^{(-)}(x_*,\varepsilon) = V^{(+)}(x_*,\varepsilon) \tag{87}$$

будет выполнено, если x_* выбрать так, чтобы в точке x_* выполнялось формальное равенство производных $\frac{dU^{(-)}}{dx}$ и $\frac{dU^{(+)}}{dx}$. Используя выражения (12) и (14), учитывая замечания 1 и 3 и умножив указанные производные в точке x_* на ε, запишем равенство в виде

$$\varepsilon\frac{d\overline{u}^{(-)}}{dx}(x_*,\varepsilon) + \frac{dQ^{(-)}u}{d\sigma}(0,\varepsilon) = \varepsilon\frac{d\overline{u}^{(+)}}{dx}(x_*,\varepsilon) + \frac{dQ^{(+)}u}{d\sigma}(0,\varepsilon).$$

Подставив в это равенство выражения для $\overline{u}^{(\pm)}$ и $Q^{(\pm)}u$ в виде рядов и учитывая, что функции $Q_i^{(\pm)}u$ зависят не только от σ, но и от x_*, т.е. $Q_i^{(\pm)}u = Q_i^{(\pm)}u(\sigma, x_*)$, перепишем равенство в виде

$$\sum_{i=0}^{\infty}\varepsilon^i\left(\frac{dQ_i^{(-)}u}{d\sigma}(0,x_*) - \frac{dQ_i^{(+)}u}{d\sigma}(0,x_*)\right) + \varepsilon\sum_{i=0}^{\infty}\varepsilon^i\left(\frac{d\overline{u}_i^{(-)}}{dx}(x_*) - \frac{d\overline{u}_i^{(+)}}{dx}(x_*)\right) = 0. \tag{88}$$

Равенство (88) является уравнением относительно x_*. Будем искать x_* в виде ряда

$$x_* = \sum_{i=0}^{\infty}\varepsilon^i x_i. \tag{89}$$

Подставим это выражение в (88), разложим левую часть уравнения в ряд по степеням ε и будем приравнивать нулю коэффициенты разложения. В нулевом приближении получим

$$\frac{dQ_0^{(-)}u}{d\sigma}(0,x_0) - \frac{dQ_0^{(+)}u}{d\sigma}(0,x_0) = 0,$$

т.е. (см. (65) и (76))

$$J(x_0) := \left[2\int_{\psi_1(x_0)}^{\psi_2(x_0)} g(u,x_0)du\right]^{1/2} - \left[2\int_{\psi_3(x_0)}^{\psi_2(x_0)} g(u,x_0)du\right]^{1/2} = 0.$$

Это уравнение эквивалентно уравнению (7) из условия A3, поэтому оно имеет корень $x_0 = \overline{x}_0 \in (0;1)$, причем $J'(\overline{x}_0) \neq 0$ в силу (8).

Для следующих коэффициентов x_i ряда (89) последовательно при $i = 1, 2, \ldots$ получаются линейные уравнения

$$J'(\overline{x}_0)x_i + k_i = 0, \tag{90}$$

где k_i – известные на i-м шаге числа, выражающиеся определенным образом через найденные уже коэффициенты x_j с номерами $j < i$. Так как $J'(\overline{x}_0) \neq 0$, то уравнение (90) имеет единственное решение, которое обозначим $\overline{x}_i$:

$$\overline{x}_i = -\left(J'(\overline{x}_0)\right)^{-1}k_i, \quad i = 1, 2, \ldots .$$

Итак, для точки перехода x_* получено формальное разложение

$$x_* = \sum_{i=0}^{\infty}\varepsilon^i\overline{x}_i =: \overline{x}_*, \tag{91}$$

обеспечивающее выполнение формального равенства

$$\frac{dU^{(-)}}{dx}(\overline{x}_*,\varepsilon) = \frac{dU^{(+)}}{dx}(\overline{x}_*,\varepsilon). \tag{92}$$

Докажем, что для точки $x_* = \bar{x}_*$ с формальным разложением (91) выполнены также формальное равенство (87) и формальное равенство

$$\frac{dV^{(-)}}{dx}(\bar{x}_*, \varepsilon) = \frac{dV^{(+)}}{dx}(\bar{x}_*, \varepsilon). \tag{93}$$

С этой целью введем обозначения

$$u^{(-)}(\sigma, \varepsilon) = \bar{u}^{(-)}(\bar{x}_* + \varepsilon\sigma, \varepsilon) + Q^{(-)}u(\sigma, \bar{x}_*, \varepsilon),$$
$$v^{(-)}(\sigma, \varepsilon) = \bar{v}^{(-)}(\bar{x}_* + \varepsilon\sigma, \varepsilon) + Q^{(-)}v(\sigma, \bar{x}_*, \varepsilon),$$

где $\bar{u}^{(-)}(x, \varepsilon)$, $\bar{v}^{(-)}(x, \varepsilon)$ – ряды (19), $Q^{(-)}u(\sigma, \bar{x}_*, \varepsilon)$, $Q^{(-)}v(\sigma, \bar{x}_*, \varepsilon)$ – ряды (57), $\bar{x}_*$ – ряд (91).

Для $u^{(-)}(\sigma, \varepsilon)$, $v^{(-)}(\sigma, \varepsilon)$, используя (58), получаем систему уравнений

$$\begin{gathered}\frac{d^2u^{(-)}}{d\sigma^2} - \varepsilon w(\bar{x}_* + \varepsilon\sigma)\frac{du^{(-)}}{d\sigma} = F(u^{(-)}(\sigma, \varepsilon), v^{(-)}(\sigma, \varepsilon), \bar{x}_* + \varepsilon\sigma, \varepsilon), \\ \varepsilon\frac{dv^{(-)}}{d\sigma} = f(u^{(-)}(\sigma, \varepsilon), v^{(-)}(\sigma, \varepsilon), \bar{x}_* + \varepsilon\sigma, \varepsilon), \quad \sigma \le 0.\end{gathered} \tag{94}$$

Такая же система уравнений с заменой индекса (–) на (+) имеет место для

$$u^{(+)}(\sigma, \varepsilon) = \bar{u}^{(+)}(\bar{x}_* + \varepsilon\sigma, \varepsilon) + Q^{(+)}u(\sigma, \bar{x}_*, \varepsilon),$$
$$v^{(+)}(\sigma, \varepsilon) = \bar{v}^{(+)}(\bar{x}_* + \varepsilon\sigma, \varepsilon) + Q^{(+)}v(\sigma, \bar{x}_*, \varepsilon)$$

при $\sigma \ge 0$.

Напишем формальные разложения $u^{(\pm)}(\sigma, \varepsilon)$ и $v^{(\pm)}(\sigma, \varepsilon)$ в ряды по степеням ε:

$$u^{(-)}(\sigma, \varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i u_i^{(-)}(\sigma), \quad v^{(-)}(\sigma, \varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i v_i^{(-)}(\sigma), \quad \sigma \le 0, \tag{95}$$

$$u^{(+)}(\sigma, \varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i u_i^{(+)}(\sigma), \quad v^{(+)}(\sigma, \varepsilon) = \sum_{i=0}^{\infty}\varepsilon^i v_i^{(+)}(\sigma), \quad \sigma \ge 0. \tag{96}$$

Главные члены этих разложений имеют вид

$$u_0^{(\pm)}(\sigma) = u_0^{(\pm)}(\bar{x}_0) + Q_0^{(\pm)}u(\sigma, \bar{x}_0), \quad v_0^{(\pm)}(\sigma) = v_0^{(\pm)}(\bar{x}_0) + Q_0^{(\pm)}v(\sigma, \bar{x}_0). \tag{97}$$

Из (86) при $x_* = \bar{x}_*$ и (92) следуют равенства

$$u_i^{(-)}(0) = u_i^{(+)}(0), \quad \frac{du_i^{(-)}}{d\sigma}(0) = \frac{du_i^{(\pm)}}{d\sigma}(0), \quad i = 0, 1, 2, \ldots. \tag{98}$$

Докажем, что аналогичные равенства имеют место для $v_i^{(\pm)}(\sigma)$, т.е.

$$v_i^{(-)}(0) = v_i^{(+)}(0), \quad \frac{dv_i^{(-)}}{d\sigma}(0) = \frac{dv_i^{(+)}}{d\sigma}(0), \quad i = 0, 1, 2, \ldots. \tag{99}$$

Отсюда последуют формальные равенства (87) при $x_* = \bar{x}_*$ и (93). Введем функции для $i = 0, 1, 2, \ldots$

$$u_i(\sigma) = \begin{cases} u_i^{(-)}(\sigma), & \sigma \le 0, \\ u_i^{(+)}(\sigma), & \sigma \ge 0, \end{cases} \quad v_i(\sigma) = \begin{cases} v_i^{(-)}(\sigma), & \sigma \le 0, \\ v_i^{(-)}(\sigma), & \sigma > 0, \end{cases}$$

в частности (см. (97))

$$u_0(\sigma) = \begin{cases} \bar{u}_0^{(-)}(\bar{x}_0) + Q_0^{(-)}u(\sigma, \bar{x}_0), & \sigma \le 0, \\ \bar{u}_0^{(+)}(\bar{x}_0) + Q_0^{(+)}u(\sigma, \bar{x}_0), & \sigma \ge 0, \end{cases} \quad v_0(\sigma) = \begin{cases} \bar{v}_0^{(-)}(\bar{x}_0) + Q_0^{(-)}v(\sigma, \bar{x}_0), & \sigma \le 0, \\ \bar{v}_0^{(-)}(\bar{x}_0) + Q_0^{(-)}v(\sigma, \bar{x}_0), & \sigma > 0. \end{cases}$$

Из (60) и (71) следует, что функция $u_0(\sigma)$ является решением дифференциального уравнения

$$\frac{d^2u_0}{d\sigma^2} = g(u_0, \bar{x}_0), \quad -\infty < \sigma < \infty,$$

а равенства (98) при $i = 0$ показывают, что это решение удовлетворяет начальным условиям

$$u_0(0) = \psi_2(\bar{x}_0), \quad \frac{du_0}{d\sigma}(0) = \left[2\int\limits_{\psi_1\bar{x}_0}^{\psi_2(\bar{x}_0)} g(u, \bar{x}_0)du\right]^{1/2}.$$

Следовательно, $u_0(\sigma)$ – бесконечно гладкая функция при $-\infty < \sigma < \infty$.

Функцию $v_0(\sigma)$ можно записать в виде (см. (59) и (70))

$$v_0(\sigma) = \varphi(u_0(\sigma), \bar{x}_0),$$

откуда следует, что $v_0(\sigma)$ также бесконечно гладкая функция при $-\infty < \sigma < \infty$, и, значит, выполнены равенства (99) для $i = 0$.

Далее по индукции докажем, что $u_i(\sigma)$ и $v_i(\sigma)$ – бесконечно гладкие функции при $\infty < \sigma < \infty$ для всех $i \geq 1$. Пусть $u_i(\sigma)$, $v_i(\sigma)$ – бесконечно гладкие функции для $i = 0, 1, \ldots, k-1$. Покажем, что тогда $u_k(\sigma)$, $v_k(\sigma)$ также будут бесконечно гладкими функциями при $-\infty < \sigma < \infty$.

Подставим выражения (95) для $u^{(-)}$, $v^{(-)}$ в систему уравнений (94), а выражения (96) – в аналогичную систему уравнений для $u^{(+)}$, $v^{(+)}$, и приравняем коэффициенты при ε^k в разложениях левой и правой части каждого уравнения. Получим систему уравнений

$$\frac{d^2u_k^{(\pm)}}{d\sigma^2} = F_u(\sigma)u_k^{(\pm)} + F_v(\sigma)v_k^{(\pm)} + r_k(\sigma), \tag{100}$$

$$f_u(\sigma)u_k^{(\pm)} + f_v(\sigma)v_k^{(\pm)} + \gamma_k(\sigma) = 0, \tag{101}$$

где $F_u(\sigma) := \dfrac{\partial F}{\partial u}(u_0(\sigma), v_0(\sigma), \bar{x}_0, 0)$, обозначения $F_v(\sigma)$, $f_u(\sigma)$, $f_v(\sigma)$ имеют аналогичный смысл, а $r_k(\sigma)$ и $\gamma_k(\sigma)$ выражаются через $u_i(\sigma)$, $v_i(\sigma)$ с номерами $i \leq k-1$ и являются бесконечно гладкими функциями при $-\infty < \sigma < \infty$ в силу индуктивного предположения. Так как $f_v(\sigma) \leq -\kappa < 0$ (это следует из условия A5 аналогично тому, как было получено неравенство (43)), то из (101) имеем

$$v_k^{(\pm)} = -f_v^{-1}(\sigma)[f_u(\sigma)u_k^{(\pm)} + \gamma_k(\sigma)]. \tag{102}$$

Подставляя это выражение в (100), приходим к уравнению для $u_k^{(\pm)}$:

$$\frac{d^2u_k^{(\pm)}}{d\sigma^2} = g_u(\sigma)u_k^{(\pm)} + h_k(\sigma),$$

где $g_u(\sigma) := \dfrac{\partial g}{\partial u}(u_0(\sigma), \bar{x}_0)$ – бесконечно гладкая функция при $-\infty < \sigma < \infty$. Следовательно, функция $u_k(\sigma)$ является решением уравнения

$$\frac{d^2u_k}{d\sigma^2} = g_u(\sigma)u_k + h_k(\sigma), \quad -\infty < \sigma < \infty,$$

с начальными условиями (см. (98))

$$u_k(0) = u_k^{(-)}(0) = u_k^{(+)}(0), \quad \frac{du_k}{d\sigma}(0) = \frac{du_k^{(-)}}{d\sigma}(0) = \frac{du_k^{(+)}}{d\sigma}(0).$$

Поэтому $u_k(\sigma)$ – бесконечно гладкая функция. Из (102) следует теперь, что $v_k(\sigma) = -f_v^{-1}(\sigma)[f_u(\sigma)u_k(\sigma) + \gamma_k(\sigma)]$ – также бесконечно гладкая функция при $-\infty < \sigma < \infty$, и, следовательно, равенства (99) выполнены для $i = k$.

Таким образом, для $x_* = \bar{x}_*$ с разложением (91) выполнены формальные равенства (87) и (93).

3. ПЕРВАЯ ВСПОМОГАТЕЛЬНАЯ ЗАДАЧА

Возьмем какое-нибудь целое число $n \geq 0$, положим

$$x_\delta := X_{n+1} + \varepsilon^{n+1}\delta, \tag{103}$$

где $X_{n+1} = \sum_{i=0}^{n+1} \varepsilon^i \bar{x}_i$, $\bar{x}_i$ – коэффициенты ряда (91), а δ – является величиной порядка $O(\varepsilon)$, и рассмотрим краевую задачу для системы (1) на отрезке $[0, x_\delta]$ с краевыми условиями:

$$u(0,\varepsilon) = u^0, \quad v(0,\varepsilon) = v^0, \quad u(x_\delta,\varepsilon) = \psi_2(x_\delta). \tag{104}$$

Для достаточно малых ε точка x_δ сколь угодно близка к $\bar{x}_0$, поэтому для задачи (1), (104) можно построить формальные асимптотические ряды $U^{(-)}(x,\varepsilon)$ и $V^{(-)}(x,\varepsilon)$ (см. (12) и (13)), в которых $x_* = x_\delta$. Составим частичные суммы $U^{(-)}_{n+1}(x,\varepsilon)$ и $V^{(-)}_{n+1}(x,\varepsilon)$ этих рядов по формулам (67) и (68). Аргумент σ функций $Q_i^{(-)}u$ и $Q_i^{(-)}v$ в этих суммах равен $(x - x_\delta)/\varepsilon$.

Теорема 1. *Если выполнены условия* A1–A8, *то для достаточно малых* ε *задача* (1), (104) *имеет решение* $u = u^{(-)}(x,\varepsilon,\delta)$, $v = v^{(-)}(x,\varepsilon,\delta)$, *для которого справедливы асимптотические равенства*

$$\begin{aligned} u^{(-)}(x,\varepsilon,\delta) &= U^{(-)}_{n+1}(x,\varepsilon) + O(\varepsilon^{n+2}), \\ v^{(-)}(x,\varepsilon,\delta) &= V^{(-)}_{n+1}(x,\varepsilon) + O(\varepsilon^{n+2}), \quad x \in [0, x_\delta]. \end{aligned} \tag{105}$$

Доказательство. Доказательство теоремы проводится с помощью асимптотического метода дифференциальных неравенств (см. [10]), суть которого состоит в том, что нижнее и верхнее решения задачи (1), (104) конструируются на основе построенной в разд. 2 формальной асимптотики. Это делается во многом так же, как в аналогичной задаче в [1], поэтому ограничимся кратким изложением схемы доказательства. Напомним понятия нижнего и верхнего решений применительно к задаче (1), (104).

Определение 1. Две пары функций $\underline{U}(x,\varepsilon)$, $\underline{V}(x,\varepsilon)$ и $\overline{U}(x,\varepsilon)$, $\overline{V}(x,\varepsilon)$ называются *упорядоченными нижним и верхним решениями задачи* (1), (104), если они удовлетворяют следующим условиям:

$$1^\circ. \quad \underline{U}(x,\varepsilon) \leq \overline{U}(x,\varepsilon), \quad \underline{V}(x,\varepsilon) \leq \overline{V}(x,\varepsilon), \quad x \in [0, x_\delta]$$

(условие упорядоченности).

$$2^\circ. \quad L_\varepsilon(\underline{U}, v) := \varepsilon^2\left(\frac{d^2\underline{U}}{dx^2} - w(x)\frac{d\underline{U}}{dx}\right) - F(\underline{U}, v, x, \varepsilon) \geq 0 \geq L_\varepsilon(\overline{U}, v)$$

при $\underline{V}(x,\varepsilon) \leq v \leq \overline{V}(x,\varepsilon)$, $0 < x < x_\delta$;

$$M_\varepsilon(\underline{V}, u) := \varepsilon^2\frac{d\underline{V}}{dx} - f(u, \underline{V}, x, \varepsilon) \leq 0 \leq M_\varepsilon(\overline{V}, u)$$

при $\underline{U}(x,\varepsilon) \leq u \leq \overline{U}(x,\varepsilon)$, $0 < x \leq x_\delta$.

$$3^\circ. \quad \underline{U}(0,\varepsilon) \leq u^0 \leq \overline{U}(0,\varepsilon), \quad \underline{V}(0,\varepsilon) \leq v^0 \leq \overline{V}(0,\varepsilon),$$

$$\underline{U}(x_\delta,\varepsilon) \leq \psi_2(x_\delta) \leq \overline{U}(x_\delta,\varepsilon).$$

Если существуют упорядоченные нижнее и верхнее решения задачи (1), (104), то эта задача имеет решение $u = u^{(-)}(x,\varepsilon,\delta)$, $v = v^{(-)}(x,\varepsilon,\delta)$ (возможно, не единственное), удовлетворяющее неравенствам

$$\begin{aligned} \underline{U}(x,\varepsilon) \leq u^{(-)}(x,\varepsilon,\delta) \leq \overline{U}(x,\varepsilon), \\ \underline{V}(x,\varepsilon) \leq v^{(-)}(x,\varepsilon,\delta) \leq \overline{V}(x,\varepsilon), \quad x \in [0, x_\delta]. \end{aligned} \tag{106}$$

Видно, что если функция $F(u,v,x,\varepsilon)$ является невозрастающей функцией аргумента v, а функция $f(u,v,x,\varepsilon)$ – неубывающей функцией аргумента u в области

$$G_0 = \{(u,v,x,\varepsilon) : \underline{U}(x,\varepsilon) \leq u \leq \overline{U}(x,\varepsilon),\ \underline{V}(x,\varepsilon) \leq v \leq \overline{V}(x,\varepsilon),\ 0 \leq x \leq x_\delta,\ 0 \leq \varepsilon \leq \varepsilon_0\} \tag{107}$$

(в таком случае говорят, что функции F и f удовлетворяют условию квазимонотонности в области G_0), то для выполнения условия 2° из определения 1 достаточно, чтобы были выполнены неравенства

$$L_\varepsilon(\underline{U},\underline{V}) \ge 0 \ge L_\varepsilon(\overline{U},\overline{V}), \quad x \in (0, x_\delta), \tag{108}$$

$$M_\varepsilon(\underline{V},\underline{U}) \le 0 \le M_\varepsilon(\overline{V},\overline{U}), \quad x \in (0, x_\delta]. \tag{109}$$

Это очевидное утверждение используется ниже при доказательстве теоремы 1.

Нижнее и верхнее решения задачи (1), (104) строятся в виде

$$\begin{aligned} \underline{U}(x,\varepsilon) &= U_{n+1}^{(-)}(x,\varepsilon) - (\alpha(x,\xi) + \gamma(\xi) + \tilde\gamma(\sigma))\varepsilon^{n+2} + G(\zeta)\varepsilon^{n+4}, \\ \underline{V}(x,\varepsilon) &= V_{n+1}^{(-)}(x,\varepsilon) - (\beta(x,\xi) + \varphi_u^{(-)}(\xi)\gamma(\xi) + \tilde\varphi_u^{(-)}(\sigma)\tilde\gamma(\sigma) + H(\zeta))\varepsilon^{n+2}, \end{aligned} \tag{110}$$

$$\begin{aligned} \overline{U}(x,\varepsilon) &= U_{n+1}^{(-)}(x,\varepsilon) + \left(\alpha(x,\xi) + \gamma(\xi) + \tilde\gamma(\sigma)\right)\varepsilon^{n+2} - G(\zeta)\varepsilon^{n+4}, \\ \overline{V}(x,\varepsilon) &= V_{n+1}^{(-)}(x,\varepsilon) + (\beta(x,\xi) + \varphi_u^{(-)}(\xi)\gamma(\xi) + \tilde\varphi_u^{(-)}(\sigma)\tilde\gamma(\sigma) + H(\zeta))\varepsilon^{n+2}. \end{aligned} \tag{111}$$

Здесь $\alpha(x,\xi)$, $\beta(x,\xi)$ – решение линейной системы уравнений

$$\hat F_u(x,\xi)\alpha + \hat F_v(x,\xi)\beta = A, \quad \hat f_u(x,\xi)\alpha + \hat f_v(x,\xi)\beta = -kA, \tag{112}$$

где

$$\hat F_u(x,\xi) := \frac{\partial F}{\partial u}(\overline{u}_0^{(-)}(x) + \Pi_0^{(-)}u(\xi), \varphi(\overline{u}_0(x) + \Pi_0^{(-)}u(\xi), x), x, 0) = \overline{F}_u^{(-)}(x) + F_u^{(-)}(\xi) - \overline{F}_u^{(-)}(0) + O(\varepsilon), \tag{113}$$

$\overline{F}_u^{(-)}(x)$ и $F_u^{(-)}(\xi)$ определены в (10) и (41), $\hat F_v(x,\xi)$, $\hat f_u(x,\xi)$, $\hat f_v(x,\xi)$ имеют выражения, аналогичные выражениям для $\hat F_u(x,\xi)$, A и k – независящие от ε положительные числа, выбор которых уточняется ниже.

В силу условий A6, A7, A5 справедливы неравенства

$$\hat F_v(x,\xi) \le -c < 0, \quad \hat f_u(x,\xi) \ge c > 0, \quad \hat f_v(x,\xi) \le -c < 0, \quad x \in [0; x_\delta], \tag{114}$$

а в силу условия A4 – неравенство

$$\hat g_u(x,\xi) := \frac{\partial g}{\partial u}\left(\overline{u}_0^{(-)}(x) + \Pi_0^{(-)}u(\xi), x\right) \ge c > 0, \quad x \in [0, x_\delta]. \tag{115}$$

Так как

$$\hat g_u(x,\xi) = \hat F_u(x,\xi) - \hat F_v(x,\xi)\hat f_v^{-1}(x,\xi)\hat f_u(x,\xi),$$

то из (114) и (115) следует, что

$$\hat F_u(x,\xi) \ge c > 0, \quad x \in [0, x_\delta].$$

Из (114) и (115) следует также, что определитель $\Delta(x,\xi)$ линейной системы (112) удовлетворяет неравенству

$$\Delta(x,\xi) = \hat F_u(x,\xi)\hat f_v(x,\xi) - \hat F_v(x,\xi)\hat f_u(x,\xi) = \hat f_v(x,\xi)\hat g_u(x,\xi) \le -c < 0, \quad x \in [0, x_\delta]. \tag{116}$$

Поэтому система (112) имеет единственное решение

$$\begin{aligned} \alpha(x,\xi) &= (\hat f_v(x,\xi) + k\hat F_v(x,\xi))\Delta^{-1}(x,\xi)A \ge c(1+k)A, \\ \beta(x,\xi) &= -(\hat f_u(x,\xi) + k\hat F_u(x,\xi))\Delta^{-1}(x,\xi)A \ge c(1+k)A. \end{aligned} \tag{117}$$

Функция $\varphi_u^{(-)}(\xi)$, входящая в выражения для $\underline{V}$ и $\overline{V}$, определена в (45), функция $\tilde\varphi_u^{(-)}(\sigma)$ имеет аналогичное выражение:

$$\hat\varphi_u^{(-)}(\sigma) = -(\tilde f_v^{(-)}(\sigma))^{-1}\tilde f_u^{(-)}(\sigma) = \frac{\partial\varphi}{\partial u}(\overline{u}_0^{(-)}(x_\delta) + Q_0^{(-)}u(\sigma), x_\delta), \tag{118}$$

где

$$\tilde{f}_v^{(-)}(\sigma) := \frac{\partial f}{\partial v}(\bar{u}_0^{(-)}(x_\delta) + Q_0^{(-)}u(\sigma), \bar{v}_0^{(-)}(x_\delta) + Q_0^{(-)}v(\sigma), x_\delta, 0),$$
$$\tilde{f}_u^{(-)}(\sigma) := \frac{\partial f}{\partial u}(\bar{u}_0^{(-)}(x_\delta) + Q_0^{(-)}u(\sigma), \bar{v}_0^{(-)}(x_\delta) + Q_0^{(-)}v(\sigma), x_\delta, 0). \tag{119}$$

Ниже нам понадобится также оценка для $\tilde{f}_v^{(-)}(\sigma)$, аналогичная (43):

$$\tilde{f}_v^{(-)}(\sigma) \le -\varkappa < 0 \quad \text{при} \quad \sigma \le 0. \tag{120}$$

Функции $\gamma(\xi)$, $\tilde{\gamma}(\sigma)$, $G(\zeta)$, $H(\zeta)$, входящие в выражения (110) и (111), выбираются в точности так же, как в [1], и имеют оценки

$$0 \le \gamma(\xi) \le c(1+k)A\exp(-\varkappa\xi), \quad \xi \ge 0; \quad 0 \le \tilde{\gamma}(\sigma) \le c(1+k)A\exp(\varkappa\sigma), \quad \sigma \le 0,$$
$$0 \le G(\zeta) \le c(1+k)A\exp(-\varkappa\zeta), \quad 0 \le H(\zeta) \le c(1+k)A\exp(-\varkappa\zeta), \quad \zeta \ge 0. \tag{121}$$

Кроме того, эти функции умножаются на срезающие функции, в результате чего

$$\gamma(\xi) = 0, \quad G(\zeta) = 0, \quad H(\zeta) = 0 \quad \text{на отрезке} \quad [x_\delta/2; x_\delta],$$
$$\tilde{\gamma}(\sigma) = 0 \quad \text{на отрезке} \quad [0; x_\delta/2]. \tag{122}$$

Заметим теперь, что кривая

$$L_\varepsilon^{(-)} := \{(u, v, x) : u = U_n^{(-)}(x, \varepsilon),\ v = V_n^{(-)}(x, \varepsilon),\ x \in [0; x_\delta]\}$$

для достаточно малых ε расположена в такой малой окрестности кривой $L^{(-)}$ (см. (9)), в которой выполняются неравенства $\frac{\partial F}{\partial v}(u, v, x, \varepsilon) < 0$ и $\frac{\partial f}{\partial u}(u, v, x, \varepsilon) > 0$ в силу условий A6 и A7. Поэтому для достаточно малого ε_0 функции F и f удовлетворяют условию квазимонотонности в области G_0, определенной в (107), и, следовательно, для выполнения условия 2° из определения 1 достаточно, чтобы были выполнены неравенства (108) и (109). Проверка выполнения этих неравенств для достаточно больших A, k и достаточно малых ε проводится раздельно на промежутках $(0; x_\delta/2]$ и $[x_\delta/2; x_\delta]$, причем выполнение неравенств (108) на обоих промежутках и неравенств (109) на промежутке $(0; x_\delta/2]$ проверяется в точности так же, как в [1]. При этом число k можно считать фиксированным, а число A выбирается достаточно большим.

Отличие от работы [1] возникает при проверке выполнения неравенств (109) на отрезке $[x_\delta/2; x_\delta]$. Именно здесь выбор числа k играет важную роль. Рассмотрим отрезок $[x_\delta/2; x_\delta]$, учитывая, что на этом отрезке $\Pi_i^{(-)}u = \Pi_i^{(-)}v = P_i^{(-)}u = P_i^{(-)}v = 0$ (см. (66)), $\gamma(\xi) = 0$, $G(\zeta) = 0$, $H(\zeta) = 0$ (см. (122)), $\hat{F}_u(x, \xi) = \bar{F}_u^{(-)}(x)$, $\hat{F}_v(x, \xi) = \bar{F}_v^{(-)}(x)$, $\hat{f}_u(x, \xi) = \bar{f}_u^{(-)}(x)$, $\hat{f}_v(x, \xi) = \bar{f}_v^{(-)}(x)$, и, следовательно, (см. (117))

$$\alpha(x, \xi) = \bar{\alpha}(x) := (\bar{f}_v^{(-)}(x) + k\bar{F}_v^{(-)}(x))\Delta^{-1}(x)A,$$
$$\beta(x, \xi) = \bar{\beta}(x) := -(\bar{f}_u^{(-)}(x) + k\bar{F}_u^{(-)}(x))\Delta^{-1}(x)A, \tag{123}$$

где

$$\Delta^{-1} = (\bar{f}_v^{(-)}(x)\bar{g}_u^{(-)}(x))^{-1} \le -c < 0, \quad x \in [x_\delta/2; x_\delta], \tag{124}$$

а формулы (110) и (111) принимают вид

$$\underline{U}(x, \varepsilon) = U_{n+1}^{(-)}(x, \varepsilon) - (\bar{\alpha}(x) + \tilde{\gamma}(\sigma))\varepsilon^{n+2},$$
$$\underline{V}(x, \varepsilon) = V_{n+1}^{(-)}(x, \varepsilon) - (\bar{\beta}(x) + \tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma}(\sigma))\varepsilon^{n+2},$$
$$\overline{U}(x, \varepsilon) = U_{n+1}^{(-)}(x, \varepsilon) + (\bar{\alpha}(x) + \tilde{\gamma}(\sigma))\varepsilon^{n+2},$$
$$\overline{V}(x, \varepsilon) = V_{n+1}^{(-)}(x, \varepsilon) + (\bar{\beta}(x) + \tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma}(\sigma))\varepsilon^{n+2}.$$

Рассмотрим выражение для $M_\varepsilon(\underline{V},\underline{U})$ на отрезке $[x_\delta/2; x_\delta]$:

$$M_\varepsilon(\underline{V},\underline{U}) = M_\varepsilon(V_{n+1}^{(-)},U_{n+1}^{(-)}) - \varepsilon^2\frac{d\overline{\beta}}{dx}\varepsilon^{n+2} - \varepsilon\frac{d}{d\sigma}(\tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma})\varepsilon^{n+2} - [f(U_{n+1}^{(-)} - (\overline{\alpha}+\tilde{\gamma})\varepsilon^{n+2},$$
$$V_{n+1}^{(-)} - (\overline{\beta}+\tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma})\varepsilon^{n+2}, x, \varepsilon) - f(U_{n+1}^{(-)}, V_{n+1}^{(-)}, x, \varepsilon)] = M_\varepsilon(V_{n+1}^{(-)},U_{n+1}^{(-)}) + \quad (125)$$
$$+ O((1+k)A)\varepsilon^{n+3} + f_a(x,\varepsilon)(\overline{\alpha}+\tilde{\gamma})\varepsilon^{n+2} + f_v(x,\varepsilon)(\overline{\beta}+\tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma})\varepsilon^{n+2} + O((1+k)^2A^2)\varepsilon^{2n+4},$$

где

$$f_u(x,\varepsilon) := \frac{\partial f}{\partial u}(U_{n+1}^{(-)}, V_{n+1}^{(-)}, x, \varepsilon) = \tilde{f}_u(x,\sigma) + O(\varepsilon),$$
$$f_v(x,\varepsilon) := \frac{\partial f}{\partial v}(U_{n+1}^{(-)}, V_{n+1}^{(-)}, x, \varepsilon) = \tilde{f}_v(x,\sigma) + O(\varepsilon),$$
$$\tilde{f}_u(x,\sigma) := \frac{\partial f}{\partial u}(\tilde{u}_0^{(-)}(x) + Q_0^{(-)}u(\sigma), \overline{v}_0^{(-)}(x) + Q_0^{(-)}v(\sigma), x, 0),$$

$\tilde{f}_v(x,\sigma)$ имеет аналогичное выражение. Производные $\tilde{f}_u(x,\sigma)$ и $\tilde{f}_v(x,\sigma)$ представим в виде, аналогичном (113):

$$\tilde{f}_u(x,\sigma) = \overline{f}_u^{(-)}(x) + \tilde{f}_u^{(-)}(\sigma) - \overline{f}_u^{(-)}(x_*) + O(\varepsilon),$$
$$\tilde{f}_v(x,\sigma) = \overline{f}_v^{(-)}(x) + \tilde{f}_v^{(-)}(\sigma) - \overline{f}_v^{(-)}(x_*) + O(\varepsilon),$$

где $\overline{f}_u^{(-)}(x)$ и $\overline{f}_v^{(-)}(x)$ выражаются формулами типа (10), а $\tilde{f}_u^{(-)}(\sigma)$ и $\tilde{f}_v^{(-)}(\sigma)$ определены в (119).

Используя написанные выражения для производных, получаем

$$f_u(x,\varepsilon)(\overline{\alpha}+\tilde{\gamma}) + f_v(x,\varepsilon)(\overline{\beta}+\tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma}) = \overline{f}_u^{(-)}(x)\overline{\alpha}(x) + (\tilde{f}_u^{(-)}(\sigma) - \overline{f}_u^{(-)}(x_\delta))[\overline{\alpha}(x_\delta) + (\overline{\alpha}(x) - \overline{\alpha}(x_\delta))] +$$
$$+ [(\overline{f}_u^{(-)}(x) - \overline{f}_u^{(-)}(x_\delta)) + \tilde{f}_u^{(-)}(\sigma)]\tilde{\gamma}(\sigma) + \overline{f}_v^{(-)}(x)\overline{\beta}(x) + (\tilde{f}_v^{(-)}(\sigma) - \overline{f}_v^{(-)}(x_\delta))[\overline{\beta}(x_\delta) + (\overline{\beta}(x) - \overline{\beta}(x_\delta))] + \quad (126)$$
$$+ [(\overline{f}_v^{(-)}(x) - \overline{f}_v^{(-)}(x_\delta)) + \tilde{f}_v^{(-)}(\sigma)]\tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma}(\sigma) + O((1+k)A)\varepsilon =$$
$$= \tilde{f}_u^{(-)}(\sigma)\overline{\alpha}(x_\delta) + \tilde{f}_v^{(-)}(\sigma)\overline{\beta}(x_\delta) + O((1+k)A)\varepsilon,$$

мы воспользовались здесь равенствами

$$\overline{f}_u^{(-)}(x)\overline{\alpha}(x) + \overline{f}_v^{(-)}(x)\overline{\beta}(x) = -kA, \quad -\overline{f}_u^{(-)}(x_\delta)\overline{\alpha}(x_\delta) - \overline{f}_v^{(-)}(x_\delta)\overline{\beta}(x_\delta) = kA,$$
$$\tilde{f}_u^{(-)}(\sigma) + \tilde{f}_v^{(-)}(\sigma)\tilde{\varphi}_u^{(-)}(\sigma) = 0 \quad \text{(см. (118))},$$
$$(\tilde{f}_u^{(-)}(\sigma) - \overline{f}_u^{(-)}(x_\delta))(\overline{\alpha}(x) - \overline{\alpha}(x_\delta)) + (\overline{f}_u^{(-)}(x) - \overline{f}_u^{(-)}(x_\delta))\tilde{\gamma}(\sigma) + (\tilde{f}_v^{(-)}(\sigma) - \overline{f}_v^{(-)}(x_\delta))(\overline{\beta}(x) - \overline{\beta}(x_\delta)) +$$
$$+ (\overline{f}_v^{(-)}(x) - \overline{f}_v^{(-)}(x_\delta))\tilde{\varphi}_u^{(-)}(\sigma)\tilde{\gamma}(\sigma) = O((1+k)A)\varepsilon,$$

последнее равенство имеет место в силу экспоненциальных оценок типа (64) для разностей $(\tilde{f}_u^{(-)}(\sigma) - \overline{f}_u^{(-)}(x_\delta))$, $(\tilde{f}_v^{(-)}(\sigma) - \overline{f}_v^{(-)}(x_\delta))$ и оценки (121) для функции $\tilde{\gamma}(\sigma)$.

Введем обозначение

$$T(\sigma, x_\delta) := \tilde{f}_u^{(-)}(\sigma)\overline{\alpha}(x_\delta) + \tilde{f}_v^{(-)}(\sigma)\overline{\beta}(x_\delta).$$

Используя равенство (126) и учитывая, что

$$M_\varepsilon(V_{n+1}^{(-)},U_{n+1}^{(-)}) = O(\varepsilon^{n+2}) \quad \text{(см. (84))},$$

запишем равенство (125) в виде

$$M_\varepsilon(\underline{V},\underline{U}) = O(\varepsilon^{n+2}) + O((1+k)A)\varepsilon^{n+3} + T(\sigma, x_\delta)\varepsilon^{n+2} + O((1+k)^2A^2)\varepsilon^{2n+4}, \quad (127)$$

где первое слагаемое в правой части не зависит от A и k.

Докажем, что для достаточно большого k функция $T(\sigma, x_\delta)$ удовлетворяет неравенству

$$T(\sigma, x_\delta) \le -A, \quad \sigma \le 0. \quad (128)$$

С этой целью, используя выражения (123) для $\bar{\alpha}(x)$ и $\bar{\beta}(x)$ и равенство $\tilde{\varphi}_u^{(-)}(\sigma) = -(\tilde{f}_v^{(-)}(\sigma))^{-1}\tilde{f}_u^{(-)}(\sigma)$, запишем $T(\sigma, x_\delta)$ в виде

$$\begin{aligned} T(\sigma, x_\delta) &= [\tilde{f}_u^{(-)}(\sigma)(\bar{f}_v^{(-)}(x_\delta) + k\bar{F}_v^{(-)}(x_\delta)) - \tilde{f}_v^{(-)}(\sigma)(\bar{f}_u^{(-)}(x_\delta) + k\bar{F}_u^{(-)}(x_\delta))]\Delta^{-1}(x_\delta)A = \\ &= -[k(\bar{F}_u^{(-)}(x_\delta) + \bar{F}_v^{(-)}(x_\delta)\tilde{\varphi}_u^{(-)}(\sigma)) + (\bar{f}_u^{(-)}(x_\delta) + \bar{f}_v^{(-)}(x_\delta)\tilde{\varphi}_u^{(-)}(\sigma))]\tilde{f}_v^{(-)}(\sigma)\Delta^{-1}(x_\delta)A. \end{aligned} \tag{129}$$

Заметим, что (см. (118))

$$\tilde{\varphi}_u^{(-)}(\sigma) = \frac{\partial\varphi}{\partial u}(\bar{u}_0^{(-)}(x_\delta) + Q_0^{(-)}u(\sigma), x_\delta) = \frac{\partial\varphi}{\partial u}(u, x_\delta)$$

при

$$u = \bar{u}_0^{(-)}(x_\delta) + Q_0^{(-)}u(\sigma) = \psi_1(x_\delta) + Q_0^{(-)}u(\sigma),$$

причем

$$\psi_1(x_\delta) \le u = \psi_1(x_\delta) + Q_0^{(-)}u(\sigma) \le \psi_2(x_\delta) \quad \text{при} \quad \sigma \in (-\infty, 0]$$

в силу того, что $Q_0^{(-)}u(\sigma)$ – монотонная функция на полупрямой $-\infty < \sigma \le 0$. Поэтому

$$\bar{F}_u^{(-)}(x_\delta) + \bar{F}_v^{(-)}(x_\delta)\tilde{\varphi}_u^{(-)}(\sigma) = R^{(-)}(\psi_1(x_\delta) + Q_0^{(-)}u(\sigma), x_\delta) \ge c > 0$$

при $\sigma \le 0$ для достаточно малых ε в силу условия A8.

Так как $\tilde{f}_v^{(-)}(\sigma) \le -\kappa < 0$ (см. (120)) и $\Delta^{-1}(x_\delta) \le -c < 0$ (см. (124)), то для достаточно большого k из (129) получаем неравенство (128). В силу (128) из (127) следует неравенство

$$M_\varepsilon(\underline{V}, \underline{U}) \le O(\varepsilon^{n+2}) + O((1+k)A)\varepsilon^{n+3} - A\varepsilon^{n+2} + O((1+k)^2A^2)\varepsilon^{2n+4},$$

где первое слагаемое в правой части не зависит от A. Следовательно, для достаточно большого A и достаточно малых ε слагаемое $(-A\varepsilon^{n+2})$ в правой части обеспечит выполнение неравенства

$$M_\varepsilon(\underline{V}, \underline{U}) < 0, \quad x \in [x_\delta/2; x_\delta].$$

Аналогично доказывается, что для достаточно больших k, A и достаточно малых ε выполняется неравенство

$$M_\varepsilon(\overline{V}, \overline{U}) > 0, \quad x \in [x_\delta/2; x_\delta].$$

Выполнение условий 1° и 3° из определения 1 проверяется в точности так же, как в [1].

Таким образом, пары функций $\underline{U}$, $\underline{V}$ и $\overline{U}$, $\overline{V}$, определенные в (110) и (111), для достаточно больших чисел A и k и достаточно малых ε являются упорядоченными нижним и верхним решениями задачи (1), (104).

Отсюда следует, что эта задача имеет для достаточно малых ε решение $u = u^{(-)}(x, \varepsilon, \delta)$, $v = v^{(-)}(x, \varepsilon, \delta)$, удовлетворяющее неравенствам (106). В свою очередь, из этих неравенств, учитывая вид (110) и (111) нижнего и верхнего решений, получаем асимптотические равенства

$$\begin{aligned} u^{(-)}(x, \varepsilon, \delta) &= U_{n+1}^{(-)}(x, \varepsilon) + O\big((1+k)A\big)\varepsilon^{n+2}, \\ v^{(-)}(x, \varepsilon, \delta) &= V_{n+1}^{(-)}(x, \varepsilon) + O((1+k)A)\varepsilon^{n+2}, \quad x \in [0, x_\delta], \end{aligned} \tag{130}$$

откуда следуют равенства (105).

Теорема 1 доказана.

Следствие 1. Так как

$$U_{n+1}^{(-)}(x, \varepsilon) = U_n^{(-)}(x, \varepsilon) + O(\varepsilon^{n+1}), \quad V_{n+1}^{(-)}(x, \varepsilon) = V_n^{(-)}(x, \varepsilon) + O(\varepsilon^{n+1}),$$

то из (105) получаем

$$\begin{aligned} u^{(-)}(x, \varepsilon, \delta) &= U_n^{(-)}(x, \varepsilon) + O(\varepsilon^{n+1}), \\ v^{(-)}(x, \varepsilon, \delta) &= V_n^{(-)}(x, \varepsilon) + O(\varepsilon^{n+1}), \quad x \in [0, x_\delta]. \end{aligned} \tag{131}$$

Следствие 2. Имеет место равенство

$$v^{(-)}(x_\delta, \varepsilon, \delta) = V_{n+1}^{(-)}(X_{n+1}, \varepsilon) + O((1+k)A)\varepsilon^{n+2}. \tag{132}$$

Для доказательства справедливости этого равенства воспользуемся равенством (130) при $x = x_\delta$:

$$v^{(-)}(x_\delta, \varepsilon, \delta) = V_{n+1}^{(-)}(x_\delta, \varepsilon) + O((1+k)A)\varepsilon^{n+2}. \tag{133}$$

Так как функции $\Pi_i^{(-)}u$ и $P_i^{(-)}u$ равны нулю в точке x_δ (см. (66)) и $x_\delta = X_{n+1} + O(\varepsilon^{n+2})$ (см. (103)), то

$$\begin{aligned} V_{n+1}^{(-)}(x_\delta, \varepsilon) = \sum_{i=0}^{n+1} \varepsilon^i (\overline{v}_i^{(-)}(x_\delta) + Q_i^{(-)}v(0, x_\delta)) = \sum_{i=0}^{n+1} \varepsilon^i (\overline{v}_i^{(-)}(X_{n+1}) + Q_i^{(-)}v(0, X_{n+1})) + O(\varepsilon^{n+2}) = \\ = V_{n+1}^{(-)}(X_{n+1}, \varepsilon) + O(\varepsilon^{n+2}). \end{aligned} \tag{134}$$

Из (133) и (134) следует (132).

Следствие 3. Нетрудно доказать, что для производной $\dfrac{du^{(-)}}{dx}(x, \varepsilon, \delta)$ в точке x_δ имеет место равенство

$$\frac{du^{(-)}}{dx}(x_\delta, \varepsilon, \delta) = \sum_{i=0}^{n+1} \varepsilon^i \left(\frac{d\overline{u}_i^{(-)}}{dx}(x_\delta) + \frac{1}{\varepsilon}\frac{dQ_i^{(-)}u}{d\sigma}(0, x_\delta) \right) + O(\varepsilon^{n+1}). \tag{135}$$

Это равенство понадобится в разд. 5.

4. ВТОРАЯ ВСПОМОГАТЕЛЬНАЯ ЗАДАЧА

Рассмотрим теперь краевую задачу для системы (1) на отрезке $[x_\delta, 1]$, где x_δ имеет вид (103), т.е.

$$\begin{gathered} x_\delta = X_{n+1} + \varepsilon^{n+1}\delta = X_{n+1} + O(\varepsilon^{n+2}), \\ X_{n+1} = \sum_{i=0}^{n+1} \varepsilon^i \overline{x}_i, \end{gathered} \tag{136}$$

с краевыми условиями

$$u(x_\delta, \varepsilon) = \psi_2(x_\delta), \quad v(x_\delta, \varepsilon) = v^{(-)}(x_\delta, \varepsilon, \delta), \quad u(1, \varepsilon) = u^1, \tag{137}$$

$v^{(-)}(x_\delta, \varepsilon, \delta)$ выражается формулой (132).

Построим частичные суммы $U_{n+1}^{(+)}(x, \varepsilon)$ и $V_{n+1}^{(+)}(x, \varepsilon)$ рядов (14) и (15), в которых $x_* = x_\delta$. Отметим, что хотя в этом построении, описанном в п. 2.3, совсем не используется второе краевое условие из (137), тем не менее частичная сумма $V_{n+1}^{(+)}(x, \varepsilon)$ в точке x_δ отличается от $v^{(-)}(x_\delta, \varepsilon, \delta)$ на величину порядка $O((1+k)A)\varepsilon^{n+2}$, т.е.

$$v^{(-)}(x_\delta, \varepsilon, \delta) = V_{n+1}^{(+)}(x_\delta, \varepsilon) + O((1+k)A)\varepsilon^{n+2}. \tag{138}$$

Чтобы убедиться в этом, напишем для $V_{n+1}^{(+)}(x_\delta, \varepsilon)$ равенство, аналогичное равенству (134) для $V_{n+1}^{(-)}(x_\delta, \varepsilon)$:

$$V_{n+1}^{(+)}(x_\delta, \varepsilon) = V_{n+1}^{(+)}(X_n, \varepsilon) + O(\varepsilon^{n+2}). \tag{139}$$

Величины $V_{n+1}^{(\mp)}(X_{n+1}, \varepsilon)$ представим в виде (см. (95) и (96))

$$V_{n+1}^{(\mp)}(X_{n+1}, \varepsilon) = \sum_{i=0}^{n+1} \varepsilon^i v_i^{(\mp)}(0) + O(\varepsilon^{n+2}),$$

а так как для X_{n+1} вида (136) справедливы равенства

$$v_i^{(-)}(0) = v_i^{(+)}(0) \quad \text{при} \quad i = 0, 1, \ldots, n+1,$$

то

$$V_{n+1}^{(-)}(X_{n+1},\varepsilon) = V_{n+1}^{(+)}(X_{n+1},\varepsilon) + O(\varepsilon^{n+2}). \tag{140}$$

Из (134), (140) и (139) получаем равенство

$$V_{n+1}^{(-)}(x_\delta,\varepsilon) = V_{n+1}^{(+)}(x_\delta,\varepsilon) + O(\varepsilon^{n+2}),$$

в силу которого из (133) следует искомое равенство (138).

Второе краевое условие в (137) можно теперь записать в виде

$$v(x_\delta,\varepsilon) = V_{n+1}^{(+)}(x_\delta,\varepsilon) + O((1+k)A\varepsilon^{n+2}). \tag{141}$$

Теорема 2. *Если выполнены условия* A1–A8, *то для достаточно малых* ε *задача* (1), (137) *имеет решение* $u = u^{(+)}(x,\varepsilon,\delta)$, $v = v^{(+)}(x,\varepsilon,\delta)$, *для которого справедливы асимптотические равенства*

$$\begin{aligned} u^{(+)}(x,\varepsilon,\delta) &= U_{n+1}^{(+)}(x,\varepsilon) + O(\varepsilon^{n+2}),\\ v^{(+)}(x,\varepsilon,\delta) &= V_{n+1}^{(+)}(x,\varepsilon) + O(\varepsilon^{n+2}), \quad x \in [x_\delta,1]. \end{aligned} \tag{142}$$

Доказательство теоремы 2 проводится аналогично доказательству теоремы 1 с помощью асимптотического метода дифференциальных неравенств. Нижнее решение задачи (1), (137) строится в виде

$$\underline{U}(x,\varepsilon) = U_{n+1}^{(+)}(x,\varepsilon) - (\alpha(x,\tilde\xi) + \gamma(\tilde\xi) + \tilde\gamma(\sigma))\varepsilon^{n+2},$$

$$\underline{V}(x,\varepsilon) = V_{n+1}^{(+)}(x,\varepsilon) - (\beta(x,\tilde\xi) + \varphi_u^{(+)}(\tilde\xi)\gamma(\tilde\xi) + \tilde\varphi_u^{(+)}(\sigma)\tilde\gamma(\sigma))\varepsilon^{n+2},$$

где $\varphi_u^{(+)}(\tilde\xi) = \dfrac{\partial\varphi}{\partial u}(u_0^{(+)}(1) + \Pi_0^{(+)}u(\tilde\xi),1)$, $\tilde\varphi_u^{(+)}(\sigma) = \dfrac{\partial\varphi}{\partial u}(\overline{u}_0^{(+)}(x_\delta) + Q_0^{(+)}u(\sigma),x_\delta)$, а функции α, β, γ, $\tilde\gamma$ определяются так же, как аналогичные функции в (110) с заменой в системе (112) числа A на число B, которое выбирается столь большим, чтобы было выполнено неравенство

$$\underline{V}(x_\delta,\varepsilon) = V_{n+1}^{(+)}(x_\delta,\varepsilon) - \overline{\beta}(x_\delta)\varepsilon^{n+2} \le v(x_\delta,\varepsilon);$$

здесь $\overline{\beta}(x_\delta)$ выражается формулой (123) с заменой A на B, а $v(x_\delta,\varepsilon)$ – формулой (141).

Верхнее решение имеет вид, аналогичный нижнему решению, нужно только знак минус перед суммами в круглых скобках заменить на знак плюс.

В процессе доказательства используются формулы (83)–(85) для $k = n+1$.

Следствие 1. Так как $U_{n+1}^{(+)}(x,\varepsilon) = U_n^{(+)}(x,\varepsilon) + O(\varepsilon^{n+1})$, $V_{n+1}^{(+)}(x,\varepsilon) = V_n^{(+)}(x,\varepsilon) + O(\varepsilon^{n+1})$, то из (142) получаем

$$\begin{aligned} u^{(+)}(x,\varepsilon,\delta) &= U_n^{(+)}(x,\varepsilon) + O(\varepsilon^{n+1}),\\ v^{(+)}(x,\varepsilon,\delta) &= V_n^{(+)}(x,\varepsilon) + O(\varepsilon^{n+1}), \quad x \in [x_\delta,1]. \end{aligned} \tag{143}$$

Следствие 2. Для производной $\dfrac{du^{(+)}}{dx}(x,\varepsilon,\delta)$ в точке x_δ имеет место асимптотическое равенство

$$\frac{du^{(+)}}{dx}(x_\delta,\varepsilon,\delta) = \sum_{i=0}^{n+1}\varepsilon^i\left(\frac{d\overline{u}_i^{(+)}}{dx}(x_\delta) + \frac{1}{\varepsilon}\frac{dQ_i^{(+)}u}{d\sigma}(0,x_\delta)\right) + O(\varepsilon^{n+1}). \tag{144}$$

Это равенство понадобится в следующем разделе.

5. ТЕОРЕМА О СУЩЕСТВОВАНИИ И АСИМПТОТИКЕ КСТС

Решения $u^{(-)}(x,\varepsilon,\delta)$, $v^{(-)}(x,\varepsilon,\delta)$ и $u^{(+)}(x,\varepsilon,\delta)$, $v^{(+)}(x,\varepsilon,\delta)$ непрерывно сшиваются в точке x_δ, так как (см. (104) и (137))

$$u^{(+)}(x_\delta,\varepsilon,\delta) = u^{(-)}(x_\delta,\varepsilon,\delta) = \psi_2(x_\delta), \quad v^{(+)}(x_\delta,\varepsilon,\delta) = v^{(-)}(x_\delta,\varepsilon,\delta). \tag{145}$$

Поэтому функции

$$u(x,\varepsilon)=\begin{cases}u^{(-)}(x,\varepsilon,\delta), & x\in[0,x_\delta],\\ u^{(+)}(x,\varepsilon,\delta), & x\in[x_\delta,1],\end{cases}\qquad v(x,\varepsilon)=\begin{cases}v^{(-)}(x,\varepsilon,\delta), & x\in[0,x_\delta],\\ v^{(+)}(x,\varepsilon,\delta), & x\in[x_\delta,1],\end{cases}$$

будут решением исходной задачи (1), (2), представляющим собой контрастную структуру типа ступеньки, если в точке x_δ непрерывно сшиваются производные $\frac{du^{(+)}}{dx}$ и $\frac{du^{(-)}}{dx}$, т.е. если имеет место равенство

$$\frac{du^{(-)}}{dx}(x_\delta,\varepsilon,\delta)-\frac{du^{(+)}}{dx}(x_\delta,\varepsilon,\delta)=0. \tag{146}$$

Заметим, что аналогичное равенство для $\frac{dv^{(-)}}{dx}$ и $\frac{dv^{(+)}}{dx}$ заведомо выполняется, так как в силу второго уравнения (1) и равенств (145) справедливы равенства

$$\frac{dv^{(-)}}{dx}(x_\delta,\varepsilon,\delta)=\varepsilon^{-2}f(u^{(-)}(x_\delta,\varepsilon,\delta),v^{(-)}(x_\delta,\varepsilon,\delta),x_\delta,\varepsilon)=$$
$$=\varepsilon^{-2}f(u^{(+)}(x_\delta,\varepsilon,\delta),v^{(+)}(x_\delta,\varepsilon,\delta),x_\delta,\varepsilon)=\frac{dv^{(+)}}{dx}(x_\delta,\varepsilon,\delta).$$

Докажем, что существует $\delta=O(\varepsilon)$, для которого выполнено (146). Используя асимптотические формулы (135) и (144) для производных в левой части равенства (146) и умножив его на ε, получим уравнение относительно δ, которое запишем в виде

$$\sum_{i=0}^{n+1}\varepsilon^i\left(\frac{dQ_i^{(-)}u}{d\sigma}(0,x_\delta)-\frac{dQ_i^{(+)}u}{d\sigma}(0,x_\delta)\right)+\varepsilon\sum_{i=0}^{n+1}\varepsilon^i\left(\frac{d\overline{u}_i^{(-)}}{dx}(x_\delta)-\frac{d\overline{u}_i^{(+)}}{dx}(x_\delta)\right)=O(\varepsilon^{n+2}).$$

Раскладывая левую часть уравнения по степеням ε и учитывая выражение (103) для x_δ, получаем:

$$J(\overline{x}_0)+\sum_{i=1}^{n}\varepsilon^i(J'(\overline{x}_0)\overline{x}_j+k_j)+\varepsilon^{n+1}(J'(\overline{x}_0)(\overline{x}_{n+1}+\delta)+k_{n+1})=O(\varepsilon^{n+2}),$$

где правая часть зависит от δ, но имеет указанный порядок малости равномерно относительно δ в фиксированной окрестности точки $\delta=0$.

Так как $\overline{x}_0$ и $\overline{x}_i$ $(i=1,2,\ldots)$ являются решениями уравнений (7) и (90), то уравнение относительно δ принимает вид

$$J'(\overline{x}_0)\delta=O(\varepsilon).$$

Оно имеет решение $\delta=\overline{\delta}(\varepsilon)=O(\varepsilon)$, так как $J'(\overline{x}_0)\neq 0$.

Следовательно, функции

$$u(x,\varepsilon)=\begin{cases}u^{(-)}(x,\varepsilon,\overline{\delta}), & x\in[0,x_{\overline{\delta}}],\\ u^{(+)}(x,\varepsilon,\overline{\delta}), & x\in[x_{\overline{\delta}},1],\end{cases}\qquad v(x,\varepsilon)=\begin{cases}v^{(-)}(x,\varepsilon,\overline{\delta}), & x\in[0,x_{\overline{\delta}}],\\ v^{(+)}(x,\varepsilon,\overline{\delta}), & x\in[x_{\overline{\delta}},1],\end{cases}$$

являются решением задачи (1), (2) с внутренним переходным слоем в окрестности точки $x_{\overline{\delta}}$. При этом в силу (131) и (143) справедливы равенства

$$u(x,\varepsilon)=U_n(x,\varepsilon)+O(\varepsilon^{n+1}),$$
$$v(x,\varepsilon)=V_n(x,\varepsilon)+O(\varepsilon^{n+1}),\quad x\in[0;1], \tag{147}$$

где

$$U_n(x,\varepsilon)=\begin{cases}U_n^{(-)}(x,\varepsilon), & x\in[0,x_{\overline{\delta}}],\\ U_n^{(+)}(x,\varepsilon), & x\in[x_{\overline{\delta}},1],\end{cases}\qquad V_n(x,\varepsilon)=\begin{cases}V_n^{(-)}(x,\varepsilon), & x\in[0,x_{\overline{\delta}}],\\ V_n^{(+)}(x,\varepsilon), & x\in[x_{\overline{\delta}},1],\end{cases} \tag{148}$$

причем в формулах для $U_n^{(\pm)}$, $V_n^{(\pm)}$ внутрислойная переменная $\sigma=\sigma_{\overline{\delta}}:=(x-x_{\overline{\delta}})/\varepsilon$.

Формулы (147) и (148) имеют тот недостаток, что величина $\bar{\delta}$ точно не известна, известен только ее порядок ($\bar{\delta} = O(\varepsilon)$), и, следовательно, $x_{\bar{\delta}}$ и $\sigma_{\bar{\delta}}$ тоже не определены точно. Заменим $x_{\bar{\delta}}$ в формулах (148) на $X_{n+1} := \sum_{i=0}^{n+1} \varepsilon^i \bar{x}_i$, т.е. в выражении для $x_{\bar{\delta}}$ отбросим последнее слагаемое $\varepsilon^{n+1}\bar{\delta} = O(\varepsilon^{n+2})$. Тогда аргумент $Q_i^{(\pm)}$-функций, т.е. $\sigma_{\bar{\delta}} = (x - x_{\bar{\delta}})/\varepsilon$, изменится на величину порядка $O(\varepsilon^{n+1})$, и, значит, эти функции изменятся на величину того же порядка. Поэтому при замене $x_{\bar{\delta}}$ на X_{n+1} в формулах (148) формулы (147) не изменятся.

Таким образом, мы доказали следующую основную теорему.

Теорема 3. *Если выполнены условия* A1–A8, *то для достаточно малых* ε *задача* (1), (2) *имеет решение* $u(x,\varepsilon)$, $v(x,\varepsilon)$, *для которого справедливы асимптотические равенства* (147), *где*

$$U_n(x,\varepsilon) = \begin{cases} U_n^{(-)}(x,\varepsilon), & x \in [0, X_{n+1}], \\ U_n^{(+)}(x,\varepsilon), & x \in [X_{n+1}, 1], \end{cases} \quad V_n(x,\varepsilon) = \begin{cases} V_n^{(-)}(x,\varepsilon), & x \in [0, X_{n+1}], \\ V_n^{(+)}(x,\varepsilon), & x \in [X_{n+1}, 1], \end{cases}$$

$X_{n+1} = \sum_{i=0}^{n+1} \varepsilon^i \bar{x}_i$, *а функции* $U_n^{(-)}$, $V_n^{(-)}$ *и* $U_n^{(+)}$, $V_n^{(+)}$ *определены формулами* (67), (68) *и* (82) *при* $k = n$, *причем внутрислойная переменная* σ *имеет вид*

$$\sigma = \sigma_{n+1} := (x - X_{n+1})/\varepsilon.$$

6. ЗАКЛЮЧИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

6.1. Из (147) следуют предельные равенства

$$\lim_{\varepsilon\to 0} u(x,\varepsilon) = \begin{cases} \psi_1(x), & x \in (0, \bar{x}_0), \\ \psi_3(x), & x \in (\bar{x}_0, 1), \end{cases}$$

$$\lim_{\varepsilon\to 0} v(x,\varepsilon) = \begin{cases} \varphi(\psi_1(x), x), & x \in (0, \bar{x}_0), \\ \varphi(\psi_3(x), x), & x \in (0, \bar{x}_0). \end{cases}$$

Можно сказать, что в пределе при $\varepsilon \to 0$ в точке $\bar{x}_0$ происходит скачок решения.

6.2. Предельным положением при $\varepsilon \to 0$ кривой

$$L_\varepsilon = \{(u, v, x) : u = u(x,\varepsilon),\ v = v(x,\varepsilon),\ x \in [0;1]\}$$

(ее можно назвать графиком решения задачи (1), (2)) является кривая $L = L^{(-)} \cup L^{(+)}$ (см. (9)).

6.3. Представляет интерес рассмотрение задачи (1), (2) в случае, когда корень $v = \varphi(u, x)$ уравнения $f(u, v, x, 0) = 0$ является кратным. Некоторые задачи о контрастных структурах в случаях кратного корня вырожденного уравнения рассмотрены в [6]–[8]. В этих случаях асимптотика решения в переходном слое имеет свои характерные особенности.

6.4. Построенное в задаче (1), (2) решение $u(x,\varepsilon)$, $v(x,\varepsilon)$ типа контрастной структуры является стационарным решением нестационарной частично диссипативной системы уравнений, которая получается из системы (1) добавлением в левую часть первого уравнения слагаемого $\left(-\varepsilon^2 w(x)\dfrac{\partial u}{\partial t}\right)$ и в левую часть второго уравнения слагаемого $\varepsilon^2 \dfrac{dv}{dt}$. Встают вопросы об устойчивости при $t \to \infty$ построенного стационарного решения нестационарной системы уравнений и о его области притяжения, т.е. о множестве начальных функций, для которых решение начально-краевой задачи для нестационарной системы стремится при $t \to \infty$ к стационарному решению. Такая задача для стационарного погранслойного решения нестационарной частично диссипативной системы рассмотрена в [1].

СПИСОК ЛИТЕРАТУРЫ

1. *Бутузов В.Ф.* Асимптотика и устойчивость стационарного погранслойного решения частично диссипативной системы уравнений // Ж. вычисл. матем. и матем. физ. 2019. V. 59. № 7. С. 1201–1229.
2. *Васильева А.Б., Бутузов В.Ф., Нефедов Н.Н.* Контрастные структуры в сингулярно возмущенных задачах // Фундаментальная и приклад. матем. 1998. V. 4. № 3. С. 799–851.

3. *Бутузов В.Ф., Неделько И.В.* Контрастная структура типа ступеньки в сингулярно возмущенной системе эллиптических уравнений с разными степенями малого параметра // Ж. вычисл. матем. и матем. физ. 2000. V. 40 № 6. С. 877–899.
4. *Васильева А.Б., Бутузов В.Ф., Нефедов Н.Н.* Сингулярно возмущенные задачи с пограничными и внутренними слоями // Тр. МИАН. 2010. V. 268. № 2. С. 268–283.
5. *Бутузов В.Ф., Левашова Н.Т., Мельникова А.А.* Контрастная структура типа ступеньки в сингулярно возмущенной системе уравнений с различными степенями малого параметра // Ж. вычисл. матем. и матем. физ. 2012. V. 52. № 11. С. 1983–2003.
6. *Бутузов В.Ф.* Сингулярно возмущенная краевая задача с многозонным внутренним переходным слоем // Моделирование и анализ информ. систем. 2015. V. 22. № 1. С. 5–22.
7. *Бутузов В.Ф.* Об асимптотике решения сингулярно возмущенной параболической задачи с многозонным внутренним переходным слоем // Ж. вычисл. матем. и матем. физ. 2019. 2018. V. 58. № 6. С. 961–987.
8. *Бутузов В.Ф.* Асимптотика контрастной структуры типа всплеска в задаче с кратным корнем вырожденного уравнения // Дифференц. ур-ния. 2019. V. 55. № 6. С. 774–791.
9. *Васильева А.Б., Бутузов В.Ф.* Асимптотические методы в теории сингулярных возмущений. М.: Высшая школа, 1990. 208 с.
10. *Нефедов Н.Н.* Метод дифференциальных неравенств для некоторых классов нелинейных сингулярно возмущенных задач с внутренними слоями // Дифференц. ур-ния. 1995. V. 31. № 7. С. 1132–1139.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 85–94

УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

УДК 517.958

ТЕОРЕМЫ ЕДИНСТВЕННОСТИ И СУЩЕСТВОВАНИЯ РЕШЕНИЯ ЗАДАЧ РАССЕЯНИЯ ЭЛЕКТРОМАГНИТНЫХ ВОЛН НА ТРЕХМЕРНЫХ АНИЗОТРОПНЫХ ТЕЛАХ В ДИФФЕРЕНЦИАЛЬНОЙ И ИНТЕГРАЛЬНОЙ ПОСТАНОВКЕ[1)]

© 2021 г. А. Б. Самохин[1,*], Ю. Г. Смирнов[2]

[1] 119454 Москва, пр-т Вернадского, 78, МИРЭА, Российский технологический университет, Россия

[2] 440026 Пенза, ул. Красная, 40, Пензенский государственный университет, Россия

*e-mail: absamokhin@yandex.ru

Поступила в редакцию 15.03.2020 г.
Переработанный вариант 26.07.2020 г.
Принята к публикации 18.09.2020 г.

Доказаны теоремы о единственности решения уравнений Максвелла для задач рассеяния электромагнитных волн на ограниченных трехмерных неоднородных анизотропных телах, в том числе без потерь и с разрывами параметров среды. Доказаны теоремы о существовании и единственности решений объемных сингулярных интегральных уравнений, отвечающих задачам рассеяния электромагнитных волн на ограниченных трехмерных неоднородных анизотропных телах, в том числе без потерь и с разрывами параметров. Библ. 14.

Ключевые слова: задачи рассеяния электромагнитных волн, уравнения Максвелла, среды без потерь, анизотропные среды, объемные сингулярные интегральные уравнения.

DOI: 10.31857/S0044466921010075

ВВЕДЕНИЕ

Вопросы существования и единственности решения задач рассеяния электромагнитных волн на ограниченных диэлектрических трехмерных телах Q, находящихся в свободном пространстве, имеют большое значение как с теоретической, так и с практической точек зрения. Как правило, при доказательстве единственности используются уравнения Максвелла с соответствующими условиями сопряжения и условием излучения на бесконечности. Для доказательства существования решения исходная задача сводится к интегральному уравнению. Тогда, если оператор интегрального уравнения является фредгольмовым в соответствующем функциональном пространстве, доказательство завершается и формулируется теорема существования и единственности решения. По такой схеме были доказаны теоремы для следующих классов задач рассеяния:

– неоднородная среда характеризуется гладкой во всем пространстве $\mathbb{R}^3$ скалярной функцией диэлектрической проницаемости $\varepsilon(x)$ или тензор-функцией диэлектрической проницаемости $\hat{\varepsilon}(x)$, а магнитная проницаемость всюду постоянна, при этом среда в Q может не иметь потерь (см. [1]–[3]);

– неоднородная среда характеризуется всюду гладкими в $\mathbb{R}^3$ тензор-функциями диэлектрической $\hat{\varepsilon}(x)$ и магнитной $\hat{\mu}(x)$ проницаемостями, при этом среда в Q может не иметь потерь (см. [4], [5]);

– область неоднородности является поглощающей, при этом тензор-функции $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ могут иметь разрывы, в том числе на границе Q (см. [4], [5]).

Случай, когда область неоднородности среды не имеет потерь, а параметры среды являются разрывными, в том числе на ∂Q, является наиболее сложным. Результаты в этом направлении получены в [6]–[9]. В [6] доказана теорема о единственности решения задачи рассеяния в обоб-

[1)]Работа выполнена при финансовой поддержке РНФ (проект № 20-11-20087).

щенной постановке (условия сопряжения на границе тела не ставятся) в анизотропной среде, которая характеризуется вещественными тензор-функциями $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$, имеющими разрывы, в частности, на ∂Q. В [7]–[9] доказана теорема о существовании и единственности решения задачи рассеяния на изотропном теле, которое характеризуется вещественной функцией диэлектрической проницаемости $\varepsilon(x) > 0$, также имеющей разрывы на ∂Q.

В настоящей работе доказаны теоремы единственности решения задач рассеяния в классической постановке (с условиями сопряжения на границе тел) на анизотропных телах, в том числе без потерь и имеющих разрывы параметров – диэлектрической и магнитной проницаемостей. Предложенный метод доказательства является оригинальным и отличается от используемых в перечисленных выше работах. Также доказаны теоремы о существовании и единственности решения системы объемных сингулярных интегральных уравнений, к которой сводится исходная краевая задача.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть Q – ограниченная область в пространстве $\mathbb{R}^3$. Будем предполагать, что граница ∂Q области Q кусочно-гладкая (точное определение см. в [10]).

Будем рассматривать следующий класс задач электродинамики. В области Q среда характеризуется тензорами диэлектрической и магнитной проницаемости $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ (матрицы-функции размерности 3×3), причем компоненты этих тензоров являются кусочно-дифференцируемыми функциями координат. Точнее, пусть область Q состоит из конечного числа подобластей Q_i с кусочно-гладкой границей ∂Q_i; $\overline{Q} = \bigcup_i \overline{Q}_i$, $Q_i \cap Q_j = \varnothing$ при $i \neq j$. Предположим, что $\hat{\varepsilon} \in C^3(\overline{Q_i})$, $\hat{\mu} \in C^3(\overline{Q_i})$ для всех i. Точнее, будем предполагать, что $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ являются сужениями на Q_i функций, заданных на более широком множестве, т.е. $\hat{\varepsilon}(x) = \hat{\varepsilon}_i(x)$, $\hat{\mu}(x) = \hat{\mu}_i(x)$ при $x \in Q_i$, $\hat{\varepsilon}_i \in C^3(\overline{B})$, $\hat{\mu}_i \in C^3(\overline{B})$, где B – (открытый) шар, содержащий Q, $\overline{Q} \subset B$. На ∂Q_i будем определять только предельные значения функций $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ с разных сторон в точках гладкости поверхности.

Вне области Q (в $\mathbb{R}^3 \backslash \overline{Q}$) среда изотропна с постоянными параметрами, $\varepsilon = \varepsilon_0$ и $\mu = \mu_0$. Требуется определить электромагнитное поле, возбуждаемое в данной среде внешним полем с временной зависимостью в виде множителя $\exp(-i\omega t)$, источником которого может быть как падающая плоская волна, так и сторонний ток $\vec{J}^0$.

В такой постановке соответствующая математическая задача формулируется следующим образом: найти векторные непрерывно дифференцируемые в Q_i и вне $\overline{Q}$ функции электромагнитного поля, удовлетворяющие в областях гладкости параметров среды уравнениям Максвелла

$$\operatorname{rot} \mathbf{H} = -i\omega\hat{\varepsilon}\mathbf{E} + \mathbf{J}^0, \quad \operatorname{rot} \mathrm{E} - i\omega\hat{\mu}\mathbf{H} \tag{1}$$

и условию излучения на бесконечности

$$\lim_{r\to\infty}\left(r\frac{\partial u}{\partial r} - ik_0 ru\right) = 0, \quad r := |x| = \sqrt{x_1^2 + x_2^2 + x_3^2}, \tag{2}$$

где $k_0 = \omega\sqrt{\varepsilon_0\mu_0}$, ($\operatorname{Im}\varepsilon_0 = 0$, $\operatorname{Im}\mu_0 = 0$, $\operatorname{Re}\varepsilon_0 > 0$, $\operatorname{Re}\mu_0 > 0$), а u – любая из декартовых компонент полей $\mathbf{E}$ или $\mathbf{H}$. В (1) $\mathbf{J}^0$ – заданный электрический ток, создающий внешнее поле $\mathbf{E}^0$, $\mathbf{H}^0$. Далее, на гладких частях поверхностей разрыва проницаемостей ∂Q_i функции $\mathbf{E}$ и $\mathbf{H}$ должны быть непрерывны вплоть до ∂Q_i (с каждой стороны) и удовлетворять условию непрерывности тангенциальных компонент полей:

$$[\mathbf{E}_\tau]\big|_{\partial Q_i} = 0, \quad [\mathbf{H}_\tau]\big|_{\partial Q_i} = 0, \tag{3}$$

где $[\,\cdot\,]_{\partial Q_i}$ означает разность следов с разных сторон ∂Q_i, τ – касательный вектор к ∂Q_i. Мы не будем вводить новые обозначения именно для гладких частей ∂Q_i, а будем в случае необходимости оговаривать особо это обстоятельство.

Кроме того, поля $\mathbf{E}$ и $\mathbf{H}$ должны удовлетворять условию ограниченности энергии в любом конечном объеме пространства, т.е. условию

$$\mathbf{E}, \mathbf{H} \in L_{2,loc}(\mathbb{R}^3). \tag{4}$$

Решения задачи (1)–(4) будем называть *классическими*.

2. ЕДИНСТВЕННОСТЬ РЕШЕНИЯ ЗАДАЧИ ДЛЯ ИЗОТРОПНОЙ СРЕДЫ

Для большей прозрачности доказательств сначала рассмотрим задачу, для которой магнитная проницаемость всюду постоянна и равна μ_0. В этом разделе предположим, что $\hat{\varepsilon}(x) = \varepsilon(x)$ – скалярная функция, $\operatorname{Re}\varepsilon_i(x) > 0$, $\operatorname{Im}\varepsilon_i(x) \geq 0$.

Однородные уравнения Максвелла для такой задачи имеют вид

$$\operatorname{rot}\mathbf{H} = -i\omega\varepsilon\mathbf{E}, \quad \operatorname{rot}\mathbf{E} = i\omega\mu_0\mathbf{H}. \tag{5}$$

С помощью леммы Реллиха, используя условия на бесконечности (2), стандартным способом доказывается (см. [6]–[9]), что в области $\mathbb{R}^3\backslash\overline{Q}$ электромагнитное поле, удовлетворяющее (2)–(5), равно нулю. Тогда из (5), учитывая теорему Стокса, следует, что тангенциальные компоненты электрического и магнитного полей на ∂Q равны нулю.

Рассмотрим однородную задачу (2)–(5). Обозначим эту задачу через A. Предположим, что в области Q электромагнитное поле не равно тождественно нулю. Не ограничивая общности, можно считать, что во всех подобластях Q_i (для всех i) поле $\mathbf{E}$, $\mathbf{H}$ не равно тождественно нулю. Действительно, в противном случае области Q_i, в которых поля тождественно равны нулю, можно исключить из Q и рассмотреть задачу в оставшейся части Q с теми же граничными условиями.

Определим функцию-срезку со следующими свойствами: $\zeta(t;a,b) \in C^\infty(\mathbf{R}^1)$, $0 \leq \zeta(t;a,b) \leq 1$, $\zeta(t;a,b) = 1$ при $t \leq a$, $\zeta(t;a,b) = 0$ при $t \geq b$, $\zeta(t;a,b) > 0$ при $a < t < b$ $(0 < a < b)$. Пусть шар B, содержащий Q (см. разд. 1), имеет центр в $x_0 \in Q$ и радиус $2d$, где d – диаметр области Q, т.е. максимальное расстояние между точками границы ∂Q. Определим функции $\varepsilon_i^c(x) := (\varepsilon_i(x) - \varepsilon_0)\zeta(|x - x_0|; d, 2d) + \varepsilon_0$. Из свойств функции-срезки следует, что $\varepsilon_i^c \in C^3(\mathbb{R}^3)$, $\varepsilon_i^c(x) = \varepsilon(x)$ при $x \in Q_i$, $\varepsilon_i^c(x) = \varepsilon_0$ при $x \notin B$. Кроме того, записывая функцию в виде $\varepsilon_i^c(x) = \varepsilon_i(x)\zeta(|x - x_0|; d, 2d) + \varepsilon_0(1 - \zeta(|x - x_0|; d, 2d))$, получаем, что если $\operatorname{Re}\varepsilon_i(x) > 0$, то и $\operatorname{Re}\varepsilon_i^c(x) > 0$ при $x \in B$.

Так как, по предположению, для задачи A во всех подобластях Q_i поле $\mathbf{E}$, $\mathbf{H}$ не равно тождественно нулю, то найдется такое j, что $\partial Q_j \cap \partial Q \neq \varnothing$, т.е. у Q_j и Q имеется общий гладкий кусок границы.

Теперь рассмотрим вспомогательную задачу C. Задача C описывается уравнениями (2)–(5), в которых диэлектрическая проницаемость $\varepsilon^c(x)$ в области Q совпадает с диэлектрической проницаемостью из задачи A, а вне области Q диэлектрическая проницаемость $\varepsilon^c(x)$ равна $\varepsilon_j^c(x)$ для указанного выше j.

Теперь опишем множество решений задачи C. Поскольку $\varepsilon_j^c(x) = \varepsilon_0$ при $x \notin B$, то, аналогично изложенному выше, получим, что в области $\mathbb{R}^3\backslash\overline{B}$ электромагнитное поле равно нулю. Обозначим $Q^j := Q\backslash\overline{Q_j}$. Из построения функции $\varepsilon_j^c(x)$ следует, что $\varepsilon_j^c \in C^3(\mathbb{R}^3\backslash\overline{Q^j})$. Тогда из однородных уравнений (5) следует уравнение относительно электрического поля в области $\mathbb{R}^3\backslash\overline{Q^j}$

$$\operatorname{rot}\operatorname{rot}\mathbf{E} - \varepsilon_j^c\mu_0\omega^2\mathbf{E} = 0. \tag{6}$$

Далее, поскольку $\operatorname{div}\operatorname{rot} \equiv 0$, из первого уравнения (5) следует, что $\operatorname{div}(\varepsilon_j^c\mathbf{E}) = 0$, поэтому получим

$$\operatorname{div}\mathbf{E} = -\frac{1}{\varepsilon_j^c}(\operatorname{grad}\varepsilon_j^c, \mathbf{E}), \tag{7}$$

где круглые скобки обозначают скалярное произведение векторов. Теперь из (6), (7), учитывая тождество rot rot $= -\Delta + \operatorname{grad}\operatorname{div}$, получаем следующее уравнение:

$$\Delta\mathbf{E} + \operatorname{grad}\frac{1}{\varepsilon_j^c}(\operatorname{grad}\varepsilon_j^c, \mathbf{E}) + \varepsilon_j^c\mu_0\omega^2\mathbf{E} = 0. \tag{8}$$

Поскольку $\varepsilon_j^c(x) \neq 0$, то уравнение (8) является эллиптическим в $\mathbb{R}^3\backslash\overline{Q^j}$. Далее, все коэффициенты уравнения (8) являются функциями из $C^2(\mathbb{R}^3\backslash\overline{Q^j})$. Поэтому можно применить принцип продолжения решения уравнения (8) по непрерывности (см. [1], [12], [13]) из области $\mathbb{R}^3\backslash\overline{B}$ в область $\mathbb{R}^3\backslash\overline{Q^j}$. Учитывая, что в области $\mathbb{R}^3\backslash\overline{B}$ поле $\mathbf{E} \equiv 0$, получаем, что в области $\mathbb{R}^3\backslash\overline{Q^j}$ поле $\mathbf{E} \equiv 0$ и (как следует из второго уравнения (5)) $\mathbf{H} \equiv 0$. Значит, электромагнитное поле в подобласти Q_j равно нулю, и из условий сопряжения (3) следует, что тангенциальные компоненты полей $\mathbf{E}$ и $\mathbf{H}$ на ∂Q равны нулю.

Далее, электромагнитное поле в задачах A и C удовлетворяет одним и тем же уравнениям (5) в области Q и одинаковым граничным условиям на ∂Q (тангенциальные компоненты электрического и магнитного полей на ∂Q равны нулю). Поэтому множества решений задач A и C совпадают в области Q. Но решение задачи C в области Q_j только тривиальное, что противоречит нетривиальности решения задачи A в этой же области.

Таким образом, из изложенного выше следует

Теорема 1. *Пусть область Q состоит из конечного числа подобластей Q_i с кусочно-гладкой границей ∂Q_i; $\overline{Q} = \bigcup_i \overline{Q}_i$, $Q_i \cap Q_j = \varnothing$ при $i \neq j$. Предположим, что $\mu = \mu_0$ в $\mathbb{R}^3$, $\varepsilon = \varepsilon_0$ в $\mathbb{R}^3\backslash\overline{Q}$, а $\varepsilon(x)$ в Q задается сужениями на Q_i функций $\varepsilon_i \in C^3(\overline{B})$ ($\varepsilon(x) = \varepsilon_i(x)$ при $x \in Q_i$), где B – (открытый) шар, содержащий Q, $\overline{Q} \subset B$. Пусть $\operatorname{Re}\varepsilon_i(x) > 0$, $\operatorname{Im}\varepsilon_i(x) \geq 0$ в B. Тогда задача (2)–(5) имеет только тривиальное решение.*

Аналогично можно получить теорему единственности для изотропных магнитодиэлектрических сред.

3. ЕДИНСТВЕННОСТЬ РЕШЕНИЯ ЗАДАЧИ ДЛЯ АНИЗОТРОПНОЙ СРЕДЫ

Теперь рассмотрим анизотропный случай. Наложим следующие ограничения на тензор-функции $\hat{\varepsilon}_i(x)$ и $\hat{\mu}_i(x)$ в области B, которые определяют значения $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ в области Q. Будем полагать, что эрмитовы тензор-функции $(\hat{\varepsilon}_i(x) + \hat{\varepsilon}_i^*(x))/2$ и $(\hat{\mu}_i(x) + \hat{\mu}_i^*(x))/2$ положительно определены (аналог условия $\operatorname{Re}\varepsilon_i(x) > 0$, $\operatorname{Re}\mu_i(x) > 0$ для изотропной среды); тензор-функции $(\hat{\varepsilon}_i(x) - \hat{\varepsilon}_i^*(x))/(2i)$ и $(\hat{\mu}_i(x) - \hat{\mu}_i^*(x))/(2i)$ неотрицательно определены (аналог условия $\operatorname{Im}\varepsilon_i(x)$, $\operatorname{Im}\mu_i(x) \geq 0$ для изотропной среды); $\hat{\varepsilon}_i, \hat{\mu}_i \in C^3(\overline{B})$. Символ * обозначает сопряженный тензор, т.е. транспонированный тензор с комплексно-сопряженными элементами.

Будем рассматривать однородные уравнения

$$\operatorname{rot}\mathbf{H} = -i\omega\hat{\varepsilon}\mathbf{E}, \quad \operatorname{rot}\mathbf{E} = i\omega\hat{\mu}\mathbf{H} \tag{9}$$

и соответственно краевую задачу (2)–(4), (9). Доказательство единственности решения задачи проводится по той же схеме, что и в разд. 2, поэтому остановимся только на отличиях анизотропного случая от изотропного.

Снова рассматриваются задачи A и C с заменой ε и μ на $\hat{\varepsilon}$ и $\hat{\mu}$. Построение тензоров $\hat{\varepsilon}^c$ и $\hat{\mu}^c$ для задачи C такое же, как и в разд. 2, т.е. с использованием функции-срезки:

$$\begin{aligned}\hat{\varepsilon}^c(x) &:= (\varepsilon_j(x) - \varepsilon_0)\zeta(|x - x_0|; d, 2d)\hat{I} + \varepsilon_0\hat{I},\\ \hat{\mu}^c(x) &:= (\mu_j(x) - \mu_0)\zeta(|x - x_0|; d, 2d)\hat{I} + \mu_0\hat{I}.\end{aligned} \tag{10}$$

Из [2], [11] следует, что уравнения (9) в области гладкости $\hat{\varepsilon}^c$ и $\hat{\mu}^c$ сводятся в декартовой системе координат к следующей системе дифференциальных уравнений:

$$\begin{aligned}
&\varepsilon_{lm}\frac{\partial^2 E_k}{\partial x_l \partial x_m}+\frac{\partial}{\partial x_k}\left(\frac{\partial \varepsilon_{lm}}{\partial x_l}E_m\right)+\frac{\partial \varepsilon_{lm}}{\partial x_k}\frac{\partial E_m}{\partial x_l}+i\omega\varepsilon_{lm}L_{kmn}\frac{\partial}{\partial x_l}(\mu_{np}H_p)=0,\quad k=1,2,3,\\
&\mu_{lm}\frac{\partial^2 H_k}{\partial x_l \partial x_m}+\frac{\partial}{\partial x_k}\left(\frac{\partial \mu_{lm}}{\partial x_l}H_m\right)+\frac{\partial \mu_{lm}}{\partial x_k}\frac{\partial H_m}{\partial x_l}-i\omega\mu_{lm}L_{kmn}\frac{\partial}{\partial x_l}(\varepsilon_{np}E_p)=0,\quad k=1,2,3,
\end{aligned}\tag{11}$$

где ε_{lm} и μ_{lm} – компоненты тензоров $\hat{\varepsilon}^c$ и $\hat{\mu}^c$. В (11) используется правило суммирования по повторяющимся индексам, а L_{kmn} – символ Леви–Чивита, который определяется формулой

$$L_{kmn}=\begin{cases}1, & \text{если} \quad kmn=123,\ 231,\ 312,\\ -1, & \text{если} \quad kmn=321,\ 213,\ 132,\\ 0 & \text{в остальных случаях.}\end{cases}$$

Из уравнений (11) следует, что детерминант главного символа дифференциального оператора определяется формулой

$$\det P^0(\xi)=\left[\sum_{l,m=1}^{3}\varepsilon_{lm}\xi_l\xi_m\right]^3\left[\sum_{l,m=1}^{3}\mu_{lm}\xi_l\xi_m\right]^3.$$

Поскольку $(\hat{\varepsilon}_i(x)+\hat{\varepsilon}_i^*(x))$ и $(\hat{\mu}_i(x)+\hat{\mu}_i^*(x))$ положительно определены, то получим, что $\det P^0(\xi)\neq 0$ при $|\xi|\neq 0$ в $\mathbb{R}^3\backslash\overline{Q^j}$. Значит, система уравнений (11) является эллиптической в этой области. Далее, все коэффициенты в (11) являются дифференцируемыми функциями в $\mathbb{R}^3\backslash\overline{Q^j}$. Поэтому можно применить принцип продолжения решения уравнений (11) по непрерывности (см. [12], [13]). Дальнейшие рассуждения повторяют аналогичные из разд. 2. Из изложенного выше следует

Теорема 2. *Пусть область Q состоит из конечного числа подобластей Q_i с кусочно-гладкой границей ∂Q_i; $\overline{Q}=\bigcup_i \overline{Q}_i$, $Q_i\cap Q_j=\varnothing$ при $i\neq j$. Предположим, что $\hat{\varepsilon}=\varepsilon_0\hat{I}$ $\hat{\mu}=\mu_0\hat{I}$ в $\mathbb{R}^3\backslash\overline{Q}$, а $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ в Q задаются сужениями на Q_i функций $\hat{\varepsilon}_i\in C^3(\overline{B})$ и $\hat{\mu}_i\in C^3(\overline{B})$ ($\hat{\varepsilon}(x)=\hat{\varepsilon}_i(x)$ и $\hat{\mu}(x)=\hat{\mu}_i(x)$ при $x\in Q_i$), где B – (открытый) шар, содержащий Q, $\overline{Q}\subset B$. Пусть эрмитовы тензор-функции $(\hat{\varepsilon}_i(x)+\hat{\varepsilon}_i^*(x))/2$ и $(\hat{\mu}_i(x)+\hat{\mu}_i^*(x))/2$ положительно определены, а тензор-функции $(\hat{\varepsilon}_i(x)-\hat{\varepsilon}_i^*(x))/(2i)$ и $(\hat{\mu}_i(x)-\hat{\mu}_i^*(x))/(2i)$ неотрицательно определены в B. Тогда задача* (2)–(4), (9) *имеет только тривиальное решение.*

4. ТЕОРЕМЫ СУЩЕСТВОВАНИЯ И ЕДИНСТВЕННОСТИ РЕШЕНИЯ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

Будем рассматривать задачи в анизотропной среде, которые удовлетворяют условиям теоремы 2. Рассматриваемые задачи могут быть сведены к системе объемных сингулярных интегральных уравнений относительно электромагнитного поля в области Q (см. [3]–[5]):

$$\begin{aligned}
&\mathbf{E}(x)+\frac{1}{3}(\hat{\varepsilon}_r(x)-\hat{I})\mathbf{E}(x)-p.v.\int_Q((\hat{\varepsilon}_r(y)-\hat{I})\mathbf{E}(y),\operatorname{grad})\operatorname{grad}G(R)dy-k_0^2\int_Q(\hat{\varepsilon}_r(y)-\hat{I})\mathbf{E}(y)G(R)dy-\\
&\qquad-i\omega\mu_0\int_Q(\hat{\mu}_r(y)-\hat{I})\mathbf{H}(y)\times\operatorname{grad}G(R)dy=\mathbf{E}^0(x),\quad x\in Q,
\end{aligned}\tag{12}$$

$$\begin{aligned}
&\mathbf{H}(x)+\frac{1}{3}(\hat{\mu}_r(x)-\hat{I})\mathbf{H}(x)-p.v.\int_Q((\hat{\mu}_r(y)-\hat{I})\mathbf{H}(y),\operatorname{grad})\operatorname{grad}G(R)dy-k_0^2\int_Q(\hat{\mu}_r(y)-\hat{I})\mathbf{H}(y)G(R)dy+\\
&\qquad+i\omega\varepsilon_0\int_Q(\hat{\varepsilon}_r(y)-\hat{I})\mathbf{E}(y)\times\operatorname{grad}G(R)dy=\mathbf{H}^0(x),\quad x\in Q.
\end{aligned}\tag{13}$$

В (12), (13) $\hat{\varepsilon}_r = \hat{\varepsilon}/\varepsilon_0$, $\hat{\mu}_r = \hat{\mu}/\mu_0$; G – функция Грина (фундаментальное решение) для уравнения Гельмгольца

$$G(R) = \frac{\exp(ik_0R)}{4\pi R}, \tag{14}$$

где $R = |x - y|$; $x = (x_1, x_2, x_3)$; $y = (y_1, y_2, y_3)$, $\times$ – векторное произведение. $\mathbf{E}^0$ и $\mathbf{H}^0$ – внешнее электромагнитное поле, создаваемое сторонним током $\mathbf{J}^0$ и удовлетворяющее уравнениям Максвелла для свободного пространства

$$\operatorname{rot}\mathbf{H}^0 = -i\omega\varepsilon_0\mathbf{E}^0 + \mathbf{J}^0, \quad \operatorname{rot}\mathbf{E}^0 = i\omega\mu_0\mathbf{H} \tag{15}$$

и условию излучения на бесконечности. Если источником поля является плоская волна, то в (15) $\mathbf{J}^0 = 0$.

Для ответа на вопрос о существовании решения рассматриваемых задач необходимо выбрать подходящее для анализа функциональное пространство. Интегралы от квадрата модуля комплексных амплитуд электромагнитного поля присутствуют в законе сохранения энергии. Значит, можно полагать, что пространство интегрируемых с квадратом вектор-функций является наиболее “физичным” для исследования интегральных уравнений задач электромагнитного рассеяния. Ниже будем использовать гильбертово пространство шестимерных вектор-функций $\mathbf{L}_2(Q)$ со скалярным произведением, определяемым формулой

$$(\mathbf{U}, \mathbf{V}) = \int_Q \vec{U}(x)\mathbf{V}^*(x)dx.$$

Отметим, что оператор уравнений (12), (13) определен в $\mathbf{L}_2(Q)$ (см. [4], [5]).

Систему уравнений (12), (13) можно записать в эквивалентной форме, в виде интегродифференциальных уравнений (см. [4]):

$$\begin{aligned}\mathbf{E}(x) - \operatorname{grad}\operatorname{div}\int_Q G(R)(\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)dy - k_0^2\int_Q G(R)(\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)dy - \\ - i\omega\mu_0\operatorname{rot}\int_Q G(R)(\hat{\mu}_r(y) - \hat{I})\mathbf{H}(y)dy = \mathbf{E}^0(x), \quad x \in Q,\end{aligned} \tag{16}$$

$$\begin{aligned}\mathbf{H}(x) - \operatorname{grad}\operatorname{div}\int_Q G(R)(\hat{\mu}_r(y) - \hat{I})\mathbf{H}(y)dy - k_0^2\int_Q G(R)(\hat{\mu}_r(y) - \hat{I})\mathbf{H}(y)dy + \\ + i\omega\varepsilon_0\operatorname{rot}\int_Q G(R)(\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)dy = \vec{H}^0(x), \quad x \in Q.\end{aligned} \tag{17}$$

Выражения (16), (17) справедливы и при $x \in \mathbb{R}^3\backslash\overline{Q}$. В этом случае они являются интегральными представлениями и определяют электромагнитное поле вне области Q по найденному значению полей в Q. Отметим, что, поскольку $\hat{\varepsilon}_r(x) = \hat{I}$, $\hat{\mu}_r(x) = \hat{I}$, $x \in \mathbb{R}^3\backslash\overline{Q}$, то в этой области интегральные представления полей (16), (17) не будут иметь сингулярности.

Для существования и единственности решения уравнений (16), (17) в $\mathbf{L}_2(Q)$ достаточно, чтобы оператор уравнений был фредгольмов и однородные уравнения имели только тривиальное решение в этом пространстве функций.

Рассмотрим сначала вопрос о единственности решения уравнений (16), (17). Непосредственное использование теоремы 2 затруднительно, поскольку пространство функций из $\mathbf{L}_2$ шире класса функций, описывающих классические решения уравнений Максвелла. Для применения теоремы 2 необходимо получать результаты о гладкости решений интегральных уравнений (см. [14]). Мы же для доказательства единственности решения системы (16), (17) будем использовать идеи, описанные в разд. 2 и 3.

Рассмотрим однородные уравнения (16), (17):

$$\begin{aligned}\mathbf{E}(x) - \operatorname{grad}\operatorname{div}\int_Q G(R)(\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)dy - k_0^2\int_Q G(R)(\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)dy - \\ - i\omega\mu_0\operatorname{rot}\int_Q G(R)(\hat{\mu}_r(y) - \hat{I})\mathbf{H}(y)dy = 0, \quad x \in Q,\end{aligned} \tag{18}$$

$$\mathbf{H}(x) - \operatorname{grad}\operatorname{div}\int_Q G(R)(\hat{\mu}_r(y) - \hat{I})\mathbf{H}(y)dy - k_0^2\int_Q G(R)(\hat{\mu}_r(y) - \hat{I})\mathbf{H}(y)dy + \\ + i\omega\varepsilon_0 \operatorname{rot}\int_Q G(R)(\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)dy = 0, \quad x \in Q. \tag{19}$$

Предположим, что в области Q решение $\mathbf{E}$, $\mathbf{H}$ уравнений (18), (19) не равно тождественно нулю. Обозначим эту систему уравнений через A1. Не ограничивая общности, можно считать, что во всех подобластях Q_i (для всех i) решение $\mathbf{E}$, $\mathbf{H}$ не равно тождественно нулю. Действительно, в противном случае области Q_i, в которых решения тождественно равны нулю, исключим из Q и рассмотрим интегральные уравнения в оставшейся части Q. Так как во всех Q_i поле $\mathbf{E}$, $\mathbf{H}$ не равно тождественно нулю, то найдется такое j, что $\partial Q_j \cap \partial Q \neq \varnothing$, т.е. у Q_j и Q имеется общий гладкий кусок границы. Электромагнитное поле вне Q определяется только значениями полей в Q с помощью интегральных представлений полей в $\mathbb{R}^3\backslash\overline{Q}$, удовлетворяющих условию излучения.

Рассмотрим решение однородных уравнений (18), (19) в $\mathbb{R}^3$. Из интегральной формулировки теоремы Пойнтинга следует соотношение для электромагнитного поля рассматриваемой задачи

$$\omega\operatorname{Im}\int_Q (\mathbf{E}^*, \hat{\varepsilon}\mathbf{E})dv + \omega\operatorname{Im}\int_Q (\mathbf{H}^*, \hat{\mu}\mathbf{H})dv + \sqrt{\frac{\varepsilon_0}{\mu_0}}\lim_{r\to\infty}\int_{S_r} |\mathbf{E}_S|^2 ds = 0,$$

где S_r – сфера радиуса r с центром в нуле, а $\mathbf{E}_S$ – касательные составляющие электрического поля на поверхности сферы. Поскольку параметры среды в Q удовлетворяют условиям теоремы 2, то первые два интеграла в этом соотношении неотрицательны, поэтому третий поверхностный интеграл тоже равен нулю. Далее, интегральные представления (18), (19) удовлетворяют однородным уравнениям Максвелла для свободного пространства в $\mathbb{R}^3\backslash\overline{Q}$ и условиям излучения. Поэтому из леммы Реллиха следует, что $\mathbf{E}$, $\mathbf{H}$ равны нулю в $\mathbb{R}^3\backslash\overline{Q}$.

Теперь рассмотрим вспомогательную систему интегральных уравнений C1 в функциональном пространстве $\mathbf{L}_2(D)$, где D – шар радиуса $3d$ с центром в точке $x_0 \in Q$ (d – диаметр области Q, $x_0 \in Q$ – центр шара B). Параметры среды в системе уравнений C1 определим следующим образом: диэлектрическая и магнитная проницаемости $\hat{\varepsilon}^c(x)$, $\hat{\mu}^c(x)$ в области Q совпадают с проницаемостями из системы A1, а вне области Q определяются формулами (10).

Система C1 описывается следующими уравнениями:

$$\mathbf{E}_c(x) - \operatorname{grad}\operatorname{div}\int_D G(R)(\hat{\varepsilon}_r^c(y) - \hat{I})\mathbf{E}_c(y)dy - k_0^2\int_D G(R)(\hat{\varepsilon}_r^c(y) - \hat{I})\mathbf{E}_c(y)dy - \\ - i\omega\mu_0 \operatorname{rot}\int_D G(R)(\hat{\mu}_r^c(y) - \hat{I})\mathbf{H}_c(y)dy = 0, \quad x \in D, \tag{20}$$

$$\mathbf{H}_c(x) - \operatorname{grad}\operatorname{div}\int_D G(R)(\hat{\mu}_r^c(y) - \hat{I})\mathbf{H}_c(y)dy - k_0^2\int_D G(R)(\hat{\mu}_r^c(y) - \hat{I})\mathbf{H}_c(y)dy + \\ + i\omega\varepsilon_0 \operatorname{rot}\int_D G(R)(\hat{\varepsilon}_r^c(y) - \hat{I})\mathbf{E}_c(y)dy = 0, \quad x \in D. \tag{21}$$

В (20), (21) $\hat{\varepsilon}_r^c = \hat{\varepsilon}^c/\varepsilon_0$, $\hat{\mu}_r^c = \hat{\mu}^c/\mu_0$ – относительные диэлектрическая и магнитная проницаемости соответственно.

Нетривиальное решение $\mathbf{E}$, $\mathbf{H}$ однородной системы A1 (18), (19) продолжим нулем из области Q на область $D\backslash\overline{Q}$ и обозначим получившиеся функции через $\mathbf{E}_1$, $\mathbf{H}_1$. Получим $\mathbf{E}_1, \mathbf{H}_1 \in \mathbf{L}_2(D)$. Тогда $\mathbf{E}_1$, $\mathbf{H}_1$ будет решением (20), (21). Действительно, поскольку $\mathbf{E}_1$, $\mathbf{H}_1$ равны нулю в $D\backslash\overline{Q}$, то при $x \in Q$ уравнения (20), (21) совпадают с (18), (19). Вне Q интегральные представления (18), (19) дают $\mathbf{E} \equiv 0$, $\mathbf{H} \equiv 0$, т.е. при $x \in D\backslash\overline{Q}$ снова совпадают с (20), (21). Получаем, что система (20), (21), рассматриваемая в $\mathbf{L}_2(D)$, имеет нетривиальное (в Q) решение.

Теперь опишем множество решений системы C1. Поскольку $\hat{\varepsilon}^c(x) = \varepsilon_0\hat{I}$ и $\hat{\mu}^c(x) = \mu_0\hat{I}$ при $x \notin B$, то, аналогично изложенному выше, получим, что в $D\backslash\overline{B}$ электромагнитное поле равно ну-

лю. Далее, из определения $\hat{\varepsilon}^c$, $\hat{\mu}^c$ следует, что $\hat{\varepsilon}^c, \hat{\mu}^c \in C^3(D\backslash\overline{Q^j})$, где $Q^j = Q\backslash\overline{Q_j}$. В областях гладкости параметров среды решение (20), (21) из $\vec{L}_2(D)$ удовлетворяет уравнениям Максвелла (9) (см. [4]). Значит, в области $D\backslash\overline{Q^j}$ решение (20), (21) удовлетворяет уравнениям (11), которые, поскольку $\hat{\varepsilon}_j, \hat{\mu}_j$ подчиняются условиям теоремы 2, являются эллиптическими (см. разд. 3). Все коэффициенты в уравнениях (11) являются дифференцируемыми функциями. Теперь, применяя принцип продолжения (см. [12], [13]) решения (11) по непрерывности из области $D\backslash\overline{B}$ в область $D\backslash\overline{Q^j}$, получим, что электромагнитное поле в подобласти Q_j для системы A1 равно нулю, что противоречит предположению о нетривиальности решения A1 в Q_j. Таким образом, имеет место

Теорема 3. *При выполнении условий теоремы* 2 *система однородных сингулярных интегральных уравнений* (12), (13) (*или* (16), (17)) *имеет единственное решение в* $\mathbf{L}_2(Q)$.

Теперь рассмотрим вопрос о фредгольмовости интегральных уравнений. Приведем несколько определений, которые используются в дальнейшем изложении.

Определение 1. Пусть A – линейный ограниченный оператор, действующий в гильбертовом пространстве H. Тогда оператор A^*, который также определен в H, называется *сопряженным к* A, если равенство $(Af, g) = (f, A^*g)$ выполняется для всех $f, g \in H$.

Решения однородного уравнения $Au = 0$ будем называть нулями оператора A. Обозначим размерность подпространства нулей через $n(A)$. Тогда $n(A^*)$ – размерность подпространства нулей сопряженного оператора A^*. Разность $\operatorname{Ind} A = n(A) - n(A^*)$ называется индексом оператора A.

Определение 2. Линейный ограниченный оператор A, действующий в гильбертовом пространстве H, называется *фредгольмовым оператором*, если значения $n(A)$ и $n(A^*)$ конечны и индекс равен нулю.

Сначала рассмотрим интегральное уравнение в анизотропной диэлектрической среде, т.е. магнитная проницаемость всюду в $\mathbb{R}^3$ постоянна и равна μ_0. Тогда система интегральных уравнений (12), (13) сводится к объемному сингулярному интегральному уравнению относительно электрического поля в области Q

$$\begin{aligned}\mathbf{E}(x) + \frac{1}{3}(\hat{\varepsilon}_r(x) - \hat{I})\mathbf{E}(x) - p.v.\int_Q ((\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y), \operatorname{grad})\operatorname{grad} G(R)dy - \\ - k_0^2\int_Q (\hat{\varepsilon}_r(y) - \hat{I})\mathbf{E}(y)G(R)dy = \mathbf{E}^0(x), \quad x \in Q, \quad \hat{\varepsilon}_r = \hat{\varepsilon}/\varepsilon_0.\end{aligned} \tag{22}$$

Обозначим через $\hat{B}_0$ оператор уравнения (22). Тогда

$$(\hat{B}_0\mathbf{W})(x) = \left(\hat{I} + \frac{1}{3}\hat{\eta}(x)\right)\mathbf{W}(x) - \int_Q \hat{G}_0(x,y)\left(\hat{\eta}(y)\mathbf{W}(y)\right)dy - p.v.\int_Q \hat{G}_1(x,y)\left(\hat{\eta}(y)\mathbf{W}(y)\right)dy, \quad x \in Q. \tag{23}$$

В (23) тензор-функция $\hat{\eta}(x) = (\hat{\varepsilon}_r(x) - \hat{I})$, а $\hat{G}_0(x,y)$ и $\hat{G}_1(x,y)$ – матричные функции, очевидным образом определяемые из (22).

Рассмотрим оператор в пространстве $\mathbf{L}_2$ вида

$$(\hat{A}\mathbf{W})(x) = \int_Q \hat{G}(x,y)\mathbf{W}(y)dy, \tag{24}$$

где $\hat{G}(x,y)$ – тензор-функция. Сопряженный оператор определяется формулой

$$(\hat{A}^*\mathbf{V})(x) = \int_Q \hat{G}^*(y,x)\mathbf{V}(y)dy, \tag{25}$$

где $\hat{G}^*$ – сопряженный к $\hat{G}$ тензор.

Тогда оператор, сопряженный к $\hat{B}_0$ в пространстве $\mathbf{L}_2(Q)$, будет иметь следующий вид:

$$(\hat{B}_0^*\mathbf{W})(x) = \left(\hat{I} + \frac{1}{3}\hat{\eta}^*(x)\right)\mathbf{W}(x) - \hat{\eta}^*(x)\int_Q \hat{G}_0^*(y,x)\mathbf{W}(y)dy - \hat{\eta}^*(x)p.v.\int_Q \hat{G}_1^*(y,x)\mathbf{W}(y)dy, \quad x \in Q. \tag{26}$$

Ниже будем полагать, что тензор-функция $\hat{\eta}(x) = \left(\hat{\varepsilon}_r(x) - \hat{I}\right)$ имеет обратную в каждой точке из $\overline{Q}$. Из (22), (23), (25) следует, что $\hat{G}_0(x,y) = \hat{G}_0(y,x)$, $\hat{G}_0 = \hat{G}_0^t$ и $\hat{G}_1 = \hat{G}_1^t$, $\hat{G}_1(x,y) = \hat{G}_1(y,x)$. Учитывая эти свойства тензоров, возьмем комплексное сопряжение от выражения (26):

$$(\hat{B}_0^*\mathbf{W})^*(x) = \left(\hat{I} + \frac{1}{3}\hat{\eta}^t(x)\right)\mathbf{W}^*(x) - \hat{\eta}^t(x)\int\limits_Q \hat{G}_0(x,y)\mathbf{W}^*(y)dy - \hat{\eta}^t(x)p.v.\int\limits_Q \hat{G}_1(x,y)\mathbf{W}^*(y)dy, \quad x \in Q. \quad (27)$$

В (27) символы t и * обозначают соответственно транспонированный тензор и комплексно-сопряженный вектор.

Пусть $\mathbf{W}$ – нуль оператора (26), т.е. $\hat{B}_0^*\mathbf{W} = 0$. Пусть $\mathbf{V} = (\hat{\eta}^t)^{-1}\mathbf{W}^*$. Тогда из (23), (26), (27) имеем

$$(\hat{B}_0^*\mathbf{W})^* = \hat{\eta}^t\hat{B}_0(\hat{\varepsilon}^t)\mathbf{V} = 0,$$

где $\hat{B}_0(\hat{\varepsilon}^t)$ – оператор уравнения (22) с тензором диэлектрической проницаемости в Q, равным $\hat{\varepsilon}^t$. Значит, размерности подпространств нулей операторов $\hat{B}_0^*(\hat{\varepsilon})$ и $\hat{B}_0(\hat{\varepsilon}^t)$ связаны неравенством $n(\hat{B}_0^*(\hat{\varepsilon})) \le n(\hat{B}_0(\hat{\varepsilon}^t))$. Теперь пусть $\mathbf{W}$ – нуль оператора (23) с диэлектрической проницаемостью $\hat{\varepsilon}^t$, т.е. $\hat{B}_0(\hat{\varepsilon}^t)\mathbf{W} = 0$. Обозначим $\mathbf{V}^* = \hat{\eta}^t\mathbf{W}$. Тогда из (23), (26), (27) имеем

$$\hat{\eta}^t\hat{B}_0(\hat{\varepsilon}^t)\mathbf{W} = (\hat{B}_0^*\mathbf{V})^* = 0,$$

откуда следует, что $n(\hat{B}_0(\hat{\varepsilon}^t)) \le n(\hat{B}_0^*(\hat{\varepsilon}))$. Значит,

$$n(\hat{B}_0(\hat{\varepsilon}^t)) = n(\hat{B}_0^*(\hat{\varepsilon})), \quad (28)$$

т.е. размерности подпространств нулей операторов $\hat{B}_0^*(\hat{\varepsilon})$ и $\hat{B}_0(\hat{\varepsilon}^t)$ равны. Если $\hat{\varepsilon} = \hat{\varepsilon}^t$ выполняется, например, в изотропных средах, то $n(\hat{B}_0) = n(\hat{B}_0^*)$ и, значит, $\mathrm{Ind}(\hat{B}_0) = 0$.

Если какой-либо эрмитов тензор $\hat{\delta}$ неотрицательно/положительно определен, то и эрмитов тензор $\hat{\delta}^t$ будет также неотрицательно/положительно определен. Поэтому при выполнении условий теоремы 2 получим

$$n(\hat{B}_0(\hat{\varepsilon})) = n(\hat{B}_0(\hat{\varepsilon}^t)) = 0. \quad (29)$$

Далее, из (28), (29) следует, что

$$n(\hat{B}_0(\hat{\varepsilon})) = n(\hat{B}_0^*(\hat{\varepsilon})) = 0. \quad (30)$$

Значит, при выполнении приведенных выше условий оператор $\hat{B}_0$ будет фредгольмовым в пространстве $\mathbf{L}_2(Q)$. Таким образом, имеем следующее утверждение.

Теорема 4. *Пусть выполняются условия теоремы* 2 *для тензор-функции* $\hat{\varepsilon}(x)$, *а* $\hat{\mu} = \mu_0$ *в* $\mathbb{R}^3$, *и, кроме того, тензор-функция* $(\hat{\varepsilon}_r(x) - \hat{I})$ *имеет обратную в каждой точке из* $\overline{Q}$. *Тогда существует и единственно решение сингулярного интегрального уравнения* (22) *в пространстве* $\mathbf{L}_2(Q)$.

Теперь рассмотрим задачи рассеяния на магнитодиэлектрическом теле, диэлектрическая и магнитная проницаемости которого являются кусочно-дифференцируемыми функциями координат в Q, а поверхности разрыва параметров удовлетворяют приведенным выше условиям.

Запишем систему сингулярных интегральных уравнений (12), (13) в символическом виде

$$\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} + \begin{pmatrix}\hat{S} & -i\omega\mu_0\hat{F}\\ i\omega\varepsilon_0\hat{F} & \hat{S}\end{pmatrix}\begin{pmatrix}(\hat{\varepsilon}_r - \hat{I})\mathbf{E}\\ (\hat{\mu}_r - \hat{I})\mathbf{H}\end{pmatrix} = \begin{pmatrix}\mathbf{E}^0\\ \mathbf{H}^0\end{pmatrix}, \quad (31)$$

где вид операторов $\hat{S}$ и $\hat{F}$ ясен из (12), (13). Очевидно, что оператор $\hat{S}$ является сингулярным оператором в $\mathbf{L}_2(Q)$, а оператор $\hat{F}$ – компактным. Здесь $\mathbf{L}_2(Q)$ – гильбертово пространство интегрируемых с квадратом шестимерных вектор-функций.

Рассмотрим следующее уравнение в $\mathbf{L}_2(Q)$:

$$\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix}+\begin{pmatrix}\hat{S} & 0\\ 0 & \hat{S}\end{pmatrix}\begin{pmatrix}(\hat{\varepsilon}_r-\hat{I})\mathbf{E}\\ (\hat{\mu}_r-\hat{I})\mathbf{H}\end{pmatrix}=\begin{pmatrix}\mathbf{E}^0\\ \mathbf{H}^0\end{pmatrix}. \tag{32}$$

Из вида (32) следует, что при выполнении условий теоремы 4 для тензор-функций $\hat{\varepsilon}(x)$ и $\hat{\mu}(x)$ оператор уравнения (32) будет фредгольмовым в $\mathbf{L}_2(Q)$. Оператор уравнения (31) отличается от оператора (32) прибавлением компактных операторов. Значит, оператор уравнения (31) является фредгольмовым оператором в $\mathbf{L}_2(Q)$. Получаем следующее утверждение.

Теорема 5. *Пусть выполняются условия теоремы* 2, *и, кроме того, тензор-функции* $(\hat{\varepsilon}_r(x)-\hat{I})$ *и* $(\hat{\mu}_r(x)-\hat{I})$ *имеют обратные в каждой точке из* $\overline{Q}$. *Тогда существует и единственно решение системы сингулярных интегральных уравнений* (12), (13) *в пространстве* $\mathbf{L}_2(Q)$.

ЗАКЛЮЧЕНИЕ

Вопросы существования и единственности решения задач рассеяния электромагнитных волн на прозрачных телах (диэлектрических или магнитодиэлектрических) имеют не только теоретический, но и практический интерес. Единственность решения уравнений Максвелла, удовлетворяющих соответствующим условиям, означает, что в области неоднородности среды Q не может быть ненулевое электромагнитное поле, которое не излучает энергию в окружающее пространство.

В настоящей статье доказаны теоремы единственности решения задач рассеяния электромагнитных волн на трехмерных ограниченных неоднородных анизотропных телах в общей дифференциальной постановке, в том числе для тел без потерь и имеющих разрывы диэлектрической и магнитной проницаемости. Также доказаны теоремы о существовании и единственности решения объемных сингулярных интегральных уравнений, отвечающих задачам рассеяния электромагнитных волн на ограниченных трехмерных неоднородных анизотропных магнитодиэлектрических телах, т.е. для задач рассеяния в интегральной постановке, в том числе для сред без потерь и с разрывами параметров.

СПИСОК ЛИТЕРАТУРЫ

1. *Colton D., Kress R.* Inverse acoustic and electromagnetic scattering theory. Applied Mathematical Sciences. V. 93. Berlin: Springer-Verlag, 1992.
2. *Potthast R.* Integral equation methods in electromagnetic scattering from anisotropic media // J. Integral Equat. Appl. 1999. V. 11. № 2. P. 197–215.
3. *Самохин А.Б.* Объемные сингулярные интегральные уравнения для задач рассеяния на трехмерных диэлектрических структурах // Дифференц. ур-ния. 2014. Т. 50. № 9. С. 1215–1230.
4. *Самохин А.Б.* Интегральные уравнения и итерационные методы в электромагнитном рассеянии. М.: Радио и связь, 1998.
5. *Samokhin A.B.* Integral equations and iteration methods in electromagnetic scattering. Utrecht: VSP, 2001.
6. *Ball J.M., Capdeboscq Y., Tsering Xiao B.* On uniqueness for time harmonic anisotropic Maxwell's equations with piecewise regular coefficients // Math. Models Meth. Appl. Sci. 2012. V. 22. № 11. P. 1–34.
7. *Смирнов Ю.Г., Цупак А.А.* Математическая теория дифракции акустических и электромагнитных волн на системе экранов и неоднородных тел. М.: Русайнс, 2016.
8. *Смирнов Ю.Г., Цупак А.А.* О существовании и единственности классического решения задачи дифракции электромагнитной волны на неоднородном диэлектрическом теле без потерь // Ж. вычисл. матем. и матем. физ. 2017. Т. 57. № 4. С. 702–709.
9. *Smirnov Yu.G., Tsupak A.A.* Integro-differential equations of the vector problem of electromagnetic wave diffraction by a system of nonintersecting screens and inhomogeneous bodies // Adv. Math. Phys. 2015. Article ID 945965.
10. *Бирман М.Ш., Соломяк М.З.* L_2-теория оператора Максвелла в произвольных областях // Успехи матем. наук. 1987. Т. 42. Вып. 6. С. 61–75.
11. *Самохин А.Б.* Исследование интегральных уравнений для задач электромагнитного рассеяния на трехмерных прозрачных структурах // Дифференц. ур-ния. 2001. Т. 37. № 10. С. 1357–1363.
12. *Хёрмандер Л.* Анализ линейных дифференциальных операторов с частными производными. Т. 3. Псевдодифференциальные операторы. М.: Мир, 1987.
13. *Protter M.H.* Unique continuation for elliptic equations // Trans. Am. Math. Soc. 1960. V. 95. P. 81–90.
14. *Смирнов Ю.Г.* Об эквивалентности электромагнитной задачи дифракции не неоднородном ограниченном диэлектрическом теле объемному сингулярному интегро-дифференциальному уравнению // Ж. вычисл. матем. и матем. физ. 2016. Т. 56. № 9. С. 1657–1666.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 95–107

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.63

МОНОТОННЫЕ СХЕМЫ ДЛЯ ЗАДАЧ КОНВЕКЦИИ-ДИФФУЗИИ С КОНВЕКТИВНЫМ ПЕРЕНОСОМ В РАЗЛИЧНОЙ ФОРМЕ[1)]

© 2021 г. П. Н. Вабищевич[1,2]

[1] *115191 Москва, Б. Тульская ул., 52, ИБРАЭ РАН, Россия*

[2] *677000 Якутск, ул. Белинского, 58, СВФУ им. М.К. Аммосова, Россия*

e-mail: vabishchevich@gmail.com

Поступила в редакцию 10.03.2020 г.
Переработанный вариант 18.06.2020 г.
Принята к публикации 18.09.2020 г.

В задачах конвекции-диффузии конвективный перенос записывается в различных формах. Обычно ориентируются на использование конвективных слагаемых в недивергентной и дивергентной формах. Для таких задач строятся монотонные и устойчивые схемы в банаховых пространствах: в равномерной и интегральной нормах соответственно. Монотонность связывается с диагональным преобладанием по строкам или столбцам. При записи конвективных слагаемых в симметричной форме (полусумма недивергентной и дивергентной форм) устойчивость устанавливается в гильбертовых пространствах сеточных функций. Сформулированы условия диагонального преобладания, которые обеспечивают монотонность двухслойных схем для нестационарных уравнений конвекции-диффузии и устойчивость в соответствующих пространствах. Библ. 27.

Ключевые слова: задачи конвекции-диффузии, двухслойные разностные схемы, логарифмическая норма, монотонные схемы.

DOI: 10.31857/S0044466920120157

ВВЕДЕНИЕ

В задачах механики сплошных сред базовыми являются краевые задачи для нестационарных уравнений конвекции-диффузии. При численном решении основное внимание уделяется аппроксимациям по пространству конвективных слагаемых, для сохранения ключевых свойств решений дифференциальной задачи. В частности, помимо устойчивости в тех или иных сеточных пространствах, особое внимание уделяется монотонности приближенного решения [1]–[3].

При рассмотрении параболических уравнений второго порядка конвективные слагаемые чаще всего берутся в недивергентной (характеристической) или дивергентной (консервативной) формах [4]. В этом случае устойчивость наиболее естественно рассматривается в соответствующих банаховых пространствах: в равномерной норме при использовании недивергентной формы и интегральной норме для дивергентной формы. Большие возможности в вычислительной гидродинамике предоставляет [5] запись конвективных слагаемых в так называемой симметричной форме, когда берется полусумма недивергентной и дивергентной форм. В этом случае обеспечивается безусловная кососимметричность оператора конвективного переноса, его энергетическая нейтральность. Сама краевая задача конвекции-диффузии рассматривается в гильбертовом пространстве, что позволяет ориентироваться на стандартные конечно-элементные аппроксимации по пространству [6], [7] и использование общих результатов теории устойчивости (корректности) операторно-разностных схем [8], [9].

При численном решении стационарных и нестационарных задач конвекции-диффузии большое внимание уделяется монотонным аппроксимациям – выполнению принципа максимума на дискретном уровне, наследованию такого свойства для решений эллиптических и параболических уравнений второго порядка [10]. На основе принципа максимума получены [11] априорные оценки для задач конвекции-диффузии в $L_\infty(\Omega)$, когда конвективные слагаемые записываются в недивергентной форме. Аналогично [4] формулируется принцип максимумадля задач с конвек-

[1)]Работа выполнена при финансовой поддержке Правительства РФ (соглашение № 14.Y26.31.0013).

тивными слагаемыми в дивергентной форме. В этом случае соответствующие априорные оценки имеют место в пространстве $L_1(\Omega)$.

Безусловно монотонные схемы для задач конвекции-диффузии строятся на основе аппроксимации конвективных слагаемых направленными разностями [8]. Хорошо известен основной недостаток таких схем, который связан с первым порядком аппроксимации. В вычислительной гидродинамике [3], [12] для построения монотонных схем второго порядка аппроксимации применяются различные подходы. В основном, явно или неявно используется идея перехода к схемам с направленными разностями в области, где нарушается условие монотонности схем с обычными центрально разностными аппроксимациями конвективных слагаемых. Подобные регуляризованные разностные схемы для задач конвекции-диффузии с недивергентными и дивергентными конвективными слагаемыми рассмотрены в книге [4]. Второй класс безусловно монотонных схем строится на основе трансформации исходного уравнения конвекции-диффузии, когда конвективный и диффузионный перенос записывается в виде диффузионного переноса вспомогательной величины. В этом случае мы приходим к экспоненциальным схемам, которые предложены в работах [13], [14], и в различных вариантах широко используются в вычислительной практике.

Для задач конвекции-диффузии с конвективными слагаемыми в недивергентной и дивергентной формах естественно ориентироваться на использование банаховых пространств $L_\infty(\Omega)$ и $L_1(\Omega)$) соответственно. Исследование устойчивости в банаховых пространствах $L_\infty(\omega)$ и $L_1(\omega)$ (сеточных аналогах $L_\infty(\Omega)$ и $L_1(\Omega)$) обычно проводится на основе принципа максимума. В работе [15] условия устойчивости двухслойных разностных схем для нестационарных задач конвекции-диффузии формулируются с использованием понятия логарифмической нормы [16], [17]. Устойчивость и монотонность этих схем обеспечивается диагональным преобладанием по строкам (в $L_\infty(\omega)$) или по столбцам (в $L_1(\omega)$).

Аналогичные проблемы устойчивости и монотонности двухслойных схем применительно к задачам конвекции-диффузии, в которых конвективный перенос записан в симметричной форме, обсуждаются в настоящей работе. Рассмотрение ведется в гильбертовом пространстве $L_2(\omega)$, а достаточные условия устойчивости и монотонности формулируются в виде специального варианта диагонального преобладания. Для того чтобы выделить ключевые моменты без усложнения нашего исследования непринципиальными техническими деталями, в качестве основного объекта рассматривается одномерное по пространству уравнение конвекции-диффузии. Общие комментарии относительно обобщения результатов на многомерные задачи даны в конце работы.

1. УРАВНЕНИЯ КОНВЕКЦИИ-ДИФФУЗИИ

Будем рассматривать модельные нестационарные задачи конвекции-диффузии с постоянным (не зависящим от времени, но зависящим от точки расчетной области) коэффициентом диффузионного переноса. В прикладных задачах естественно ориентироваться на случай, когда коэффициенты конвективного переноса зависят также и от времени.

Сформулируем простейшие одномерные задачи конвекции-диффузии. Уравнение конвекции-диффузии с конвективными слагаемыми в недивергентном виде [4] имеет вид:

$$\frac{\partial u}{\partial t} + v(x,t)\frac{\partial u}{\partial x} - \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) = f(x,t) \tag{1.1}$$

при $0 < x < l$, $t > 0$. Это уравнение дополним простейшими однородными граничными условиями Дирихле и начальным условием

$$u(0,t) = 0, \quad u(l,t) = 0, \quad t > 0, \tag{1.2}$$

$$u(x,0) = u^0(x), \quad 0 < x < l. \tag{1.3}$$

Вторым важнейшим примером является нестационарное уравнение конвекции-диффузии с конвективным переносом в дивергентной форме:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(v(x,t)u) - \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) = f(x,t). \tag{1.4}$$

Основным объектом нашего рассмотрения являются задачи конвекции-диффузии, в которых конвективный перенос берется в симметричной форме:

$$\frac{\partial u}{\partial t} + \frac{1}{2}\left(\frac{\partial}{\partial x}(v(x,t)u) + v(x,t)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) = f(x,t). \tag{1.5}$$

Будем рассматривать множество функций $u(x,t)$, удовлетворяющих граничным условиям (1.2). Нестационарную задачу конвекции-диффузии запишем в виде дифференциально-операторного уравнения

$$\frac{du}{dt} + \mathscr{A}u = f(t), \quad \mathscr{A} = \mathscr{A}(t) = \mathscr{C}(t) + \mathscr{D}, \tag{1.6}$$

где $\mathscr{C}(t)$ – оператор конвективного переноса, а $\mathscr{D}$ – оператор диффузионного переноса. Рассматривается задача Коши для эволюционного уравнения (1.6), т.е. уравнение дополняется условием

$$u(0) = u^0. \tag{1.7}$$

Приведем простейшие априорные оценки для рассматриваемых задач конвекции-диффузии (1.1)–(1.3) и (1.2)–(1.4) и (1.2), (1.3), (1.5). Рассматриваются пространства $L_\infty(0,l)$, $L_1(0,l)$ и $L_2(0,l)$, нормы в которых есть

$$\|v\|_\infty = \max_{0<x<l}|v(x)|, \quad \|v\|_1 = \int_0^l |v(x)|dx, \quad \|v\|_2 = \left(\int_0^l v^2(x)dx\right)^{1/2}.$$

Для решения нестационарной задачи конвекции-диффузии (1.1)–(1.3) (конвективный перенос в недивергентной форме) справедлива [18] априорная оценка в $L_\infty(0,l)$, которая следует из принципа максимума:

$$\|u(x,t)\|_\infty \le \|u^0(x)\|_\infty + \int_0^t \|f(x,\theta)\|_\infty d\theta. \tag{1.8}$$

Для решения задачи (1.2)–(1.4) имеет (см. [4]) место аналогичная априорная оценка, но в $L_1(0,l)$:

$$\|u(x,t)\|_1 \le \|u^0(x)\|_1 + \int_0^t \|f(x,\theta)\|_1 d\theta. \tag{1.9}$$

Наиболее просто устанавливается оценка для задачи конвекции-диффузии с конвективными слагаемыми в симметричной форме (1.2), (1.3), (1.5) в гильбертовом пространстве $L_2(0,l)$:

$$\|u(x,t)\|_2 \le \|u^0(x)\|_2 + \int_0^t \|f(x,\theta)\|_2 d\theta. \tag{1.10}$$

Для доказательства (1.10) привлекаются свойства положительности самосопряженного оператора диффузионного переноса ($\mathscr{D} = \mathscr{D}^* > 0$) и кососимметричности оператора конвективного переноса ($\mathscr{C} = -\mathscr{C}^*$). Мы хотим иметь аналогии априорных оценок (1.8)–(1.10) при рассмотрении дискретных задач.

2. УСТОЙЧИВОСТЬ И МОНОТОННОСТЬ ДВУХСЛОЙНЫХ СХЕМ

Сформулируем достаточные условия устойчивости двухслойных разностных схем для задачи Коши для линейной системы обыкновенных дифференциальных уравнений. Приведем также аналогичные условия монотонности в виде тех или иных условий диагонального преобладания.

Ищется решение следующей системы линейных обыкновенных уравнений первого порядка:

$$\frac{dw_i}{dt} + \sum_{j=1}^m a_{ij}(t)w_j = \varphi_i(t), \quad i = 1,2,\ldots,m, \quad t > 0. \tag{2.1}$$

Полагая $w = w(t) = \{w_1, w_2, \ldots, w_m\}$, $A = \{a_{ij}\}$, запишем (2.1) в матричном (операторном) виде:

$$\frac{dw}{dt} + A(t)w = \varphi(t). \tag{2.2}$$

Будем строить разностные схемы для приближенного решения задачи Коши, когда (2.2) дополняется начальными условиями

$$w(0) = u^0. \tag{2.3}$$

Нас интересует устойчивость разностного решения задачи (2.2), (2.3) в пространствах L_∞, L_1 и L_2. Для нормы вектора и подчиненной ей нормы матрицы [19] в L_∞ имеем

$$\|w\|_\infty = \max_{1\le i\le m}|w_i|, \quad \|A\|_\infty = \max_{1\le i\le m}\sum_{j=1}^{m}|a_{ij}|. \tag{2.4}$$

Аналогично в L_1

$$\|w\|_1 = \sum_{i=1}^{m}|w_i|, \quad \|A\|_1 = \max_{1\le j\le m}\sum_{i=1}^{m}|a_{ij}|. \tag{2.5}$$

Для L_2 имеем

$$\|w\|_2 = \left(\sum_{i=1}^{m}|w_i|^2\right)^{1/2}, \quad \|A\|_2 = \max_{1\le i\le m}(\lambda_i(AA^*))^{1/2}, \tag{2.6}$$

где $\lambda_i(G)$ – собственные значения симметричной матрицы G.

Устойчивость решения задачи Коши (2.2), (2.3) удобно формулировать с привлечением понятия логарифмической нормы. Логарифмическая норма матрицы A есть (см. [20], [21]) число

$$\mu(A) = \lim_{\delta\to 0+}\frac{\|I+\delta A\|-1}{\delta},$$

где I – единичная матрица. Для решения задачи (2.2), (2.3) имеет место априорная оценка

$$\|w(t)\| \le \exp(\mu(-A)t)\left(\|u^0\| + \int_0^t \exp(-\mu(-A)(t-\theta))\|\varphi(\theta)\|\,d\theta\right), \tag{2.7}$$

которая обеспечивает устойчивость по начальным данным и правой части.

Для логарифмической нормы матрицы в L_∞ (согласованной с (2.4)), в L_1 (согласованной с (2.5)) и в L_2 (согласованной с (2.6)) имеют место выражения

$$\mu_\infty(A) = \max_{1\le i\le m}\left(a_{ii} + \sum_{j\ne i,\,j=1}^{m}|a_{ij}|\right),$$

$$\mu_1(A) = \max_{1\le i\le m}\left(a_{ii} + \sum_{j\ne i,\,j=1}^{m}|a_{ji}|\right),$$

$$\mu_2(A) = \max_{1\le i\le m}\left(\lambda_i\left(\frac{A+A^*}{2}\right)\right).$$

При рассмотрении задач конвекции-диффузии мы ориентируемся на априорные оценки (1.8)–(1.10). Принимая во внимание (2.7), это соответствует тому, что

$$\mu_\alpha(-A) \le 0, \quad \alpha = \infty, 1, 2. \tag{2.8}$$

Сформулируем достаточные условия выполнения (2.8), которые имеют место для $\alpha = \infty, 1$.

Задачу (2.2), (2.3) будем рассматривать при следующих ограничениях. Диагональные элементы матрицы A предполагаются неотрицательными и имеется диагональное преобладание в том или ином варианте. Будем считать, например, что справедливы неравенства

$$a_{ii} \ge \sum_{j\ne i,\,j=1}^{m}|a_{ij}|, \quad i = 1, 2, \ldots, m, \tag{2.9}$$

т.е. имеет место нестрогое диагональное преобладание по строкам. Второй вариант связан с нестрогим диагональным преобладанием по столбцам:

$$a_{ii} \geq \sum_{j\neq i,\, j=1}^{m} |a_{ji}|, \quad i = 1, 2, \ldots, m. \tag{2.10}$$

Непосредственно убеждаемся, что при выполнении (2.9) или (2.10) имеет место неравенство (2.8) при $\alpha = \infty$ или $\alpha = 1$ соответственно.

При $\alpha = 2$ логарифмическая норма определяется по симметричной части A, причем

$$\mu_2(-A) = \max_{1\leq i\leq m}\left(\lambda_i\left(-\frac{A + A^*}{2}\right)\right) = -\min_{1\leq i\leq m}\left(\lambda_i\left(\frac{A + A^*}{2}\right)\right).$$

Тем самым $\mu_2(-A) \leq 0$, если A – неотрицательно определенная матрица:

$$A \geq 0. \tag{2.11}$$

Матрица с нестрогим диагональным преобладанием является неотрицательно-определенной. Это достаточное условие для выполнения (2.11) запишем для симметричной части матрицы A:

$$a_{ii} \geq \frac{1}{2}\sum_{j\neq i,\, j=1}^{m} |a_{ij} + a_{ji}|, \quad i = 1, 2, \ldots, m. \tag{2.12}$$

В силу этого при выполнении (2.12) для решения задачи (2.2), (2.3) справедлива оценка (1.10).

Для приближенного решения задачи (2.2), (2.3) будем использовать двухслойные схемы с весами. Обозначим приближенное решение на момент времени $t^n = n\tau$, где τ – шаг по времени, через y^n. Оно определяется из уравнения

$$\frac{y^{n+1} - y^n}{\tau} + A(\sigma y^{n+1} + (1 - \sigma)y^n) = \varphi^n, \tag{2.13}$$

где, например, $A = A^{n+\sigma} = A(\sigma t^{n+1} + (1 - \sigma)t^n)$, при начальном условии

$$y^0 = u^0. \tag{2.14}$$

Достаточные условия устойчивости разностной схемы (2.13), (2.14) в L_∞ и в L_1 формулируются в виде следующего утверждения.

Теорема 1. *Для задачи Коши* (2.2), (2.3) *с матрицей* A, *удовлетворяющей условиям* (2.9) (*или* (2.10)), *разностная схема с весами* (2.13), (2.14) *безусловно устойчива в* L_∞ (*или в* L_1) *при* $\sigma \geq 1$. *Если матрица* A *удовлетворяет условиям* (2.12), *то схема* (2.13), (2.14) *безусловно устойчива в* L_2 *при* $\sigma \geq 0.5$. *При этом для разностного решения верна априорная оценка*

$$\left\|y^{n+1}\right\| \leq \left\|u^0\right\| + \sum_{k=0}^{n} \tau\left\|\varphi^k\right\|. \tag{2.15}$$

Доказательство. При рассмотрении устойчивости в L_∞ и L_1 будем следовать [15]. Схему с весами (2.13), (2.14) удобно рассматривать как чисто неявную схему

$$\frac{y^{n+\sigma} - y^n}{\sigma\tau} + Ay^{n+\sigma} = \varphi^n \tag{2.16}$$

для функции $y^{n+\sigma} = \sigma y^{n+1} + (1 - \sigma)y^n$. Из (2.16) непосредственно следует

$$\left\|(I + \sigma\tau A)y^{n+\sigma}\right\| \leq \left\|y^n\right\| + \tau\sigma\left\|\varphi^n\right\|.$$

Для логарифмической нормы имеют место оценки

$$\|Au\| \geq -\mu(A)\|u\|, \quad \|Au\| \geq -\mu(-A)\|u\|.$$

В силу этого с учетом того, что при диагональном преобладании справедливо (2.8), имеем

$$\left\|(I + \sigma\tau A)y^{n+\sigma}\right\| \geq -\mu(-I - \sigma\tau A)\left\|y^{n+\sigma}\right\| = (1 + \sigma\tau\mu(-A))\left\|y^{n+\sigma}\right\| \geq \left\|y^{n+\sigma}\right\|.$$

Тем самым приходим к оценке

$$\left\|y^{n+\sigma}\right\| \le \left\|y^{n}\right\| + \tau\sigma\left\|\varphi^{n}\right\|. \tag{2.17}$$

При ограничениях $\sigma \ge 1$ имеем

$$\left\|y^{n+\sigma}\right\| \ge \sigma\left\|y^{n+1}\right\| - (\sigma-1)\left\|y^{n}\right\|.$$

В этих условиях из (2.17) получим неравенство

$$\left\|y^{n+1}\right\| \le \left\|y^{n}\right\| + \tau\left\|\varphi^{n}\right\|, \tag{2.18}$$

из которого следует оценка (2.15).

Для гильбертова пространства L_2 безусловная устойчивость схемы с весами (2.13), (2.14) при выполнении (2.11) имеет место [8], [9] при более слабых ограничениях на вес. Домножим скалярно уравнение (2.13) на $y^{n+\sigma}$, что с учетом (2.11) дает

$$(y^{n+1} - y^{n}, y^{n+\sigma}) \le \tau(y^{n+\sigma}, \varphi^{n}). \tag{2.19}$$

Для оценки левой части (2.19) воспользуемся следующим вспомогательным результатом [22].

Лемма 1. *Пусть* $w = \sigma u + (1-\sigma)v$ *и постоянная* $\sigma \ge 0.5$ *для* u, v *из некоторого гильбертового пространства* H_G ($G = G^* > 0$). *Тогда имеем*

$$(G(u-v), w) \ge (\|u\|_G - \|v\|_G)\|w\|_G. \tag{2.20}$$

В нашем случае $u = y^{n+1}$, $v = y^{n}$, $w = y^{n+\sigma}$, $G = I$ и поэтому для левой части (2.19) при $\sigma \ge 0.5$ неравенство (2.20) дает

$$(y^{n+1} - y^{n}, y^{n+\sigma}) \ge \left(\left\|y^{n+1}\right\| - \left\|y^{n}\right\|\right)\left\|y^{n+\sigma}\right\|.$$

С учетом неравенства

$$(y^{n+\sigma}, \varphi^{n}) \le \left\|y^{n+\sigma}\right\|\left\|\varphi^{n}\right\|$$

получаем (2.19).

Замечание 1. При малых σ мы можем рассчитывать на условную устойчивость. Например, для матриц A с диагональным преобладанием (2.9) (или (2.10)) разностная схема с весами (2.13), (2.14) условно устойчива [15] при $0 \le \sigma < 1$ в L_∞ (в L_1), если

$$\tau \le \frac{1}{1-\sigma}\left(\max_{1\le i\le m} a_{ii}\right)^{-1}. \tag{2.21}$$

С приведенными выше условиями устойчивости в виде диагонального преобладания непосредственно связывается свойство монотонности разностного решения задачи (2.13), (2.14) при дополнительном предположении о неположительности внедиагональных элементов матрицы A. Докажем следующее утверждение о безусловной монотонности схемы с весами (2.13), (2.14).

Теорема 2. *Пусть в схеме* (2.13), (2.14) *выполнены условия диагонального преобладания* (2.9) (*или* (2.10), *или* (2.12)) *при*

$$a_{ij} \le 0, \quad i \ne j, \quad i, j = 1, 2, \ldots, m, \tag{2.22}$$

и пусть

$$u^{0} \ge 0, \quad \varphi^{n} \ge 0, \quad n = 0, 1, \ldots,$$

тогда

$$y^{n+1} \ge 0, \quad n = 1, 2, \ldots,$$

при любых $\tau > 0$ и $\sigma \ge 1$.

Доказательство. Для перехода с одного временного слоя на другой из (2.16) имеем

$$By^{n+\sigma} = g^{n}, \quad n = 0, 1, \ldots, \tag{2.23}$$

где

$$B = I + \sigma\tau A, \quad g^n = y^n + \tau\varphi^n.$$

Доказательство неотрицательности решения проводится по индукции. Предположим, что $y^n \geq 0$ (при $n = 0$ это имеет место в силу предположений теоремы). Покажем, что при этом неотрицательным будет и $y^{n+1} \geq 0$.

Сначала установим, что в условиях теоремы матрица системы линейных алгебраических уравнений (2.23) является М-матрицей. Для элементов матрицы B имеем

$$b_{ii} = 1 + \sigma\tau a_{ii}, \quad b_{ij} = \sigma\tau a_{ij}, \quad j = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, m.$$

Тем самым обеспечена положительность диагональных и неположительность внедиагональных элементов матрицы.

При (2.9) имеем строгое диагональное преобладание по строкам:

$$b_{ii} > \sum_{j \neq i,\, j=1}^{m} |b_{ij}|, \quad i = 1, 2, \ldots, m.$$

Аналогично, при ограничениях (2.10) имеем строгое диагональное преобладание по столбцам:

$$b_{ii} > \sum_{j \neq i,\, j=1}^{m} |b_{ji}|, \quad i = 1, 2, \ldots, m.$$

В силу этого (см. [19]) матрица B является М-матрицей.

Отдельно рассматривается случай ограничений (2.12). Сформулируем следующее вспомогательное утверждение [23] (см. также [24], [25]) для матриц с α-диагональным преобладанием.

Лемма 2. *Пусть элементы матрицы A*

$$a_{ii} > 0, \quad a_{ij} \leq 0, \quad i \neq j, \quad i, j = 1, 2, \ldots, m,$$

и имеет место строгое α-диагональное преобладание

$$a_{ii} > \alpha R_i(A) + (1 - \alpha) S_i(A), \quad i = 1, 2, \ldots, m, \tag{2.24}$$

где

$$R_i(A) = \sum_{j \neq i,\, j=1}^{m} |a_{ij}|, \quad S_i(A) = \sum_{j \neq i,\, j=1}^{m} |a_{ji}|, \quad i = 1, 2, \ldots, m.$$

Тогда матрица A является М*-матрицей при всех* $\alpha \in [0, 1]$.

В условиях (2.12) и (2.22) для матрицы имеем

$$b_{ii} > \frac{1}{2}(R_i(B) + S_i(B)), \quad i = 1, 2, \ldots, m,$$

т.е. она является матрицей с α-диагональным преобладанием при $\alpha = 0.5$. В силу леммы 2 матрица B является М-матрицей.

Для правой части (2.23) имеем $g^n \geq 0$. Поэтому для решения уравнения (2.23) с М-матрицей B имеем $y^{n+\sigma} \geq 0$. Принимая во внимание, что $\sigma y^{n+1} = y^{n+\sigma} + (\sigma - 1)y^n$, получаем $y^{n+1} \geq 0$ при $\sigma \geq 1$, что и завершает доказательство теоремы.

Замечание 2. При $0 \leq \sigma < 1$ монотонность обеспечивается при малых шагах по времени. Например, для матрицы A с внедиагональными неположительными элементами и диагональным преобладанием по строкам или столбцам монотонность имеет место [15] при ограничениях (2.21).

Теперь мы можем использовать установленные результаты для исследования безусловной устойчивости и монотонности разностных схем для нестационарных задач конвекции-диффузии при записи конвективного переноса в различных формах.

3. РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЙ КОНВЕКЦИИ-ДИФФУЗИИ

Будем использовать на отрезке $[0,l]$ равномерные сетки, когда

$$\bar{\omega} \equiv \omega \cup \partial\omega = \{x \,|\, x = x_i = ih,\ i = 0,1,\ldots,N,\ Nh = l\},$$

и ω есть множество внутренних узлов:

$$\omega = \{x \,|\, x = x_i = ih,\ i = 1,2,\ldots,N-1,\ Nh = l\}.$$

При граничных условиях Дирихле (1.2) аппроксимация уравнения конвекции-диффузии (1.1) (или (1.4), (1.6)) и начальных условий (1.3) дает задаче (2.2), (2.3), при этом $m = N-1$ и приближенное решение $w_i(t) = w(x,t)$, $x \in \omega$.

Будем использовать стандартные аппроксимации на трехточечном шаблоне [8] при рассмотрении задач конвекции-диффузии [4]. Разностный оператор диффузии определим, например, в виде

$$Dw = -\frac{1}{h^2}k(x+0.5h)(w(x+h,t) - w(x,t)) + \frac{1}{h^2}k(x-0.5h)(w(x,t) - w(x-h,t)), \quad x \in \omega, \tag{3.1}$$

причем

$$w(x,t) = 0, \quad x \in \partial\omega. \tag{3.2}$$

Наиболее интересная возможность аппроксимации со вторым порядком h членов конвективного переноса связана с заданием скорости $v(x,t)$ в полуцелых узлах сетки $\bar{\omega}$. Оператор конвективного переноса в недивергентной форме (уравнение (1.1)) на множестве сеточных функций (3.2) зададим в виде

$$C_\infty w = \frac{1}{2h}v(x+0.5h,t)(w(x+h,t) - w(x,t)) + \frac{1}{2h}v(x-0.5h,t)(w(x,t) - w(x-h,t)), \quad x \in \omega. \tag{3.3}$$

Аналогично сеточный оператор конвективного переноса в дивергентной форме (уравнение (1.4)) возьмем в виде

$$C_1 w = \frac{1}{2h}v(x+0.5h,t)(w(x+h,t) + w(x,t)) - \frac{1}{2h}v(x-0.5h,t)(w(x,t) + w(x-h,t)), \quad x \in \omega. \tag{3.4}$$

Для оператора конвективного переноса из уравнения (1.5) положим

$$C_2 = \frac{1}{2}(C_1 + C_\infty),$$

так что

$$C_2 w = \frac{1}{2h}(v(x+0.5h,t)w(x+h,t) - v(x-0.5h,t)w(x-h,t))), \quad x \in \omega. \tag{3.5}$$

После аппроксимации по пространству придем к задаче (2.2), (2.3), когда

$$A = C + D, \tag{3.6}$$

при задании D согласно (3.1) на множестве функций (3.2), а $C = C_\infty, C_1, C_2$ – согласно (3.3), (3.4), (3.5) соответственно.

Сформулируем условия устойчивости и монотонности схем с весами (2.13), (2.14) для задачи (2.2), (2.3), (3.6). Свое рассмотрение начнем с задачи конвекции-диффузии (1.1)–(1.3) (аппроксимации (3.1)–(3.3), (3.6)).

Для того чтобы воспользоваться теоремами 1, 2, выпишем явно элементы матрицы A. Для (3.1)–(3.3), (3.6) ($C = C_\infty$) для диагональных элементов получим

$$a_{ii} = \frac{1}{h^2}(k_{i+1/2} + k_{i-1/2}) - \frac{1}{2h}v_{i+1/2} + \frac{1}{2h}v_{i-1/2},$$

а для внедиагональных имеем

$$a_{i,i-1} = -\frac{1}{h^2}k_{i-1/2} - \frac{1}{2h}v_{i-1/2}, \quad a_{i,i+1} = -\frac{1}{h^2}k_{i+1/2} + \frac{1}{2h}v_{i+1/2} \tag{3.7}$$

при использовании обозначений $k_{i\pm1/2} = k(x \pm 0.5h), x \in \omega$.

Ключевое для монотонности условие неположительности внедиагональных элементов (2.22) дает

$$\frac{1}{h^2}k_{i-1/2}+\frac{1}{2h}v_{i-1/2}\geq 0,\quad \frac{1}{h^2}k_{i+1/2}-\frac{1}{2h}v_{i+1/2}\geq 0. \tag{3.8}$$

При неположительности внедиагональных элементов матрицы A нестрогое диагональное преобладание по строкам (см. (2.9)) выполнено всегда. Неравенства (3.8) запишем в виде ограничений на шаг сетки по пространству:

$$\frac{h|v(x\pm 0.5h,t)|}{k(x\pm 0.5h)}\leq 2,\quad x\in\omega. \tag{3.9}$$

В случае конвективного переноса в дивергентной форме (3.1), (3.2), (3.4), (3.6) ($C=C_1$) для диагональных элементов матрицы A имеем

$$a_{ii}=\frac{1}{h^2}(k_{i+1/2}+k_{i-1/2})+\frac{1}{2h}v_{i+1/2}-\frac{1}{2h}v_{i-1/2}.$$

Внедиагональные элементы снова определяются согласно (3.7), а условия их неотрицательности (3.8) дают ограничения (3.9). Но теперь имеет место слабое диагональное преобладание по столбцам (см. (2.10)).

В наиболее интересном случае конвективного переноса в симметричной форме (3.1), (3.2), (3.5), (3.6) ($C=C_2$) для диагональных элементов матрицы A получим выражение

$$a_{ii}=\frac{1}{h^2}(k_{i+1/2}+k_{i-1/2}),$$

с внедиагональными элементами (3.7). Условие диагонального преобладания выполнено в форме (2.12).

Итогом нашего рассмотрения является следующее утверждение, которое является следствием теорем 1, 2.

Теорема 3. *Пусть в схеме* (2.13), (2.14), (3.6) *с сеточным оператором конвективного переноса* $C=C_\infty,C_1,C_2$, *определяемым согласно* (3.3)–(3.5), *выполнены условия* (3.9). *Тогда схема монотонна при всех* $\tau>0$, *если* $\sigma\geq 1$, *а для разностного решения имеет место априорная оценка* (2.15) *в пространстве* $L_\infty(\omega)$, $L_1(\omega)$, $L_2(\omega)$ соответственно.

Ограничение на шаг сетки по пространству можно снять, используя направленные разности для аппроксимации конвективных слагаемых. При этом мы жертвуем точностью: ранее рассмотренные центрально-разностные аппроксимации конвективных слагаемых имели второй порядок аппроксимации по h, а аппроксимации конвективного переноса направленными разностями – первый. Введем обозначения

$$v(x,t)=v^+(x,t)+v^-(x,t),$$

$$v^+(x,t)=\frac{1}{2}(v(x,t)+|v(x,t)|)\geq 0,\quad v^-(x,t)=\frac{1}{2}(v(x,t)-|v(x,t)|)\leq 0.$$

Ориентируясь на задание скорости в полуцелых узлах, вместо (3.3) положим

$$C_\infty w=\frac{1}{h}v^-(x+0.5h,t)(w(x+h,t)-w(x,t))+\frac{1}{h}v^+(x-0.5h,t)(w(x,t)-w(x-h,t)),\quad x\in\omega. \tag{3.10}$$

Удобно записать (3.10), выделив явно диагональную часть оператора конвективного переноса:

$$\begin{aligned}C_\infty w=&\frac{1}{h}(v^+(x-0.5h,t)-v^-(x+0.5h,t))w(x,t)+\\&+\frac{1}{h}(v^-(x+0.5h,t)(w(x+h,t)-v^+(x-0.5h,t)w(x-h,t)),\quad x\in\omega.\end{aligned} \tag{3.11}$$

Для сеточного оператора конвективного переноса в дивергентной форме имеем

$$C_1 w = \frac{1}{h}(v^+(x+0.5h,t) - v^-(x-0.5h,t))w(x,t) + \\ + \frac{1}{h}(v^-(x+0.5h,t)(w(x+h,t) - v^+(x-0.5h,t)w(x-h,t)), \quad x \in \omega. \tag{3.12}$$

Сеточный оператор конвективного переноса в симметричной форме аппроксимируется направленными разностями следующим образом:

$$C_2 w = \frac{1}{2h}(|v(x+0.5h,t)| + |v(x-0.5h,t)|)w(x,t) + \\ + \frac{1}{h}(v^-(x+0.5h,t)(w(x+h,t) - v^+(x-0.5h,t)w(x-h,t)), \quad x \in \omega. \tag{3.13}$$

При таких аппроксимациях конвективных слагаемых внедиагональные элементы всегда неположительны, а диагональные – неотрицательны. Условия слабого диагонального преобладания в формах (2.9), (2.10) и (2.12) при аппроксимациях направленными разностями (3.11)–(3.13) проверяются непосредственно.

Теорема 4. *Пусть в схеме* (2.13), (2.14), (3.6) *сеточный оператор конвективного переноса* $C = C_\infty, C_1, C_2$ *определяется согласно* (3.11)–(3.13). *Тогда схема монотонна при всех* $\tau > 0$ и $\sigma \geq 1$, *а для разностного решения имеет место априорная оценка* (2.15) *в пространстве* $L_\infty(\omega)$, $L_1(\omega)$, $L_2(\omega)$ *соответственно*.

4. ЭКСПОНЕНЦИАЛЬНЫЕ СХЕМЫ

Безусловно монотонные схемы наиболее просто конструируются на основе трансформации исходных уравнений конвекции-диффузии с исключением членов конвективного переноса [4]. От (1.1) можно прийти к уравнению

$$\frac{\partial u}{\partial t} - \frac{1}{\chi(x,t)}\frac{\partial}{\partial x}\left(k(x)\chi(x,t)\frac{\partial u}{\partial x}\right) = f(x,t), \tag{4.1}$$

в котором

$$\chi(x,t) = \exp\left(-\int_0^x \frac{v(s,t)}{k(s)}ds\right). \tag{4.2}$$

С использованием функции $\chi(x,t)$ уравнение (1.4) можно записать в виде

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\frac{k(x)}{\chi(x,t)}\frac{\partial(\chi(x,t)u)}{\partial x}\right) = f(x,t). \tag{4.3}$$

Для уравнения (1.5) имеем

$$\frac{\partial u}{\partial t} - \frac{1}{2\chi(x,t)}\frac{\partial}{\partial x}\left(k(x)\chi(x,t)\frac{\partial u}{\partial x}\right) - \frac{1}{2}\frac{\partial}{\partial x}\left(\frac{k(x)}{\chi(x,t)}\frac{\partial(\chi(x,t)u)}{\partial x}\right) = f(x,t). \tag{4.4}$$

Отменим некоторые возможности по разностной аппроксимации таких диффузионно-конвективных слагаемых.

Будем отталкиваться от уравнения (4.1). Для сеточных функций, удовлетворяющих (3.2), положим

$$Aw = -\frac{1}{h^2\chi(x,t)}k(x+0.5h)\chi(x+0.5h,t)(w(x+h,t) - w(x,t)) + \\ + \frac{1}{h^2\chi(x,t)}k(x-0.5h)\chi(x-0.5h,t)(w(x,t) - w(x-h,t)). \tag{4.5}$$

Для вычисления $\chi(x \pm 0.5h,t)$ учитываем то, что

$$\chi(x \pm 0.5h,t) = \chi(x)\exp\left(-\int_x^{x\pm 0.5h} \frac{v(s,t)}{k(s)}ds\right),$$

и сеточные функции $k(x)$, $v(x,t)$ заданы в полуцелых узлах. С точностью до $\mathbb{O}(h^2)$ положим

$$\chi(x \pm 0.5h, t) = \chi(x)\exp(\mp\theta(x \pm 0.5h, t)h)$$

при использовании обозначений

$$\theta(x,t) = \frac{v(x,t)}{2k(x)}.$$

Это позволяет вместо (4.5) использовать следующую аппроксимацию для $A = A_\infty$:

$$\begin{aligned} A_\infty w = &-\frac{1}{h^2}k(x+0.5h)\exp(-\theta(x+0.5h,t)h)(w(x+h,t)-w(x,t)) + \\ &+\frac{1}{h^2}k(x-0.5h)\exp(\theta(x-0.5h,t)h)(w(x,t)-w(x-h,t)). \end{aligned} \tag{4.6}$$

По аналогии с (4.5) для уравнения (4.3) используем

$$\begin{aligned} Aw = &-\frac{k(x+0.5h)}{h^2\chi(x+0.5h,t)}(\chi(x+h,t)w(x+h,t)-\chi(x,t)w(x,t)) + \\ &+\frac{k(x-0.5h)}{h^2\chi(x-0.5h,t)}(\chi(x,t)w(x,t)-\chi(x-h,t)w(x-h,t)). \end{aligned}$$

Упрощая это выражение, получаем для $A = A_1$:

$$\begin{aligned} A_1 w = &-\frac{1}{h^2}k(x+0.5h)\exp(-\theta(x+0.5h,t)h)w(x+h,t) + \frac{1}{h^2}k(x+0.5h)\exp(\theta(x+0.5h,t)h)w(x,t) + \\ &+\frac{1}{h^2}k(x-0.5h)\exp(-\theta(x-0.5h,t)h)w(x,t) - \frac{1}{h^2}k(x-0.5h)\exp(\theta(x-0.5h,t)h)w(x-h,t). \end{aligned} \tag{4.7}$$

Для уравнения (4.4) используем аппроксимацию для $A_2 = 0.5(A_1 + A_\infty)$. С учетом (4.6) и (4.7) получим

$$\begin{aligned} A_2 w = &-\frac{1}{h^2}k(x+0.5h)\exp(-\theta(x+0.5h,t)h)w(x+h,t) + \\ &+\frac{1}{2h^2}k(x+0.5h)(\exp(\theta(x+0.5h,t)h)+\exp(-\theta(x+0.5h,t)h))w(x,t) + \\ &+\frac{1}{2h^2}k(x-0.5h)(\exp(\theta(x-0.5h,t)h)+\exp(-\theta(x-0.5h,t)h))w(x,t) - \\ &-\frac{1}{h^2}k(x-0.5h)\exp(\theta(x-0.5h,t)h)w(x-h,t). \end{aligned} \tag{4.8}$$

Для внедиагональных элементов $A = A_1, A_2, A_\infty$ имеем

$$a_{i,i-1} = -\frac{1}{h^2}k_{i-1/2}\exp(\theta_{i-1/2}), \quad a_{i,i+1} = -\frac{1}{h^2}k_{i+1/2}\exp(-\theta_{i+1/2}),$$

т.е. они все неотрицательны. Для диагональных элементов матрицы $A = A_\infty$ из (4.6) следует слабое диагональное преобладание по строкам (2.9). Аналогично, из (4.7) для $A = A_1$ имеем слабое диагональное преобладание по столбцам (2.10). Для $A = A_2$ из (4.8) получим слабое диагональное преобладание в форме (2.12). Тем самым можно сформулировать утверждение.

Теорема 5. *Пусть в схеме* (2.13), (2.14) *сеточный оператор* $A = A_\infty, A_1, A_2$ *определяется согласно* (4.6)–(4.8). *Тогда схема монотонна при всех* $\tau > 0$ *и* $\sigma \geq 1$, *а для разностного решения имеет место априорная оценка* (2.15) *в пространстве* $L_\infty(\omega)$, $L_1(\omega)$, $L_2(\omega)$ *соответственно*.

Как и для схем с аппроксимацией конвективного переноса направленными разностями (теорема 4) мы имеем безусловно устойчивые и монотонные в различных пространствах. Преимущество экспоненциальных схем состоит в том, что как и для схем с центрально-разностными аппроксимациями (теорема 3), погрешность аппроксимации по пространству имеет второй порядок. Платой за повышение точности является более сложное вычисление коэффициентов разностного оператора.

5. МНОГОМЕРНЫЕ ЗАДАЧИ

Отметим возможности построения монотонных схем для нестационарных уравнений конвекции-диффузии на примере модельных двумерных задач. В прямоугольнике

$$\Omega = \{\mathbf{x} | \mathbf{x} = (x_1, x_2),\ 0 < x_\alpha < l_\alpha,\ \alpha = 1, 2\}$$

рассматривается нестационарное уравнение конвекции-диффузии с конвективными слагаемыми в недивергентном виде

$$\frac{\partial u}{\partial t} + \sum_{\alpha=1}^{2} v_\alpha(\mathbf{x}, t)\frac{\partial u}{\partial x_\alpha} - \sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}\left(k(\mathbf{x})\frac{\partial u}{\partial x_\alpha}\right) = f(\mathbf{x}, t). \tag{5.1}$$

Это уравнение дополняется однородными граничными условиями Дирихле

$$u(\mathbf{x}, t) = 0, \quad \mathbf{x} \in \partial\Omega, \quad t > 0. \tag{5.2}$$

Кроме того, задается начальное условие

$$u(\mathbf{x}, 0) = u^0(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{5.3}$$

Как и при рассмотрении одномерных задач, вторым примером является нестационарное уравнение конвекции-диффузии с конвективным переносом в дивергентной форме:

$$\frac{\partial u}{\partial t} + \sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}(v_\alpha(\mathbf{x}, t)u) - \sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}\left(k(\mathbf{x})\frac{\partial u}{\partial x_\alpha}\right) = f(\mathbf{x}, t).$$

Основным является уравнение конвекции-диффузии с конвективным переносом в симметричной форме, когда

$$\frac{\partial u}{\partial t} + \frac{1}{2}\sum_{\alpha=1}^{2}\left(\frac{\partial}{\partial x_\alpha}(v_\alpha(\mathbf{x}, t)u) + v_\alpha(\mathbf{x}, t)\frac{\partial u}{\partial x_\alpha}\right) - \sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}\left(k(\mathbf{x})\frac{\partial u}{\partial x_\alpha}\right) = f(\mathbf{x}, t).$$

Операторы конвекции-диффузии в многомерных задачах представляются в виде суммы одномерных операторов конвекции-диффузии. В силу этого при построении монотонных схем для многомерных задач мы можем опираться на рассмотренные выше аппроксимации одномерных операторов конвекции-диффузии. В частности, можно строить экспоненциальные схемы на основе записи уравнений аналогично (4.1), (4.3), (4.4), например, для задачи (5.1)–(5.3) будем использовать уравнение

$$\frac{\partial u}{\partial t} - \sum_{\alpha=1}^{2}\frac{1}{\chi_\alpha(\mathbf{x}, t)}\frac{\partial}{\partial x_\alpha}\left(k(\mathbf{x})\chi_\alpha(\mathbf{x}, t)\frac{\partial u}{\partial x_\alpha}\right) = f(\mathbf{x}, t),$$

где теперь

$$\chi_1(\mathbf{x}, t) = \exp\left(-\int_0^{x_1}\frac{v_1(s, x_2, t)}{k(s, x_2)}ds\right), \quad \chi_2(\mathbf{x}, t) = \exp\left(-\int_0^{x_2}\frac{v_2(x_1, s, t)}{k(x_1, s)}ds\right).$$

При использовании равномерных по каждому направлению сеток исследование устойчивости и монотонности разностных схем для многомерных уравнений конвекции-диффузии в соответствующих пространствах сеточных функций проводится полностью аналогично одномерному случаю.

При расщеплении по направлениям монотонность обеспечивается использованием локально-одномерных схем покомпонентного расщепления [26]. Такое рассмотрение проведено в [15] для многомерных задач конвекции-диффузии, когда конвективный перенос берется в недивергентной или дивергентной формах.

Не принципиальным является также обобщение результатов на случай неравномерных прямоугольных сеток. Более интересные проблемы порождаются использованием общих неструктурированных сеток. В этой связи отметим, что монотонные схемы метода конечных объемов строятся [27] при использовании сеток Делоне и разбиения Вороного, как контрольного объема. В методе конечных элементов мы можем использовать линейные конечные элементы на таких сетках и специальные аппроксимации коэффициента при производной по времени – lumping процедуры (см., например, [7]).

СПИСОК ЛИТЕРАТУРЫ

1. *Hundsdorfer W.H.* Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Berlin: Springer, 2003.
2. *Morton K.W.* Numerical Solution of Convection-Diffusion Problems. London: Chapman & Hall, 1996.
3. *Wesseling P.* Principles of Computational Fluid Dynamics. Berlin: Springer, 2001.
4. *Самарский А.А., Вабищевич П.Н.* Численные методы решения задач конвекции–диффузии. М.: URSS, 1999.
5. *Vabishchevich P.N.* On the form of the hydrodynamics equations // High Speed Flow Field conference, 19–22, November 2007, Moscow, Russia. Moscow, 2007. P. 1–9.
6. *Brenner S., Scott R.*The Mathematical Theory of Finite Element Methods. New York: Springer, 2007.
7. *Thomée V.* Galerkin Finite Element Methods for Parabolic Problems. Berlin: Springer, 2006.
8. *Самарский А.А.* Теория разностных схем. М.: Наука, 1989.
9. *Самарский А.А., Гулин А.В.* Устойчивость разностных схем. М.: Наука, 1973.
10. *Protter M.H., Weinberger H.F.* Maximum Principles in Differential Equations. New York: Springer, 1967.
11. *Ладыженская О.А., Солонников В.А., Уральцева Н.Н.* Линейные и квазилинейные уравнения параболического типа М.: Наука, 1967.
12. *Tannehill J.C., Anderson D.A., Pletcher R. H.* Computational Fluid Mechanics and Heat Transfer. Philadelphia: Taylor & Francis, 1997.
13. *de G.Allen D.N.*, *Southwell R.V.* Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder // Quart. J. Mech. Appl. Math. 1955. V. 8. P. 129–145.
14. *Scharfetter D.L., Gummel H.K.* Large-signal analysis of a silicon read diode oscillator // IEEE Trans. Electron Devices. 1969. V. ED-16. P. 4–77.
15. *Afanas'eva N.M., Churbanov A.G., Vabishchevich P.N.* Unconditionally monotone schemes for unsteady convection-diffusion problems // Comput. Meth. in Appl. Math. 2013. V. 13. № 2. P. 185–205.
16. *Dekker K., Verwer J.G.* Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations. Amsterdam: North-Holland, 1984.
17. *Desoer C.A., Vidyasagar M.* Feedback Systems: Input-Output Properties. Philadelphia: SIAM, 2008.
18. *Friedman A.* Partial Differential Equations of Parabolic Type. Englewood: Prentice-Hall, 1964.
19. *Horn R.A., Johnson C.R.* Matrix Analysis. Cambridge: Cambridge University Press, 1990.
20. *Dekker K., Verwer J.G.* Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations. Amsterdam: North-Holland, 1984.
21. *Hairer E., Norsett S.P., Wanner G.* Solving Ordinary Differential Equations. I. Nonstiff Problems. Berlin: Springer, 1987.
22. *Вабищевич П.Н.* Потоковые схемы расщепления для параболических задач со смешанными производными // Ж. вычисл. матем. и матем. физ. 2013. Т. 53. № 8. С. 1314–1328.
23. *Sun Y.X.* Sufficient conditions for generalized diagonally dominant matrices // Numerical Math. (A Journal of Chinese Universities). 1997. V. 19. № 3. P. 216–223. in Chinese.
24. *Wang G., Hong Z., Gao Z.* Sufficient conditions of nonsingular H-matrices // J. of Shanghai University (English Edition). 2004. V. 8. № 1. P. 35–37.
25. *Guo Z.J., Guo Z.J., Yang J.G.* A new criteria for a matrix is not generalized strictly diagonally dominant matrix // Applied Mathematical Sciences. 2011. V. 5. № 6. P. 273–278.
26. *Вабищевич П.Н.* Аддитивные операторно-разностные схемы (схемы расщепления). М.: URSS, 2013.
27. *Vabishchevich P.N.* Finite-difference approximation of mathematical physics problems on irregular grids // Computational Methods in Applied Mathematics. 2005. V. 5. № 3. P. 294–330.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 108–123

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 517.958

МЕТОД ВОЗМУЩЕНИЙ В ТЕОРИИ РАСПРОСТРАНЕНИЯ ДВУХЧАСТОТНЫХ ЭЛЕКТРОМАГНИТНЫХ ВОЛН В НЕЛИНЕЙНОМ ВОЛНОВОДЕ I: ТЕ-ТЕ ВОЛНЫ[1)]

© 2021 г. Д. В. Валовик

440026 Пенза, ул. Красная, 40, Пензенский гос. ун-т, Россия

e-mail: dvalovik@mail.ru

Поступила в редакцию 06.02.2020 г.
Переработанный вариант 06.02.2020 г.
Принята к публикации 18.09.2020 г.

В статье рассмотрена задача о распространении двухчастотной электромагнитной волны в волноводе, заполненном нелинейной средой. Двухчастотная волна является суммой двух монохроматических ТЕ-волн, характеризующихся разными частотами. Диэлектрическая проницаемость волновода характеризуется весьма общей функцией нелинейности, отвечающей эффектам самовоздействия. В работе показано, что при некоторых условиях рассматриваемая двухчастотная волна является собственной модой волновода. С математической точки зрения изучаемая задача сводится к нелинейной двухпараметрической задаче на собственные значения для системы (нелинейных) уравнений Максвелла. Основным результатом статьи является доказательство существования нелинеаризуемых решений указанной задачи. Библ. 36. Фиг. 2.

Ключевые слова: уравнения Максвелла, нелинейная диэлектрическая проницаемость, плоский диэлектрический волновод, нелинейная задача типа Штурма–Лиувилля, двухпараметрическая задача на собственные значения, метод возмущений, интегральное характеристическое уравнение.

DOI: 10.31857/S0044466921010099

1. МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ И ВВОДНЫЕ ЗАМЕЧАНИЯ

Классический метод возмущений [1] основан на следующей идее. Рассмотрим нелинейную задачу $R(\boldsymbol{\alpha})$, где $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)$ – вектор числовых параметров. В операторном виде такую задачу можно записать так: $\mathcal{R}(\mathbf{u}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = 0$, где $\mathcal{R}$ – некоторая нелинейная по $\mathbf{u}$ оператор-функция, $\boldsymbol{\lambda}$ – еще один вектор параметров (например, спектральных). Пусть существует решение $\mathbf{u} \equiv \mathbf{u}_0(x; \boldsymbol{\lambda})$ задачи $R(\mathbf{0})$, где $\mathbf{0} = (0, \ldots, 0)$, которая является линейной. При достаточно широких предположениях относительно $\mathcal{R}$ можно доказать, что как только величина $|\boldsymbol{\alpha}|$ достаточно мала, задача $R(\boldsymbol{\alpha})$ имеет решение $\mathbf{u} \equiv \mathbf{u}_{\boldsymbol{\alpha}}(x; \boldsymbol{\lambda})$ и $\mathbf{u}_{\boldsymbol{\alpha}}(x; \boldsymbol{\lambda}) \to \mathbf{u}_0(x; \boldsymbol{\lambda})$ при $|\boldsymbol{\alpha}| \to \mathbf{0}$. Этот подход реализован, например, в [2]–[4].

Пусть задача $R(\boldsymbol{\alpha}')$, где $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_{l'}, 0, \ldots, 0)$ и $\alpha_1, \ldots, \alpha_{l'} \neq 0$, остается нелинейной, но является в некотором смысле более простой, чем задача $R(\boldsymbol{\alpha})$. Предположим, что разрешимость задачи $R(\boldsymbol{\alpha}')$ может быть установлена и пусть $\mathbf{u} \equiv \mathbf{u}_{\boldsymbol{\alpha}'}(x; \boldsymbol{\lambda})$ – ее решение. Тогда при определенных условиях на $\mathcal{R}$ можно доказать, что если достаточно мала величина $|\boldsymbol{\alpha} - \boldsymbol{\alpha}'|$, то разрешима будет и задача $R(\boldsymbol{\alpha})$ и $\mathbf{u}_{\boldsymbol{\alpha}}(x; \boldsymbol{\lambda}) \to \mathbf{u}_{\boldsymbol{\alpha}'}(x; \boldsymbol{\lambda})$ при $|\boldsymbol{\alpha} - \boldsymbol{\alpha}'| \to 0$.

Обсуждаемый подход может оказаться полезным, если:

(i) найдется вектор $\boldsymbol{\alpha}'$ такой, что $R(\boldsymbol{\alpha}')$, будучи более простой нелинейной задачей, может быть исследована каким-либо методом;

(ii) решения задачи $R(\boldsymbol{\alpha}')$ существуют не только для “малых” $|\boldsymbol{\alpha}'|$;

(iii) задача $R(\boldsymbol{\alpha}')$ имеет решения, не связанные с решениями задачи $R(\mathbf{0})$.

[1)]Работа выполнена при финансовой поддержке РНФ (код проекта 18-71-10015).

Очевидно, что в случаях (ii), (iii) рассмотрение задачи $R(\mathbf{0})$ в качестве "невозмущенной" даст весьма скудную информацию о решениях задачи $R(\boldsymbol{\alpha})$. Более того, задача $R(\mathbf{0})$ может не иметь решений.

В электродинамике нелинейных волноведущих структур возникают задачи, для которых справедливы ситуации (i)–(iii). Например, задачи, зависящие от скалярного параметра α и удовлетворяющие условиям (ii), (iii), изучались в работах [5], [6]. В работе [4] введен широкий класс задач, зависящих от (векторного) параметра $\boldsymbol{\alpha}$, к которым применим изложенный выше подход. В работах [7], [8] изучались многопараметрические задачи, для которых справедливы условия (i)–(iii).

Перейдем к постановке задачи. Всюду ниже индекс j принимает значения 1, 2. Пусть $\mathrm{I} = (0, h)$, $\overline{\mathrm{I}} = [0, h]$, где $h > 0$ – фиксированная постоянная,

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2), \quad \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4), \quad \boldsymbol{\alpha}' = (\alpha_1, \alpha_2, 0, 0)$$

являются наборами положительных параметров. Считаем, что

$$\boldsymbol{\Lambda}^* = \Lambda_1^* \times \Lambda_2^*, \quad \mathbf{A}^* = \mathrm{A}_1^* \times \mathrm{A}_2^* \times \mathrm{A}_3^* \times \mathrm{A}_4^*,$$

где $\Lambda_j^* = (b_j, \lambda_j^*)$, $\mathrm{A}_j^* = (0, \alpha_j^*)$, $\mathrm{A}_{j+2}^* = (0, \alpha_{j+2}^*)$; $b_j > 0$ – фиксированные постоянные; положительные числа λ_j^*, α_j^*, α_{j+2}^* фиксированы, при этом числа λ_j^* и α_{j+2}^* будут определены позднее (см. следствие 1 и утверждение 5). Необходимо отметить, что параметры α_j^* не предполагаются малыми.

Наконец, введем в рассмотрение вещественнозначные неотрицательные функции $f_j \equiv f_j(s_j)$ и функции произвольного знака $g_j \equiv g_j(s_1, s_2)$ одного и двух аргументов соответственно, такие, что $f_j(0) = 0$ и $g_j(0,0) = 0$. При этом функции $f_j(s_j) \in C[0, +\infty)$, $s_j f_j'(s_j) \in C[0, +\infty)$, а g_j являются однократно непрерывно дифференцируемыми при $s_1, s_2 \in [0, +\infty)$. Кроме этого, предполагается, что функции f_j монотонно возрастают и на бесконечности характеризуются следующим поведением:

$$f_j(t) = t^{q_j} + \tilde{f}_j(t) \quad \text{при} \quad t \to +\infty,$$

где $\lim_{t \to +\infty} t^{-q_j} \tilde{f}_j(t) = 0$, а $q_j > 0$ – некоторые постоянные.

Задача $P = P(\boldsymbol{\alpha})$ заключается в нахождении тех значений параметра $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$, для которых существуют решения $u_1 \equiv u_1(x; \overline{\boldsymbol{\lambda}}, \boldsymbol{\alpha})$ и $u_2 \equiv u_2(x; \overline{\boldsymbol{\lambda}}, \boldsymbol{\alpha})$ системы

$$\begin{aligned} u_1'' &= -(a_1 - \lambda_1 + \alpha_1 f_1(u_1^2) + \alpha_3 g_1(u_1^2, u_2^2))u_1, \\ u_2'' &= -(a_2 - \lambda_2 + \alpha_2 f_2(u_2^2) + \alpha_4 g_2(u_1^2, u_2^2))u_2, \end{aligned} \tag{1.1}$$

удовлетворяющие краевым условиям

$$u_j\big|_{x=0} = A_j, \quad u_j'\Big|_{x=0} = \kappa_{s,j} A_j, \tag{1.2}$$

$$(\kappa_{c,j} u_j + u_j')\Big|_{x=h} = 0, \tag{1.3}$$

где $(x, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \in \overline{\mathrm{I}} \times \boldsymbol{\Lambda}^* \times \mathbf{A}^*$, $\kappa_{c,j} = \sqrt{\lambda_j - c_j} > 0$, $\kappa_{s,j} = \sqrt{\lambda_j - b_j} > 0$, a_j, b_j, c_j, $A_j > 0$ – вещественные постоянные такие, что $c_j \le b_j < a_j$, и

$$u_1, u_2 \in C^2[0, h]. \tag{1.4}$$

Определение 1. Вектор $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}^*$, где $\overline{\boldsymbol{\lambda}} = (\overline{\lambda}_1, \overline{\lambda}_2)$ такой, что существуют функции $u_1 \equiv u_1(x; \overline{\boldsymbol{\lambda}}, \boldsymbol{\alpha})$, $u_2 \equiv u_2(x; \overline{\boldsymbol{\lambda}}, \boldsymbol{\alpha})$, удовлетворяющие (1.1)–(1.4), называется векторным собственным значением задачи $P(\boldsymbol{\alpha})$, а отвечающие ему функции u_1, u_2 – собственными функциями задачи $P(\boldsymbol{\alpha})$.

Рассмотрим **задачу** $P(\boldsymbol{\alpha}')$. Эта задача распадается на две (независимые) нелинейные задачи, обозначаемые P_j, удовлетворяющие условиями (ii), (iii).

Пусть $\Lambda_j = (b_j, +\infty)$ и $\mathrm{A} = (0, +\infty)$. Задача P_j заключается в нахождении тех значений параметра $\lambda_j = \hat{\lambda}_j$, для которых существуют решения $v_j \equiv v_j(x; \hat{\lambda}_j, \alpha_j)$ уравнения

$$v_j'' = -(a_j - \lambda_j)v_j - \alpha_j f_j(v_j^2)v_j, \tag{1.5}$$

удовлетворяющие краевым условиям

$$v_j\big|_{x=0} = A_j, \quad v_j'\Big|_{x=0} = \kappa_{s,j}A_j, \tag{1.6}$$

$$(\kappa_{c,j}v_j + v_j')\Big|_{x=h} = 0, \tag{1.7}$$

где $(x, \lambda_j, \alpha_j) \in \overline{\mathrm{I}} \times \Lambda_j \times \mathrm{A}$, постоянные и параметры a_j, A_j, $\kappa_{s,j}$, $\kappa_{c,j}$ определены в (1.1)–(1.3), и

$$v_j \in C^2[0, h]. \tag{1.8}$$

Определение 2. Число $\lambda_j = \hat{\lambda}_j \in \Lambda_j$, для которого существует функция $v_j \equiv v_j(x; \hat{\lambda}_j, \alpha_j)$, удовлетворяющая (1.5)–(1.8), называется собственным значением задачи P_j, а отвечающая ему функция v_j – собственной функцией задачи P_j.

Задача $P(\mathbf{0})$, где $\mathbf{0} = (0, 0, 0, 0)$, распадается на две (независимые) линейные задачи, которые обозначим P_j^0. Задача P_j^0 заключается в нахождении тех значений параметра $\lambda_j = \tilde{\lambda}_j$, для которых существуют нетривиальные решения $w_j \equiv w_j(x; \tilde{\lambda}_j)$ уравнения $w_j'' = -(a_j - \lambda_j)w_j$, удовлетворяющие краевым условиям $(\kappa_{s,j}w_j - w_j')\Big|_{x=0} = 0$, $(\kappa_{c,j}w_j + w_j')\Big|_{x=h} = 0$, где $(x, \lambda_j) \in \overline{\mathrm{I}} \times (b_j, a_j)$, постоянные и параметры a_j, $\kappa_{s,j}$, $\kappa_{c,j}$ определены в (1.1)–(1.3), и $w_j \in C^2[0, h]$.

Основным методом исследования задач P_j является метод интегрального характеристического уравнения [9]. Отметим, что широко известные подходы нелинейного анализа, такие как вариационный метод [10], [11] и методы теории ветвления решений [12], [13], не применимы для исследования задач $P(\boldsymbol{\alpha})$ и P_j, см. также комментарии в работе [14].

Результаты и методы, представленные в настоящей статье, могут быть полезны и в других областях нелинейной математической физики, где возникают многопараметрические задачи на собственные значения, например, в теории связанных осцилляторов с нелинейным взаимодействием, см. [15]–[19] и библиографию там; экспериментальные наблюдения представлены в [17], [19]. Основная идея предлагаемого здесь метода заключается в сведении (когда таковое возможно) нелинейной многопараметрической задачи к нескольким нелинейным задачам с меньшим числом параметров. В рассматриваемом случае это однопараметрические задачи. Если эти (нелинейные) однопараметрические задачи могут быть эффективно исследованы и их решения существенно отличаются от решений соответствующих линеаризованных задач, то предлагаемый метод, вероятно, позволит распространить результаты, полученные для (нелинейных) однопараметрических задач, на случай многопараметрической задачи.

Заметим, что предложенный ниже подход может быть применен для задач с различными краевыми условиями, см., например, [3], где используются нелинейные краевые условия и [8], где используются краевые условия I рода.

Статья написана в соответствии со следующим планом: формулировки результатов приведены в разд. 2; численные результаты представлены в разд. 3; обсуждение результатов и заключительные комментарии даны в разд. 4; доказательства представлены в разд. 5; физическая задача о распространении волн, которая приводит к задаче $P(\boldsymbol{\alpha})$, сформулирована в разд. 6.

2. РЕЗУЛЬТАТЫ

Собственные значения задач $P(\boldsymbol{\alpha})$, P_j и P_j^0 обозначим через $\overline{\boldsymbol{\lambda}}_{k,l} = (\overline{\lambda}_{1,k}, \overline{\lambda}_{2,l})$, $\hat{\lambda}_{j,k}$ и $\tilde{\lambda}_{j,k}$ соответственно, где $k, l = 1, 2, \ldots$, – некоторые целочисленные индексы; также будут использоваться обозначения $\overline{\boldsymbol{\lambda}}$, $\hat{\lambda}_j$ и $\tilde{\lambda}_j$. Если собственные значение снабжены дополнительным индексом, то предполагается, что они упорядочены по возрастанию.

2.1. Задача P_j^0

Решение задачи P_j^0 хорошо известно из классической электродинамики волноводов [20], [21]. А именно, все собственные значения (постоянные распространения) задачи P_j^0 и только они являются (однократными) корнями трансцендентного уравнения

$$\tan\sqrt{a_j-\lambda_j}h=\frac{\sqrt{a_j-\lambda_j}(\sqrt{\lambda_j-b_j}+\sqrt{\lambda_j-c_j})}{a_j-\lambda_j-\sqrt{\lambda_j-b_j}\sqrt{\lambda_j-c_j}}. \tag{2.1}$$

Уравнение (2.1) в электродинамике носит название дисперсионного уравнения [20], [21]; с точки зрения теории Штурма–Лиувилля это уравнение естественно называть характеристическим уравнением задачи P_j^0 [22].

Элементарное исследование уравнения (2.1) позволяет сформулировать

Утверждение 1. *Существует постоянная $h_0>0$ такая, что задача P_j^0 имеет конечное число* (*не менее одного*) *положительных и простых* (*кратности* 1) *собственных значений $\tilde\lambda_{j,k}$ и все $\tilde\lambda_{j,k}\in(b_j,a_j)$; при этом, если $b_j=c_j$, то $h_0=0$. Если $a_j\le b_j$, то задача P_j^0* (*положительных*) *решений не имеет.*

Доказательство опустим в силу его элементарности.

Используя в качестве "невозмущенной" линейную задачу P_j^0, методом возмущений можно доказать, что при достаточно малых α_j, α_{j+2} в некоторой окрестности всякого собственного значения $\tilde\lambda_{j,k}$ задачи P_j^0 будет лежать собственное значение $\hat\lambda_{j,k}$ задачи P_j [23], а в окрестности пары $(\tilde\lambda_{1,k},\tilde\lambda_{2,l})$ – собственное значение $\overline{\boldsymbol\lambda}_{k,l}$ задачи $P(\boldsymbol\alpha)$ [2]–[4]. Очевидно, что при $a_j\le b_j$ упомянутый метод возмущений неприменим для нахождения решений задач P_j и P, поскольку задача P_j^0 не имеет (положительных) решений.

2.2. Задача P_j

В дальнейшем нам понадобится

Утверждение 2. *Задача Коши* (1.5), (1.6) *глобально однозначно разрешима при $x\in\overline{\mathrm{I}}$, а ее* (*классическое*) *решение $v_j\equiv v_j(x;\lambda_j,\alpha_j)$ непрерывно зависит от точки $(x,\lambda_j,\alpha_j)\in\overline{\mathrm{I}}\times\Lambda_j\times\mathrm{A}$.*

В работе [9] получено уравнение относительно λ_j, которое имеет вид

$$\Phi_j(\lambda_j;n_j)\equiv\int\limits_{-\kappa_{c,j}}^{\kappa_{s,j}}\frac{ds}{w_j(s;\lambda_j)}+n_j\int\limits_{-\infty}^{+\infty}\frac{ds}{w_j(s;\lambda_j)}=h, \tag{2.2}$$

где

$$w_j(\mu_j;\lambda_j)\equiv\mu_j^2+a_j-\lambda_j+\alpha_j f_j(\theta_j(\mu_j;\lambda_j)),\quad n_j=0,1,\ldots,$$

величины μ_j и θ_j связаны соотношением

$$(\mu_j^2+a_j-\lambda_j)\theta_j+\alpha_j F_j(\theta_j^2)=C_j, \tag{2.3}$$

здесь $F_j(v_j^2)=\int_0^{v_j^2}f_j(t)dt$, а $C_j=(a_j-b_j)A_j^2+\alpha_j F_j(A_j^2)>0$ – постоянная. Уравнение (2.2) является семейством уравнений для различных n_j.

Имеет место следующая

Теорема 1 (об эквивалентности). *Число $\hat\lambda_j\in\Lambda_j$ является собственным значением задачи P_j если и только если существует целое число $\hat n_j\ge0$ такое, что $\lambda_j=\hat\lambda_j$ удовлетворяет уравнению* (2.2) *при $n_j=\hat n_j$; при этом собственная функция $v_j\equiv v_j(x;\hat\lambda_j\alpha_j)$, отвечающая собственному значению $\hat\lambda_j$, имеет $\hat n_j$ простых нулей $x_i\in\mathrm{I}$.*

Пусть $\Delta = (0, +\infty)$. Обозначим через R_Δ (открытую) окрестность множества Δ на комплексной плоскости $\mathbb{C}$, не содержащую точки $z = 0$.

Разрешимость задачи P_j устанавливает

Теорема 2. *Задача P_j имеет бесконечное число собственных значений $\hat\lambda_{j,k} \in \Lambda_j$ с точкой накопления на бесконечности. Кроме того, верно следующее*:

(i) *если задача P_j^0 имеет k' решений $\tilde\lambda_{j,k}$, $k = 1, \ldots, k'$, то существует постоянная $\alpha_j'' > 0$ такая, что для любого положительного $\alpha_j = \alpha_j' < \alpha_j''$ верно, что*

$$\hat\lambda_{j,k} \in (b_j, a_j) \quad \text{и} \quad \lim_{\alpha_j' \to +0} \hat\lambda_{j,k} = \tilde\lambda_{j,k}, \quad k = 1, \ldots, k',$$

где $\hat\lambda_{j,1} < \hat\lambda_{j,2} < \ldots < \hat\lambda_{j,k'}$ – первые k' решений задачи P_j при $\alpha_j = \alpha_j'$;

(ii) *для больших $\hat\lambda_j$ справедлива оценка* $\max_{x\in[0,h]} \left|u_j(x;\hat\lambda_j)\right| = O(\hat\gamma^{1/q_j})$.

Если дополнительно предположить, что функция $f_j(z)$ является аналитической функцией z при $z \in R_\Delta \subset \mathbb{C}$, то множество собственных значений задачи P_j является дискретным на Λ_j, т.е. на каждом отрезке $\Lambda_j' \subset \Lambda_j$ содержится не более конечного числа (изолированных) собственных значений.

Замечание 1. Если $k > k'$, то собственные значения $\hat\lambda_{j,k}$ не связаны с решениями (линейной) задачи P_j^0, в том числе при $\alpha_j \to +0$ [9].

Используя полученные результаты, можно показать, что справедливо

Утверждение 3. *Если функция $f_j(z)$ является аналитической функцией z при $z \in R_\Delta \subset \mathbb{C}$, то из последовательности $\{\hat\lambda_{j,k}\}$ собственных значений задачи P_j можно выделить такую (бесконечную) подпоследовательность $\{\hat\lambda_{j,k'}\}$ с точкой накопления на бесконечности, что для каждого элемента $\hat\lambda_{j,k'}$ указанной подпоследовательности найдется такое число $\delta_{j,k'} > 0$, что*

$$(\Phi_j(\hat\lambda_{j,k'} - \delta_{j,k'}; \hat n_j) - h) \cdot (\Phi_j(\hat\lambda_{j,k'} + \delta_{j,k'}; \hat n_j) - h) < 0 \tag{2.4}$$

и отрезок $[\hat\lambda_{j,k'} - \delta_{j,k'}, \hat\lambda_{j,k'} + \delta_{j,k'}]$ не содержит других элементов последовательности $\{\hat\lambda_{j,k}\}$.

Из утверждения 3 получаем

Следствие 1. Можно выбрать такую постоянную $\lambda_j^* > a_j$, что интервал (a_j, λ_j^*) содержит собственные значения $\hat\lambda_j$ задачи P_j, которые не связаны с решениями $\tilde\lambda_j$ задачи P_j^0 при $\alpha_j \to +0$.

В силу теоремы 2 для всякого достаточно малого $\alpha_j > 0$ все собственные значения $\hat\lambda_j \in (b_j, a_j)$ задачи P_j удовлетворяют свойству, сформулированному в утверждении 3. Однако если $\alpha_j > 0$ не мало, то интервал (a_j, λ_j^*) может содержать в том числе такие собственные значения $\hat\lambda_j$, что $\lim_{\alpha_j \to +0} \hat\lambda_j = \tilde\lambda_j$.

Свойство функции Φ_j, определяемое неравенством (2.4), имеет место без предположения об аналитичности функции $f_j(z)$. Однако в этом случае вопрос о том, являются ли собственные значения изолированными, остается открытым. Если рассмотреть задачу P_j с краевыми условиями I рода и функцией $f_j(z) = z$, то можно доказать, что отвечающая такой задаче функция Φ_j является монотонной при достаточно больших $\lambda_j > 0$ [8]. Аналогичный результат можно получить и для некоторых других функций f_j, например для $f_j(z) = P_k(z^2)$, где P_k – многочлен степени $k \geq 1$ с неотрицательными коэффициентами. Из указанного свойства монотонности следует, что все достаточно большие собственные значения являются изолированными.

Принимая во внимание теорему 1, уравнение (2.2) естественно называть *интегральным характеристическим уравнением* (ИХУ) задачи P_j.

Уравнения типа уравнения (2.2) являются мощным инструментом исследования нелинейных задач на собственные значения [6], [14], [24]. В частности, на основании исследования уравне-

ния (2.2) в работе [9] получены глубокие результаты о разрешимости задачи P_j и установлены многие свойства собственных значений и собственных функций.

В классической (линейной) теории Штурма–Лиувилля используется понятие характеристической функции [22]. В случае задачи P_j можно ввести характеристическую функцию аналогичным образом.

Имеет место следующее

Утверждение 4. *Число $\hat{\lambda}_j \in \Lambda_j$ является собственным значением задачи P_j, если и только если $\lambda_j = \hat{\lambda}_j$ удовлетворяет уравнению*

$$\varphi_j(\lambda_j, \alpha_j) = 0, \tag{2.5}$$

где $\varphi_j(\lambda_j, \alpha_j) := (\kappa_{c,j} v_j + v_j')\Big|_{x=h}$, а $v_j \equiv v_j(x; \lambda_j, \alpha_j)$ – решение задачи Коши (1.5), (1.6).

Естественно называть уравнение (2.5) *характеристическим уравнением, а его левую часть – характеристической функцией задачи P_j.*

Теоремы 1, 2 и утверждение 4 дают

Следствие 2. Всякое решение $\lambda_j = \hat{\lambda}_j \in \Lambda_j$ уравнения (2.2) является решением уравнения (2.5) и наоборот; при этом, если $\hat{\lambda}_j \in \Lambda_j$ такое, что для него выполняется условие (2.4), то для него также выполняется условие

$$\varphi_j(\hat{\lambda}_j - \delta_j, \alpha_j) \cdot \varphi_j(\hat{\lambda}_j + \delta_j, \alpha_j) < 0, \tag{2.6}$$

где функция $\varphi_j(\lambda_j, \alpha_j)$ определена в (2.5), а $\delta_j > 0$ – некоторая постоянная. В силу теоремы 2 для каждого собственного значения $\hat{\lambda}_j$ задачи P_j, удовлетворяющего свойству (2.4), постоянную δ_j можно выбрать так, что отрезок $[\hat{\lambda}_j - \delta_j, \hat{\lambda}_j + \delta_j]$ не содержит других собственных значений задачи P_j.

2.3. Задача $P(\boldsymbol{\alpha})$

Имеет место

Утверждение 5. *Существуют такие $\alpha_{j+2}^* > 0$, что при $0 < \alpha_{j+2} < \alpha_{j+2}^*$ задача Коши* (1.1), (1.2) *глобально однозначно разрешима при $x \in \overline{\mathrm{I}}$, а ее* (*классическое*) *решение*

$$u_1 \equiv u_1(x; \boldsymbol{\lambda}, \boldsymbol{\alpha}), \quad u_2 \equiv u_2(x; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \tag{2.7}$$

непрерывно зависит от точки $(x, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \in \overline{\mathrm{I}} \times \boldsymbol{\Lambda}^* \times \mathbf{A}^*$. Кроме того, при $\alpha_{j+2} \to +0$ функции u_j и u_j', определенные формулами (2.7) *равномерно при $x \in \overline{\mathrm{I}}$ стремятся к функциям v_j и v_j' соответственно, где $v_j \equiv v_j(x; \lambda_j, \alpha_j)$ – решение задачи Коши* (1.5), (1.6), *а $v_j' \equiv v_j'(x; \lambda_j, \alpha_j)$.*

Далее считаем, что $(x, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \in \overline{\mathrm{I}} \times \boldsymbol{\Lambda}^* \times \mathbf{A}^*$, где λ_j^* выбраны в соответствии со следствием 1, а α_{j+2}^* выбраны по λ_j^*, α_j^* в смысле утверждения 5. Подчеркнем, что числа λ_j^*, α_j^* не предполагаются малыми.

Ввиду того, что утверждение 5 дает существование глобально определенного непрерывного решения задачи Коши (1.1), (1.2), можно ввести понятие системы характеристических уравнений задачи $P(\boldsymbol{\alpha})$, аналогично тому, как это сделано в утверждении 4 для задачи P_j.

Имеем место следующее

Утверждение 6. *Величина $\overline{\boldsymbol{\lambda}} = (\overline{\lambda}_1, \overline{\lambda}_2)$ является* (*векторным*) *собственным значением задачи $P(\boldsymbol{\alpha})$, если и только если $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}^*$ удовлетворяет системе уравнений*

$$\begin{aligned} \psi_1(\boldsymbol{\lambda}, \boldsymbol{\alpha}) &= 0, \\ \psi_2(\boldsymbol{\lambda}, \boldsymbol{\alpha}) &= 0, \end{aligned} \tag{2.8}$$

где $\psi_j(\boldsymbol{\lambda}, \boldsymbol{\alpha}) := (\kappa_{c,j} u_j + u_j')\Big|_{x=h}$, а $u_j \equiv u_j(x; \boldsymbol{\lambda}, \boldsymbol{\alpha})$ – решение задачи Коши (1.1), (1.2), *определенное в утверждении* 5.

Рассмотрим систему уравнений (2.8). Вычитая из левых и правых частей уравнений (2.8) левые части уравнений (2.5), получаем следующую систему уравнений:

$$\begin{aligned}\psi_1(\boldsymbol{\lambda},\boldsymbol{\alpha}) - \varphi_1(\lambda_1,\alpha_1) &= -\varphi_1(\lambda_1,\alpha_1),\\ \psi_2(\boldsymbol{\lambda},\boldsymbol{\alpha}) - \varphi_2(\lambda_2,\alpha_2) &= -\varphi_2(\lambda_2,\alpha_2),\end{aligned} \tag{2.9}$$

где все входящие в формулу (2.9) выражения определены формулами (2.5) и (2.8). Доказательство существования векторных собственных значений $\overline{\boldsymbol{\lambda}}$ задачи $P(\boldsymbol{\alpha})$ основано на изучении системы (2.9) при $\alpha_{j+2} \to +0$.

Левые части уравнений (2.9) зависят от параметров α_j, α_{j+2}, в то время как правые части только от α_j. Кроме того, при предельном переходе $\alpha_{j+2} \to +0$ система уравнений (2.9) распадается на два независимых уравнения (2.5). Принимая во внимание следствие 2 о существовании собственных значений задачи P_j, при переходе через которые характеристическая функция φ_j задачи P_j меняет знак, получаем основной результат настоящей работы.

Теорема 3. *Пусть каждая из задач P_j имеет n_j' собственных значений*

$$\hat{\lambda}_{j,1},\ldots,\hat{\lambda}_{j,n_j'} \in (b_j,\lambda_j^*) \subset \Lambda_j$$

соответственно; при этом постоянные λ_j^* выбраны так, что множество (a_j, λ_j^*) содержит собственные значения $\hat{\lambda}_j$ задачи P_j, удовлетворяющие свойству, сформулированному в следствии 2.

Тогда найдутся такие постоянные $\alpha_{j+2}^0 > 0$, что для любых $\alpha_{j+2} \in (0, \alpha_{j+2}^0)$ задача $P(\boldsymbol{\alpha})$ имеет, по крайней мере, $n_1' \times n_2'$ векторных собственных значений $\overline{\boldsymbol{\lambda}}_{i_1,i_2} = (\overline{\lambda}_{1,i_1}, \overline{\lambda}_{2,i_2})$, где $i_j = \overline{1, n_2'}$; при этом каждое $\overline{\boldsymbol{\lambda}}_{i_1,i_2}$ содержится в некоторой окрестности точки $(\overline{\lambda}_{1,i_1}, \overline{\lambda}_{2,i_2})$.

Замечание 2. Числа λ_j^*, α_j^* и α_{j+2}^* введены при постановке задачи $P(\boldsymbol{\alpha})$. При этом числа λ_j^* выбираются достаточно большими, чтобы в интервал $(\alpha_j^*, \lambda_j^*)$ попали собственные значения $\hat{\lambda}_{j,i}$ задачи P_j (см. следствие 1), а числа α_j^* (>0) могут быть выбраны без каких-либо дополнительных условий. После того как зафиксированы параметры λ_j^* и α_j^*, в соответствии с утверждением 5 фиксируются (достаточно малые) параметры α_{j+2}^*.

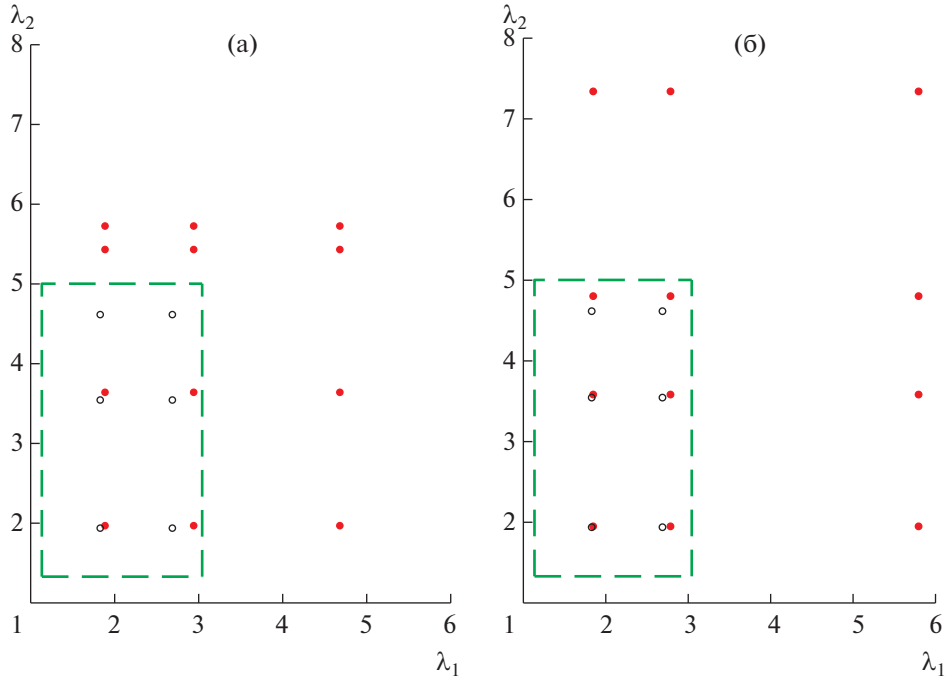
3. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

В вычислениях использованы следующие значения параметров: $n = 2$, $\varepsilon_{l,1} = 3$, $\varepsilon_{l,2} = 5$, $\varepsilon_{c,1} = \varepsilon_{s,1} = 1$, $\varepsilon_{c,2} = \varepsilon_{s,2} = 1.2$, $A_1 = A_2 = 1$, $h = 4$. Параметры α_j и α_{j+2} выбраны различными в различных экспериментах.

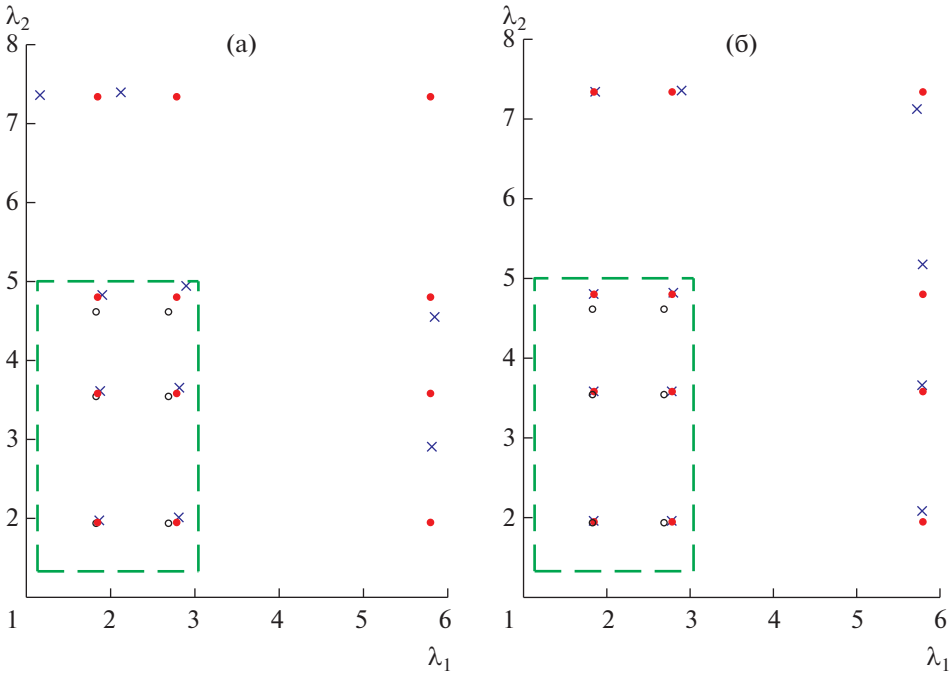
Несмотря на то что собственные значения задач P_1^0, P_2^0 и P_1, P_2 являются скалярными величинами, мы рассматриваем их как двумерные наборы $\tilde{\boldsymbol{\lambda}}_{k,k'} = (\tilde{\lambda}_{i,k}, \tilde{\lambda}_{j,k'})$ и $\hat{\boldsymbol{\lambda}}_{k,k'} = (\hat{\lambda}_{i,k}, \hat{\lambda}_{j,k'})$, где $i, j = \{1, 2\}$ и k, k' – неотрицательные целые индексы. Объединяя эти собственные значения в пары, мы имеем возможность построить их как точки на плоскости $O\lambda_1\lambda_2$. Мы не перегружаем графики излишними обозначениями: все необходимые пояснения даны в тексте и подрисуночных подписях.

На каждом графике пары $\tilde{\boldsymbol{\lambda}}_{k,k'} = (\tilde{\lambda}_{i,k}, \tilde{\lambda}_{j,k'})$, которые составлены из решений задач P_1^0, P_2^0, существуют только внутри прямоугольников на плоскости $O\lambda_1\lambda_2$, границы которых отмечены штриховой линией.

На фиг. 1, 2 пары $\tilde{\boldsymbol{\lambda}}_{k,k'} = (\tilde{\lambda}_{i,k}, \tilde{\lambda}_{j,k'})$ решений задач P_1^0, P_2^0 обозначены окружностями; пары $\hat{\boldsymbol{\lambda}}_{k,k'} = (\hat{\lambda}_{i,k}, \hat{\lambda}_{j,k'})$ решений задач P_1, P_2 – красными кружками. На фиг. 2 векторные собственные значения $\overline{\boldsymbol{\lambda}}_{k,k'} = (\overline{\lambda}_{i,k}, \overline{\lambda}_{j,k'})$ задачи $P(\boldsymbol{\alpha})$ обозначены синими крестиками.



Фиг. 1. Выбраны следующие значения параметров: для (а) – $\alpha_1 = 0.045$, $\alpha_2 = 0.055$, и $\alpha_3 = \alpha_{21} = 0$; здесь $\hat{\lambda}_{1,0} \approx 1.798$, $\hat{\lambda}_{1,1} \approx 2.904$, $\hat{\lambda}_{1,2} \approx 4.727$ и $\hat{\lambda}_{2,0} \approx 1.851$, $\hat{\lambda}_{2,1} \approx 3.581$, $\hat{\lambda}_{2,2} \approx 5.433$, $\hat{\lambda}_{2,3} \approx 5.738$; для (б) – $\alpha_1 = 0.019$, $\alpha_2 = 0.022$, и $\alpha_3 = \alpha_4 = 0$, где $\hat{\lambda}_{1,0} \approx 1.774$, $\hat{\lambda}_{1,1} \approx 2.760$, $\hat{\lambda}_{1,2} \approx 5.902$ и $\hat{\lambda}_{2,0} \approx 1.843$, $\hat{\lambda}_{2,1} \approx 3.527$, $\hat{\lambda}_{2,2} \approx 4.797$, $\hat{\lambda}_{2,3} \approx 7.407$.



Фиг. 2. Выбраны следующие значения параметров: для (а) – $\alpha_1 = 0.019$, $\alpha_2 = 0.022$, $\alpha_3 = 0.01$, $\alpha_4 = 0.02$, где $\bar{\boldsymbol{\lambda}}_{00} \approx (1.778, 1.857)$, $\bar{\boldsymbol{\lambda}}_{01} \approx (1.789, 3.552)$, $\bar{\boldsymbol{\lambda}}_{02} \approx (1.815, 4.810)$, $\bar{\boldsymbol{\lambda}}_{03} \approx (1.041, 7.429)$, $\bar{\boldsymbol{\lambda}}_{10} \approx (2.766, 1.896)$, $\bar{\boldsymbol{\lambda}}_{11} \approx (2.773, 3.595)$, $\bar{\boldsymbol{\lambda}}_{3} \approx (2.858, 4.929)$, $\bar{\boldsymbol{\lambda}}_{13} \approx (2.046, 7.465)$, $\bar{\boldsymbol{\lambda}}_{20} \approx (5.912, 2.822)$, $\bar{\boldsymbol{\lambda}}_{21} \approx (5.947, 4.522)$; для (б) – $\alpha_1 = 0.019$, $\alpha_2 = 0.022$, $\alpha_3 = 0.001$, $\alpha_4 = 0.002$, где $\bar{\boldsymbol{\lambda}}_{00} \approx (1.774, 1.843)$, $\bar{\boldsymbol{\lambda}}_{01} \approx (1.775, 3.530)$, $\bar{\boldsymbol{\lambda}}_{02} \approx (1.777, 4.799)$, $\bar{\boldsymbol{\lambda}}_{03} \approx (1.797, 7.409)$, $\bar{\boldsymbol{\lambda}}_{10} \approx (2.761, 1.848)$, $\bar{\boldsymbol{\lambda}}_{11} \approx (2.761, 3.534)$, $\bar{\boldsymbol{\lambda}}_{12} \approx (2.768, 4.809)$, $\bar{\boldsymbol{\lambda}}_{20} \approx (5.901, 1.977)$, $\bar{\boldsymbol{\lambda}}_{21} \approx (5.903, 3.602)$, $\bar{\boldsymbol{\lambda}}_{22} \approx (5.908, 5.176)$, $\bar{\boldsymbol{\lambda}}_{23} \approx (5.858, 7.190)$.

Для выбранных значений параметров существует 2 и 3 собственных значения задач P_1^0 и P_2^0 соответственно. Отсюда следует, что можно составить 6 пар, а именно $\tilde{\boldsymbol{\lambda}}_{k,k'} = (\tilde{\lambda}_{i,k}, \tilde{\lambda}_{j,k'})$, где $\tilde{\lambda}_{1,0} \approx 1.754$, $\tilde{\lambda}_{1,1} \approx 2.669$ и $\tilde{\lambda}_{2,0} \approx 1.828$, $\tilde{\lambda}_{2,1} \approx 3.488$, $\tilde{\lambda}_{2,2} \approx 4.612$.

На фиг. 1 пары $\hat{\boldsymbol{\lambda}}_{k,k'}$ собственных значений задач P_1, P_2 построены вместе с парами $\tilde{\boldsymbol{\lambda}}_{k,k'}$ собственных значений задач $P_{0,1}$, $P_{0,2}$ для того чтобы можно было сравнить их. Каждая из задач P_1, P_2 имеет бесконечное число положительных собственных значений; на фигурах представлены только по несколько первых собственных значений в каждом случае.

Фигуры 1а, 1б иллюстрируют утверждение (i) теоремы 2. Действительно, уменьшая α_1 и α_2, можно видеть, что существует, по крайней мере, одна пара $\hat{\boldsymbol{\lambda}}_{k,k'} = (\hat{\lambda}_{i,k}, \hat{\lambda}_{j,k'})$ в окрестности каждой пары $\tilde{\boldsymbol{\lambda}}_{k,k'} = (\tilde{\lambda}_{i,k}, \tilde{\lambda}_{j,k'})$ и, следовательно, существует, по крайней мере, одно собственное значение $\hat{\lambda}_{i,k}$ в окрестности всякого собственного значения $\tilde{\lambda}_{i,k}$ как только $\alpha_j > 0$ достаточно малы.

Фигуры 2а, 2б частично иллюстрируют теорему 3. Действительно, как видно существует по крайней мере одна пара $\hat{\boldsymbol{\lambda}}_{k,k'} = (\hat{\lambda}_{i,k}, \hat{\lambda}_{j,k'})$ в окрестности каждой пары $\tilde{\boldsymbol{\lambda}}_{k,k'} = (\tilde{\lambda}_{i,k}, \tilde{\lambda}_{j,k'})$ как только $\alpha_3 > 0$ и $\alpha_4 > 0$ достаточно малы.

4. ОБСУЖДЕНИЕ И ЗАКЛЮЧЕНИЕ

В связи с теоремой 3 уместно дать два комментария. Во-первых, поскольку числа λ_j^* и Λ_j^* заданы произвольно, а значит, могут быть выбраны достаточно большими, то теорема 3 утверждает существование векторных собственных значений задачи $P(\boldsymbol{\alpha})$ в том числе в области, в которой отсутствуют собственные значения задач P_j^0. Во-вторых, теорема 3 также дает существование тех собственных значений $\overline{\boldsymbol{\lambda}}_{\mathbf{k,l}} = (\overline{\lambda}_{1,k}, \overline{\lambda}_{2,l})$, которые являются возмущениями пар решений $(\overline{\lambda}_{1,k}, \overline{\lambda}_{2,l})$ линейных задач P_j^0. Результат, аналогичный последнему, в некоторых нелинейных задачах ранее был получен с помощью интегральных уравнений, см. [2]–[4] и библиографию там.

Теорема 3 утверждает существование лишь конечного числа векторных собственных значений задачи $P(\boldsymbol{\alpha})$. Учитывая, что каждая из задач P_j имеет бесконечное число собственных значений (при любом $\alpha_j > 0$), можно предположить, что задача $P(\boldsymbol{\alpha})$ также имеет бесконечное число векторных собственных значений. Доказательство этого результата хотя бы для малых α_{j+2}^* явилось бы следующим существенным продвижением. Первое и очевидное препятствие к получению такого результата – отсутствие однозначной глобальной разрешимости задачи Коши (1.1), (1.2) при $x \in \overline{\mathrm{I}}$ для всех $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ таких, что $(\boldsymbol{\lambda}, \boldsymbol{\alpha}) \in \boldsymbol{\Lambda} \times \mathbf{A}^*$, где $\boldsymbol{\Lambda} = \Lambda_1 \times \Lambda_2$. Следующим препятствием является неограниченный рост максимумов собственных функций v_j при $\hat{\lambda}_j \to +\infty$, см. утверждение (ii) теоремы 2. Вероятно, что последнее затруднение можно преодолеть, используя подходящую "нормировку" собственных функций, как это сделано при доказательстве утверждения 2.1 в работе [14].

Также обратим внимание читателя, что техника, развитая в настоящей статье, существенно отличается от техники, предложенной в работах [7], [8]. В цитированных работах для многопараметрической задачи на собственные значения получена система ИХУ, являющихся многомерным обобщением уравнения (2.2). Исследование такой системы уравнений позволяет получить результаты, аналогичные представленным в настоящей работе. Важным, однако, является то, что здесь результаты получены при помощи простых вспомогательных средств (утверждение 5 и 6), а выкладки технически более просты, чем выкладки в работах [7], [8]. В то же время необходимо отметить, что техника, развитая в цитированных работах, позволяет получить некоторые глубокие результаты о собственных значениях в "неэлектродинамическом" случае задачи $P(\boldsymbol{\alpha})$. А именно, если поставить задачу $P(\boldsymbol{\alpha})$ в экранированном волноводе (в этом случае условия III рода (1.2), (1.3) надо заменить на условия I рода [8]), то в такой задаче существует бесконечное число векторных собственных значений с отрицательными компонентами, такие собственные значения не имеют электродинамического смысла, но могут представлять интерес с точки зрения теории задач на собственные значения. В указанном случае метод ИХУ позволяет не только доказать существование бесконечного числа векторных собственных значений, но и выяснить их асимптотику и другие свойства [8].

5. ДОКАЗАТЕЛЬСТВА

Доказательство утверждения 2. Это утверждение доказано в работе [9] (см. утверждение 2 в [9]). Двукратная непрерывная дифференцируемость решения v_j по переменной x при $x \in \overline{\mathrm{I}}$ следует из гладкости правой части уравнения (1.5).

Доказательство теоремы 1. Доказательство этой теоремы следует из доказательства теоремы 1 и следствия 1, представленных в работе [9].

Доказательство теоремы 2. Доказательство этой теоремы следует из доказательств теорем 4, 5, 6, представленных в работе [9].

Доказательство утверждения 3. Легко показать, что функция w_j, определенная в (2.2), является положительной. Действительно, приравняв w_j к нулю, выразив из этого соотношения $\mu_j^2 + a_j - \lambda_j$ и подставив результат в (2.3), получим противоречие. Теперь достаточно выяснить знак w_j в какой-либо одной точке. Легко видеть, что для любого $\lambda_j < a_j$ функция w_j положительна, а значит, она (если существует) положительна для всех $\lambda_j \in \Lambda_j$. Существование функции w_j для всех $\lambda_j \in \Lambda_j$ элементарно следует из анализа формулы (2.3).

Непрерывность функции Φ_j относительно $\lambda_j \in \Lambda_j$ элементарно следует из анализа формул (2.2) и (2.3).

Принимая во внимание положительность функции w_j, получаем следующие неравенства:

$$n_j T_j(\lambda_j) < \Phi_j(\lambda_j; n_j) < (n_j + 1) T_j(\lambda_j), \tag{5.1}$$

где $n_j = 0,1,\ldots$, $\lambda_j \in \Lambda_j$, а $T_j(\lambda_j) = \int_{-\infty}^{+\infty} \dfrac{ds}{w_j(s;\lambda_j)}$.

В работе [9] доказано, что $\lim_{\lambda_j \to +\infty} T_j(\lambda_j) = 0$. Отсюда следует существование бесконечного числа решений уравнения (2.2). Действительно, для любого $h > 0$ можно подобрать такой номер $n_j = n_j^0 \geq 0$, что $h < n_j^0 \max_{\lambda_j \in \Lambda_j} T_j(\lambda_j) < \max_{\lambda_j \in \Lambda_j} \Phi_j(\lambda_j; n_j^0)$. Поскольку $\lim_{\lambda_j \to +\infty} T_j(\lambda_j) = 0$, то всегда можно выбрать такое значение $\lambda_j \geq \lambda_j^0$, что $\Phi_j(\lambda_j^0; n_j^0) < (n_j^0 + 1) T_j(\lambda_j^0) < h$. Но тогда между значением λ_j, на котором функция $\Phi_j(\lambda_j; n_j^0)$ достигает максимального значения, и значением λ_j^0 найдется такое $\lambda_j = \hat{\lambda}_j$, что $\Phi_j(\hat{\lambda}_j; n_j^0) = h$.

Далее, поскольку $\lim_{\lambda_j \to +\infty} T_j(\lambda_j) = 0$, то среди собственных значений $\hat{\lambda}_{j,k}$ встречается бесконечное множество таких, при переходе через которые функция $\Phi_j(\lambda_j; n_j) - h$, см. формулу (2.2), меняет знак. Таким образом, мы доказали, что существует бесконечное число собственных значений задачи P_j, для которых выполняется неравенство (2.4). Так как $f_j(z)$ является аналитической функцией z при $z \in R_\Delta \subset \mathbb{C}$, то в силу следствий теоремы 4 работы [9] все указанные собственные значения изолированные, а значит, каждое из них можно заключить внутрь некоторой окрестности, в замыкание которой не попадает других собственных значений задачи P_j.

Доказательство утверждения 4. Пусть $v_j \equiv v_j(x; \hat{\lambda}_j, \alpha_j)$ – решение задачи Коши (1.5), (1.6). Если $\lambda_j = \hat{\lambda}_j$ удовлетворяет уравнению (2.5), то, очевидно, $\hat{\lambda}_j$ является собственным значением, а $v_j(x; \hat{\lambda}_j, \alpha_j)$ – собственной функцией задачи P_j.

Пусть $\lambda_j = \hat{\lambda}_j^* \in \Lambda_j$ – некоторое решение уравнения (2.5). Рассмотрим задачу Коши (1.5), (1.6), где $\lambda_j = \hat{\lambda}_j^*$. В силу утверждения 2 нетривиальное решение $v_j \equiv v_j(x; \hat{\lambda}_j^*, \alpha_j)$ указанной задачи Коши существует, единственно и непрерывно зависит от точки $(x, \lambda_j, \alpha) \in \overline{\mathrm{I}} \times \Lambda_j \times \mathrm{A}$. Если уравнение (2.5) выполняется при $\lambda_j = \hat{\lambda}_j^*$ для указанного решения задачи Коши, то $\lambda_j = \hat{\lambda}_j^*$ является собственным значением задачи P_j. Предположим, что для указанного решения задачи Коши уравнение (2.5) не выполняется при $\lambda_j = \hat{\lambda}_j^*$. Но отсюда следует, что существует неединственное решение задачи Коши (1.5), (1.6), где $\lambda_j = \hat{\lambda}_j^*$; для одного из них уравнение (2.5) выполняется, а для другого – нет. Такой вывод противоречит утверждению 2. А значит, предположение о существовании решения задачи Коши (1.5), (1.6), где $\lambda_j = \hat{\lambda}_j^*$, для которого не выполняется уравнение (2.5), неверно.

Доказательство утверждения 5. Указанное утверждение является следствием “интегральной” теоремы непрерывности [25]. Следуя [25], рассмотрим нормальную систему уравнений

$$u_i' = f_i(x, u_1, \ldots, u_n, \gamma_1, \ldots, \gamma_l), \quad i = \overline{1,n}, \tag{5.2}$$

правые части которой зависят от параметров $\gamma_1, \ldots, \gamma_l$. В векторной форме систему (5.2) запишем в виде

$$\mathbf{u}' = \mathbf{f}(x, \mathbf{u}, \boldsymbol{\gamma}). \tag{5.3}$$

Будем предполагать, что правые части системы (5.2) определены и непрерывны вместе с их частными производными $\dfrac{\partial}{\partial u_k} f_i(x, \mathbf{u}, \boldsymbol{\gamma})$ в некоторой области $\widetilde{\Gamma}$ пространства $\widetilde{R}$ переменных $x, u_1, \ldots, u_n, \gamma_1, \ldots, \gamma_l$.

Справедлива

Теорема 4 (см. [25]). *Пусть* $(x_0, \mathbf{u}_0, \boldsymbol{\gamma}_0)$ *– некоторая точка области* $\widetilde{\Gamma}$ *и* $\mathbf{u} = \boldsymbol{\varphi}(x, \boldsymbol{\gamma})$ *– решение уравнения* (5.3), *удовлетворяющее начальному условию* $\boldsymbol{\varphi}(x_0, \boldsymbol{\gamma}) = \mathbf{u}_0$. *Если решение* $\mathbf{u} = \boldsymbol{\varphi}(x, \boldsymbol{\gamma}_0)$ *определено при* $x \in \overline{\mathrm{I}}$, *то существует такое число* $\gamma_0 > 0$, *что при* $|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0| < \gamma_0$ *решение* $\mathbf{u} = \boldsymbol{\varphi}(x, \boldsymbol{\gamma})$ *определено на том же отрезке* $\overline{\mathrm{I}}$, *а функция* $\boldsymbol{\varphi}(x, \boldsymbol{\gamma})$ *непрерывна по совокупности переменных* x, $\boldsymbol{\gamma}$ *при* $x \in \overline{\mathrm{I}}$ *и* $|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0| < \gamma_0$. *Кроме того,* $\boldsymbol{\varphi}(x, \boldsymbol{\gamma}) \to \boldsymbol{\varphi}(x, \boldsymbol{\gamma}^*)$ *равномерно по* $x \in \overline{\mathrm{I}}$ *при* $\boldsymbol{\gamma} \to \boldsymbol{\gamma}^*$ *как только* $|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*| < \gamma_0$.

Необходимо отметить, что эта теорема сформулирована в [25] без дополнительного утверждения о равномерности стремления $\boldsymbol{\varphi}(x, \boldsymbol{\gamma})$ к $\boldsymbol{\varphi}(x, \boldsymbol{\gamma}^*)$. Однако этот факт следует из доказательства, данного в [25].

Теперь утверждение 5 является прямым следствием применения указанной выше теоремы к задаче Коши (1.1), (1.2) при дополнительном предположении о глобальной однозначной разрешимости задач Коши (1.5), (1.6) при $(x, a_j, \alpha_j) \in \overline{\mathrm{I}} \times \Lambda_j \times \mathrm{A}$. Справедливость этого дополнительного предположения гарантируется утверждением 2.

Двукратная непрерывная дифференцируемость решений u_1, u_2 по x при $x \in \overline{\mathrm{I}}$ следует из гладкости правых частей уравнений (1.1).

Доказательство утверждения 6. Пусть $u_j \equiv u_j(x; \overline{\boldsymbol{\lambda}}, \boldsymbol{\alpha})$ – решение задачи Коши (1.1), (1.2). Если $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$ удовлетворяет системе уравнений (2.8), то, очевидно, $\overline{\boldsymbol{\lambda}}$ является собственным значением, а $u_j \equiv u_j(x; \overline{\boldsymbol{\lambda}}, \boldsymbol{\alpha})$ – собственной функцией задачи $P(\boldsymbol{\alpha})$.

Пусть $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^* \in \boldsymbol{\Lambda}^*$ – некоторое решение системы уравнений (2.8). Рассмотрим задачу Коши (1.1), (1.2), где $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^*$. В силу теоремы 5 нетривиальное решение $u_j \equiv u_j(x; \overline{\boldsymbol{\lambda}}^*, \boldsymbol{\alpha})$ указанной задачи Коши существует, единственно и непрерывно зависит от точки $(x, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \in \overline{\mathrm{I}} \times \boldsymbol{\Lambda}^* \times \mathbf{A}^*$. Если уравнения (2.8) выполняются при $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^*$ для указанного решения задачи Коши, то $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^*$ является собственным значением задачи $P(\boldsymbol{\alpha})$. Предположим, что для указанного решения задачи Коши уравнение (2.8) не выполняется при $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^*$. Но отсюда следует, что существует неединственное решение задачи Коши (1.1), (1.2), где $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^*$: для одного из них уравнения (2.8) выполняются, а для другого – нет. Такой вывод противоречит утверждению 5. А значит, предположение о существовании решения задачи Коши (1.1), (1.2), где $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}^*$, для которого не выполняется уравнение (2.8), неверно.

Доказательство теоремы 3. Искомый результат получается из таких рассуждений. Из утверждения 5 следует, что если $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}^*$ и $\boldsymbol{\alpha} \to \boldsymbol{\alpha}'$, то

$$u_1(x; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \to v_1(x; \lambda_1, \alpha_1), \quad u_2(x; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \to v_2(x; \lambda_2, \alpha_2), \tag{5.4}$$

$$u_1'(x; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \to v_1'(x; \lambda_1, \alpha_1), \quad u_2'(x; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \to v_2'(x; \lambda_2, \alpha_2) \tag{5.5}$$

равномерно при $x \in \overline{\mathrm{I}}$; $u_j(x; \lambda_j, \alpha_j)$ – решение задачи Коши (1.1), (1.2), $v_j(x; \lambda_j, \alpha_j)$ – решение задачи Коши (1.5), (1.6).

Учитывая формулы (5.4), (5.5), ясно, что если $\boldsymbol{\alpha} \to \boldsymbol{\alpha}'$, то

$$\psi_1(\boldsymbol{\lambda}, \boldsymbol{\alpha}) \to \varphi_1(\lambda_1, \alpha_1), \tag{5.6}$$

$$\psi_2(\boldsymbol{\lambda}, \boldsymbol{\alpha}) \to \varphi_2(\lambda_2, \alpha_2) \tag{5.7}$$

равномерно при $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}^*$, где φ_j и ψ_j определены формулами (2.5) и (2.8) соответственно.

В силу формул (5.6), (5.7) получаем, что для любого $\epsilon > 0$ найдется такое $\epsilon' > 0$, что для всех $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}^*$ левые части формул (2.9) по абсолютному значению будут меньше ϵ как только $|\boldsymbol{\alpha} \to \boldsymbol{\alpha}'| < \epsilon'$.

Нули правых частей формул (2.9) являются собственными значениями задач P_j. Из следствия 1 известно, что существует такая постоянная λ_j^* $(>a_j)$, что интервал (b_j, λ_j^*) содержит собственные значение задачи P_j, удовлетворяющие следствию 2. При этом предполагается, что λ_j^* достаточно велико и, таким образом, интервал (b_j, λ_j^*) содержит собственные значения задачи P_j, которые не переходят в соответствующие собственные значения задачи P_j^0 при $\alpha_{j+2} \to +0$. Отсюда следует, что для всякого собственного значения $\hat{\lambda}_{j,k}$, удовлетворяющего следствию 2, найдется содержащий его отрезок такой, что правая часть формулы в (2.9), отвечающая $\hat{\lambda}_{j,k}$, принимает значения разных знаков на противоположных концах этого отрезка. Поскольку левые части могут быть сделаны как угодно малыми, а правые части не зависят от α_{j+2}, непрерывны по λ_j и меняют знак при переходе через $\hat{\lambda}_{j,k}$, то при достаточно малых α_{j+2} в указанных отрезках найдутся числа $\overline{\lambda}_{1,i_1}$ и $\overline{\lambda}_{2,i_2}$, удовлетворяющие системе (2.9).

6. ФИЗИЧЕСКАЯ ФОРМУЛИРОВКА ЗАДАЧИ О РАСПРОСТРАНЕНИИ ВОЛН

Задача $P(\boldsymbol{\alpha})$ описывает распространение двухчастотной электромагнитной ТЕ-ТЕ-волны в плоском экранированном немагнитном диэлектрическом волноводе, заполненном нелинейной средой. Нелинейный отклик среды отвечает эффектам самовоздействия для сред с центром инверсии [26]–[29].

Рассмотрим эту задачу подробнее. Пусть $\Sigma = \{(x, y, z) \in \mathbb{R}^3 : 0 \le x \le h, (y, z) \in \mathbb{R}^2\}$ – диэлектрический слой, расположенный между полупространствами $x < 0$ и $x > h$ в декартовой системе координат $Oxyz$. Полупространства заполнены немагнитными средами, характеризующимися вещественными диэлектрическими проницаемостями $\epsilon = \epsilon_s \ge \epsilon_0 > 0$ и $\epsilon = \epsilon_c \ge \epsilon_0 > 0$ соответственно, где ϵ_0 – диэлектрическая проницаемость вакуума. Диэлектрическая проницаемость ϵ_l слоя Σ будет введена ниже; магнитная проницаемость μ во всем пространстве есть положительная постоянная.

В соответствии с [2], [4], введем двухчастотное электромагнитное поле

$$\mathbf{E}_\omega = \mathbf{E}_1 e^{-i\omega_1 t} + \mathbf{E}_2 e^{-i\omega_2 t}, \quad \mathbf{H}_\omega = \mathbf{H}_1 e^{-i\omega_1 t} + \mathbf{H}_2 e^{-i\omega_2 t}, \tag{6.1}$$

где

$$\begin{aligned} \mathbf{E}_1 &= (0, e_y, 0)^\top e^{i\gamma_1 z}, \quad \mathbf{H}_1 = (h_x^{(1)}, 0, h_z)^\top e^{i\gamma_1 z}, \\ \mathbf{E}_2 &= (0, 0, e_z)^\top e^{i\gamma_2 y}, \quad \mathbf{H}_2 = (h_x^{(2)}, h_y, 0)^\top e^{i\gamma_2 y}; \end{aligned} \tag{6.2}$$

здесь компоненты e_y, e_z, $h_x^{(1)}$, $h_x^{(2)}$, h_y, h_z зависят только от одной пространственной координаты x, а γ_j – подлежащие определению вещественные постоянные. Частоты ω_1, ω_2 различны, но их выбор подчинен некоторым ограничениям, связанным с выбранным законом нелинейности для ϵ_l [4], [5], [26]. Величины $\mathbf{E}_j = \mathbf{E}_j^+ + i\mathbf{E}_j^-$, $\mathbf{H}_j = \mathbf{H}_j^+ + i\mathbf{H}_j^-$ – называются комплексными амплитудами [30]. Другими словами, мы рассматриваем сумму двух ТЕ-волн, распространяющихся в направлениях Oz и Oy соответственно. Поле (6.1), (6.2) носит название ТЕ-ТЕ-волны и представляет собой частный случай многочастотной ТЕ-волны [4]. Вещественное (физическое) поле $\tilde{\mathbf{E}}_\omega$, $\tilde{\mathbf{H}}_\omega$ имеет вид $\operatorname{Re}\mathbf{E}_\omega$, $\operatorname{Re}\mathbf{H}_\omega$ соответственно.

Считаем, что диэлектрическая проницаемость ϵ_l слоя Σ описывается диагональным (3×3)-тензором

$$\epsilon_l(\tilde{\mathbf{E}}_\omega) \equiv \begin{pmatrix} * & 0 & 0 \\ 0 & \epsilon_{yy} & 0 \\ 0 & 0 & \epsilon_{zz} \end{pmatrix}, \tag{6.3}$$

где $\epsilon_{yy} = \epsilon_y + \beta_1 f_1(s_1) + \beta_3 g_1(s_1, s_2)$, $\epsilon_{zz} = \epsilon_z + \beta_2 f_2(s_2) + \beta_4 g_2(s_1, s_2)$, элемент * тензора ϵ_l не оказывает влияния на распространение TE-TE-волны; ϵ_y, ϵ_z, β_j, $\beta_{j+2} > 0$ – вещественные постоянные; $s_1 = |(\mathbf{E}_1, \mathbf{e}_y)|^2$, $s_2 = |(\mathbf{E}_2, \mathbf{e}_z)|^2$, $\mathbf{E}_1$, $\mathbf{E}_2$ – векторы электрических полей, образующих TE-TE-волну; $\mathbf{e}_y$, $\mathbf{e}_z$ – единичные орты осей Oy, Oz декартовых координат $Oxyz$; $(\cdot,\cdot)$ – евклидово скалярное произведение.

Предполагается, что выполняются неравенства

$$\min\{\epsilon_y, \epsilon_z\} > \epsilon_s \geq \epsilon_c > 0.$$

Неравенство $\min\{\epsilon_y, \epsilon_z\} > \max\{\epsilon_s, \epsilon_c\}$ является необходимым условием для существования распространяющихся TE-волн в слое, заполненном линейной средой, т.е. при $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

Диэлектрическая проницаемость вида (6.3) соответствует некоторым важным прикладным случаям [5], [24], [30]–[36]; кроме того, свойства функций f_j и g_j позволяют изучать широкий класс нелинейностей, возникающих в оптике; керровская, полиномиальная, степенная нелинейности и т.д.

Теоретически существование двухчастотных распространяющихся волн было доказано недавно: (связанные) TE-TM- и TE-TE-волны в слое с керровской нелинейностью введены и изучены в [3] и [2] соответственно. Задача о TE-TE-волне, распространяющейся на одной частоте в слое с керровской нелинейностью, была рассмотрена впервые в [31]; позже эта задача также привлекала внимание исследователей [5], [33], [34]. Случай многочастотной распространяющейся волны в нелинейных фотонных кристаллах указан в работе [35]. Общие формулировки задач о распространении многочастотных волн различных конфигураций в плоских и круглых цилиндрических волноводах, заполненных нелинейными средами, впервые представлены в [4].

Подставляя (6.1) в уравнении Максвелла и принимая во внимание линейность оператора rot, получаем, что $\mathbf{E}_k$, $\mathbf{H}_k$ удовлетворяют следующим (связанным) уравнениям:

$$\begin{aligned} e^{-i\omega_1 t}\,\mathrm{rot}\,\mathbf{H}_1 + e^{-i\omega_2 t}\,\mathrm{rot}\,\mathbf{H}_2 &= -i\epsilon(\omega_1\mathbf{E}_1 e^{-i\omega_1 t} + \omega_2\mathbf{E}_2 e^{-i\omega_2 t}), \\ e^{-i\omega_1 t}\,\mathrm{rot}\,\mathbf{E}_1 + e^{-i\omega_2 t}\,\mathrm{rot}\,\mathbf{E}_2 &= i\mu(\omega_1\mathbf{H}_1 e^{-i\omega_1 t} + \omega_2\mathbf{H}_2 e^{-i\omega_2 t}), \end{aligned} \tag{6.4}$$

где i – мнимая единица. Поскольку полученная система справедлива для всех t, то приходим к системе 4 (векторных) уравнений

$$\begin{aligned} \mathrm{rot}\,\mathbf{H}_j &= -i\epsilon\omega_j\mathbf{E}_j, \\ \mathrm{rot}\,\mathbf{E}_j &= i\mu\omega_j\mathbf{H}_j, \end{aligned} \tag{6.5}$$

где $j = 1, 2$.

Итак, поля $\mathbf{E}_j$, $\mathbf{H}_j$ удовлетворяют уравнениям (6.5). Поскольку мы ищем распространяющиеся волны, то искомые решения затухают как $O(|x|^{-1})$ при $|x| \to \infty$. Кроме этого, классическая теория электромагнитного поля утверждает, что касательные компоненты полей $\mathbf{E}_j$, $\mathbf{H}_j$ являются непрерывными на границах $x = 0$, $x = h$ [20], [21]. Дополнительно мы требуем, чтобы величины $e_y(x)\big|_{x=0}$ и $e_z(x)\big|_{x=0}$ имели фиксированные (известные) значения.

Подставляя (6.2) в (6.5) и используя обозначения $u_1 := e_y$, $u_2 := e_z$, после некоторых преобразований приходим к системе

$$\begin{aligned} u_1 &= -\mu\omega_1^2(\epsilon_1 - \gamma_1^2)u_1, \\ u_2 &= -\mu\omega_2^2(\epsilon_2 - \gamma_2^2)u_2, \end{aligned} \tag{6.6}$$

$h_y = -(i\mu\omega_2)^{-1}e_z'$, $h_z = (i\mu\omega_1)^{-1}e_y'$, где

$$\epsilon_1 = \begin{cases} \epsilon_s, & x < 0, \\ \epsilon_{yy}, & 0 \le x \le h, \\ \epsilon_c, & x > h, \end{cases} \quad \epsilon_2 = \begin{cases} \epsilon_s, & x < 0, \\ \epsilon_{zz}, & 0 \le x \le h, \\ \epsilon_c, & x > h, \end{cases} \tag{6.7}$$

а $\epsilon_{yy} = \epsilon_y + \beta_1 f_1(u_1^2) + \beta_3 g_1(u_1^2, u_2^2)$, $\epsilon_{zz} = \epsilon_z + \beta_2 f_2(u_2^2) + \beta_4 g_2(u_1^2, u_2^2)$.

В соответствии с условием на бесконечности решения системы (6.6) в полупространствах имеют вид

$$\begin{aligned} u_j(x) &= A_j \exp(\omega_j\sqrt{\mu(\gamma_j^2 - \epsilon_s)}x) \quad \text{для} \quad x < 0, \\ u_j(x) &= B_j \exp(-\omega_j\sqrt{\mu(\gamma_j^2 - \epsilon_c)}(x - h)) \quad \text{для} \quad x > h, \end{aligned} \tag{6.8}$$

где $A_j, B_j \ne 0$ – вещественные постоянные и без потери общности $A_j > 0$.

Внутри слоя Σ система (6.6) принимает вид

$$\begin{aligned} u_1'' &= -\mu\omega_1^2(\epsilon_y - \gamma_1^2)u_1 - \mu\omega_1^2\beta_1 f_1(u_1^2)u_1 + \mu\omega_1^2\beta_3 g_1(u_1^2, u_2^2)u_1, \\ u_2'' &= -\mu\omega_2^2(\epsilon_z - \gamma_2^2)u_2 - \mu\omega_2^2\beta_2 f_2(u_2^2)u_2 + \mu\omega_2^2\beta_4 g_2(u_1^2, u_2^2)u_2. \end{aligned} \tag{6.9}$$

Касательными компонентами поля (6.1), (6.2) являются e_y, h_z и e_z, h_y. Таким образом, функции u_j и u_j' непрерывны на границах $x = 0$ и $x = h$. Принимая во внимание указанную непрерывность и используя решения в полупространствах (6.8), приходим к условиям сопряжения

$$\omega_j\sqrt{\mu(\gamma_j^2 - \epsilon_s)}u_j\Big|_{x=0} - u_j'\Big|_{x=0} = 0, \quad -\omega_j\sqrt{\mu(\gamma_j^2 - \epsilon_c)}u_j\Big|_{x=h} - u_j'\Big|_{x=h} = 0. \tag{6.10}$$

Упомянутое условие фиксированных значений поля на границе имеет вид

$$u_j\big|_{x=0} = A_j,$$

где A_j совпадает с одноименной постоянной в (6.8).

Сформулированная физическая задача о распространении волн есть ни что иное как задача $P(\boldsymbol{\alpha})$, где использованы обозначения: $\lambda_j = \omega_j^2\mu_0\gamma_j^2$, $a_1 = \omega_1^2\mu_0\epsilon_y$, $a_2 = \omega_2^2\mu_0\epsilon_z$, $b_j = \omega_j^2\mu_0\epsilon_s$, $c_j = \omega_j^2\mu_0\epsilon_c$; $\alpha_j = \omega_j^2\mu_0\beta_j$, $\alpha_{j+2} = \omega_j^2\mu_0\beta_{j+2}$, где μ_0 – магнитная проницаемость вакуума. Постановку задачи для закрытого волновода см. в [4], [8].

Условия, перечисленные в этом пункте, приводят к условиям (1.2), (1.3) при $n = 2$ и условиям (1.6), (1.7) при $n = 1$.

Поле (6.1), (6.2) распространяется в слое Σ только для выделенных значений пар (γ_1, γ_2). Эти значения будем называть *(векторными) постоянными распространения*. С математической точки зрения, сформулированная задача является нелинейной двухпараметрической задачей на собственные значения для системы (6.9) с перечисленными начальными и краевыми условиями. Векторные собственные значения этой задачи являются векторными постоянными распространения.

Поскольку γ_j в (6.2) вещественные, то $\lambda_j = \mu\omega_j^2\gamma_j^2 > 0$. Если $n = 1$ или $\beta_3 = \beta_4 = 0$, то с точностью до обозначений получаем задачи P_j. Если $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, то получаем 2 линейных задачи об определении постоянных распространения ТЕ-волн, распространяющихся в волноводе, заполненном линейной средой. Эти задачи эквиваленты задачам P_j^0, сформулированным в разд. 1.

Нелинейные законы, используемые в оптике нелинейных волноводов, часто содержат множители, которые являются малыми параметрами [5], [26], [27]. Это позволяет применять методы теории возмущений, основанные на поиске возмущенных решений соответствующих линеаризованных задач, см., например, [2]–[4]. Такой подход не всегда оправдан, поскольку могут существовать нелинеаризуемые решения (см. теорему 3).

СПИСОК ЛИТЕРАТУРЫ

1. *Малкин И.Г.* Некоторые задачи теории нелинейных колебаний. М.: ГИТТЛ, 1956.
2. *Smirnov Yu.G., Valovik D.V.* Problem of nonlinear coupled electromagnetic TE-TE wave propagation // J. Math. Phys. 2013. V. 54. № 8. Art. no. 083502 (13 pages).
3. *Valovik D.V.* On the problem of nonlinear coupled electromagnetic TE-TM wave propagation // J. Math. Phys. 2013. V. 54. № 4. Art. no. 042902 (14 pages).
4. *Valovik D.V.* Nonlinear multi-frequency electromagnetic wave propagation phenomena // J. Opt. 2017. V. 18. № 11. Art. no. 115502 (16 pages).
5. *Boardman A.D., Egan P., Lederer F., Langbein U., Mihalache D.* Third-Order Nonlinear Electromagnetic TE and TM Guided Waves. Elsevier Sci. Publ. North-Holland, Amsterdam London New York Tokyo, 1991. Reprinted from Nonliear Surface Electromagnetic Phenomena, Eds. Ponath H.-E., Stegeman G. I.
6. *Valovik D.V.* On the existence of infinitely many nonperturbative solutions in a transmission eigenvalue problem for nonlinear Helmholtz equation with polynomial nonlinearity // Appl. Math. Modelling. 2018. V. 52. P. 296–309.
7. *Tikhov S.V., Valovik D.V.* Perturbation of nonlinear operators in the theory of nonlinear multifrequency electromagnetic wave propagation // Communications in Nonlinear Science and Numerical Simulation. 2019. V. 75. P. 76–93.
8. *Kurseeva V.Yu., Tikhov S.V., Valovik D.V.* Nonlinear multiparameter eigenvalue problems. Linearised and nonlinearised solutions // J. of Differential Equations. 2019. V. 267. № 4. P. 2357–2384.
9. *Валовик Д.В.* Распространение электромагнитных волн в открытом плоском диэлектрическом волноводе, заполненном нелинейной средой I: TE-волны // Ж. вычисл. матем. и матем. физ. 2019. Т. 59. № 5. С. 838–858.
10. *Вайнберг М.М.* Вариационные методы исследования нелинейных операторов. М.: ГИТТЛ, 1956.
11. *Ambrosetti A., Rabinowitz P.H.* Dual variational methods in critical point theory and applications // J. of Functional Analysis. 1973. V. 14. № 4. P. 349–381.
12. *Красносельский М.А.* Топологические методы в теории нелинейных интегральных уравнений. М.: ГИТТЛ, 1956.
13. *Вайнберг М.М., Треногин В.А.* Теория ветвления решений нелинейных уравнений. М.: Наука, 1969.
14. *Валовик Д.В.* О нелинейной задаче на собственные значения, связанной с теорией распространения электромагнитных волн // Дифференц. ур-ния. 2018. Т. 54. № 2. С. 168–179.
15. *Theotokoglou E.E., Panayotounakos D.E.* Nonlinear asymptotic analysis of a system of two free coupled oscillators with cubic nonlinearities // Appl. Math. Modelling. 2017. V. 43. P . 509–520.
16. *Cuevas J., Kevrekidis P.G., Saxena A., Khare A.* Pt-symmetric dimmer of coupled nonlinear oscillators // Phys. Rev. A. 2013. V. 88. № 3. Art. no. 032108 (11 pages).
17. *Borkowski L., Perlikowski P., Kapitaniak T., Stefanski A.* Experimental observation of three-frequency quasiperiodic solution in a ring of unidirectionally coupled oscillators // Phys. Rev. E. 2015. V. 91. № 6. Art. no. 062906 (7 pages).
18. *Tao M.* Simply improved averaging for coupled oscillators and weakly nonlinear waves // Communications in Nonlinear Science and Numerical Simulation. 2019. V. 71. P. 1–21.
19. I*n V., Kho A., Neff J.D., Palacios A., Longhini P., Meadows B.K.* Experimental observation of multifrequency patterns in arrays of coupled nonlinear oscillators // Phys. Rev. Lett. 2003. V. 91. № 24. Art. no. 244101 (4 pages).
20. *Вайнштейн Л.А.* Электромагнитные волны. М.: Радио о связь, 1988.
21. *Адамс М.* Введение в теорию оптических волноводов. М.: Мир, 1984.
22. *Марченко В.А.* Спектральная теория операторов Штурма–Лиувилля. Наук. думка, 1972.
23. *Schüurmann H.W., Smirnov Yu.G., Shestopalov Yu.V.* Propagation of TE-waves in cylindrical nonlinear dielectric waveguides // Phys. Rev. E. 2005. V. 71. № 1. Art. no. 016614 (10 pages).
24. *Smirnov Yu.G., Valovik D.V.* Guided electromagnetic waves propagating in a plane dielectric waveguide with nonlinear permittivity // Phys. Rev. A. 2015. V. 91. № 1. Art. no. 013840 (6 pages).
25. *Понтрягин Л.С.* Обыкновенные дифференциальные уравнения. М.: Физматлит, 1961.
26. *Ландау Л.Д., Лившиц Е.М.* Электродинамика сплошных сред. М.: Наука, 1982.
27. *Шен И.Р.* Принципы нелинейной оптики. М.: Наука, 1989.
28. *Ахмедиев Н.Н., Анкевич А.* Солитоны. М.: Физматлит, 2003.
29. *Маныкин Э.А.* Взаимодействие излучения с веществом. Феноменология нелинейной оптики. М.: МИФИ, 1996.
30. *Eleonskii P.N., Oganes'yants L.G., Silin V.P.* Cylindrical nonlinear waveguides // Soviet Physics JETP. 1972. V. 35. № 1. P. 44–47.
31. *Eleonskii P.N., Oganes'yants L.G., Silin V.P.* Structure of three-component vector fields in self-focusing waveguides // Soviet Physics JETP. 1973. V. 36. № 2. P. 282–285.

32. *Akhmediev N.N., Ankevich A.* Solitons, Nonlinear Pulses and Beams. London: Chapman and Hall, 1997.
33. *Boardman A.D., Twardowski T.* Theory of nonlinear interaction between TE and TM waves // J. of the Optical Society of America B. 1988. V. 5. № 2. P. 523–528.
34. *Boardman A.D., Twardowski T.* Transverse-electric and transverse-magnetic waves in nonlinear isotropic waveguides // Phys. Rev. A. 1989. V. 39. № 5. P. 2481–2492.
35. *Ping Xie, Zhao-Qing Zhang.* Multifrequency gap solitons in nonlinear photonic crystals // Phys. Rev. Lett. 2003. V. 91. № 21. Art. no. 213904 (4 pages).
36. *Skryabin D.V., Biancalana F., Bird D.M., Benabid F.* Effective kerr nonlinearity and two-color solitons in photonic band-gap fibers filled with a Raman active gas // Phys. Rev. Lett. 2004. V. 93. № 14. Art. no. 143907 (4 pages).

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 124–135

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.634

ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ ГАЗОВЫХ СМЕСЕЙ В РАМКАХ КВАЗИГАЗОДИНАМИЧЕСКОГО ПОДХОДА НА ПРИМЕРЕ ВЗАИМОДЕЙСТВИЯ УДАРНОЙ ВОЛНЫ С ПУЗЫРЬКОМ ГАЗА

© 2021 г. Т. Г. Елизарова[1,*], Е. В. Шильников[1,]**

[1] *125047 Москва, Миусская пл., 4, ИПМ им. М.В. Келдыша РАН, Россия*

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

***e-mail: shilnikov@imamod.ru*

Поступила в редакцию 21.04.2020 г.
Переработанный вариант 02.07.2020 г.
Принята к публикации 18.09.2020 г.

Представлен новый численный алгоритм для моделирования течений нереагирующих газовых смесей в трансзвуковых режимах. Алгоритм основан на методе конечного объема, записанного для регуляризованных, или квазигазодинамических, уравнений. Уравнения для описания течения смеси выведены феноменологически на базе существующей регуляризованной системы для однокомпонентного газа и классических уравнений для газовой смеси. Примеры численного моделирования включают в себя расчет задачи о нестационарном взаимодействии газового потока с тяжелой и легкой каплями газа. Библ. 23. Фиг. 9. Табл. 1.

Ключевые слова: смесь газов, квазигазодинамические уравнения, метод конечного объема, трансзвуковые течения, газовые капли.

DOI: 10.31857/S004446692101004X

ВВЕДЕНИЕ

Полученные более тридцати лет назад квазигазодинамические (КГД), или регуляризованные, уравнения для описания течений газов различной природы успешно применялись для построения численных алгоритмов как для традиционных вычислительных систем, так и для систем с массивным параллелизмом (см. [1]–[3]). Оценки показывают, что КГД-подход особенно эффективен для моделирования нестационарных или неустановившихся течений газа. В настоящее время распространение этого подхода дополнительно поддерживается тем фактом, что КГД-алгоритм включен в открытый программный комплекс OpenFOAM и доступен широкому кругу пользователей как в России, так и за рубежом (см. [4]–[6]).

Представляется естественным расширение имеющегося семейства КГД-алгоритмов применительно к задачам неустановившихся течений газовых смесей. В данной работе построен и протестирован КГД-алгоритм для моделирования течения газовой смеси, структура которого повторяет структуру имеющихся в комплексе OpenFOAM модулей с КГД-методикой. Этот факт позволит естественным образом подключить новый модуль для расчета смесевых течений к платформе OpenFOAM и объединить его с уже имеющимися в системе многочисленными расширениями газодинамических методик. В частности, это модули для расчета турбулентных течений, блоки моделирования химических реакций между компонентами, постоянно совершенствуемые способы распараллеливания всего алгоритма на современных вычислительных системах.

Большое число инженерных задач, включающих в себя моделирование течений флюидов разного состава и разной природы, не поддается перечислению. В частности, это задачи о смешении газовых струй в реакторах, задачи о вытекании струй при авариях газопроводов, задачи, связанные со смешением газов для поддержания их эффективного горения, и множество других. Поэтому практическая важность моделирования течений газовых смесей не вызывает сомнений, что определяет актуальность изложенных далее результатов.

Структура работы имеет следующий вид. В разд. 1 приведена одна из имеющихся математических моделей для описания газовой смеси без химических реакций. За основу принята извест-

ная одножидкостная модель, описывающая смесь газов в рамках системы уравнений для плотностей отдельных компонент совместно с общими для обоих компонент уравнениями переноса импульса и полной энергии, без явного выделения межфазных границ. Для простоты изложения считаем, что смесь состоит из двух идеальных газов. В разд. 2 на основе феноменологических соображений выписан регуляризованный аналог этой системы уравнений. В разд. 3 описана постановка задачи о взаимодействии ударной волны с каплей газа. Здесь же кратко изложены принципы численной реализации алгоритма. Раздел 4 посвящен подробному описанию результатов численных экспериментов. Рассмотрены процессы обтекания легкой и тяжелой капель газа воздухом с распространяющейся в нем ударной волной. Заключительные замечания к изложенным результатам приведены в конце работы.

1. СИСТЕМА УРАВНЕНИЙ ДЛЯ ОПИСАНИЯ ТЕЧЕНИЯ ГАЗОВОЙ СМЕСИ

Рассмотрим газовую смесь без химических реакций в рамках одной из известных моделей. В этой модели каждая из компонент смеси удовлетворяет отдельному уравнению неразрывности, записанному относительно плотности соответствующей компоненты. Полагается, что взаимодействие между молекулами разного сорта происходит быстро по сравнению с гидродинамическими временами, поэтому скорость и температура в течении для компонент смеси одинаковы. Тем самым уравнения импульса и полной энергии записываются для смеси в целом, без их расщепления на уравнения для каждой из компонент. Процессы вязкости и теплопроводности рассматриваются на уровне всей смеси в целом и определяются коэффициентами вязкости и теплопроводности для смеси.

Уравнение для плотности смеси расщеплено на уравнения для плотностей отдельных компонент без явного выделения межфазных границ. Тем самым фазовые границы между компонентами смеси формируются автоматически в областях больших градиентов концентраций компонент, и в модель не включено отдельное уравнение для определения границ между компонентами. Взаимная диффузия между компонентами смеси не учитывается.

Такая модель описания течений смеси газов является весьма распространенной. В частности, она применяется в ряде работ Р. Абгралла и С. Карни, например, в [7]–[10]. При использовании этого вида моделей предпринималось много усилий для подавления нефизичных осцилляций на границе раздела флюидов при расчетах по схемам повышенного порядка аппроксимации. Для этого применялась так называемая Double flux модификация широко распространенных численных методов, таких как TVD-MUSCL схемы (см. [9]), разрывный метод Галеркина (см. [11]), схемы Годунова (см. [12], [13]). Отметим, что в отличие от описанной выше модели, в [12] рассматривается многожидкостная модель, когда каждый флюид имеет не только свою собственную плотность, но и свои скорости и давления. Это приводит к появлению обменных членов в правых частях уравнений. Такая модель хорошо справляется с моделированием течений смеси жидкостей и газов. Однако наличие обменных членов приводит к необходимости интегрирования на каждом шаге, вообще говоря, жесткой системы ОДУ, что сильно усложняет алгоритм. В случае моделирования течений реагирующих газов жесткую систему ОДУ приходится решать по причине использования в расчетах сильно различающихся скоростей химических реакций (см. [13], [14]), но при отсутствии химических реакций представляется более желательным обойтись без этого усложнения.

Для краткости рассмотрим смесь двух газов a и b. В этом случае описанная выше модель представляется в виде следующей системы уравнений:

$$\frac{\partial \rho_a}{\partial t} + \operatorname{div}(\rho_a \mathbf{u}) = 0, \tag{1}$$

$$\frac{\partial \rho_b}{\partial t} + \operatorname{div}(\rho_b \mathbf{u}) = 0, \tag{2}$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \operatorname{div}(\rho(\mathbf{u} \otimes \mathbf{u})) + \nabla p = \operatorname{div} \Pi_{\mathrm{NS}} + \rho \mathbf{F}, \tag{3}$$

$$\frac{\partial E}{\partial t} + \operatorname{div}((E + p)\mathbf{u}) = -\operatorname{div} \mathbf{q}_{\mathrm{NS}} + \operatorname{div}(\Pi_{\mathrm{NS}} \cdot \mathbf{u}) + \rho(\mathbf{u} \cdot \mathbf{F}) + Q. \tag{4}$$

В приведенной системе использованы общепринятые обозначения газодинамических величин. Дополнительно введены вектор $\mathbf{F}$ и скаляр Q, которые обозначают удельную внешнюю силу и внешний источник или сток энергии соответственно. Последнее будет существенно для случая

реагирующих газов, который пока не включен в рассмотрение. Тензор вязких напряжений Навье–Стокса Π_{NS} и вектор теплового потока $\mathbf{q}_{NS}$ в форме закона Фурье имеют традиционный вид

$$\Pi_{NS} = \mu\left((\nabla \otimes \mathbf{u}) + (\nabla \otimes \mathbf{u})^{T} - \frac{2}{3}\mathrm{I}\,\mathrm{div}\,\mathbf{u}\right), \quad \mathbf{q}_{NS} = -\kappa\nabla T,$$

где I – единичная матрица. В выписанной модели предполагается, что газовая смесь имеет единую скорость $\mathbf{u}$ и температуру T, а плотность смеси, ее давление и удельная полная энергия определяются через параметры ее компонент как

$$\rho = \rho_a + \rho_b, \quad p = p_a + p_b, \quad E = \rho\varepsilon + \rho\mathbf{u}^2/2.$$

Кроме того, газодинамическая смесь удовлетворяет обычным соотношениям для идеального политропного газа

$$p = \rho RT = \rho\varepsilon(\gamma - 1),$$

где γ – показатель адиабаты смеси, R – газовая постоянная и ε – удельная внутренняя энергия смеси:

$$R = \frac{R_a\rho_a + R_b\rho_b}{\rho} = c_p - c_V, \quad \gamma = \frac{c_p}{c_V}, \quad \gamma - 1 = \frac{R}{c_V},$$

$$\varepsilon = \frac{\varepsilon_a\rho_a + \varepsilon_b\rho_b}{\rho} = c_V T, \quad c_V = \frac{c_{Va}\rho_a + c_{Vb}\rho_b}{\rho}.$$

Подчеркнем, что выписанные выше термодинамические параметры для смеси c_p, c_V, а следовательно, и R, γ, в отличие от аналогичных величин для идеального политропного газа уже не являются постоянными и определяются через взвешенные значения параметров смеси. Они зависят от плотности каждой компоненты газа в каждой пространственно-временной точке $(\mathbf{x}, t)$.

Скорость звука для смеси может быть вычислена с использованием соотношения

$$\rho c_s^2 = \gamma_a p_a + \gamma_b p_b.$$

Очевидно, что обобщение описанной здесь модели на случай смеси с большим числом компонент не вызывает затруднений.

2. РЕГУЛЯРИЗОВАННАЯ СИСТЕМА УРАВНЕНИЙ ДЛЯ ОПИСАНИЯ ТЕЧЕНИЯ ГАЗОВОЙ СМЕСИ

Регуляризованные (или КГД) уравнения для описания течения вязкого политропного идеального газа имеют следующий вид (см., например, [2], [3]):

$$\frac{\partial\rho}{\partial t} + \mathrm{div}(\rho(\mathbf{u} - \mathbf{w})) = 0, \tag{5}$$

$$\frac{\partial\rho\mathbf{u}}{\partial t} + \mathrm{div}(\rho(\mathbf{u} - \mathbf{w}) \otimes \mathbf{u}) + \nabla p = \mathrm{div}\,\Pi + (\rho - \tau\,\mathrm{div}(\rho\mathbf{u}))\mathbf{F}, \tag{6}$$

$$\frac{\partial E}{\partial t} + \mathrm{div}((E + p)(\mathbf{u} - \mathbf{w})) = -\mathrm{div}\,\mathbf{q} + \mathrm{div}(\Pi \cdot \mathbf{u}) + \rho(\mathbf{u} - \mathbf{w}) \cdot \mathbf{F} + Q. \tag{7}$$

Эта система уравнений включает в себя регуляризирующие ее добавки вида

$$\mathbf{w} = \frac{\tau}{\rho}\left(\mathrm{div}(\rho\mathbf{u} \otimes \mathbf{u}) + \nabla p - \rho\mathbf{F}\right), \tag{8}$$

$$\hat{\mathbf{w}} = \frac{\tau}{\rho}\left(\rho(\mathbf{u}\nabla)\mathbf{u} + \nabla p - \rho\mathbf{F}\right), \tag{9}$$

$$\mathbf{q} = \mathbf{q}_{NS} + \mathbf{q}^{\tau}, \tag{10}$$

$$\Pi = \Pi_{NS} + \rho\mathbf{u} \otimes \hat{\mathbf{w}} + \tau(\mathbf{u}\nabla p + \gamma p\,\mathrm{div}\,\mathbf{u} - (\gamma - 1)Q), \tag{11}$$

$$-\mathbf{q}^{\tau} = \tau\rho\mathbf{u}\left(\mathbf{u}\nabla\varepsilon + p(\mathbf{u}\nabla)\left(\frac{1}{\rho}\right) - \frac{Q}{\rho}\right). \tag{12}$$

Все эти добавки пропорциональны малому коэффициенту τ, который имеет размерность времени и для газодинамических течений может быть представлен в виде

$$\tau = l/c_s, \tag{13}$$

где c_s – скорость звука и l – некоторый характерный размер, определяемый рассматриваемой задачей. В большинстве вычислительных задач в качестве характерного размера удобно выбрать величину шага пространственной сетки h:

$$\tau = \alpha h/c_s, \tag{14}$$

с численным коэффициентом $\alpha \le 1$, который выбирается в процессе решения задачи из соображений точности и устойчивости численного алгоритма. Коэффициент τ определяет величину подсеточной диссипации алгоритма и тесно связан с условием устойчивости алгоритма.

Подчеркнем, что дополнительные слагаемые являются добавками к вектору скорости $\mathbf{u}$, тензору вязких напряжений Π_{NS} и тепловому потоку $\mathbf{q}_{NS}$. Такое введение регуляризирующих добавок обеспечивает для системы уравнений (5)–(7) выполнение законов сохранения массы, импульса и энергии совместно с законами сохранения момента импульса и неотрицательностью диссипативной функции.

По аналогии с КГД-уравнениями для однокомпонентного газа (5)–(7) выпишем регуляризованный аналог системы для смеси. Таким же образом, как в системе (1)–(4), положим $\rho = \rho_a + \rho_b$ и расщепим уравнение неразрывности (5) на уравнения для каждой из компонент ρ_a и ρ_b. Соответствующим образом расщепим и регуляризатор, входящий в исходное уравнение для общей плотности смеси. Принимая во внимание, что компоненты смеси имеют одинаковую скорость, положим, что и регуляризирующие добавки к скоростям для компонент смеси одинаковы: $\mathbf{w}_a = \mathbf{w}_b = \mathbf{w}$. Тогда регуляризованная (или КГД) система уравнений для смеси газов будет иметь следующий вид:

$$\frac{\partial \rho_a}{\partial t} + \operatorname{div}(\rho_a(\mathbf{u} - \mathbf{w})) = 0, \tag{15}$$

$$\frac{\partial \rho_b}{\partial t} + \operatorname{div}(\rho_b(\mathbf{u} - \mathbf{w})) = 0, \tag{16}$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \operatorname{div}(\rho(\mathbf{u} - \mathbf{w}) \otimes \mathbf{u}) + \nabla p = \operatorname{div}\Pi + (\rho - \tau \operatorname{div}(\rho \mathbf{u}))\mathbf{F}, \tag{17}$$

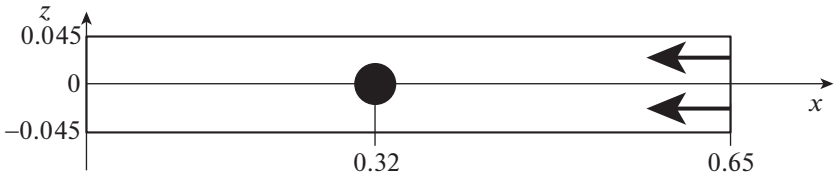
$$\frac{\partial E}{\partial t} + \operatorname{div}((E + p)(\mathbf{u} - \mathbf{w})) = -\operatorname{div}\mathbf{q} + \operatorname{div}(\Pi \cdot \mathbf{u}) + \rho(\mathbf{u} - \mathbf{w}) \cdot \mathbf{F} + Q. \tag{18}$$

Эта система дополняется регуляризирующими добавками вида (8)–(12). Для вычисления коэффициента τ через скорость звука c_s для смеси используются соотношения (13) или (14).

Регуляризованные уравнения, представленные выше, получены феноменологическим путем, и математические свойства для этих уравнений детально не исследовались. Однако первые расчеты с использованием этих уравнений, проведенные для задачи о гравитационной неустойчивости Рэлея–Тейлора (см. [15]) показали высокую работоспособность и устойчивость основанного на них численного алгоритма. Вычислительные возможности этой модели для трансзвуковых течений показаны в разд. 4.

В заключение этого раздела отметим, что ранее в рамках КГД-подхода были построены и протестированы в численных экспериментах другие регуляризованные системы уравнений для описания смесей нереагирующих газов.

Первая система уравнений из этого класса (см., например, [2], [16]) была ориентирована на описание течений разреженных газов. Эта система уравнений представляла собой так называемую двухжидкостную модель, в которой для каждой из компонент смеси выписывались свои уравнения неразрывности, импульса и полной энергии. Связь уравнений осуществлялась с помощью обменных слагаемых, обеспечивающих обмены импульсом и энергией между компонентами смеси и входящих в правые части уравнений для импульса и энергии. Существенные усовершенствования этой системы уравнений представлены в [17]. В частности, это запись системы уравнений в виде законов сохранения, вывод для нее уравнения переноса энтропии с неотрицательной диссипативной функцией, обобщение обменных слагаемых на случай многоатомных газов. Прояснение физического смысла обменных слагаемых для случая многоатомных газов представлено в [18].



Фиг. 1. Схема расчетной области.

Однако, как показали анализ и опыт численного моделирования, двухжидкостная система уравнений с обменными слагаемыми оказалась неэффективной при расчетах достаточно плотных газов. Для случая таких неразреженных газов регуляризирующие добавки были подвергнуты существенной модификации и были выписаны две новые системы уравнений для газовых смесей (см. [19]). Для них были выведены уравнения баланса массы, кинетической и внутренней энергий, а также новые уравнения баланса полной энтропии, и доказана неотрицательность производства энтропии. Однако полученные уравнения содержат дополнительные слагаемые недивергентного вида в уравнении полной энергии, что приводит к сложностям в их численной реализации.

Близкая по своей структуре более простая квазигидродинамическая модель течения смеси была построена в [20] для моделирования медленных двухфазных течений в пористых средах с учетом межфазных взаимодействий. Вид этой системы уравнений имеется в [19].

3. ПОСТАНОВКА ЗАДАЧИ О ВЗАИМОДЕЙСТВИИ ПУЗЫРЬКА ГАЗА С УДАРНОЙ ВОЛНОЙ

Тестовая задача состоит в моделировании взаимодействия плоской ударной волны, движущейся в воздухе, с цилиндрическим пузырем другого газа. Такой эксперимент описан в [21] и численно исследован в [7], [8], [22], [23]. Рассматривается прямоугольная двумерная область, заполненная воздухом (фиг. 1). Плоская ударная волна, проходя через воздух, падает на цилиндрический пузырь из гелия или хладоагента R22 ($CHClF_2$). Пузырь с радиусом $R = 0.025$ помещается в воздух с центром пузыря в точке $(x_c, y_c) = (0.32, 0)$. Пренебрегая граничными эффектами, т.е. считая цилиндр бесконечно длинным, можно моделировать такое течение в двумерной постановке. В центральном сечении течение является практически двумерным и краевыми эффектами можно пренебречь, если длина цилиндра не мала по сравнению с высотой установки.

Все газовые компоненты считаются идеальными газами. Газовая постоянная $R = R_{\text{univ}}/m$, где R_{univ} – универсальная газовая постоянная, m – молекулярная масса газа. Мы пренебрегаем физической вязкостью μ и теплопроводностью κ газов и используем уравнения смеси в формулировке Эйлера. Внешняя сила $\mathbf{F}$ и источник тепла Q считаются равными нулю. Начальные динамически равновесные параметры газов в расчетной области приведены в табл. 1.

На правой границе задается условие притока воздуха с параметрами, лежащими за ударной волной, движущейся справа налево через воздух со скоростью, соответствующей числу Маха $M_s = 1.22$:

$$(\rho, u, v, p, \gamma)\big|_{\text{right}} = (1.3764, -124.82414, 0.0, 156983.9256, 1.4).$$

Все остальные границы области рассматриваются как твердые непроницаемые стенки с граничными условиями скольжения. Расчеты проводятся на прямоугольных сетках: грубая сетка,

Таблица 1. Начальные параметры газов. Размерности всех величин соответствуют системе СИ

Газ	ρ	u	v	p	γ	m	R	c_s
Воздух	1.0	0.0	0.0	10^5	1.4	28.96	287.1	374.16
Гелий	0.182	0.0	0.0	10^5	5/3	4.003	2077	915.75
R22	3.1538	0.0	0.0	10^5	1.249	86.47	96.15	199.0

состоящая из 1300×178 ячеек с пространственным шагом $h = 5\times10^{-4}$ в обоих направлениях, и в два раза более подробная сетка, состоящая из 2600×356 ячеек.

Для численной реализации системы QGD уравнений (15)–(18) мы используем явную по времени схему, построенную на основе метода конечных объемов, с аппроксимацией всех пространственных производных центральными разностями второго порядка. Все газодинамические переменные отнесены к центрам ячеек. Их значения в центрах граней ячеек рассчитываются с использованием линейной интерполяции. Условие устойчивости для этой схемы имеет вид условия Куранта, и шаг по времени определяется по формуле

$$\Delta t = \beta \min_i \frac{h_i}{c_{si} + |\mathbf{u}_i|}, \tag{19}$$

где минимум берется по всем ячейкам сетки, i – номер ячейки, β – числовой коэффициент (число Куранта), который не зависит от шага пространственной сетки. Конечное время расчета принимается равным $t_{\text{fin}} = 1.4\times10^{-3}$ для пузырька гелия и $t_{\text{fin}} = 1.1\times10^{-3}$ для пузырька R22. Время измеряется в секундах и отсчитывается от момента начала движения ударной волны от правой границы расчетной области.

4. АНАЛИЗ РЕЗУЛЬТАТОВ РАСЧЕТОВ

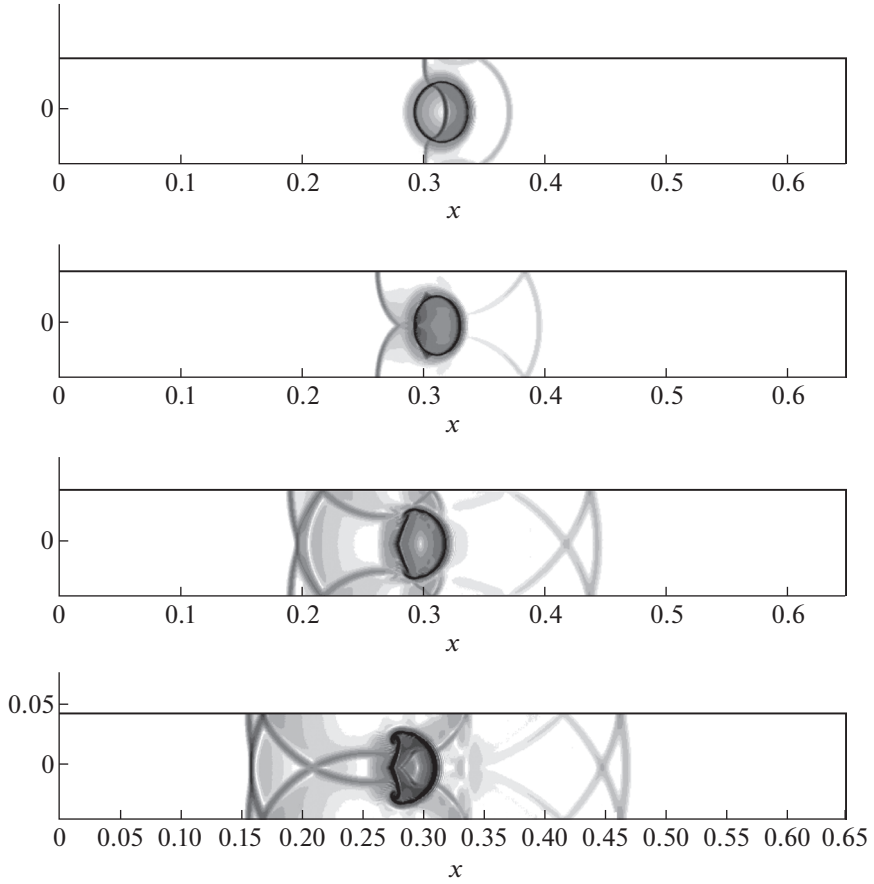
Расчеты для случая пузыря хладоагента R22 проводились с параметром схемы $\alpha = 0.4$ и числом Куранта $\beta = 0.5$. На фиг. 2 численные Шлирен-образы приведены в последовательные моменты времени (сверху вниз): $t = 7.61\times10^{-4}$, $t = 8.47\times10^{-4}$, $t = 1.01\times10^{-3}$, $t = 1.1\times10^{-3}$.

Шлирен-метод широко используется в экспериментах для визуализации различных процессов в газовой среде: в аэродинамических трубах, в механике жидкости, баллистике, при изучении распространения и смешивания газов и растворов и т.п. В связи с этим сравнение с экспериментом требует использования численного аналога Шлирен-фотографии. В численных экспериментах это достигается изображением поля логарифма модуля градиента плотности, каким-то образом отмасштабированного. В нашей работе были использованы возможности графического пакета Tecplot 360. Масштабирование достигалось выбором пределов изменения изображаемой величины таким образом, чтобы на получаемой картине течения наилучшим образом были видны ее характерные особенности.

После того как падающая ударная волна наталкивается на пузырь, последний начинает двигаться налево и деформируется (фиг. 2). На верхней картинке, соответствующей времени $t = 7.61\times10^{-4}$, падающая ударная волна состоит из двух вертикальных фрагментов сверху и снизу от пузыря. Внутри пузыря видна преломленная ударная волна, которая вследствие меньшей скорости распространения искривляется. Это связано с тем, что скорость звука в R22 почти в два раза меньше, чем в воздухе (табл. 1). При этом внешние концы преломленной волны движутся вместе с внутренними концами фрагментов падающей волны. Изогнутая отраженная волна, более слабая, чем падающая и преломленная, перемещается вправо от пузыря к правой границе области. Отраженная от пузыря волна соединена с фрагментами падающей волны двумя слабыми волнами, отраженными от горизонтальных стенок. Дифракция волн вокруг пузыря приводит к его деформации. На следующих рисунках на фиг. 2 видны формирование и развитие двух завитков в верхней и нижней частях пузыря, в которых возникает завихренность. Через некоторое время фрагменты падающей ударной волны, миновав пузырь, соединяются и в дальнейшем двигаются влево как целое. Отстающая вначале точка соединения фрагментов падающей волны постепенно догоняет их внешние концы, и фронт волны выпрямляется (нижнее изображение на фиг. 2). Все эти эффекты находятся в хорошем согласии с экспериментом и численными результатами других авторов.

Процесс прохождения преломленной волны через пузырь и происходящее при этом изменение ее формы можно проследить на фиг. 3.

Как уже было отмечено, скорость звука в газе R22 гораздо меньше, чем в воздухе, поэтому скорость преломленной волны, движущейся внутри пузыря, меньше, чем падающей, движущейся между пузырем и стенками. При этом внешние края преломленной волны движутся по поверхности пузыря вместе с внутренними краями падающей волны. В результате преломленная волна искривляется, и в процессе движения ее кривизна увеличивается. Так, на фиг. 3д она уже



Фиг. 2. Пузырь R22. Численные Шлирен-образы в последовательные моменты времени (сверху вниз): $t = 7.61\times10^{-4}$, $t = 8.47\times10^{-4}$, $t = 1.01\times10^{-3}$, $t = 1.1\times10^{-3}$.

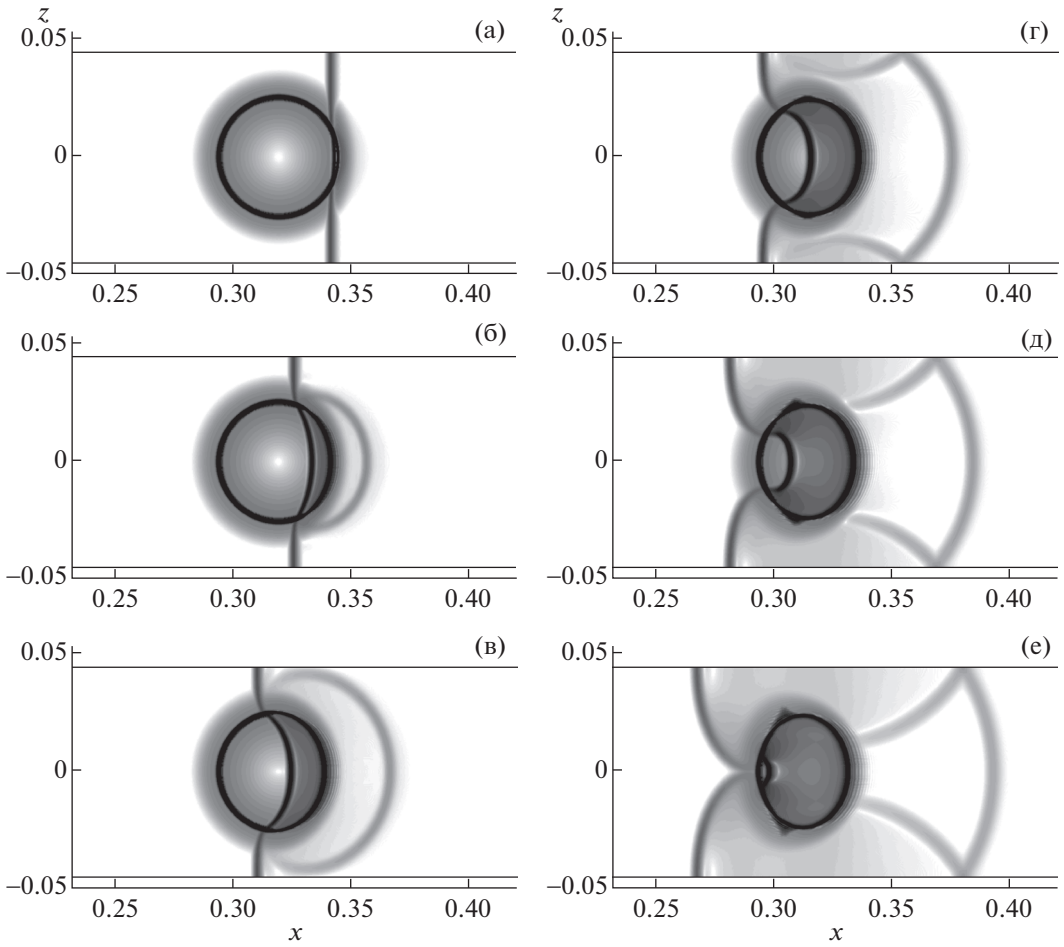
имеет форму полукруга, а на фиг. 3е – практически стянулась в круг. Таким образом, пузырь тяжелого газа R22 действует на падающую волну, как собирающая линза.

Сравнение формы пузырька в последний момент времени, полученное в наших расчетах, с численными из [7] и экспериментальными из [21] результатами показано на фиг. 4. Здесь приведена форма пузыря R22 в момент окончания расчета $t = 1.1\times10^{-3}$: (а) – наши результаты, (б) – результаты из [7], (в) – экспериментальные данные из [21]. Следует отметить, что как эволюция формы пузырька, так и общая структура сложного потока с большим количеством волн разных типов совпадают с результатами этих и других авторов.

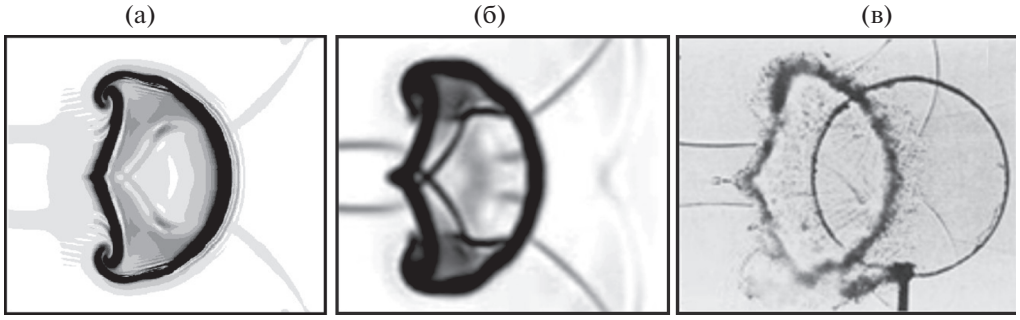
Расчеты для случая гелиевого пузыря проводились с параметром $\alpha = 0.4$ и числом Куранта $\beta = 0.2$. Они представлены на фиг. 5. Здесь приведены численные Шлирен-образы в последовательные моменты времени (сверху вниз): $t = 7.04\times10^{-4}$, $t = 7.52\times10^{-4}$, $t = 9.13\times10^{-4}$, $t = 1.342\times10^{-3}$.

В отличие от предыдущего случая пузырь гелия, с которым взаимодействует падающая ударная волна, легче окружающего воздуха и, таким образом, он действует как рассеивающая акустическая линза. В результате структура волн и общая картина течения сильно отличаются от тех, которые наблюдались в случае тяжелого пузыря.

Аналогично предыдущему случаю, после того, как падающая ударная волна сталкивается с пузырем, внутри пузыря образуется изогнутая преломленная волна. Однако эта волна движется быстрее, чем падающая, поскольку скорость звука в гелии в два с половиной раза выше, чем в воздухе (табл. 1), поэтому ее выпуклость направлена влево. Дальнейшая деформация гелиевого пузыря намного сильнее, чем для R22. Уже на верхнем рисунке (фиг. 5) видно, что наветренная сторона пузыря стала сильно приплюснутой, а на втором она уже почти плоская. В дальнейшем



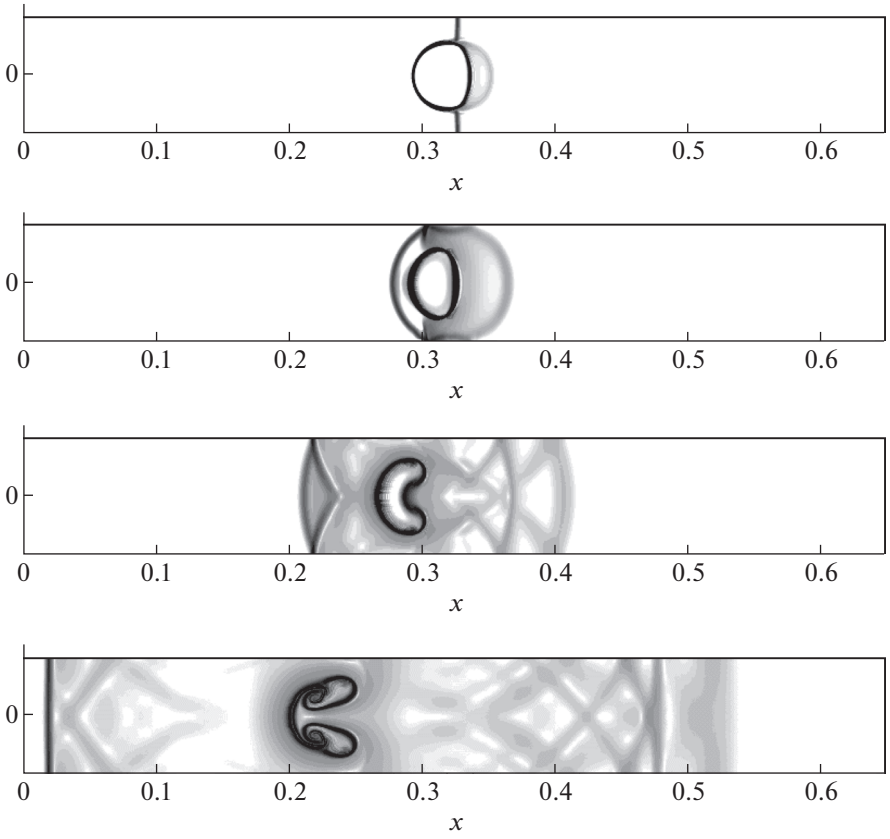
Фиг. 3. Прохождение ударной волны через пузырь R22.



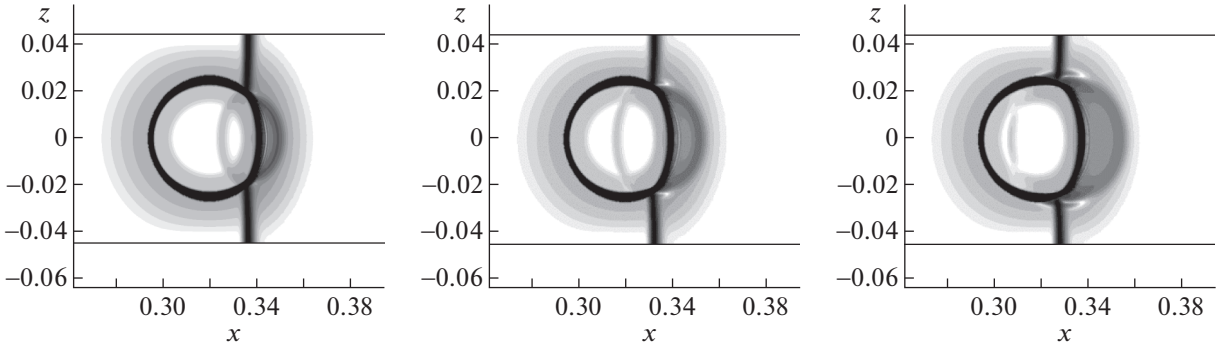
Фиг. 4. Форма пузыря R22 в момент окончания расчета $t = 1.1 \times 10^{-3}$: (а) – наши результаты, (б) – результаты из [7], (в) – экспериментальные данные из [21].

она все больше прогибается влево и практически достигает подветренной стороны. В результате пузырь фактически разделяется на два, соединенных тонкой перемычкой (последний рисунок на фиг. 5). Так же, как и в предыдущем случае, в верхней и нижней частях пузыря развиваются завитки, в которых образуется завихренность. Однако в этом случае эти завитки находятся не снаружи, а внутри пузыря. Отметим еще естественный факт, что легкий гелиевый пузырь под действием падающей волны движется влево быстрее, чем тяжелый пузырь R22.

Скорость преломленной волны настолько велика, что на фиг. 5 даже не видно, как она проходит через пузырь. Уже на верхнем рисунке фиг. 5 эта волна достигла левой границы пузыря.



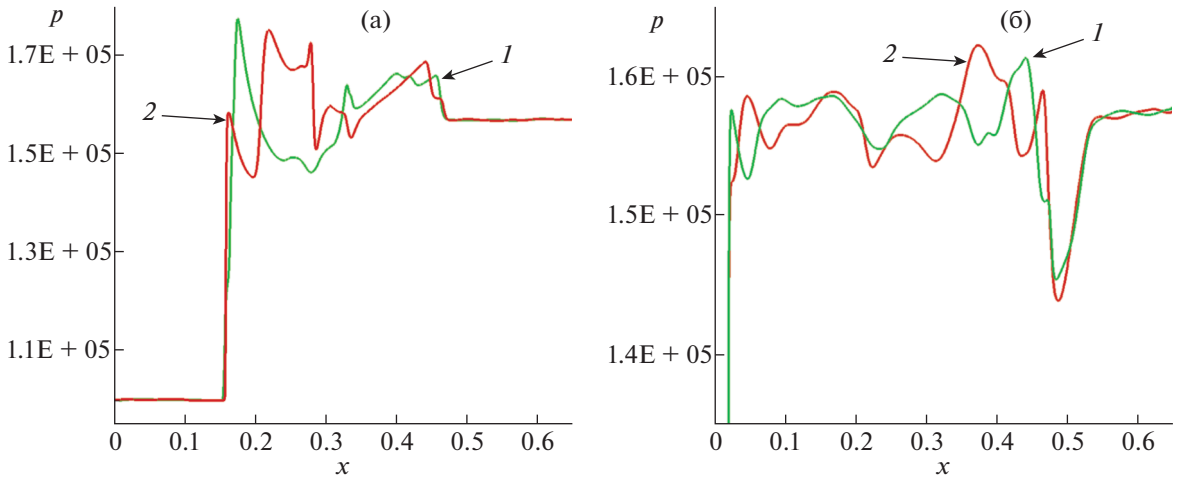
Фиг. 5. Пузырь гелия. Численные Шлирен-образы в последовательные моменты времени (сверху вниз): $t = 7.04\times10^{-4}$, $t = 7.52\times10^{-4}$, $t = 9.13\times10^{-4}$, $t = 1.342\times10^{-3}$.



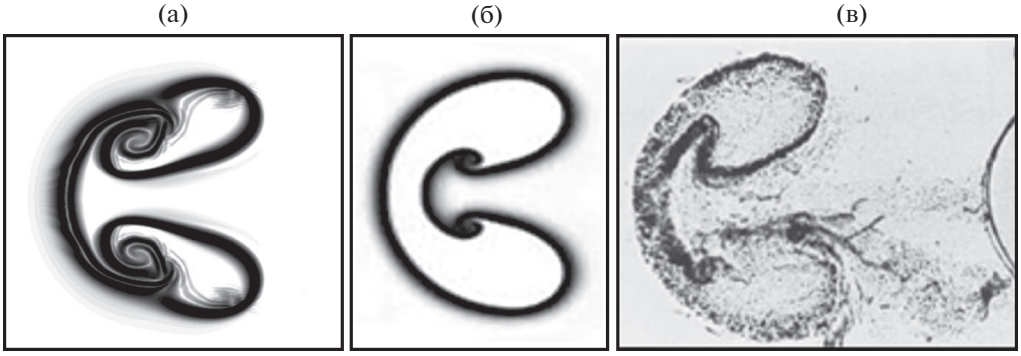
Фиг. 6. Прохождение преломленной волны через пузырь гелия. Моменты времени (слева направо): $t = 6.86\times10^{-4}$, $t = 6.95\times10^{-4}$, $t = 7.03\times10^{-4}$.

Чтобы увидеть процесс движения преломленной волны внутри пузыря, на фиг. 6 приведены фрагменты картины течения в более ранние близкие моменты времени. Моменты времени составляют (слева направо): $t = 6.86\times10^{-4}$, $t = 6.95\times10^{-4}$, $t = 7.03\times10^{-4}$. Видно, что за то время, когда преломленная волна прошла весь пузырь, падающая волна почти не продвинулась.

Еще одно отличие от предыдущего случая состоит в том, что отраженная волна здесь является не ударной волной, а волной разрежения. Это хорошо видно на фиг. 7. На ней изображены профили давления на момент окончания расчета вдоль центрального сечения $z = 0$ (линии *1*) и вдоль сечения $z = 0.4$ (линии *2*), проходящей между стенкой и пузырем для обоих вариантов расчета.



Фиг. 7. Профили давления вдоль линии $z = 0$ (линии *1*) и $z = 0.4$ (линии *2*): (а) – для пузыря R22, (б) – для пузыря гелия.



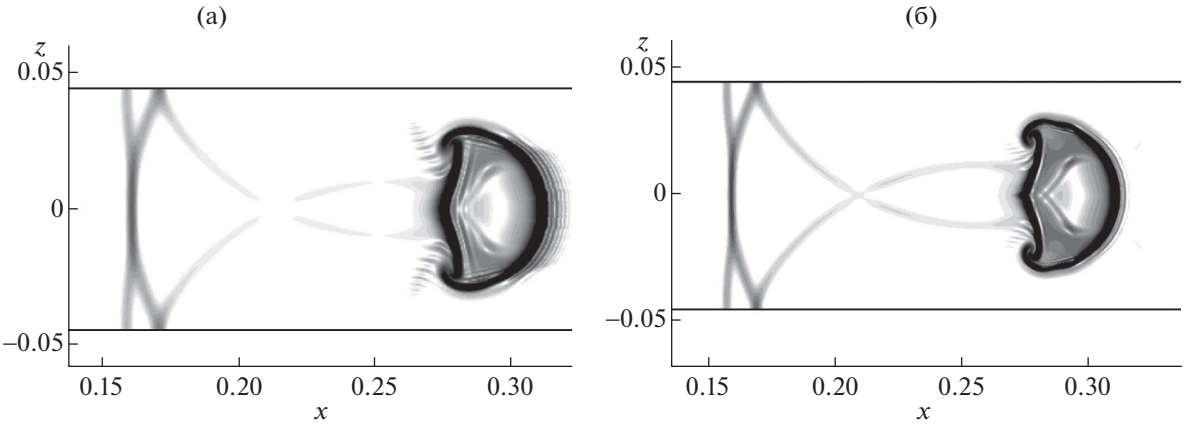
Фиг. 8. Форма пузыря гелия в момент времени $t = 1.342 \times 10^{-3}$: (а) – наши результаты, (б) – результаты из [7], (в) – экспериментальные данные из [21].

В случае R22 (на фиг. 7а) в волне, распространяющейся вверх по потоку, давление выше, чем на фоне (“полочка” давления у правой границы области и следующий за ней подъем давления), а в случае гелия (см. фиг. 7б) – ниже, где левее «полочки» давления наблюдается его падение. Это различие в поведении давления объясняется тем, что в момент падения ударной волны пузырь тяжелого R22 выступает в роли почти твердой стенки, а легкий пузырь гелия – почти вакуума.

Сравнение формы пузырька на момент окончания расчета, полученное в наших расчетах с теми же результатами, что и для случая R22, показано на фиг. 8. Здесь показана форма пузыря гелия в момент времени $t = 1.342 \times 10^{-3}$: (а) – наши результаты, (б) – результаты из [7], (в) – экспериментальные данные из [21].

Полученные результаты расчетов показывают высокую эффективность и точность использованного здесь численного подхода. Общая картина потока и основные характерные особенности сложных течений хорошо соответствуют как экспериментам, так и численным результатам других авторов. Отдельно отметим, что численное решение имеет правильную симметрию.

Сравнение результатов, полученных на разных пространственных сетках, представлены на фиг. 9. На нем приведены результаты расчетов с каплей R22 на момент времени окончания расчета на разных пространственных сетках, а именно, показаны численные Шлирен-образы на момент времени $t = 1.1 \times 10^{-3}$, полученные на сетке 1300×178 ячеек (слева) и 2600×356 ячеек (справа). Видно, что при дроблении сетки общая структура течения сохраняется, а его качество улучшается. Волны становятся тоньше и более четкими, проявляются более тонкие особенности течения, особенно внутри капли. Уменьшение параметра регуляризации α влияет на решение



Фиг. 9. Численные Шлирен-образы на момент времени $t = 1.1 \times 10^{-3}$, (а) – полученные на сетке 1300×178 ячеек и (б) – для 2600×356 ячеек.

примерно так же, как и дробление сетки. Однако при этом ухудшается устойчивость схемы и требуется уменьшение шага по времени. Поэтому расчеты проводились при минимальном значении α, обеспечивающем устойчивость при разумном значении числа Куранта β.

Отметим также, что в работе [7], с которой сравнивались наши результаты, использовалась гораздо более сложная разностная схема на динамически адаптирующейся встроенной сетке, эквивалентная равномерной сетке $16\,000 \times 800$ ячеек, т.е. на порядок более подробная, чем в настоящей работе. Для оценки эффективности работы алгоритма укажем, что, например, расчет задачи о тяжелой капле на подробной сетке 2600×356 с числом точек порядка миллиона ячеек занимал около 10 ч процессорного времени на персональном компьютере.

ЗАКЛЮЧИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

В работе представлен эффективный алгоритм для численного моделирования течений газовых смесей в отсутствие химических реакций. Регуляризованные (или КГД) уравнения, лежащие в основе алгоритма, получены феноменологическим путем, и форма их записи близка к соответствующим регуляризованным уравнениям для однокомпонентного газа. Построенный алгоритм позволяет получить тонкие эффекты, возникающие при нестационарном взаимодействии ударной волны с пузырьками легкого и тяжелого газов. На основе представленного алгоритма моделировались также задачи о гравитационной неустойчивости Рэлея–Тейлора в широком диапазоне чисел Атвуда (см. [15]). Таким образом, вычислительная практика показала высокую работоспособность представленного здесь алгоритма.

Структура алгоритма позволяет естественным образом обобщить его на случай многокомпонентной смеси идеальных газов.

Близкое родство данного метода с КГД-алгоритмами расчета однокомпонентных газов является основой того, что представленный метод естественно обобщается на неструктурированные пространственные сетки, на течения с источниками, и может по аналогии с КГД-алгоритмом быть включен в открытый пакет OpenFOAM в качестве нового оригинального вычислительного ядра.

СПИСОК ЛИТЕРАТУРЫ

1. *Четверушкин Б.Н.* Кинетические схемы и квазигазодинамическая система уравнений. М.: МАКС Пресс, 2004.
2. *Елизарова Т.Г.* Квазигазодинамические уравнения и методы расчета вязких течений. М.: Научный мир, 2007. Перевод: *Elizarova TG.* Quasi-Gas Dynamic Equations. Springer, Berlin, 2009.
3. *Шеретов Ю.В.* Динамика сплошных сред при пространственно-временном осреднении. Москва–Ижевск: РХД, 2009. Перевод названия: *Sheretov Yu.* Continuum Dynamics under Spatiotemporal Averaging. Regular and Chaotic Dynamics, Moscow-Izhevsk, 2009 (in Russian).

4. *Kraposhin M.V., Ryazanov D.A., Smirnova E.V., Elizarova T.G., Istomina M.A.* Development of OpenFOAM solver for compressible viscous flows simulation using quasi-gas dynamic equations // IEEE eXplore. 2017. V. 29. № 6. P. 117–124.
5. *Kraposhin M.V., Smirnova E.V., Elizarova T.G., Istomina M.A.* Development of a new OpenFOAM solver using regularized gas dynamic equations // Comput. Fluids. 2018. V. 166. P. 163–175.
6. OpenFOAM User Guide, version 7. https://cfd.direct/openfoam/user-guide/
7. *Quirk J., Karni S.* On the dynamics of a shock-bubble interaction // J. Fluid Mech. 1996. V. 318. P. 129–163.
8. *Abgrall R.* How to prevent pressure oscillations in multicomponent flow calculations: a quasi conservative approach // J. Comput. Phys. 1996. V. 125. P. 150–160.
9. *Abgrall R., Karni S.* Computations of compressible multifluids // J. Comput. Phys. 2001. V. 169. P. 594–623.
10. *Billet G., Abgrall R.* An adaptive shock-capturing algorithm for solving unsteady reactive flows // Comput. Fluids. 2003. V. 32. P. 1473–1495.
11. *Billet G., Ryan J.* A Runge–Kutta discontinuous Galerkin approach to solve reactive flows: The hyperbolic operator // J. Comput. Phys. 2011. V. 230. P. 1064–1083.
12. *Saurel R., Abgrall R.* A multiphase Godunov method for compressible multifluid and multiphase flows // J. Comput. Phys. 1999. V. 150. P. 425–467.
13. *Borisov V.E., Rykov Yu.G.* Modified Godunov method for multicomponent flow simulation // J. Phys.: Conf. Ser. 2019. N1250. P. 012006.
14. *Lian Y.S., Xu K.* A Gas-kinetic scheme for multimaterial flows and its application in chemical reactions // J. Comput. Phys. 2000. V. 163. P. 349–375.
15. *Shilnikov E.V., Elizarova T.G.* Numerical simulation of gravity instabilities in gas flows by use of the quasi gas dynamic equation system // IOP Conf. Ser.: Materials Science and Engng. 2019. V. 657. P. 012035.
16. *Elizarova T.G., Graur I.A., Lengrand J.C.* Two-fluid computational model for a binary gas mixture // Europ. J. Mech. B. 2001. V. 3. P. 351–369.
17. *Елизарова Т.Г., Злотник А.А., Четверушкин Б.Н.* О квазигазо- и гидродинамических уравнениях бинарных смесей газов // Докл. АН. 2014. Т. 459. № 4. С. 395–399.
18. *Elizarova T.G., Lengrand J.-C.* Free parameters for the modelization of nonmonatomic binary gas mixtures // J. Thermophys. Heat Transfer. 2016. V. 30. № 3. P. 695–697.
19. *Elizarova T.G., Zlotnik A.A., Shil'nikov E.V.* Regularized equations for numerical simulation of flows of homogeneous binary mixtures of viscous compressible gases // Comput. Math. and Math. Phys. 2019. V. 59. № 11. P. 1832–1847.
20. *Балашов В.А., Савенков Е.Б.* Многокомпонентная квазигидродинамическая модель для описания течений многофазной жидкости с учетом межфазного взаимодействия // Прикл. механ. техн. физ. 2018. Т. 59. № 3. С. 57–68.
21. *Haas J.F., Sturtevant B.* Interaction of weak shock waves with cylindrical and spherical gas inhomogeneities // J. Fluid Mech. 1987. V. 181. P. 41–76.
22. *Иванов И.Э., Крюков И.А.* Численное моделирование течений многокомпонентного газа с сильными разрывами свойств среды // Матем. моделирование. 2007. Т. 19. № 12. С. 89–100.
23. *Борисов В.Е., Рыков Ю.Г.* Численное моделирование течений многокомпонентных газовых смесей с использованием метода двойного потока // Матем. моделирование. 2020. Т. 32. № 10. С. 3–20.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 136–149

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 532.51

СПЕКТРАЛЬНЫЙ АНАЛИЗ ОПТИМАЛЬНЫХ ВОЗМУЩЕНИЙ СТРАТИФИЦИРОВАННОГО ТУРБУЛЕНТНОГО ТЕЧЕНИЯ КУЭТТА[1)]

© 2021 г. Г. В. Засько[1,*], Ю. М. Нечепуренко[1,2,]**

[1] *119333 Москва, ул. Губкина, 8, Отделение Московского центра фундаментальной и прикладной математики в ИВМ им. Г.И. Марчука РАН, Россия*

[2] *119333 Москва, ул. Губкина, 8, ИВМ им. Г.И. Марчука РАН, Россия*

**e-mail: zasko.gr@bk.ru*

***e-mail: yumnech@yandex.ru*

Поступила в редакцию 10.12.2019 г.
Переработанный вариант 27.07.2020 г.
Принята к публикации 18.09.2020 г.

Рассматриваются собственные моды и оптимальные возмущения уравнений стратифицированного турбулентного течения Куэтта, осредненных по горизонтальным пространственным переменным и линеаризованных относительно стационарного состояния. Установлено, что спектр таких уравнений симметричен относительно вещественной оси и лежит строго в левой полуплоскости, т.е. все собственные моды устойчивые, а главная часть оптимального возмущения представляет собой линейную комбинацию большого числа мод, отвечающих собственным значениям с наибольшими вещественными частями. При этом число наиболее значимых мод в этой линейной комбинации растет с ростом числа Рейнольдса. Библ. 20. Фиг. 5. Табл. 2.

Ключевые слова: стратифицированное турбулентное течение Куэтта, мелкомасштабная турбулентность, крупномасштабные структуры, собственные моды, максимальная амплификация, оптимальные возмущения.

DOI: 10.31857/S0044466921010105

1. ВВЕДЕНИЕ

В геофизических пограничных слоях на фоне мелкомасштабной турбулентности часто наблюдаются крупномасштабные структуры. Они вносят существенный вклад в обмены импульсом, теплом и влагой между свободной атмосферой и подстилающей поверхностью (см. [1]). Примером таких структур являются слоистые структуры в поле температуры, наблюдаемые в природе и при прямом (DNS) или вихреразрешающем (LES) численном моделировании устойчиво-стратифицированной атмосферной турбулентности (см. [2]–[4]).

В [3], [5] было проведено DNS-моделирование стратифицированного турбулентного течения Куэтта. Это модельное течение близко к турбулентным течениям в пограничных слоях атмосферы и океана. Установлено, что при больших числах Рейнольдса, в широком диапазоне статической устойчивости, характеризуемой различными значениями числа Ричардсона, наряду с хаотической турбулентностью течение Куэтта содержит крупные структуры, которые могут быть выделены из результатов DNS-моделирования путем разложения мгновенных полей в ряды Фурье по горизонтальным переменным и отбора крупномасштабных гармоник. В случаях, близких к нейтральной стратификации, эти структуры представляют собой крупномасштабные вихри приблизительно круглой формы в поперечном сечении канала, а в случае устойчивой стратификации – крупномасштабные наклонные слои в поле температуры в продольном сечении канала.

Иногда возникновение крупномасштабных структур удается объяснить гидродинамической неустойчивостью осредненного течения (см. [6], [7]). Однако в случае течения Куэтта среднее течение устойчиво. В [8]–[10] образование крупномасштабных структур связывается с возникновением и развитием в мелкомасштабном турбулентном потоке оптимальных возмущений. Опти-

[1)] Работа выполнена при финансовой поддержке Московского Центра фундаментальной и прикладной математики (соглашение с Минобрнауки России № 075-15-2019-1624).

мальные возмущения вычислялись с помощью технологии, предложенной в [11], [12], на основе уравнений турбулентного течения, осредненных по горизонтальным пространственным переменным и линеаризованных относительно стационарного состояния. Качественное и количественное сравнение крупномасштабных структур с соответствующими им по волновым числам оптимальными возмущениями показало совпадение их пространственных масштабов и конфигураций.

Данная работа посвящена численному исследованию собственных мод и оптимальных возмущений уравнений турбулентного течения, осредненных по горизонтальным пространственным переменным и линеаризованных относительно стационарного состояния. Установлено, что спектр таких уравнений симметричен относительно вещественной оси и лежит строго в левой полуплоскости, т.е. все собственные моды устойчивые, а глобальные оптимальные возмущения представляют собой линейную комбинацию большого числа собственных мод, отвечающих собственным значениям из ведущей части спектра. Причем число значимых мод в этой линейной комбинации растет с ростом числа Рейнольдса.

2. ОСРЕДНЕННЫЕ УРАВНЕНИЯ ТУРБУЛЕНТНОГО ТЕЧЕНИЯ

Рассмотрим в декартовых координатах x (продольная), y (вертикальная), z (поперечная) турбулентное течение вязкой несжимаемой жидкости в поле силы тяжести в бесконечном трехмерном канале

$$\{-\infty < x < +\infty,\ -h < y < h,\ -\infty < z < +\infty\}$$

полувысоты h, верхняя стенка которого движется со скоростью $(U_0/2, 0, 0)$, нижняя – со скоростью $(-U_0/2, 0, 0)$. На стенках поддерживаются постоянные значения температуры T_2 и $T_1 < T_2$ соответственно. Определим числа Рейнольдса, Ричардсона и Прандтля как

$$\mathrm{Re} = U_0 h/\nu, \quad \mathrm{Ri} = g(T_2 - T_1)U_0/T_1 h^2, \quad \mathrm{Pr} = \nu/\mu,$$

где g – ускорение свободного падения, ν и μ – кинематическая вязкость и теплопроводность соответственно.

Следуя [8]–[10], будем считать, что эволюция крупномасштабных составляющих течения в нормированных переменных определяется следующей системой уравнений:

$$\begin{gathered}
\frac{\partial U}{\partial t} + \frac{\partial U^2}{\partial x} + \frac{\partial UV}{\partial y} + \frac{\partial UW}{\partial z} + \frac{\partial P}{\partial x} - \left(\bar{\nu}\frac{\partial^2 U}{\partial x^2} + \frac{\partial}{\partial y}\bar{\nu}\frac{\partial U}{\partial y} + \bar{\nu}\frac{\partial^2 U}{\partial z^2}\right) = 0,\\
\frac{\partial V}{\partial t} + \frac{\partial UV}{\partial x} + \frac{\partial V^2}{\partial y} + \frac{\partial VW}{\partial z} + \frac{\partial P}{\partial y} - \left(\bar{\nu}\frac{\partial^2 V}{\partial x^2} + \frac{\partial}{\partial y}\bar{\nu}\frac{\partial V}{\partial y} + \bar{\nu}\frac{\partial^2 V}{\partial z^2}\right) - \mathrm{Ri}\,T = 0,\\
\frac{\partial W}{\partial t} + \frac{\partial UW}{\partial x} + \frac{\partial VW}{\partial y} + \frac{\partial W^2}{\partial z} + \frac{\partial P}{\partial z} - \left(\bar{\nu}\frac{\partial^2 W}{\partial x^2} + \frac{\partial}{\partial y}\bar{\nu}\frac{\partial W}{\partial y} + \bar{\nu}\frac{\partial^2 W}{\partial z^2}\right) = 0,\\
\frac{\partial T}{\partial t} + \frac{\partial UT}{\partial x} + \frac{\partial VT}{\partial y} + \frac{\partial WT}{\partial z} - \left(\bar{\mu}\frac{\partial^2 T}{\partial x^2} + \frac{\partial}{\partial y}\bar{\mu}\frac{\partial T}{\partial y} + \bar{\mu}\frac{\partial^2 T}{\partial z^2}\right) = 0,\\
\frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} + \frac{\partial W}{\partial z} = 0,
\end{gathered} \tag{2.1}$$

где взаимодействие с мелкомасштабной турбулентностью параметризовано с помощью коэффициентов турбулентной вязкости $\bar{\nu}$ и теплопроводности $\bar{\mu}$, зависящих только от вертикальной координаты y. Здесь x, y, z – нормированные декартовы координаты; U, V, W, T, P – нормированные компоненты вектора скорости, температура и удельное давление соответственно. Верхняя и нижняя стенки канала движутся со скоростями $1/2$ и $-1/2$ в направлении x, и на них поддерживаются постоянные значения температуры 2 и 1 соответственно.

Система уравнений (2.1) имеет стационарное решение вида

$$U = \bar{U}(y), \quad V = 0, \quad W = 0, \quad T = \bar{T}(y), \quad P = \bar{P}(y) \tag{2.2}$$

с профилями продольной компоненты скорости, давления и температуры, удовлетворяющими соотношениям

$$\bar{\nu}\frac{d\bar{U}}{dy} = \text{const}, \quad \bar{\mu}\frac{d\bar{T}}{dy} = \text{const}, \quad \frac{d\bar{P}}{dy} = \text{Ri}\,\bar{T}(y). \tag{2.3}$$

Это течение мы далее будем называть основным.

Представим произвольное решение системы (2.2) в окрестности основного течения в виде

$$\begin{gathered} U = \bar{U} + \delta u' + o(\delta), \quad V = \delta v' + o(\delta), \quad W = \delta w' + o(\delta), \\ T = \bar{T} + \delta T' + o(\delta), \quad P = \bar{P} + \delta p' + o(\delta), \end{gathered} \tag{2.4}$$

где δ – малый параметр. Требуя, чтобы для любого сколь угодно малого δ (2.4) было решением системы (2.2), получаем следующие уравнения распространения малых возмущений:

$$\begin{gathered} \frac{\partial u'}{\partial t} + \bar{U}\frac{\partial u'}{\partial x} + \frac{d\bar{U}}{dy}v' + \frac{\partial p'}{\partial x} - \left(\bar{\nu}\frac{\partial^2 u'}{\partial x^2} + \frac{\partial}{\partial y}\bar{\nu}\frac{\partial u'}{\partial y} + \bar{\nu}\frac{\partial^2 u'}{\partial z^2}\right) = 0, \\ \frac{\partial v'}{\partial t} + \bar{U}\frac{\partial v'}{\partial x} + \frac{\partial p'}{\partial y} - \left(\bar{\nu}\frac{\partial^2 v'}{\partial x^2} + \frac{\partial}{\partial y}\bar{\nu}\frac{\partial v'}{\partial y} + \bar{\nu}\frac{\partial^2 v'}{\partial z^2}\right) - \text{Ri}\,T' = 0, \\ \frac{\partial w'}{\partial t} + \bar{U}\frac{\partial w'}{\partial x} + \frac{\partial p'}{\partial z} - \left(\bar{\nu}\frac{\partial^2 w'}{\partial x^2} + \frac{\partial}{\partial y}\bar{\nu}\frac{\partial w'}{\partial y} + \bar{\nu}\frac{\partial^2 w'}{\partial z^2}\right) = 0, \\ \frac{\partial T'}{\partial t} + \bar{U}\frac{\partial T'}{\partial x} + \frac{d\bar{T}}{dy}v' - \left(\bar{\mu}\frac{\partial^2 T'}{\partial x^2} + \frac{\partial}{\partial y}\bar{\mu}\frac{\partial T'}{\partial y} + \bar{\mu}\frac{\partial^2 T'}{\partial z^2}\right) = 0, \\ \frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} + \frac{\partial w'}{\partial z} = 0. \end{gathered} \tag{2.5}$$

Уравнения (2.5) рассматриваются с нулевыми граничными условиями для u', v', w', T' при $y = \pm 1$.

Нас будут интересовать периодические по x и z возмущения вида

$$(u', v', w', T', p') = \text{Real}\{(u_{\alpha\gamma}, v_{\alpha\gamma}, w_{\alpha\gamma}, T_{\alpha\gamma}, p_{\alpha\gamma})e^{\text{i}\alpha x + \text{i}\gamma z}\}, \tag{2.6}$$

где α и γ – вещественные продольное и поперечное волновые числа соответственно, а $f_{\alpha\gamma}$ – комплексные амплитуды, зависящие только от y и t. Амплитуды таких возмущений удовлетворяют системе уравнений

$$\begin{gathered} \frac{\partial u_{\alpha\gamma}}{\partial t} + \text{i}\alpha\bar{U}u_{\alpha\gamma} + \frac{d\bar{U}}{dy}v_{\alpha\gamma} + \text{i}\alpha p_{\alpha\gamma} + \left(\alpha^2\bar{\nu}u_{\alpha\gamma} - \frac{\partial}{\partial y}\bar{\nu}\frac{\partial u_{a\gamma}}{\partial y} + \gamma^2\bar{\nu}u_{\alpha\gamma}\right) = 0, \\ \frac{\partial v_{\alpha\gamma}}{\partial t} + \text{i}\alpha\bar{U}v_{\alpha\gamma} + \frac{\partial p_{\alpha\gamma}}{\partial y} + \left(\alpha^2\bar{\nu}v_{\alpha\gamma} - \frac{\partial}{\partial y}\bar{\nu}\frac{\partial v_{\alpha\gamma}}{\partial y} + \gamma^2\bar{\nu}v_{\alpha\gamma}\right) - \text{Ri}\,T_{\alpha\gamma} = 0, \\ \frac{\partial w_{\alpha\gamma}}{\partial t} + \text{i}\alpha\bar{U}w_{\alpha\gamma} + \text{i}\gamma p_{\alpha\gamma} + \left(\alpha^2\bar{\nu}w_{\alpha\gamma} - \frac{\partial}{\partial y}\bar{\nu}\frac{\partial w_{\alpha\gamma}}{\partial y} + \gamma^2\bar{\nu}w_{\alpha\gamma}\right) = 0, \\ \frac{\partial T_{\alpha\gamma}}{\partial t} + \text{i}\alpha\bar{U}T_{\alpha\gamma} + \frac{d\bar{T}}{dy}v_{\alpha\gamma} + \left(\alpha^2\bar{\mu}T_{\alpha\gamma} - \frac{\partial}{\partial y}\bar{\mu}\frac{\partial T_{\alpha\gamma}}{\partial y} + \gamma^2\bar{\mu}T_{\alpha\gamma}\right) = 0, \\ \text{i}\alpha u_{\alpha\gamma} + \frac{\partial v_{\alpha\gamma}}{\partial y} + \text{i}\gamma w_{\alpha\gamma} = 0. \end{gathered} \tag{2.7}$$

3. СОБСТВЕННЫЕ МОДЫ И ОПТИМАЛЬНЫЕ ВОЗМУЩЕНИЯ

Решение вида

$$(u_{\alpha\gamma}, v_{\alpha\gamma}, w_{\alpha\gamma}, T_{\alpha\gamma}, p_{\alpha\gamma}) = (\tilde{u}_{\alpha\gamma}, \tilde{v}_{\alpha\gamma}, \tilde{w}_{\alpha\gamma}, \tilde{T}_{\alpha\gamma}, \tilde{p}_{\alpha\gamma})e^{\lambda t} \tag{3.1}$$

системы (2.7), где λ – комплексное число, а $\tilde{f}_{\alpha\gamma}$ – комплексные амплитуды, зависящие только от y, называется ее собственной модой, отвечающей собственному значению λ. Подставляя (3.1) в (2.7), получаем следующую проблему собственных значений:

$$\begin{gathered}
\lambda\tilde{u}_{\alpha\gamma} + i\alpha\bar{U}\tilde{u}_{\alpha\gamma} + \frac{d\bar{U}}{dy}\tilde{v}_{\alpha\gamma} + i\alpha\tilde{p}_{\alpha\gamma} + \left(\alpha^2\bar{\nu}\tilde{u}_{\alpha\gamma} - \frac{d}{dy}\bar{\nu}\frac{d\tilde{u}_{\alpha\gamma}}{dy} + \gamma^2\bar{\nu}\tilde{u}_{\alpha\gamma}\right) = 0,\\
\lambda\tilde{v}_{\alpha\gamma} + i\alpha\bar{U}\tilde{v}_{\alpha\gamma} + \frac{d\tilde{p}_{\alpha\gamma}}{dy} + \left(\alpha^2\bar{\nu}\tilde{v}_{\alpha\gamma} - \frac{d}{dy}\bar{\nu}\frac{d\tilde{v}_{\alpha\gamma}}{dy} + \gamma^2\bar{\nu}\tilde{v}_{\alpha\gamma}\right) - \mathrm{Ri}\,\tilde{T}_{\alpha\gamma} = 0,\\
\lambda\tilde{w}_{\alpha\gamma} + i\alpha\bar{U}\tilde{w}_{\alpha\gamma} + i\gamma\tilde{p}_{\alpha\gamma} + \left(\alpha^2\bar{\nu}\tilde{w}_{\alpha\gamma} - \frac{d}{dy}\bar{\nu}\frac{d\tilde{w}_{\alpha\gamma}}{dy} + \gamma^2\bar{\nu}\tilde{w}_{\alpha\gamma}\right) = 0,\\
\lambda\tilde{T}_{\alpha\gamma} + i\alpha\bar{U}\tilde{T}_{\alpha\gamma} + \frac{d\bar{T}}{dy}\tilde{v}_{\alpha\gamma} + \left(\alpha^2\bar{\mu}\tilde{T}_{\alpha\gamma} - \frac{d}{dy}\bar{\mu}\frac{d\tilde{T}_{\alpha\gamma}}{dy} + \gamma^2\bar{\mu}\tilde{T}_{\alpha\gamma}\right) = 0,\\
i\alpha\tilde{u}_{\alpha\gamma} + \frac{d\tilde{v}_{\alpha\gamma}}{dy} + i\gamma\tilde{w}_{\alpha\gamma} = 0.
\end{gathered} \tag{3.2}$$

Теорема 1. *Пусть $\bar{U}(y)$, $\bar{T}(y)$, $\bar{\nu}(y)$, $\bar{\mu}(y)$ удовлетворяют соотношениям* (2.3), *причем профили скорости и температуры являются нечетными функциями. Тогда спектр проблемы собственных значений* (3.2) *симметричен относительно вещественной оси, и если*

$$(\lambda, \tilde{u}_{\alpha\gamma}(y), \tilde{v}_{\alpha\gamma}(y), \tilde{w}_{\alpha\gamma}(y), \tilde{T}_{\alpha\gamma}(y), \tilde{p}_{\alpha\gamma}(y))$$

является решением проблемы (3.2), *то*

$$(\lambda^*, \tilde{u}_{\alpha\gamma}(-y)^*, \tilde{v}_{\alpha\gamma}(-y)^*, \tilde{w}_{\alpha\gamma}(-y)^*, \tilde{T}_{\alpha\gamma}(-y)^*, -\tilde{p}_{\alpha\gamma}(-y)^*)$$

также является решением проблемы (3.2), *где $*$ означает комплексное сопряжение*.

Отметим, что приведенная теорема доказывается комплексным сопряжением уравнений (3.2) и заменой переменных $y \to -y$. Она является непосредственным обобщением соответствующего утверждения, справедливого для классического течения Куэтта (см. [13]), т.е. ламинарного течения с $\mathrm{Ri} = 0$ и $\bar{\nu} = 1/\mathrm{Re} = \mathrm{const}$. Классическое течение Куэтта имеет линейный профиль продольной компоненты скорости $\bar{U}(y) = y/2$. Известно, что оно устойчиво при любом числе Рейнольдса, т.е. спектр, соответствующий проблеме (3.2), лежит строго в левой полуплоскости [14]. Доказательством последнего утверждения для проблемы (3.2) в условиях теоремы 1 мы не располагаем, но, как показывают приведенные ниже результаты расчетов, оно по-видимому верно.

Обозначим через $r_{\max}^{\alpha\gamma}$ максимальную вещественную часть собственных значений проблемы (3.2) при фиксированных значениях волновых чисел α, γ, а через

$$r_{\max} = \max_{\alpha,\gamma} r_{\max}^{\alpha\gamma}$$

глобальную максимальную вещественную часть соответственно, где максимум берется по всем вещественным α и γ. Собственную моду вида (3.1), на которой достигается $r_{\max}$, будем далее называть *глобальной ведущей*.

Определим среднюю плотность полной энергии возмущения вида (2.6) в момент времени t как

$$\mathscr{E}_t = \frac{1}{2}\int_{-1}^{1}\left(|u_{\alpha\gamma}|^2 + |v_{\alpha\gamma}|^2 + |w_{\alpha\gamma}|^2 + \frac{\mathrm{Ri}}{d\bar{T}/dy}|T_{\alpha\gamma}|^2\right)dy. \tag{3.3}$$

Максимально возможное увеличение

$$\Gamma^{\alpha\gamma}(t) = \max\frac{\mathscr{E}_t}{\mathscr{E}_0}$$

средней плотности полной энергии возмущения, где максимум берется по всем решениям системы (2.7), будем называть *максимальной амплификацией средней плотности* полной энергии при фиксированных значениях α, γ и t. Введем обозначения

$$\Gamma_{\max}^{\alpha\gamma} = \max_{t\geq 0} \Gamma^{\alpha\gamma}(t), \quad \Gamma_{\max} = \max_{\alpha,\gamma} \Gamma_{\max}^{\alpha\gamma}, \quad (\alpha_{\text{opt}}, \gamma_{\text{opt}}, t_{\text{opt}}) = \arg\max_{\alpha,\gamma,t} \Gamma^{\alpha\gamma}(t)$$

для максимальной амплификации средней плотности полной энергии при фиксированных волновых числах, глобальной максимальной амплификации и оптимальных значений волновых чисел и оптимального момента времени соответственно. Начальное возмущение, на котором достигается $\Gamma_{\max}$, будем называть *глобальным оптимальным возмущением*.

Для классического течения Куэтта известно (см. [15]), что глобальная максимальная амплификация достигается при значении продольного волнового числа α, близкого к нулю, а глобальное оптимальное возмущение представляет собой крупномасштабные чередующиеся по направлению вращения вихри в поперечном сечении канала.

4. ЧИСЛЕННАЯ МОДЕЛЬ

Следуя работам [8]–[10], пространственную аппроксимацию по y системы уравнений (2.7) будем выполнять методом Галеркина-коллокаций, описанным в [16], [17]. В качестве узлов сетки для давления выберем корни $y_1 < \cdots < y_n$ производной L'_{n+1} многочлена Лежандра степени $n+1$, а в качестве узлов сетки для компонент скорости и температуры – те же узлы вместе с точками $y_0 = -1$ и $y_{n+1} = 1$ (узлы Гаусса–Лобатто). В качестве базисных функций для компонент скорости и температуры с учетом нулевых граничных условий будем использовать элементарные интерполяционные многочлены Лагранжа, представимые в виде

$$\psi_i(y) = \frac{(y^2-1)L'_{n+1}(y)}{(n+1)(n+2)(y-y_i)L_{n+1}(y_i)},$$

а для давления – элементарные интерполяционные многочлены Лагранжа, представимые в виде

$$\varphi_i(y) = \frac{(y_i^2-1)L'_{n+1}(y)}{(n+1)(n+2)(y-y_i)L_{n+1}(y_i)}.$$

Таким образом, компоненты амплитуд возмущения будем аппроксимировать как

$$g(y,t) \approx \sum_{i=1}^{n} g_i(t)\psi_i(y), \quad p_{\alpha\gamma}(y,t) \approx \sum_{i=1}^{n} p_{\alpha\gamma,i}(t)\varphi_i(y),$$

где g означает $u_{\alpha\gamma}$, $v_{\alpha\gamma}$, $w_{\alpha\gamma}$ или $T_{\alpha\gamma}$. Коэффициенты $g_i(t)$, $p_i(t)$ при такой аппроксимации являются значениями аппроксимантов в узле y_i.

В качестве пробных функций для каждого из первых четырех уравнений в (2.7) будем использовать функции $\psi_i(y)$, а для уравнения неразрывности – $\varphi_i(y)$. Для расчета фигурирующих в слабой постановке скалярных произведений будем использовать квадратурную формулу с узлами и весами Гаусса–Лобатто (см. [16]):

$$\int_{-1}^{1} f(y)dy \approx \sum_{k=0}^{n+1} \kappa_i f(y_i), \quad \kappa_i = \frac{2}{(n+1)(n+2)L_{n+1}^2(y_i)},$$

точную для многочленов от y степени не выше $2n+1$.

Обозначим через K_0 положительно определенную диагональную матрицу порядка $n+2$ квадратурных коэффициентов, а через K – ее подматрицу порядка n, отвечающую внутренним узлам. Введем также следующие диагональные матрицы порядка n: U, U_y, N, M и T_y, составленные соответственно из значений профиля $\overline{U}$ и производной $d\overline{U}/dy$ профиля продольной компоненты скорости основного течения, значений коэффициентов турбулентной вязкости $\overline{\nu}$ и диффузии $\overline{\mu}$ и значений производной $d\overline{T}/dy$ профиля температуры основного течения в узлах $y_1 < \ldots < y_n$, а также диагональные матрицы N_0, M_0 порядка $n+2$, составленные из значений коэффициентов турбулентной вязкости и диффузии в узлах сетки $y_0 < \ldots < y_{n+1}$. Для вычисления значений в узлах

$y_0 < \ldots < y_{n+1}$ производной функции, заданной в узлах $y_1 < \ldots < y_n$ и удовлетворяющей нулевым граничным условиям, будем использовать матрицу дифференцирования D размера $(n+2)\times n$. Так же нам потребуется матрица проектирования P размера $(n+2)\times n$, восстанавливающая по значениям функции в узлах $y_1 < \ldots < y_n$ ее значения в узлах $y_0 < \ldots < y_{n+1}$. Эффективные методы вычисления матриц D и P на основе интерполяционных многочленов Лагранжа описаны в [18].

Выполнив дискретизацию системы уравнений (2.7) методом галеркина-коллокаций, получим систему обыкновенных дифференциальных и алгебраических уравнений, которую можно привести к следующему виду:

$$\frac{d\mathbf{v}}{dt} = J\mathbf{v} - Gp, \quad F\mathbf{v} = 0, \tag{4.1}$$

где

$$\mathbf{v} = E^{1/2}(u^{\mathrm{T}}, v^{\mathrm{T}}, w^{\mathrm{T}}, T^{\mathrm{T}})^{\mathrm{T}}, \quad E = \frac{1}{2}\operatorname{diag}(K, K, K, \operatorname{Ri} KT_y^{-1}),$$

а u, v, w, T, p суть n-компонентные векторные функции, компонентами которых являются соответственно значения амплитуд $u_{\alpha\gamma}$, $v_{\alpha\gamma}$, $w_{\alpha\gamma}$, $T_{\alpha\gamma}$, $p_{\alpha\gamma}$ в узлах $y_1 < \ldots < y_n$. Отметим, что при выбранной нормировке дискретным аналогом функционала (3.3) средней плотности полной энергии будет $\|\mathbf{v}\|_2^2$.

Матрицы в (4.1) устроены следующим образом:

$$J = \begin{bmatrix} S_{\nu} & -U_y & 0 & 0 \\ 0 & S_{\nu} & 0 & R \\ 0 & 0 & S_{\nu} & 0 \\ 0 & -R & 0 & S_{\mu} \end{bmatrix}, \quad G = \begin{bmatrix} \mathrm{i}\alpha I \\ G \\ \mathrm{i}\gamma I \\ 0 \end{bmatrix}, \quad F = -G^*,$$

где I – единичная матрица порядка n, $G = -K^{-1/2}D^{\mathrm{T}}K_0PK^{-1/2}$,

$$S_{\nu} = -\mathrm{i}\alpha U - \alpha^2 N + L_{\nu} - \gamma^2 N, \quad S_{\mu} = T_y^{-1/2}(-\mathrm{i}\alpha U - \alpha^2 M + L_{\mu} - \gamma^2 M)T_y^{1/2},$$

$$L_{\nu} = -K^{-1/2}D^{\mathrm{T}}K_0N_0DK^{-1/2}, \quad L_{\mu} = -K^{-1/2}D^{\mathrm{T}}K_0M_0DK^{-1/2}.$$

Отметим, что матрицы G и F являются дискретными аналогами оператора $\partial/\partial y$ в градиенте давления и уравнении неразрывности соответственно, L_{ν} и L_{μ} – соответственно дискретные аналоги операторов

$$\frac{\partial}{\partial y}\overline{\nu}\frac{\partial}{\partial y}, \quad \frac{\partial}{\partial y}\overline{\mu}\frac{\partial}{\partial y},$$

а $R = \sqrt{\mathrm{Ri}}T_y^{1/2}$.

Из второго уравнения системы (4.1) следует, что ее решение принадлежит ядру матрицы F. После замены переменных $\mathbf{v} = V\mathbf{u}$, где V – прямоугольная матрица, столбцы которой образуют ортонормированный базис в ядре матрицы F, и умножения полученного уравнения слева на V^*, а также с учетом того, что $G = -F^*$, мы получим следующую систему обыкновенных дифференциальных уравнений:

$$\frac{d\mathbf{u}}{dt} = H\mathbf{u}, \tag{4.2}$$

где $H = V^*JV$ – квадратная комплексная матрица порядка $3n+1$ при $\alpha = \gamma = 0$ и порядка $3n$ в остальных случаях. Подробное обоснование такого типа редукций линейных дифференциально-алгебраических систем дано в [19].

Для вычисления вектора значений амплитуды $(u_{\alpha\gamma}, v_{\alpha\gamma}, w_{\alpha\gamma}, T_{\alpha\gamma})$ возмущения вида (2.6) в узлах расчетной сетки необходимо сделать обратную замену переменных

$$(u^{\mathrm{T}}, v^{\mathrm{T}}, w^{\mathrm{T}}, T^{\mathrm{T}})^{\mathrm{T}} = E^{-1/2}V\mathbf{u}. \tag{4.3}$$

5. ВЫЧИСЛЕНИЕ СОБСТВЕННЫХ МОД И ОПТИМАЛЬНЫХ ВОЗМУЩЕНИЙ

Каждой собственной паре $(\lambda, \tilde{\mathbf{u}})$ матрицы H системы (4.2) соответствует (с точностью до погрешности аппроксимации) собственная мода (3.1) исходной системы (2.7). Ее амплитуда в узлах расчетной сетки может быть вычислена по формуле (4.3) с $\mathbf{u} = \tilde{\mathbf{u}}$.

В силу унитарной инвариантности второй нормы $\|\mathbf{u}(t)\|_2^2 = \|\mathbf{v}(t)\|_2^2$. Следовательно, $\|\mathbf{u}(t)\|_2^2$ является дискретным аналогом средней плотности полной энергии соответствующего возмущения вида (2.6). Поскольку произвольное решение системы (4.2) представимо в виде

$$\mathbf{u}(t) = \exp\{tH\}\mathbf{u}^0,$$

с точностью до погрешности аппроксимации

$$\Gamma^{\alpha\gamma}(t) = \|\exp\{tH\}\|_2^2, \quad \Gamma^{\alpha\gamma}_{\max} = \max_{t\geq 0}\Gamma^{\alpha\gamma}(t).$$

Таким образом, вычисление максимальной амплификации средней плотности полной энергии возмущений сводится к вычислению для заданной квадратной комплексной матрицы H максимума функции $\Gamma^{\alpha\gamma}(t)$ при $t \geq 0$. Для нахождения $t = t^{\alpha\gamma}_{\text{opt}}$, дающего максимум $\Gamma^{\alpha\gamma}(t)$ с заданной относительной точностью, будем использовать эффективный алгоритм, предложенный в [12] и основанный на малоранговой аппроксимации. После того как $t^{\alpha\gamma}_{\text{opt}}$ найдено, вычисляем максимальное сингулярное число σ_{opt} и отвечающий ему правый сингулярный вектор $\mathbf{u}_{\text{opt}}$ матрицы $\exp\{t^{\alpha\gamma}_{\text{opt}}H\}$. Максимальная амплификация $\Gamma^{\alpha\gamma}_{\max}$ равна σ^2_{opt}, а амплитуда оптимального возмущения в узлах расчетной сетки может быть вычислена по формуле (4.3) с $\mathbf{u} = \mathbf{u}_{\text{opt}}$.

Пусть Λ означает некоторое изолированное подмножество спектра матрицы H, $\mathcal{U}$ – инвариантное подпространство, отвечающее Λ, а $P_{\mathcal{U}}$ – соответствующий спектральный проектор, т.е. проектор на $\mathcal{U}$, коммутирующий с матрицей H. Нас будет интересовать квадрат нормы

$$c_{\mathcal{U}}(t) = \|P_{\mathcal{U}}\exp\{tH\}\mathbf{u}_{\text{opt}}\|_2^2$$

проекции глобального оптимального возмущения на подпространство $\mathcal{U}$ в моменты времени $t = 0$ и $t^{\alpha\gamma}_{\text{opt}}$ и максимальная амплификация

$$\Gamma_{\mathcal{U},\max} = \max_{t\geq 0}\max_{\mathbf{w}\in\mathcal{U},\|\mathbf{w}\|_2=1}\|\exp\{tH\}\mathbf{w}\|_2^2$$

на этом подпространстве. Эти величины мы будем вычислять на основе разложения Шура (см. [20])

$$H = [Q_1, Q_2]\begin{bmatrix} S_1 & S_{12} \\ 0 & S_2 \end{bmatrix}\begin{bmatrix} Q_1^* \\ Q_2^* \end{bmatrix}, \tag{5.1}$$

где S_1 и S_2 – верхние треугольные матрицы, причем $\lambda(S_1) = \Lambda$, а $[Q_1, Q_2]$ – унитарная матрица, разбитая на два блока в соответствии с разбиением на блоки формы Шура. Максимальная амплификация на подпространстве $\mathcal{U}$ равна

$$\Gamma_{\mathcal{U},\max} = \max_{t\geq 0}\|\exp\{tS_1\}\|_2^2,$$

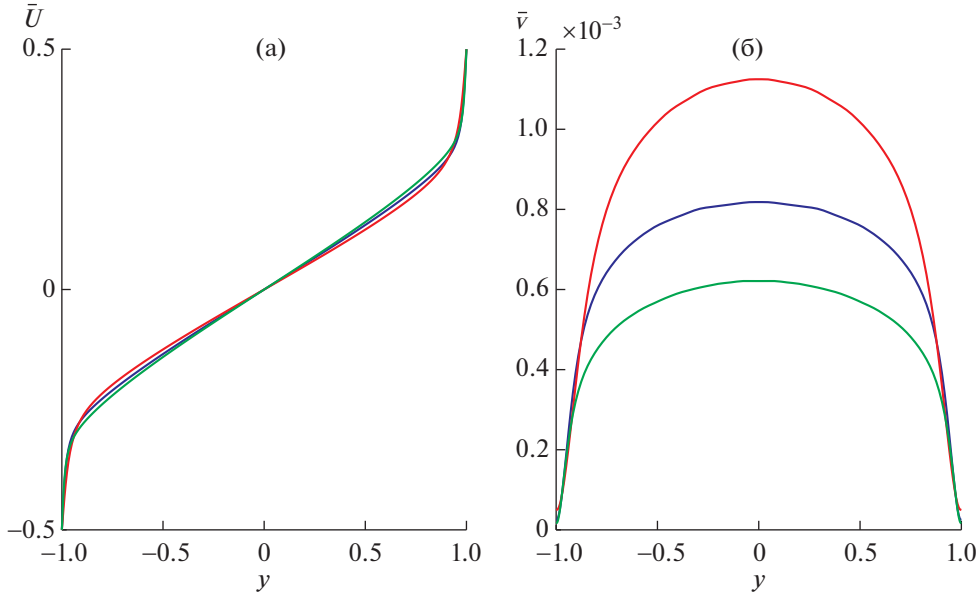
а

$$c_{\mathcal{U}}(t) = \left\|Q_1\exp\{tS_1\}(Q_1^* - MQ_2^*)\mathbf{u}_{\text{opt}}\right\|_2^2,$$

где M – решение уравнения Сильвестра

$$MS_2 - S_1M = S_{12}. \tag{5.2}$$

Отметим, что уравнение (5.2) однозначно разрешимо, так как спектры матриц S_1 и S_2 не пересекаются.



Фиг. 1. Профили $\bar{U}(y)$ (а) и $\bar{\nu}(y)$ (б) основного течения при Ri = 0.03 и $\mathrm{Re} = 2\times10^4$ (красным), 4×10^4 (синим) и 6×10^4 (зеленым).

6. РЕЗУЛЬТАТЫ РАСЧЕТОВ

Профили продольной компоненты скорости $\bar{U}(y)$ и температуры $\bar{T}(y)$ основного течения, а также соответствующие коэффициенты турбулентной вязкости $\bar{\nu}(y)$ и теплопроводности $\bar{\mu}(y)$ возьмем из результатов работы [5], где было выполнено прямое численное моделирование стратифицированного турбулентного течения Куэтта в широких диапазонах чисел Рейнольдса $1 \le \mathrm{Re}\times10^{-4} \le 6$ и Ричардсона $0 \le \mathrm{Ri} \le 0.12$. В данной работе мы ограничимся рассмотрением устойчиво-стратифицированного течения при фиксированном числе Ричардсона Ri = 0.03 и значениях числа Рейнольдса $\mathrm{Re} = 2\times10^4$, 4×10^4 и 6×10^4. Профили $\bar{U}(y)$ и $\bar{\nu}(y)$ при указанных значениях Ri и Re изображены на фиг. 1. Профили $\bar{T}(y)$ и $\bar{\mu}(y)$ выглядят аналогично.

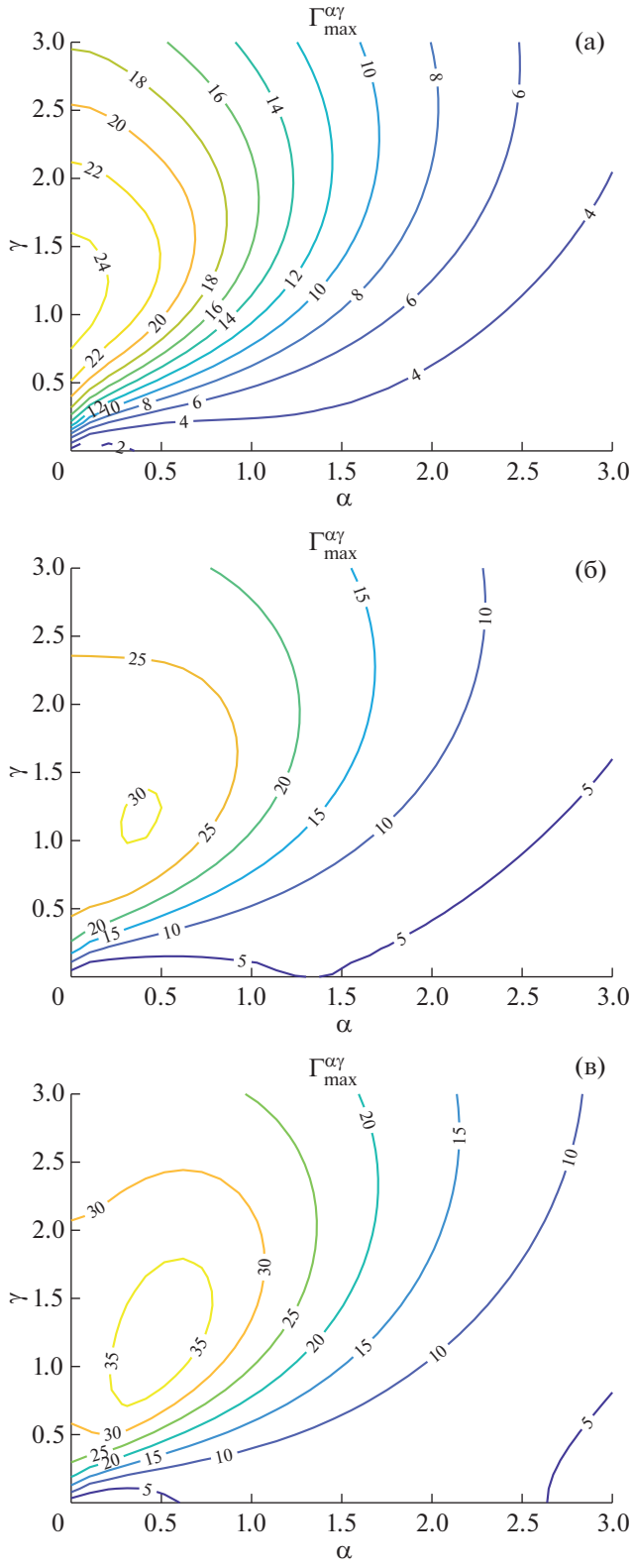
6.1. Сравнение ведущих мод и глобальных оптимальных возмущений

Линии уровня $\Gamma^{\alpha\gamma}_{\max}$ при рассматриваемых значениях чисел Ричардсона и Рейнольдса изображены на фиг. 2. Значения глобальной максимальной амплификации $\Gamma_{\max}$, оптимальных волновых чисел и оптимального времени $(\alpha_{\mathrm{opt}}, \gamma_{\mathrm{opt}}, t_{\mathrm{opt}})$ приведены в табл. 1. Видно, что с увеличением числа Рейнольдса наблюдается увеличение $\Gamma_{\max}$. Глобальная максимальная амплификация достигается при ненулевом поперечном волновом числе γ при всех рассмотренных числах Рейнольдса и ненулевом продольном волновом числе α при всех рассмотренных числах Рейнольдса, кроме минимального. Причем,

с увеличением числа Рейнольдса наблюдается увеличение оптимального продольного волнового числа α_{opt}, а оптимальное поперечное волновое число γ_{opt} почти не меняется.

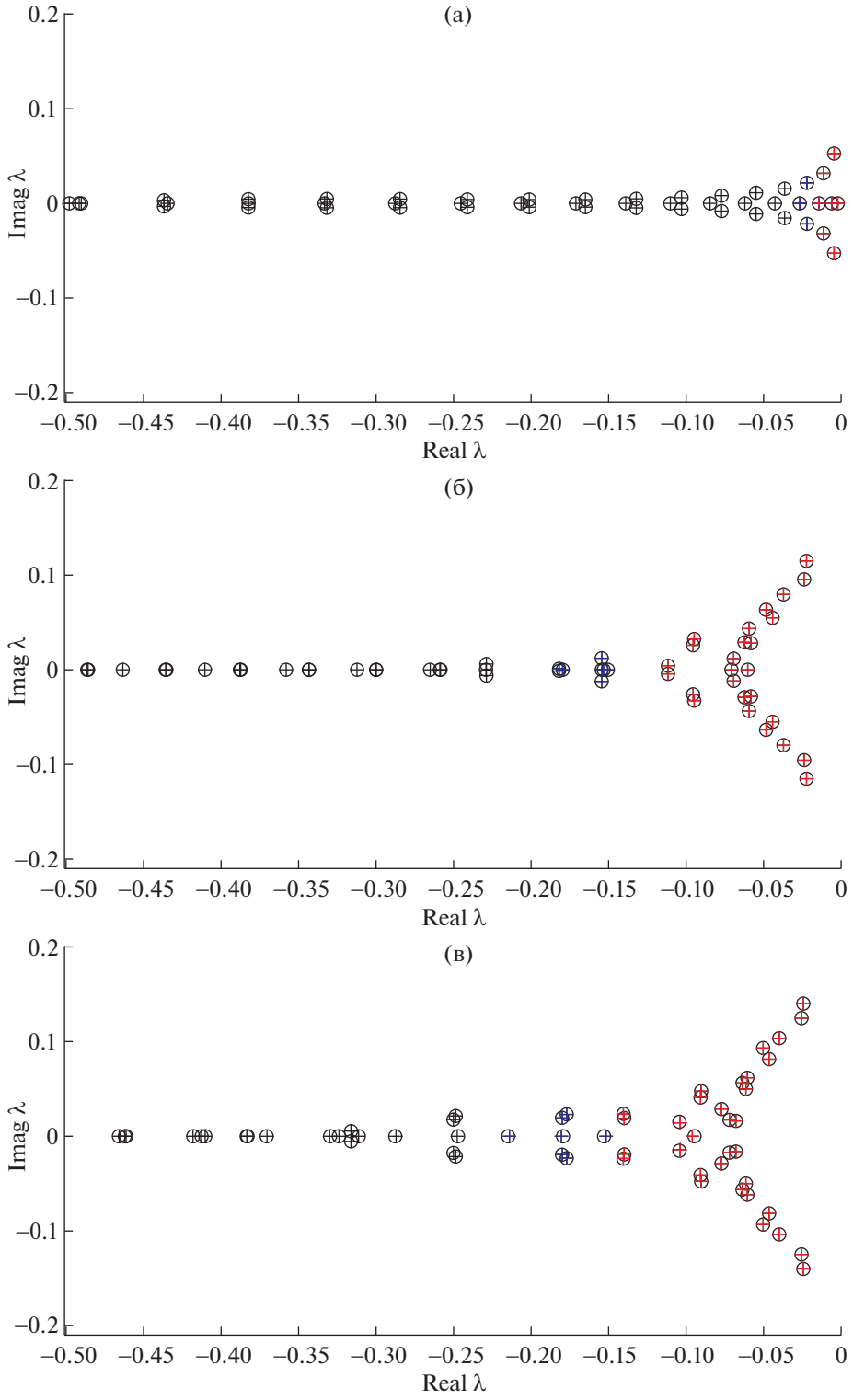
Таблица 1. Значения $\Gamma_{\max}$, $(\alpha_{\mathrm{opt}}, \gamma_{\mathrm{opt}}, t_{\mathrm{opt}})$ и $r_{\max}$ в зависимости от числа Рейнольдса Re при Ri = 0.03

Re	$\Gamma_{\max}$	$(\alpha_{\mathrm{opt}}, \gamma_{\mathrm{opt}}, t_{\mathrm{opt}})$	$r_{\max}$
2×10^4	24.8383	(0.0000, 1.1165, 57.6)	−0.000847
4×10^4	30.4001	(0.3890, 1.1606, 46.6)	−0.000683
6×10^4	38.0889	(0.4664, 1.1424, 46.8)	−0.000553



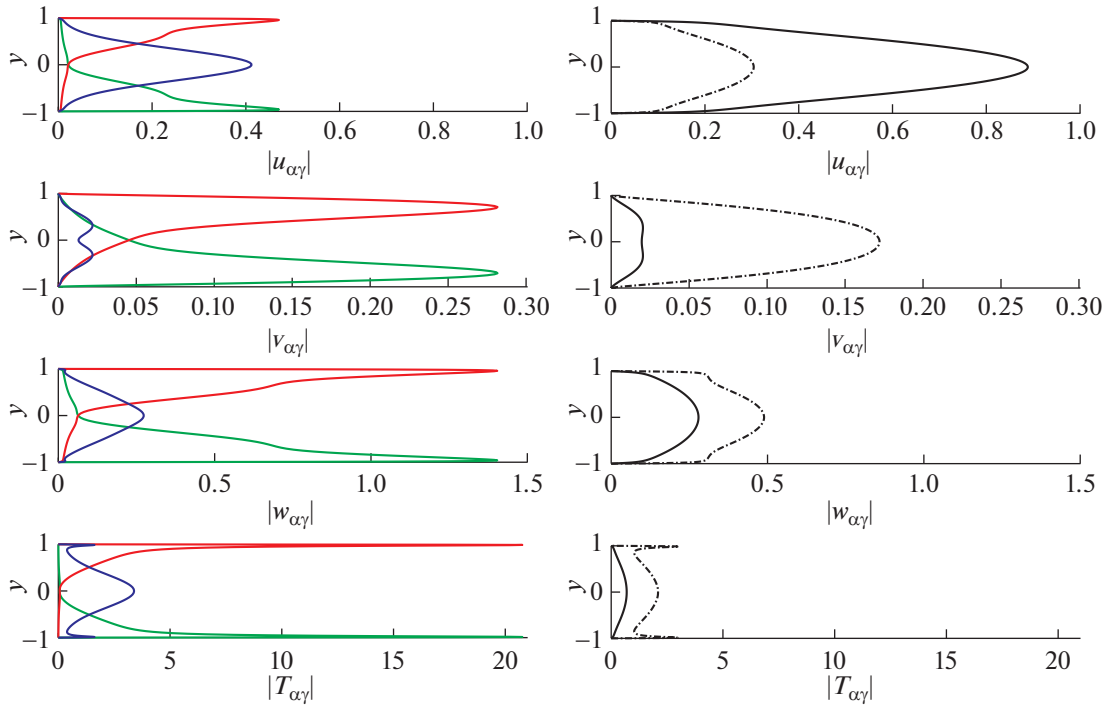
Фиг. 2. Линии уровня $\Gamma^{\alpha\gamma}_{\max}$ в плоскости (α, γ) при Ri = 0.03 и Re = 2×10^4 (а), 4×10^4 (б) и 6×10^4 (в).

Отметим, что согласно [9], [10], оптимальные возмущения при $\alpha = 0$ представляют собой ролики – крупномасштабные вихри приблизительно круглой формы в поперечном сечении канала, а оптимальные возмущения при $\alpha > 0$ – крупномасштабные наклонные слоистые структуры.



Фиг. 3. Ведущая часть спектра матрицы H при Ri = 0.03, Re = 2×10^4 (а), 4×10^4 (б), 6×10^4 (в) и оптимальных значениях волновых чисел $(\alpha_{opt}, \gamma_{opt})$, вычисленная на сетках с $n = 100$ (“+”) и 200 (“o”). Подмножества Λ_{11} и Λ_{12} выделены красным и синим цветами соответственно.

Причем ролики развиваются во времени заметно медленнее, чем слоистые структуры, что соответствует приведенным в табл. 1 оптимальным временам. Таким образом, с увеличением числа Рейнольдса наблюдается существенное изменение структуры глобального оптимального возмущения.



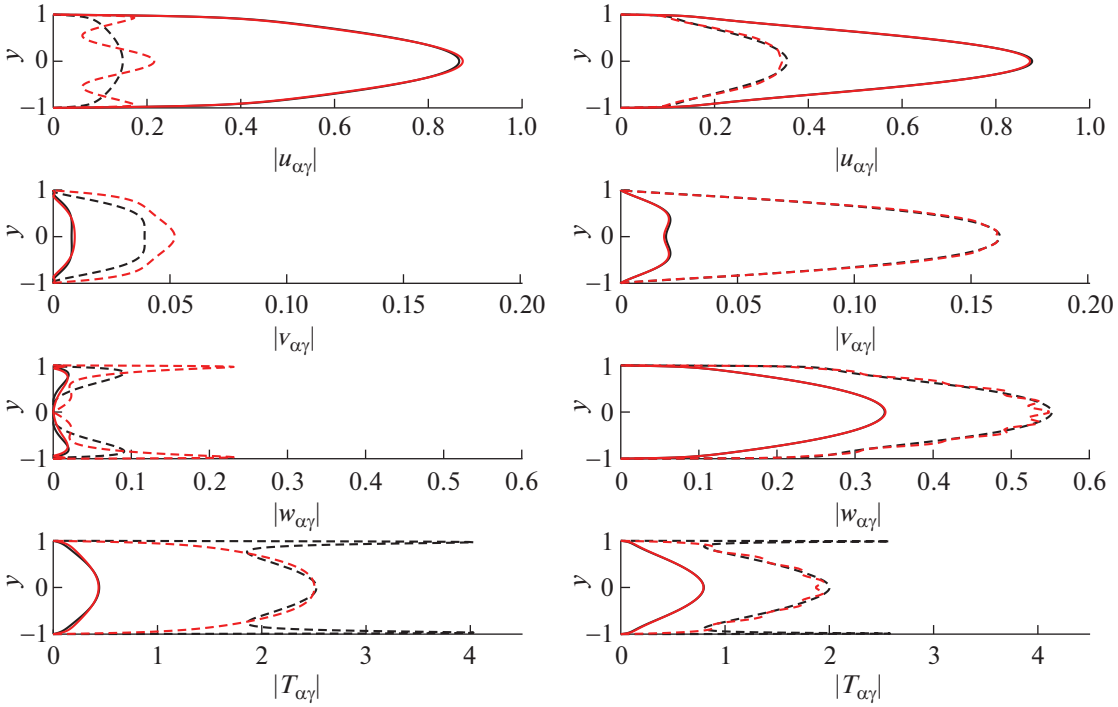
Фиг. 4. Результаты при $\mathrm{Re} = 4\times10^4$, $\mathrm{Ri} = 0.03$ и оптимальных значениях волновых чисел. Слева: абсолютные величины амплитуд компонент нормированной ведущей вещественной моды (синим) и нормированных ведущих комплексно-сопряженных мод (зеленым и красным – моды, отвечающие собственным значениям с положительными и отрицательными мнимыми частями соответственно); справа: абсолютные величины амплитуд компонент нормированного глобального оптимального возмущения в момент времени $t = 0$ (штрихпунктир) и нормированного глобального оптимального возмущения в моменты времени t_{opt} (сплошная линия).

В табл. 1 также представлены результаты вычисления глобальной максимальной вещественной части $r_{\max}$ собственных значений матрицы H, т.е. инкременты нарастания глобальных ведущих мод. Видно, что как и глобальная максимальная амплификация, величина $r_{\max}$ возрастает с ростом числа Рейнольдса. Однако она остается отрицательной. Отметим, что величина $r_{\max}$ достигалась при значениях волновых чисел $\alpha = \gamma = 0$ при всех рассмотренных числах Рейнольдса. Таким образом, глобальные ведущие моды существенно отличаются от глобальных оптимальных возмущений даже по волновым числам, т.е. глобальные ведущие моды не входят в состав глобальных оптимальных возмущений. Для сравнения ведущих мод и оптимальных возмущений при оптимальных значениях волновых чисел обратимся к фиг. 3 и 5.

На фиг. 3 изображена ведущая часть спектра матрицы H при различных числах Рейнольдса и значениях числа узлов сетки $n = 100$ и 200. В изображенных ведущих частях спектров кратных собственных значений обнаружено не было. Видно, что имеется сходимость по шагу сетки при $n = 100$. Все дальнейшие результаты расчетов мы будем приводить только для $n = 100$, поскольку расчеты при $n = 200$ давали те же результаты с точностью до приводимых значащих цифр.

На фиг. 3 видно, что при минимальном из рассмотренных чисел Рейнольдса максимальную вещественную часть имеет вещественное собственное значение, а при больших числах Рейнольдса максимальную вещественную часть имеет комплексно-сопряженная пара собственных значений.

На фиг. 4 при $\mathrm{Re} = 4\times10^4$ сравниваются абсолютные величины амплитуд компонент нормированной ведущей вещественной моды и нормированных ведущих комплексно-сопряженных мод с абсолютными величинами амплитуд компонент нормированного глобального оптимального возмущения в моменты времени $t = 0$ и t_{opt}. Видно, что как ведущая вещественная, так и ведущие комплексные моды существенно отличаются от оптимального возмущения как при $t = 0$, так и при $t = t_{\mathrm{opt}}$.



Фиг. 5. Абсолютные величины амплитуд компонент нормированного глобального оптимального возмущения (черным) и оптимального на подпространстве Λ_1 (красным) в моменты времени $t = 0$ (штриховая) и t_{opt} (сплошная линия) при $\text{Re} = 2\times10^4$ (слева) и $\text{Re} = 6\times10^4$ (справа).

6.2. Спектральный состав глобальных оптимальных возмущений

Для каждого из рассматриваемых значений числа Рейнольдса и соответствующих оптимальных значений волновых чисел будем рассматривать инваринтные подпространства, отвечающие изолированным подмножествам спектра $\lambda(H)$ матрицы H следующего вида:

$$\Lambda_1 = \{\lambda \in \lambda(H) : \text{Real}\,\lambda > r_1\}, \quad \Lambda_2 = \lambda(H)\backslash\Lambda_1,$$

$$\Lambda_{11} = \{\lambda \in \lambda(H) : \text{Real}\,\lambda > r_2\}, \quad \Lambda_{12} = \Lambda_1\backslash\Lambda_{11},$$

где $r_1 < r_2 < 0$. Подмножества Λ_{11} и Λ_{12} выделены на фиг. 3 соответственно красным и синим цветами.

Пусть Λ означает одно из описанных выше подмножеств спектра, $\mathcal{U}$ – инвариантное подпространство, отвечающее Λ, $\dim\mathcal{U}$ – его размерность, т.е. суммарная алгебраическая кратность собственных значений, входящих в Λ. Нас будет интересовать квадрат нормы $c_{\mathcal{U}}(t)$ проекции глобального оптимального возмущения на $\mathcal{U}$ в моменты времени $t = 0$ и t_{opt} и максимальная амплификация $\Gamma_{\mathcal{U},\max}$ на этом подпространстве. Результаты вычисления этих величин для введенных подмножеств спектра приведены в табл. 2.

Из табл. 2 видно, что глобальное оптимальное возмущение лежит главным образом в подпространстве $\mathcal{U}_1$, поскольку квадрат нормы его проекции на дополнительное инвариантное подпространтво $\mathcal{U}_2$ при $t = 0$ равен величине порядка 10^{-2}, а при t_{opt} становится равным величине порядка 10^{-4} либо меньше. При этом необходимая размерность подпространства $\mathcal{U}_1$ растет (21, 37 и 52) с ростом числа Рейнольдса.

Учитывая, что $\Gamma_{U,\max}$ примерно равно $c_{\mathcal{U}}(t_{\text{opt}})/c_{\mathcal{U}}(0)$ для $\mathcal{U} = \mathcal{U}_1$, можно также заключить, что проекция глобального оптимального возмущения на подпространство $\mathcal{U}_1$ является в этом подпространстве оптимальным возмущением, т.е. имеет наибольшую амплификацию. В то же время любой вектор из дополнительного подпространства $\mathcal{U}_2$ не может иметь амплификацию,

Таблица 2. Размерность инвариантного подпространства $\mathcal{U}$, максимальная амплификация $\Gamma_{\mathcal{U},\max}$ и квадрат нормы проекции глобального оптимального возмущения на $\mathcal{U}$ в моменты времени $t = 0$ и t_{opt} в зависимости от числа Рейнольдса

Параметры	$\mathcal{U}$	$\dim \mathcal{U}$	$\Gamma_{\mathcal{U},\max}$	$c_{\mathcal{U}}(0)$	$c_{\mathcal{U}}(t_{\text{opt}})$
$\text{Re} = 2\times10^4$	$\mathcal{U}_1$	21	24.6367	1.0133	24.8382
$r_1 = -0.11$	$\mathcal{U}_2$	279	1.0000	0.0133	0.0000
$r_2 = -0.025$	$\mathcal{U}_{11}$	9	24.0670	1.0584	24.8702
	$\mathcal{U}_{12}$	12	1.6596	0.0685	0.0004
$\text{Re} = 4\times10^4$	$\mathcal{U}_1$	37	30.3495	1.0216	30.4001
$r_1 = -0.24$	$\mathcal{U}_2$	263	1.0900	0.0216	0.0000
$r_2 = -0.13$	$\mathcal{U}_{11}$	26	30.0807	3.2660	30.4001
	$\mathcal{U}_{12}$	11	1.0000	2.3251	0.0000
$\text{Re} = 6\times10^4$	$\mathcal{U}_1$	52	38.0686	1.0122	38.0889
$r_1 = -0.38$	$\mathcal{U}_2$	248	1.1516	0.0122	0.0000
$r_2 = -0.15$	$\mathcal{U}_{11}$	33	37.8930	10.5533	38.0889
	$\mathcal{U}_{12}$	19	1.0000	9.5607	0.0000

большую максимальной амплификации на этом подпространстве, которая максимальна в случае $\text{Re} = 6\times10^4$ и равна 1.1516.

Для более детального спектрального анализа глобального оптимального возмущения мы разлагали инвариантное пространство $\mathcal{U}_1$ в прямую сумму двух инвариантных подпространств $\mathcal{U}_{11}$ и $\mathcal{U}_{12}$ и рассматривали проекции глобального оптимального возмущения на каждое из этих подпространств. Результаты некоторых возможных вариантов таких разбиений показаны в остальной части табл. 2. Видно, что подскок глобального оптимального возмущения достигается за счет двух различных факторов. Во-первых, величина квадрата нормы его проекции на подпространство $\mathcal{U}_{11}$ при $t = 0$ больше единицы. Во-вторых, квадрат нормы этой проекции возрастает при $t = t_{\text{opt}}$, а проекция глобального оптимального возмущения на подпространство $\mathcal{U}_{12}$ становится малозначимой. При максимальном числе Рейнольдса наиболее значим первый фактор, при минимальном – второй.

Интерес представляет и то, что максимальная амплификация векторов из подпространства $\mathcal{U}_{11}$ немногим меньше амплификации глобального оптимального возмущения. Возникает вопрос, чем отличается глобальное оптимальное возмущение от оптимального возмущения из подпространства $\mathcal{U}_{11}$. Ответ на этот вопрос дает фиг. 5, где сравниваются абсолютные величины компонент этих оптимальных возмущений при $t = 0$ и t_{opt}. Видно, что абсолютные величины сравниваемых оптимальных возмущений близки друг к другу в C-норме, причем близость увеличивается с ростом числа Рейнольдса, но глобальное оптимальное возмущение – значительно более гладкая функция (кроме его температурной компоненты в пристеночной области) даже при максимальном из рассмотренных значений числа Рейнольдса.

7. ЗАКЛЮЧЕНИЕ

В данной работе установлено, что спектр уравнений стратифицированного турбулентного течения Куэтта, осредненных по горизонтальным пространственным переменным и линеаризованных относительно стационарного состояния, симметричен относительно вещественной оси и лежит строго в левой полуплоскости, т.е. все собственные моды устойчивые, а главная часть оптимального возмущения представляет собой линейную комбинацию большого числа мод, отвечающих собственным значениям с наибольшими вещественными частями. При этом число наиболее значимых мод растет с ростом числа Рейнольдса.

Авторы благодарны А.В. Глазунову и Е.В. Мортикову за предоставление данных прямого численного моделирования, интерес к данной работе и полезные обсуждения результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Drobinski P., Brown R., Flamant P., Pelon J.* Evidence of organized large eddies by ground-based doppler lidar, sonic anemometer and sodar // Bound.-Layer Meteor. 1988. V. 88. № 3. P. 343–361.
2. *Глазунов А.В.* Численное моделирование устойчиво-стратифицированных турбулентных течений над плоской и городской поверхностями. // Изв. РАН, сер. ФАиО. 2014. Т. 50. № 3. С. 271–281.
3. *Глазунов А.В., Мортиков Е.В., Барсков К.В., Каданцев Е.В., Зилитинкевич С.С.* О слоистой структуре устойчиво-стратифицированных турбулентных течений со сдвигом скорости // Изв. РАН, сер. ФАиО. 2019. Т. 55. № 4. С. 13–26.
4. *Sullivan P.P., Weil J.C., Patton E.G., Jonker H.J., Mironov D.V.* Turbulent winds and temperature fronts in large-eddy simulations of the stable atmospheric boundary layer // J. Atmos. Sci. 2016. V. 73. № 4. P. 1815–1840.
5. *Mortikov E.V., Glazunov A.V., Lykosov V.N.* Numerical study of plane Couette flow: turbulence statistics and the structure of pressure-strain correlations // Russ. J. Num. Anal. Math. Model. 2019. V. 34. № 2. P.119–132.
6. *Lilly D.K.* On the instability of Ekman boundary flow // J. Atmos. Sci. 1966. V. 23. № 5 P. 481–494.
7. *Brown A.R.* A secondary flow model for the planetary boundary layer // J. Atmos. Sci. 1970. V. 27. № 5. P. 742–757.
8. *Glazunov A.V., Zasko G.V., Mortikov E.V., Nechepurenko Yu.M.* Optimal disturbances of stably stratified turbulent Couette flow // Doklady Physics. 2019. V. 64. № 7. P. 308–312.
9. *Засько Г.В., Глазунов А.В., Мортиков Е.В., Нечепуренко Ю.М.* Крупномасштабные структуры стратифицированного турбулентного течения Куэтта и оптимальные возмущения: Препринты ИПМ им. Келдыша. 2019. № 63.
10. *Zasko G.V., Glazunov A.V., Mortikov E.V., Nechepurenko Yu.M.* Large-scale structures in stratified turbulent Couette flow and optimal disturbances // Russ. J. Numer. Anal. Math. Model. 2020. V. 35. № 2. P. 37–53.
11. *Бойко А.В., Клюшнев Н.В., Нечепуренко Ю.М.* Устойчивость течения жидкости над оребренной поверхностью. М.: ИПМ им. Келдыша, 2016. 123 с.
12. *Nechepurenko Yu.M., Sadkane M.* A low-rank approximation for computing the matrix exponential norm // SIAM J. Matrix. Anal. Appl. 2011. V. 32. № 2. P. 349–363.
13. *Schmid P.J., Henningson D.S.* Stability and transition in shear flows. Berlin: Springer–Verlag, 2000.
14. *Romanov V.A.* Stability of plane-parallel Couette flow // Func. Anal. Appl. 1973. V. 7. P. 137–146.
15. *Butler K.M., Farrell B.F.* Optimal perturbations and streak spacing in wall-bounded turbulent shear flow // Phys. Fluids A: Fluid Dynamics. 1993. V. 5. № 3. P. 774–777.
16. *Canuto C., Hussaini M.Y., Quarteroni A., Zang T.A.* Spectral methods. Fundamentals in single domains. Berlin: Springer, 2006.
17. *Canuto C., Hussaini M.Y., Quarteroni A., Zang T.A.* Spectral methods. Evolution to complex geometries and applications to fluid dynamics. Berlin: Springer, 2007.
18. *Weideman J.A.C., Reddy S.C.* A MATLAB Differentiation matrix suite // ACM Transact. Math. Soft. 2000. V. 26. № 4. P. 465–519.
19. *Нечепуренко Ю.М.* О редукции линейных дифференциально-алгебраических систем управления // Докл. АН. 2012. Т. 445. № 1. С. 17–19.
20. *Golub G.H., van Loan C.F.* Matrix computations. London: The John Hopkins Univer. Press, 1991.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 150–161

МАТЕМАТИЧЕСКАЯ ФИЗИКА

УДК 519.63

ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ О ГАШЕНИИ КОЛЕБАНИЙ ДВИЖУЩЕГОСЯ ПОЛОТНА

© 2021 г. И. Е. Михайлов[1,*], И. А. Суворов[2,]**

[1] *119333 Москва, ул. Вавилова, 44/2, ФИЦ ИУ РАН, Россия*

[2] *125993 Москва, Волоколамское ш., 4, МАИ НИУ, Россия*

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

***e-mail: ivan.a.suv@gmail.com*

Поступила в редакцию 06.12.2019 г.
Переработанный вариант 13.06.2020 г.
Принята к публикации 18.09.2020 г.

Моделируются механические процессы, происходящие при производстве бумаги. В бумагоделательной машине бумага перемещается в виде тонкого листа. Характерная толщина листа варьируется от 0.1 мм (офисная бумага) до 1 мм (картон). Все бумагоделательные машины содержат открытые участки полотна, где бумажное полотно проходит без механической поддержки во время движения от одного опорного ролика к другому. В это время оно может потерять стабильность, начать совершать поперечные колебания и в итоге порваться. Рассматривается возможность уменьшить эти колебания с помощью различных управляющих актьюаторов. Поперечные колебания движущегося полотна с ненулевой изгибной жесткостью моделируются с помощью неоднородного дифференциального уравнения в частных производных четвертого порядка. Воздействие управляющих актьюаторов моделируется функцией в правой части уравнения. Предполагается, что амплитуда колебаний одинакова в поперечном сечении движущегося полотна. Задача гашения колебаний сводится к минимизации некоторой функции многих переменных. Решение задачи разбивается на два этапа: решение начально-краевой задачи с заданным управлением и минимизация некоторой функции многих переменных. Для решения начально-краевой задачи предлагается численный метод. Дифференциальное уравнение четвертого порядка сводится к системе двух дифференциальных уравнений второго порядка. Далее делается замена искомых функций, позволяющая упростить эти уравнения. Получившиеся уравнения аппроксимируются конечно-разностной схемой, для которой показана ее абсолютная устойчивость. Эта разностная схема решается с помощью матричной прогонки. Для минимизации функции многих переменных используется метод Хука–Дживса. Приводятся примеры расчетов для трех типов актьюаторов: точечного, действующего на участке полотна и действующего на всем протяжении полотна. Библ. 5. Фиг. 12.

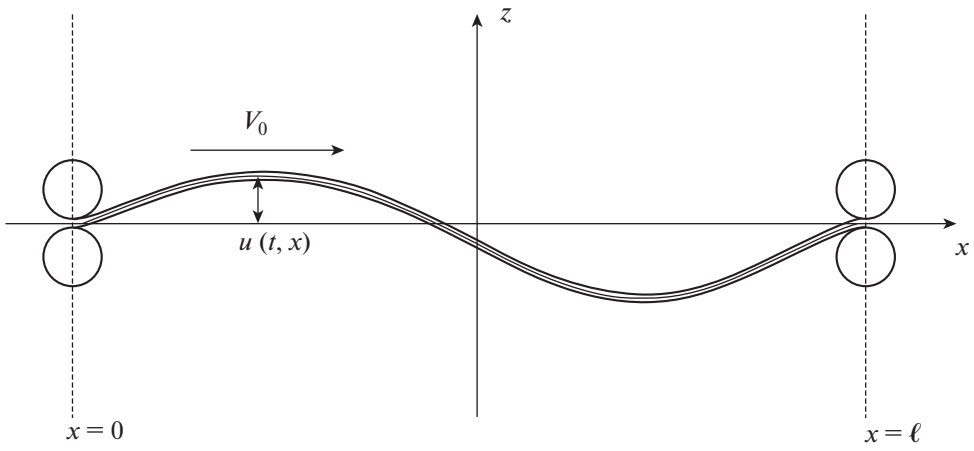
Ключевые слова: движущееся полотно, гашение колебаний, актьюаторы, метод оптимизации Хука–Дживса.

DOI: 10.31857/S004446692012011X

1. ПОСТАНОВКА ЗАДАЧИ

Целью данной работы является разработка численных методов решения задачи гашения вынужденных поперечных колебаний $u(t, x)$ движущегося с постоянной скоростью V_0 полотна толщиной H с помощью различных видов актьюаторов (фиг. 1). Предполагается, что амплитуда этих колебаний одинакова в направлении ширины полотна. Поперечные колебания движущегося полотна с заданной изгибной жесткостью под действием внешнего управления $g(t, x)$ описываются уравнением (см. [1], [2])

$$u_{tt} + 2V_0u_{tx} + (V_0^2 - c^2)u_{xx} + Du_{xxxx} = g(t,x), \quad (t,x) \in \Pi = \{0 \le x \le l,\ 0 \le t \le T\}, \tag{1}$$



Фиг. 1. Движущееся между роликами бумажное полотно.

где c – скорость распространения возмущений вдоль полотна. Время t и линейный размер x отнесены к характерным величинам t^* и x^*.Член Du_{xxxx} представляет собой силу реакции, возникающую из-за сопротивления изгибу. Константа D называется изгибной жесткостью и равна

$$D = \frac{EH^3}{12(1-v^2)},$$

где E – модуль Юнга, v – коэффициент Пуассона.

Начальные отклонение и скорость поперечного перемещения полотна

$$\begin{aligned} u(0,x) &= \varphi(x), \\ u_t(0,x) &= \psi(x) \end{aligned} \tag{2}$$

мы будем рассматривать как заданные начальные возмущения.

В качестве граничных условий при $x = 0$ и $x = l$ возьмем условия шарнирного закрепления

$$\begin{aligned} u(t,0) = u(t,l) = 0, \\ u_{xx}(t,0) = u_{xx}(t,l) = 0. \end{aligned} \tag{3}$$

Ставится следующая задача гашения: найти функцию $g(t, x)$ и время T такие, чтобы

$$\begin{aligned} u(T,x) = 0, \\ u_t(T,x) = 0. \end{aligned}$$

Отметим, что эти условия эквивалентны условию

$$J(T) = \int_0^l (u^2(T,x) + u_t^2(T,x))dx = 0. \tag{4}$$

2. ГАШЕНИЕ КОЛЕБАНИЙ И МОДЕЛИ АКТЬЮАТОРОВ

Для гашения колебаний мы будем использовать модели актьюаторов различных видов, при этом функция $g(t, x)$ будет определяться видом актьюатора.

1. Модель точечного актьюатора:

$$g(t,x) = s(t)\delta(x - x_0), \tag{5}$$

где x_0 – точка приложения актьюатора, $s(t)$ – управляющая функция, δ – дельта-функция Дирака, определенная в [3].

2. Модель актьюатора конечной ширины $[x_0, x_1] \subset [0, l]$:

$$g(t,x) = s(t)\begin{cases}1, & x \in [x_0, x_1],\\ 0, & x \notin [x_0, x_1].\end{cases} \tag{6}$$

3. Модель актьюатора, действующего одинаково по всей длине полотна:

$$g(t,x) = s(t). \tag{7}$$

Во всех трех моделях на управляющую функцию $s(t)$ могут быть наложены ограничения $s_{\min} \le s(t) \le s_{\max}$, где $s_{\min}$, $s_{\max}$ – заданные константы.

3. ЧИСЛЕННЫЙ МЕТОД РЕШЕНИЯ НАЧАЛЬНО-КРАЕВОЙ ЗАДАЧИ (1)–(3)

Введем новую вспомогательную функцию $v(t,x)$ такую, чтобы уравнение (1) можно было представить в виде двух уравнений второго порядка:

$$\begin{aligned} u_t &= v_{xx} - 2V_0 u_x,\\ v_t &= -Du_{xx} - (V_0^2 - c^2)u + f(t,x).\end{aligned} \tag{8}$$

Найдем вид функции $f(t, x)$, при которой система (8) была бы эквивалентна (1) при достаточной гладкости функций $u(t,x)$ и $v(t,x)$. Продифференцируем первое уравнение по t, а второе дважды по x:

$$\begin{aligned} u_{tt} &= v_{txx} - 2V_0 u_{tx},\\ v_{txx} &= -Du_{xxxx} - (V_0^2 - c^2)u_{xx} + f_{xx}(t,x),\end{aligned} \tag{9}$$

откуда

$$u_{tt} + 2V_0 u_{tx} + (V_0^2 - c^2)u_{xx} + Du_{xxxx} = f_{xx}(t,x).$$

Сравнивая это равенство с (1), получаем

$$f_{xx}(t,x) = g(t,x).$$

Дважды проинтегрировав левую и правую части равенства по x, найдем выражение для $f(t, x)$:

$$f(t,x) = \int_0^x \left[\int_0^\eta g(t,\xi)d\xi\right] d\eta + k_1(t)x + k_2(t),$$

где $k_1(t)$ и $k_2(t)$ – произвольные функции.

Получим начальные условия для системы (8). Подставив второе начальное условие $u_t(0,x) = \psi(x)$ из (2) в первое уравнение (8), получим

$$u_t(0,x) = (v_{xx} - 2V_0 u_x)\big|_{(0,x)} = \psi(x).$$

Откуда выражение для v_{xx} примет вид

$$v_{xx}(0,x) = 2V_0 u_x(0,x) + \psi(x).$$

Дважды проинтегрируем это выражение по x и получим выражение для $v(0,x)$:

$$v(0,x) = \Phi(x) + c_1 x + c_2.$$

Здесь c_1, c_2 – произвольные константы интегрирования, а функция $\Phi(x)$:

$$\Phi(x) = 2V_0\int_0^x \varphi(x)dx + \int_0^x\left[\int_0^\eta \psi(\xi)d\xi\right]d\eta.$$

В итоге начальные условия для системы (8) примут следующий вид:

$$\begin{aligned} u(0,x) &= \varphi(x),\\ v(0,x) &= \Phi(x) + c_1 x + c_2.\end{aligned}$$

Найдем новые граничные условия для системы (8).

Граничные условия для $u(t, x)$ будут такими

$$u(t,0) = u(t,l) = 0.$$

Найдем $v(t,0)$. Приняв во внимание, что $u(t,0) = 0$ и $u_{xx}(t,0) = 0$ из (3), и подставив эти значения во второе уравнение системы (8), получим

$$u_{xx}(t,0) = \frac{1}{D}[-v_t - (V_0^2 - c^2)u + f]\Big|_{(t,0)} = \frac{1}{D}[f - v_t]\Big|_{(t,0)} = 0.$$

Откуда

$$v_t(t,0) = f(t,0) = k_2(t),$$

$$v(t,0) = \int_0^t k_2(t)dt + c_3,$$

где c_3 – произвольная константа интегрирования. Сравнивая начальное условие и левое граничное условие для функции v в точке $(0, 0)$, получаем $c_2 = c_3$.

Найдем теперь $v(t,l)$. Подставим во второе уравнение системы (8) значения $u(t,l)$ и $u_{xx}(t,l)$ из (3), имеем:

$$u_{xx}(t,l) = \frac{1}{D}[-v_t - (V_0^2 - c^2)u + f]_{\{t,l\}} = \frac{1}{D}[f - v_t]_{(t,l)} = 0,$$

$$v_t(t,l) = f(t,l) = \int_0^l \left[\int_0^x g(t,\xi)d\xi\right]dx + k_1(t)l + k_2(t),$$

$$v(t,l) = \int_0^t \left[\int_0^l \left[\int_0^x g(t,\xi)d\xi\right]dx\right]dt + l\int_0^t k_1(t)dt + \int_0^t k_2(t)dt + c_4,$$

где c_4 – произвольная константа интегрирования. Сравнивая начальное условие и правое граничное условие для функции v в точке $(0,l)$, получаем $c_4 = \Phi(l) + c_1 l$.

Подберем теперь произвольные функции $k_1(t)$, $k_2(t)$ и константы c_1, c_2 так, чтобы для любого $t \in [0,T]$ граничные условия для функции $v(t,x)$ имели бы наиболее простой вид, а именно, равнялись нулю:

$$v(t,0) = v(t,l) \equiv 0.$$

Положим $k_2(t) \equiv 0$, $c_2 = c_3 = 0$, тогда $v(t,0) \equiv 0$ для всех $t \geq 0$.

Положим теперь

$$k_1(t) = -\frac{1}{l}\int_0^l \left[\int_0^x g(t,\xi)d\xi\right]dx, \quad c_1 = -\frac{1}{l}\Phi(l),$$

тогда $v(t,l) \equiv 0$ для всех $t \geq 0$.

Тогда выражение для $f(t,x)$ примет вид

$$f(t,x) = \int_0^x \left[\int_0^\eta g(t,\xi)d\xi\right]d\eta - \frac{x}{l}\int_0^l \left[\int_0^\eta g(t,\xi)d\xi\right]d\eta.$$

Начальные условия для системы (8) перепишутся в виде

$$\begin{aligned} u(0,\ x) &= \varphi(x), \\ v(0,\ x) &= \Phi(x) - \frac{x}{l}\Phi(l). \end{aligned} \tag{10}$$

Граничные условия имеют следующий вид:

$$u(t,0) = 0, \quad u(t,l) = 0, \quad v(t,0) = 0, \quad v(t,l) = 0. \tag{11}$$

Запишем систему (8) в матричном виде

$$\begin{pmatrix} u \\ v \end{pmatrix}_t = \begin{pmatrix} 0 & 1 \\ -D & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}_{xx} + \begin{pmatrix} -2V_0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}_x + \begin{pmatrix} 0 & 0 \\ c^2 - V_0^2 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ f \end{pmatrix}$$

или

$$W_t = AW_{xx} + BW_x + C + F, \tag{12}$$

где

$$W = \begin{pmatrix} u \\ v \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ -D & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -2V_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ c^2 - V_0^2 & 0 \end{pmatrix}, \quad F = \begin{pmatrix} 0 \\ f \end{pmatrix}.$$

Сделаем замену исходной функции $W = \mathcal{A}\hat{W}$, где $\mathcal{A} = \begin{pmatrix} 1 & 0 \\ \alpha(t,x) & 1 \end{pmatrix}$, $\alpha(t,x) = V_0 x + (c^2 - V_0^2)t$, а $\hat{W} = \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix}$ – новая искомая вектор-функция.

В итоге, уравнение (12) для новой вектор-функции $\hat{W}$ примет вид:

$$\hat{W}_t = \begin{pmatrix} \alpha & 1 \\ -D - \alpha^2 & -\alpha \end{pmatrix} \hat{W}_{xx} + \begin{pmatrix} 0 \\ f \end{pmatrix}. \tag{13}$$

Теперь для новой вектор-функции $\hat{W}$ запишем начальные условия

$$\begin{gathered} \hat{u}(0,x) = \varphi(x), \\ \hat{v}(0,x) = \Phi(x) - \frac{x}{l}\Phi(l) - V_0 x\varphi(x). \end{gathered} \tag{14}$$

Граничные условия примут вид

$$\hat{u}(t,0) = 0, \quad \hat{u}(t,l) = 0, \quad \hat{v}(t,0) = 0, \quad \hat{v}(t,l) = 0. \tag{15}$$

Построим разностную схему для численного решения (13). Зададим натуральные числа M, N и разобьем рассматриваемую область $\{0 \le t \le T,\ 0 \le x \le l\}$ на прямоугольные ячейки параллельными прямыми $x_m = mh$, $m = 0,1,\ldots,M$, $t_n = n\tau$, $n = 0,1,\ldots,N$, где $h = \dfrac{l}{M}$, $\tau = \dfrac{T}{N}$.

Рассмотрим шаблон, на котором уравнение (13) аппроксимируем конечно-разностной схемой

$$\frac{\hat{W}_m^{n+1} - \hat{W}_m^n}{\tau} = \begin{pmatrix} \alpha & 1 \\ -D - \alpha^2 & -\alpha \end{pmatrix} \frac{\hat{W}_{m+1}^{n+1} - 2\hat{W}_m^{n+1} + \hat{W}_{m-1}^{n+1}}{h^2} + \begin{pmatrix} 0 \\ f \end{pmatrix}. \tag{16}$$

Покажем, что эта разностная схема абсолютно устойчива по Нейману. Будем искать решение однородного уравнения в виде

$$\hat{W}_m^n = \lambda^n e^{ipm} W_0,$$

где $W_0 \ne 0$, p – действительное число, i – мнимая единица.

Тогда имеем

$$\frac{\lambda - 1}{\tau} W_0 = \begin{pmatrix} \alpha & 1 \\ -D - \alpha^2 & -\alpha \end{pmatrix} \frac{e^{ip} - 2 + e^{-ip}}{h^2} \lambda W_0,$$

или

$$\left((\lambda - 1)E - S \begin{pmatrix} \alpha & 1 \\ -D - \alpha^2 & -\alpha \end{pmatrix} \left(-4\lambda \left(\sin \frac{p}{2} \right)^2 \right) \right) W_0 = 0,$$

где $S = \dfrac{\tau}{h^2}$.

Таблица 1

T	Аналитическое решение	Численные решения			
		$M = 10$	$M = 20$	$M = 40$	$M = 80$
1	0.67277	0.6754	0.6717	0.6728	0.6731
2	–0.094749	–0.09466	–0.09384	–0.09499	–0.09483
5	–0.52113	–0.5219	–0.5240	–0.5218	–0.5217
10	–0.45686	–0.4594	–0.4539	–0.4561	–0.4567

Эта система имеет нетривиальное решение, если

$$\begin{vmatrix} \lambda - 1 + 4\alpha S\lambda\left(\sin\frac{p}{2}\right)^2 & 4S\lambda\left(\sin\frac{p}{2}\right)^2 \\ -4S\lambda(-D-\alpha^2)\left(\sin\frac{p}{2}\right)^2 & \lambda - 1 - 4\alpha S\lambda\left(\sin\frac{p}{2}\right)^2 \end{vmatrix} =$$

$$= (\lambda-1)^2 - 16\alpha^2 S^2\lambda^2\left(\sin\frac{p}{2}\right)^4 + 16S^2\lambda^2(D+\alpha^2)\left(\sin\frac{p}{2}\right)^4 = 0,$$

$$(\lambda-1)^2 + 16S^2\lambda^2 D\left(\sin\frac{p}{2}\right)^4 = 0.$$

Пусть

$$\beta^2 = 16S^2 D\left(\sin\frac{p}{2}\right)^4, \quad (\lambda-1)^2 + \beta^2\lambda^2 = 0, \quad \lambda = \pm\beta\lambda i + 1, \quad \lambda(1 \mp \beta i) = 1,$$

$$\lambda = \frac{1}{1\mp\beta i} = \frac{1\pm\beta i}{1+\beta^2} = \frac{1}{1+\beta^2} \pm \frac{\beta}{1+\beta^2} i, \quad |\lambda| = \sqrt{\frac{1}{(1+\beta^2)^2} + \frac{\beta^2}{(1+\beta^2)^2}} = \sqrt{\frac{1}{1+\beta^2}} \le 1,$$

т.е. схема абсолютно устойчива при любых β. Будем ее решать с помощью матричной прогонки.

Сравним аналитическое решение задачи (1)–(3) с численным решением задачи (18)–(20). При

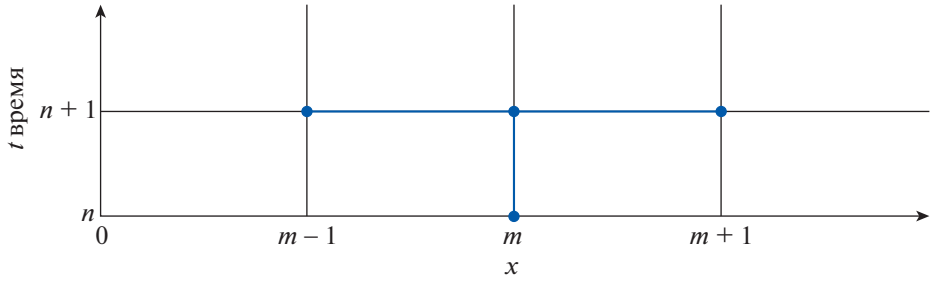
$$g(t,x) = -2V_0\frac{\pi}{l}\omega\sin(\omega t)\cos\frac{\pi x}{l},$$

где

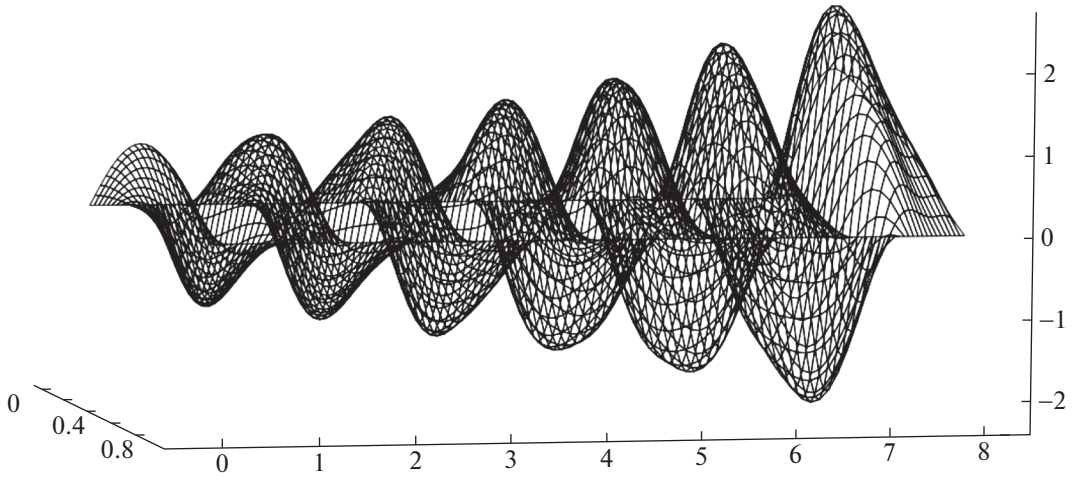
$$\omega = \frac{\pi}{l}\sqrt{(c^2 - V_0^2) + D\frac{\pi^2}{l^2}}, \quad \varphi(x) = \sin\frac{\pi x}{l}, \quad \psi(x) = 0$$

и условиях (3) аналитическим решением задачи (1)–(3) является функция $u(t,x) = \cos(\omega t)\sin\frac{\pi x}{l}$.

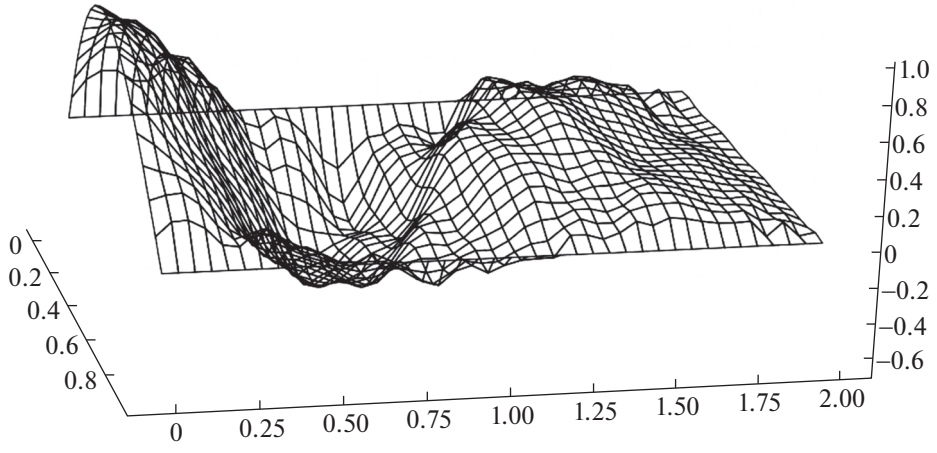
В табл. 1 при $V_0 = 1$, $c = 2$, $D = 0.005$, $l = 1$ приведены аналитическое решение и расчетные решения задачи (18)–(20) в точке $x = 0.5$ при различных T и шагах сетки $M \times N$. При этом число узловых точек по времени $N = 2MT$.



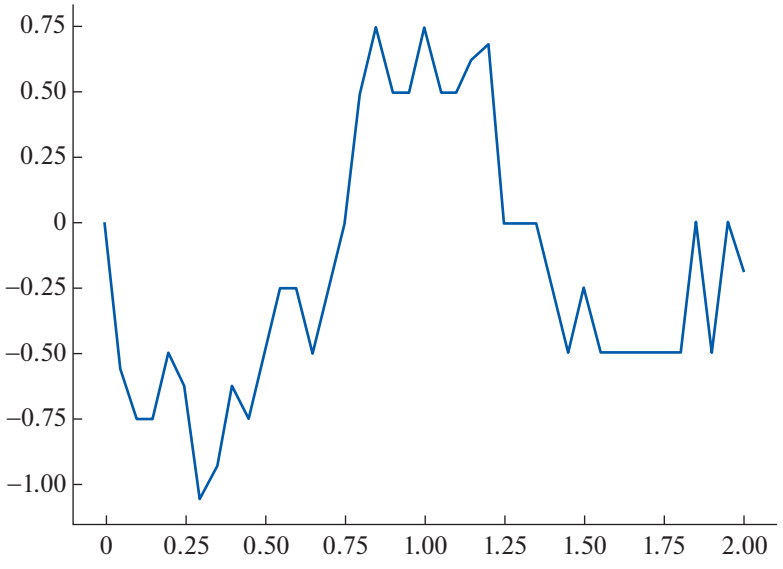
Фиг. 2.



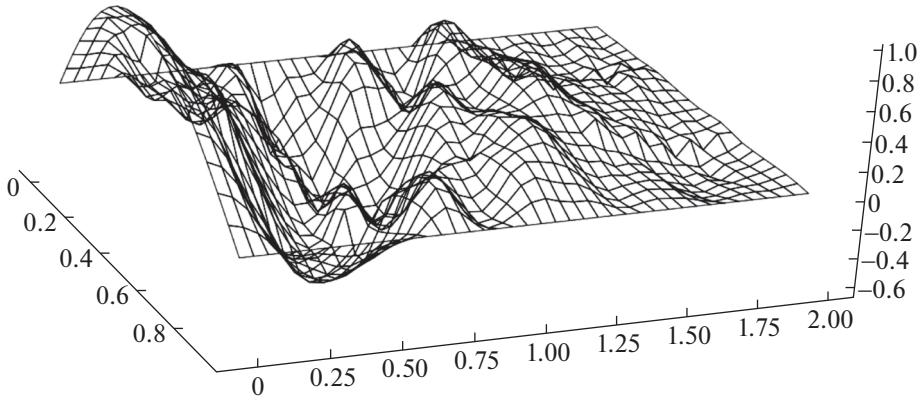
Фиг. 3.



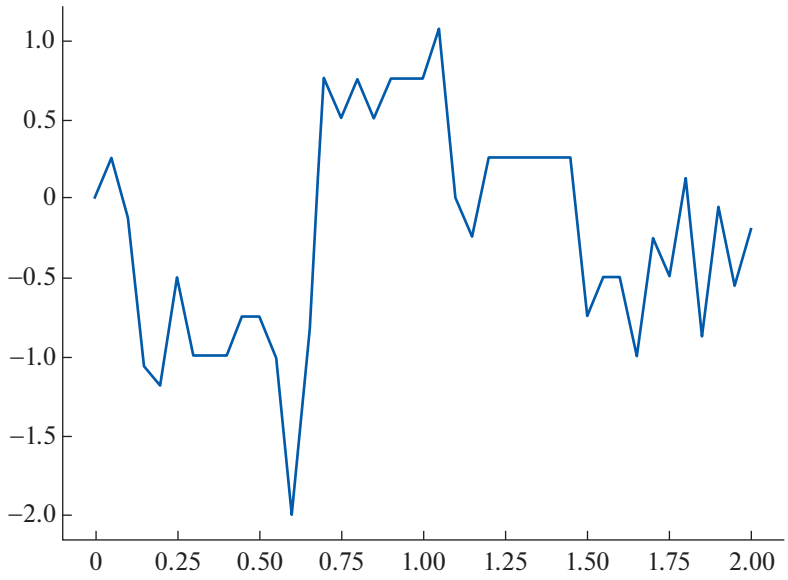
Фиг. 4.



Фиг. 5.



Фиг. 6.



Фиг. 7.

Таким образом, мы видим сходимость предложенного метода с уменьшением шагов сетки и хорошее совпадение с аналитическим решением.

4. ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ ГАШЕНИЯ КОЛЕБАНИЙ

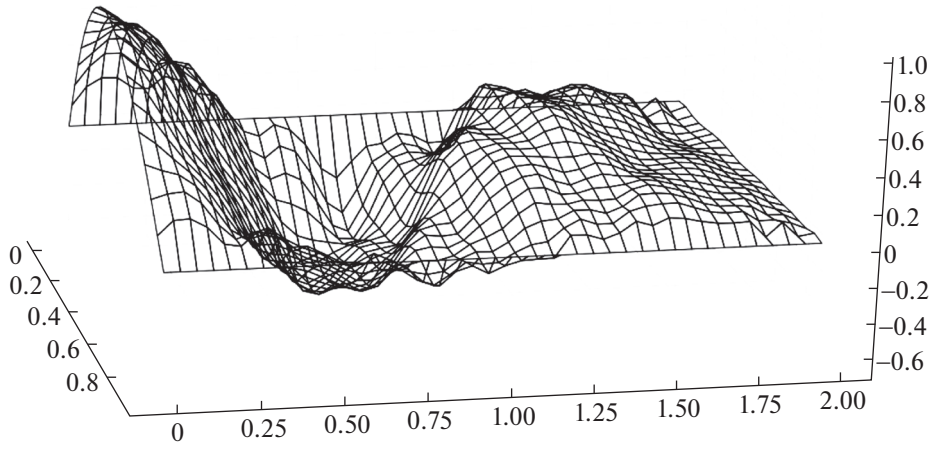
Отметим, что из условия (4) следует условие

$$J(T) = \int_0^l (\hat{u}^2(T,x) + \hat{u}_t^2(T,x))dx = 0. \tag{17}$$

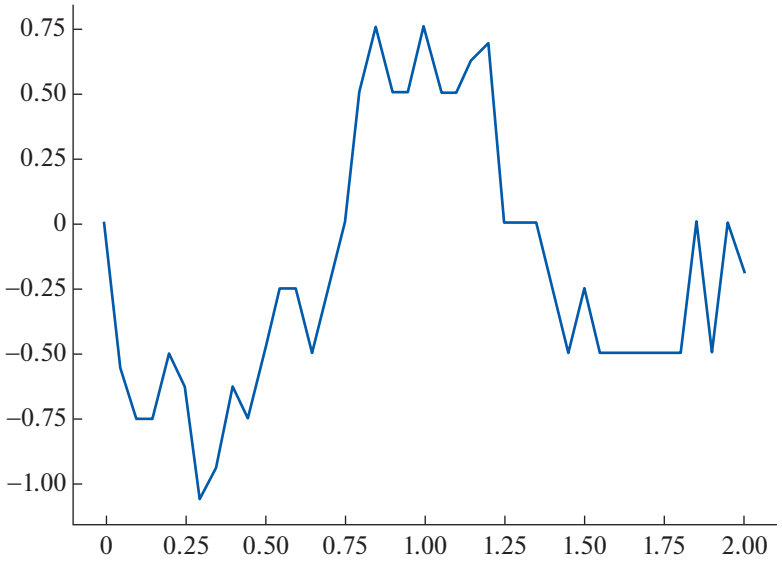
Для рассматриваемых моделей актьюаторов (5)–(7) функция $f(t,x)$ записывается в следующих видах.

1. Модель точечного актьюатора (5):

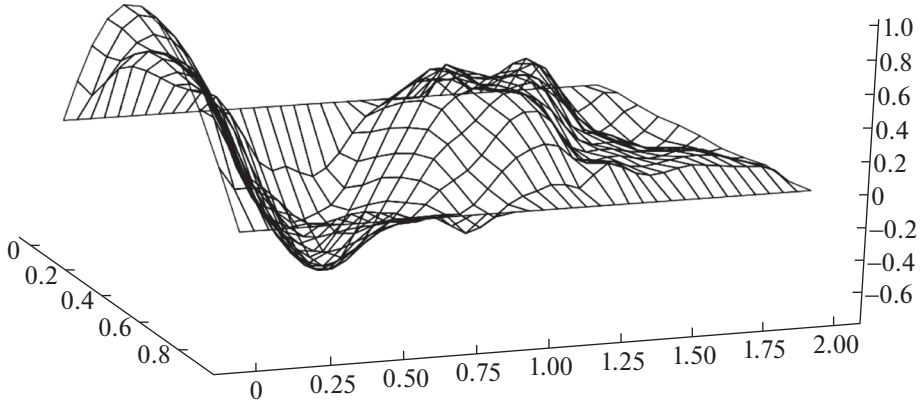
$$f(t,x) = s(t)\begin{cases} -\dfrac{x}{l}(l-x_0), & x < x_0, \\ (x-x_0) - \dfrac{x}{l}(l-x_0), & x \geq x_0. \end{cases}$$



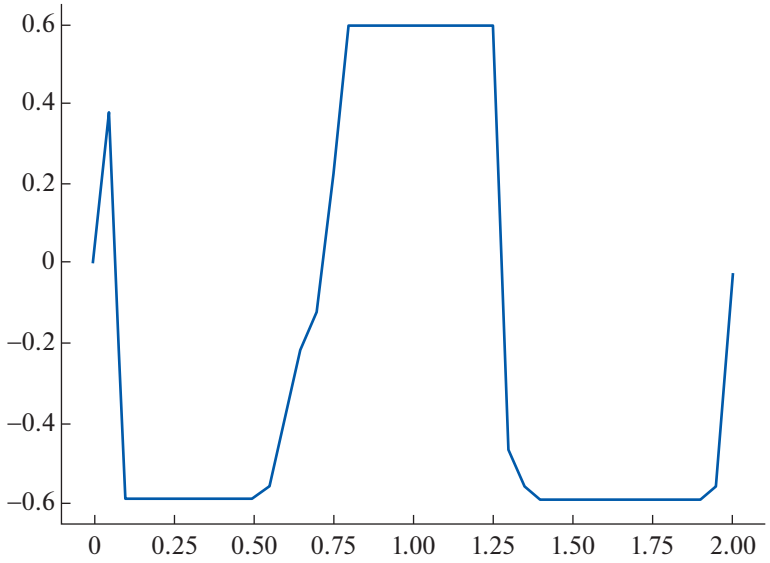
Фиг. 8.



Фиг. 9.



Фиг. 10.



Фиг. 11.

2. Модель актьюатора конечной ширины $[x_0, x_1]$ (6):

$$f(t,x) = s(t)\begin{cases} -\dfrac{x}{l}(x_1 - x_0)\left(l - \dfrac{x_0 + x_1}{2}\right), & x < x_0, \\ \dfrac{(x - x_0)^2}{2} - \dfrac{x}{l}(x_1 - x_0)\left(l - \dfrac{x_0 + x_1}{2}\right), & x_0 \le x \le x_1, \\ \left(\dfrac{x}{l} - 1\right)\dfrac{x_1^2 - x_0^2}{2}, & x_1 < x. \end{cases}$$

3. Модель актьюатора, действующего одинаково вдоль всего полотна (7):

$$f(t,x) = s(t)\frac{x(x-l)}{2}.$$

Функцию $s(t)$ на отрезке $[0,T]$ аппроксимируем ломаной, соединяющей соседние точки (t_n, s_n), (t_{n+1}, s_{n+1}), $n = 0,1,\ldots,N-1$, прямыми, где $s_0,\ldots,s_N$ – неизвестные пока постоянные.

Константы $s_0,\ldots,s_N$ будем искать, минимизируя функцию многих переменных:

$$J(s_0,\ldots,s_N) = h\sum_{m=1}^{M-1}\left[(\hat{u}_m^N)^2 + \left(\frac{\hat{u}_m^N - \hat{u}_m^{N-1}}{\tau}\right)^2\right], \tag{18}$$

представляющую собой квадратурную формулу трапеций для интеграла (17) с учетом граничных условий (15). Для минимизации будем использовать метод Хука–Дживса (см. [4]).

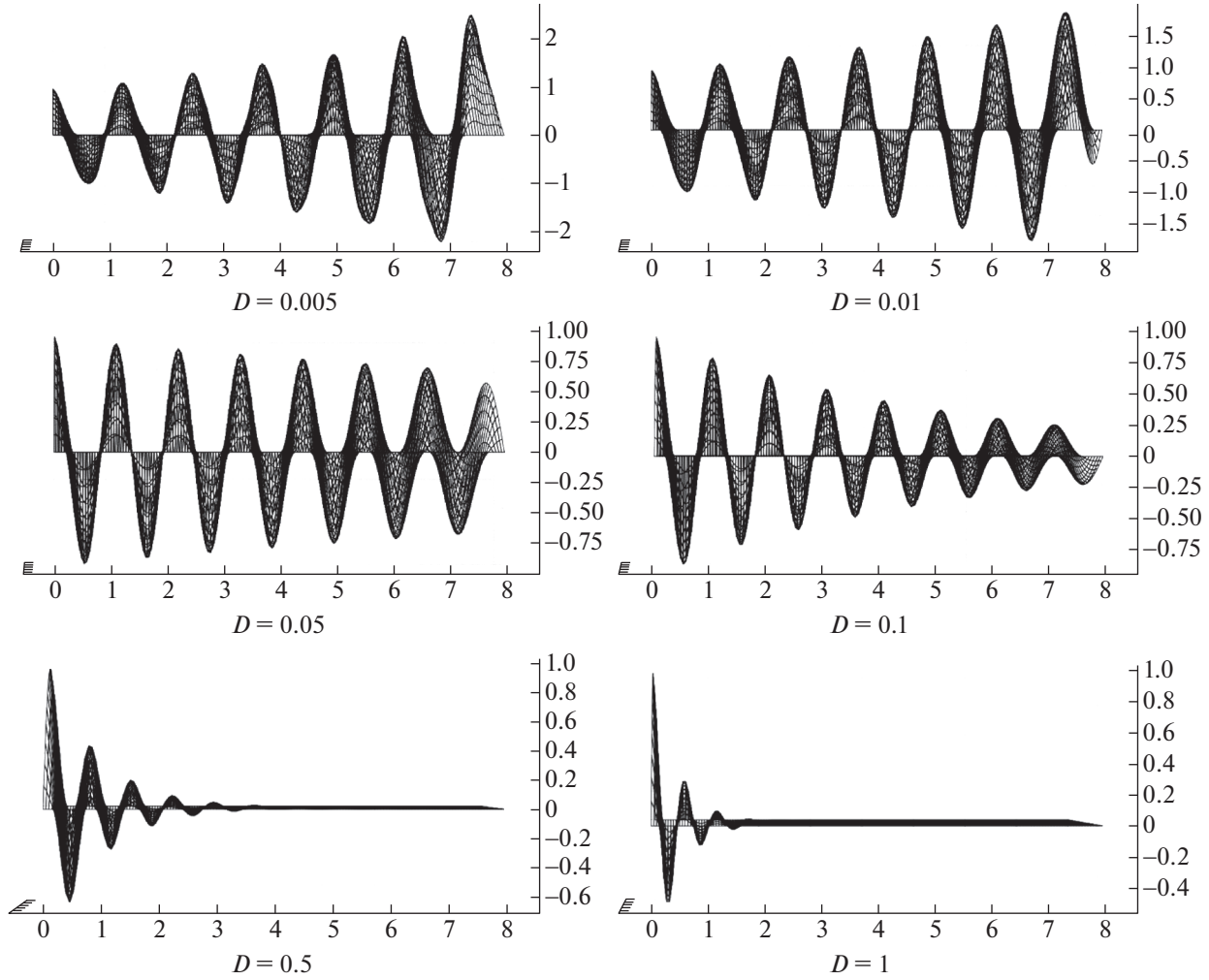
5. ПРИМЕРЫ

Во всех описанных ниже примерах время T будем выбирать так, чтобы гарантировать выполнение условия $J(s_0,\ldots,s_N) < \varepsilon$, где ε полагалось равным 0.001.

Пример 1. Рассмотрим свободные колебания движущегося полотна ($g(t,x) = 0$) с начальными условиями $\varphi(x) = \sin(\pi x)$, $\psi(x) = 0$. Значения входных данных: $V_0 = 1$, $c = 2$, $D = 0.005$, $l = 1$, $T = 8$, $M = 20$, $N = 160$.

На фиг. 3 видно, что максимальная амплитуда колебаний полотна с течением времени возрастает.

Пример 2. Рассмотрим теперь задачу гашения колебаний с использованием точечного актьюатора (5), помещенного в точку $x_0 = 0.5$. Начальные условия и значения входных данных совпа-



Фиг. 12.

дают со значениями из примера 1, $T = 2$. На фиг. 4 показан процесс гашения колебаний с помощью управляющей функции $s(t)$, изображенной на фиг. 5.

Пример 3. Погасим начальные колебания $\varphi(x) = \sin(\pi x)$, $\psi(x) = 0$ с помощью узкого актьюатора (6) с $x_0 = 0.3$, $x_1 = 0.6$. Значения входных данных совпадают со значениями из примера 1, $T = 2$. На фиг. 6 показан процесс гашения колебаний с помощью управляющей функции $s(t)$, изображенной на фиг. 7.

Пример 4. Погасим колебания с помощью актьюатора (7), действующего на протяжении всего полотна. Значения всех параметров совпадают со значениями из примера 1, $T = 2$. На фиг. 8 показан процесс гашения колебаний с помощью управляющей функции $s(t)$, изображенной на фиг. 9.

Пример 5. Будем использовать актьюатор (7) и установим ограничения на управляющую функцию $s_{\min} = -0.6$, $s_{\max} = 0.6$, $T = 2$. На фиг. 10 и 11 изображены процесс гашения $u(t,x)$ и управляющая функция $s(t)$ соответственно.

В [5] с использованием принципа максимума Понтрягина было показано, что при гашении с помощью актьюатора (7) с ограничениями $s_{\min} \le s(t) \le s_{\max}$ значение управляющей функции может частично совпадать с граничными значениями $s_{\min}$, $s_{\max}$. Фиг. 11 подтверждает этот вывод.

Пример 6. Рассмотрим влияние параметра D на свободные колебания $u(t,x)$ $(g(t,x) = 0)$ (фиг. 12). Все входные данные взяты из примера 1.

Как мы видим, при малых D свободные колебания возрастают с течением времени, а при увеличении D, начиная с некоторого значения, затухают сами собой.

СПИСОК ЛИТЕРАТУРЫ

1. *Jeronen J.* On the Mechanical Stability and Out-of-Plane Dynamics of a Travelling Panel Submerged in Axially Flowing Ideal Fluid. Univer. Jyväskylä, 2011. 243 p.
2. *Banichuk N., Barsuk A., Jeronen J., Tuovinen T., Neittaanmäki P.* Stability of Axially Moving Materials. Solid Mechanics and Its Applications. V. 259. Springer, Cham, 2020. 642 p.
3. *Владимиров В.С., Жаринов В.В.* Уравнения математической физики. М.: ФИЗМАТЛИТ, 2004. 400 с.
4. *Hooke R., Jeeves T.A.* Direct Search Solution of Numerical and Statistical Problems // J. of the ACM (JACM). 1961. V. 8. Issue 2. P. 212–229.
5. *Banichuk N., Petrov V., Sinitsyn A., Neittaanmdki P., Tuovinen T.* On the Optimality Conditions for Suppression of Vibration of Axially Moving Materials.: Rep. Depart. Math. Inform. Technolgy. Ser. B. Sci. Comp. No. B 13120L6. P. 1–19. Univer. Jyvaskyla, 2016.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 1, с. 162–176

ИНФОРМАТИКА

УДК 519.86

ЗАДАЧА АГРЕГИРОВАНИЯ МЕЖОТРАСЛЕВОГО БАЛАНСА И ДВОЙСТВЕННОСТЬ[1)]

© 2021 г. А. А. Шананин[1,2,3,4]

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, МФТИ, Россия
[2] 119333 Москва, ул. Вавилова, 44, кор. 2, ФИЦ ИУ РАН, Россия
[3] 119991 Москва, Ленинские горы, МГУ, Россия
[4] 117198 Москва, ул. Миклухо-Маклая, 6, РУДН, Россия
e-mail: alexshan@yandex.ru

Поступила в редакцию 25.06.2020 г.
Переработанный вариант 25.06.2020 г.
Принята к публикации 18.09.2020 г.

С помощью теоремы двойственности Фенхеля и преобразования Янга в работе построена операция свертки технологий и на ее основе исследована задача агрегирования модели нелинейного межотраслевого баланса с вогнутыми положительно-однородными производственными функциями. Библ. 9.

Ключевые слова: множители Лагранжа, двойственность по Фенхелю, преобразование Янга, межотраслевой баланс, продуктивность, производственная функция, агрегирование, эластичность замещения.

DOI: 10.31857/S0044466921010087

1. ВВЕДЕНИЕ

Метод межотраслевого баланса В.В. Леонтьева был удостоен премии имени Нобеля по экономике и успешно использовался в XX веке для анализа экстенсивного восстановительного роста экономики США после великой экономической депрессии и экономик европейских стран и Японии в послевоенное тридцатилетие. Модели межотраслевого баланса позволяли строить мультипликаторы, выявлять узкие места экономической динамики, определять драйверы экономического роста. В основу метода В.В. Леонтьева были положены система материальных балансов и гипотеза о постоянстве норм затрат на выпуск продукции в процессе межотраслевого взаимодействия. Однако в 90-е годы в развитых капиталистических странах изменился характер экономической динамики: экстенсивное увеличение объемов производства сменилось ростом разнообразия и качества товаров и услуг. В этих условиях гипотеза В.В. Леонтьева о постоянстве норм затрат перестала соответствовать возросшей взаимозаменяемости товаров и услуг. Модели межотраслевого баланса в этот период утратили прежнюю популярность. Их стали вытеснять модели, в которых игнорировалась отраслевая специфика, а экономическая динамика описывалась как воспроизводство валового внутреннего продукта (см., например, [1], [2]). Однако мировые экономические кризисы конца XX века и начала XXI века вновь актуализировали модели межотраслевых балансов как инструмент исследования структурных диспропорций. Стали разрабатываться сетевые модели межотраслевых связей для экономики США (см., например, [3]). В [3] гипотеза В.В. Леонтьева о постоянстве норм материальных затрат заменена гипотезой о постоянстве структуры финансовых затрат в процессе производства товаров и услуг с учетом их отраслевой дифференциации. Эта гипотеза соответствует допущению, что производитель фиксирует пропорции своих расходов, в рамках которых в зависимости от ценовой конъюнктуры осуществляет материальные затраты, варьируя качество приобретаемых товаров и услуг. В экономических условиях конца XX века и начала XXI века такая гипотеза представляется более адекватной, чем гипотеза В.В. Леонтьева. В отличие от леонтьевских отраслевых производственных функций с постоянными пропорциями, ей соответствуют производственные функции, учитывающие замещение производственных факторов. В данной работе рассматриваются новые математические задачи агрегирования и калибровки моделей нелинейного межотраслевого баланса.

[1)]Работа выполнена при финансовой поддержке РФФИ код проекта 20-51-15004.

2. НЕЛИНЕЙНЫЙ МЕЖОТРАСЛЕВОЙ БАЛАНС

Рассмотрим группу из m чистых отраслей, связанных взаимными поставками продукции в качестве производственных факторов текущего пользования (ПФТП). Обозначим через X_i^j объем продукции i-й отрасли, который используется в качестве ПФТП в процессе производства в j-й отрасли, а через $X^j = (X_1^j, \ldots, X_m^j)$ – затраты j-й отраслью ПФТП, производимых рассматриваемой группой отраслей. Будем также предполагать, что в процессе производства отрасли затрачивают в качестве ПФТП первичные ресурсы (n видов), т.е. продукты, не производимые рассматриваемой группой отраслей. Обозначим через $l^j = (l_1^j, \ldots, l_n^j)$ вектор затрат первичных ресурсов j-й отраслью, а через $F_j(X^j, l^j)$ – производственную функцию j-й отрасли, т.е. зависимость выпуска j-й отрасли от затрат ПФТП. Будем предполагать, что производственные функции отраслей обладают неоклассическими свойствами, т.е. являются вогнутыми, монотонно неубывающими, непрерывными функциями на R_+^{m+n}, обращающимися в нуле в нуль. Кроме того, будем считать, что $F_j(X^j, l^j)$ являются функциями, положительно-однородными первой степени. Будем говорить, что такие функции принадлежат классу Φ_{m+n}.

Обозначим через $X^0 = (X_1^0, \ldots, X_m^0)$ объемы поставок производимой рассматриваемыми отраслями продукции внешним потребителям. Будем считать, что спрос внешних потребителей описывается с помощью функции полезности $F_0(X^0)$. Предположим, что функция $F_0(X^0) \in \Phi_m$ и что предложение первичных ресурсов рассматриваемой группе отраслей ограничено объемом $l = (l_1, \ldots, l_n) \geq 0$. Рассмотрим задачу об оптимальном распределении этих ресурсов между отраслями в целях максимизации функции полезности внешних потребителей при балансовых ограничениях по первичным ресурсам и выпускаемой отраслями продукции:

$$F_0(X^0) \to \max, \tag{2.1}$$

$$F_j(X^j, l^j) \geq \sum_{i=0}^{m} X_j^i, \quad j = 1, \ldots, m, \tag{2.2}$$

$$\sum_{j=1}^{m} l^j \leq l, \tag{2.3}$$

$$X^0 \geq 0, \quad X^1 \geq 0, \quad \ldots, \quad X^m \geq 0, \quad l^1 \geq 0, \quad \ldots, \quad l^m \geq 0. \tag{2.4}$$

Будем считать, что рассматриваемая группа отраслей продуктивна, т.е. существуют $\hat{X}^1 \geq 0, \ldots, \hat{X}^m \geq 0, \hat{l}^1 \geq 0, \ldots, \hat{l}^m \geq 0$, такие, что

$$F_j(\hat{X}^j, \hat{l}^j) > \sum_{i=0}^{m} \hat{X}_j^i, \quad j = 1, \ldots, m.$$

Нетрудно доказать, что если группа отраслей продуктивна и $l = (l_1, \ldots, l_n) > 0$, то задача оптимизации (2.1)–(2.4) удовлетворяет условиям Слейтера.

Предложение 1 (см. [4]). *Для того чтобы набор векторов* $\{\hat{X}^0, \hat{X}^1, \ldots, \hat{X}^m, \hat{l}^1, \ldots, \hat{l}^m\}$, *удовлетворяющих ограничениям* (2.2)–(2.4), *являлся решением задачи оптимизации* (2.1)–(2.4), *необходимо и достаточно, чтобы существовали множители Лагранжа* $p_0 > 0$, $q = (q_1, \ldots, q_m) \geq 0$, $s = (s_1, \ldots, s_n) \geq 0$ *такие, что*

$$(\hat{X}^j, \hat{l}^j) \in \operatorname{Arg\,max}\left\{q_j F_j(X^j, l^j) - qX^j - sl^j \,\middle|\, X^j \geq 0, l^j \geq 0\right\}, \quad j = 1, \ldots, m, \tag{2.5}$$

$$q_j\left(F_j(\hat{X}^j, \hat{l}^j) - \hat{X}_j^0 - \sum_{i=1}^{m} \hat{X}_j^i\right) = 0, \quad j = 1, \ldots, m, \tag{2.6}$$

$$s_k\left(l_k - \sum_{j=1}^{m} \hat{l}_k^j\right) = 0, \quad k = 1, \ldots, n, \tag{2.7}$$

$$\hat{X}^0 \in \operatorname{Arg\,max}\left\{p_0F_0(X^0) - qX^0 \,\middle|\, X^0 \geq 0\right\}. \tag{2.8}$$

Будем интерпретировать множители Лагранжа $q = (q_1, \ldots, q_m)$ к балансовым ограничениям по выпускаемым отраслями продуктам как цены на эти продукты, а множители Лагранжа $s = (s_1, \ldots, s_n)$ к балансовым ограничениям по первичным ресурсам – как цены на первичные ресурсы. Тогда соотношение (2.5) означает, что предложение продукции отраслями и их спрос на ПФТП определяются из максимизации прибыли при ценах (q, s). Соотношение (2.8) описывает спрос при ценах q репрезентативного рационального конечного потребителя с функцией полезности $F_0(X^0)$, и, кроме того, $p_0 = q_0(q)$, где функция $q_0(q)$ является преобразованием Янга функции $F_0(X^0)$, т.е.

$$q_0(q) = \inf\left\{\frac{qX^0}{F_0(X^0)}\middle| X^0 \geq 0,\ F_0(X^0) > 0\right\}.$$

Соотношения (2.2) и (2.6), (2.3) и (2.7) означают, что цены $q = (q_1, \ldots, q_m)$ и $s = (s_1, \ldots, s_n)$ равновесные. Таким образом, оптимальными механизмами распределения являются равновесные рыночные механизмы.

Двойственным описанием технологии производства j-й отрасли является функция себестоимости

$$q_j(q, s) = \inf\left\{\frac{qX^j + sl^j}{F_j(X^j, l^j)}\middle| X^j \geq 0,\ l^j \geq 0,\ F_j(X^j, l^j) > 0\right\}.$$

Функция себестоимости $q_j(q, s)$ является преобразованием Янга производственной функции $F_j(X^j, l^j)$.

3. ЗАДАЧА ОБ АГРЕГИРОВАННОМ ОПИСАНИИ ГРУППЫ ОТРАСЛЕЙ

Рассмотрим задачу агрегирования межотраслевого баланса (2)–(4) с помощью функции полезности $F_0(X^0)$. Обозначим через $F^A(l)$ оптимальное значение функционала в задаче (2.1)–(2.4) в зависимости от вектора предложения первичных ресурсов l в правой части балансового ограничения (2.3). Функция $F^A(l) \in \Phi_n$ интерпретируется как агрегированная производственная функция. Агрегированной производственной функции $F^A(l)$ соответствует двойственная агрегированная функция себестоимости

$$q_A(s) = \inf\left\{\frac{sl}{F^A(l)}\middle| l \geq 0,\ F^A(l) > 0\right\}. \tag{3.1}$$

Функция себестоимости $q_A(s) \in \Phi_n$ и, кроме того, справедливо соотношение (см., например, [5])

$$F^A(l) = \inf\left\{\frac{sl}{q_A(s)}\middle| s \geq 0,\ q_A(s) > 0\right\}.$$

В силу двойственности между производственными функциями и функциями себестоимости, например, производственной функции с постоянной эластичностью замещения (constant elasticity substitution, CES) $\left(\left(\frac{X_1}{w_1}\right)^{-\rho} + \left(\frac{X_2}{w_2}\right)^{-\rho} + \left(\frac{X_n}{w_n}\right)^{-\rho}\right)^{-1/\rho}$, где $\rho \in [-1, 0) \cup (0, +\infty]$, $w_1 > 0, \ldots, w_n > 0$, в силу преобразования Янга соответствует CES-функция себестоимости $\left((s_1w_1)^{-\sigma} + (s_2w_2)^{-\sigma} + (s_nw_n)^{-\sigma}\right)^{-1/\sigma}$, где $\sigma = -\frac{\rho}{1+\rho}$. Производственной функции Кобба–Дугласа

(предельный случай при $\rho \to 0$) $F_{KD}(X_1, \ldots, X_n) = AX_1^{\alpha_1} \ldots X_n^{\alpha_n}$, где $A > 0$, $\alpha_1 > 0, \ldots, \alpha_n > 0$, $\alpha_1 + \ldots + \alpha_n = 1$, в силу преобразования Янга соответствует функция себестоимости

$$q_{KD}(p_1, \ldots p_n) = \frac{1}{F_{KD}(\alpha_1, \ldots, \alpha_n)} p_1^{\alpha_1} \ldots p_n^{\alpha_n}.$$

Теорема 1. *Агрегированная функция себестоимости представима в виде*

$$q_A(s) = \sup\{q_0(p) \,|\, p = (p_1, \ldots, p_m) \geq 0,\ q_j(p, s) \geq p_j,\ j = 1, \ldots, m\}. \tag{3.2}$$

Доказательство. Построим двойственную задачу к задаче (2.1)–(2.4), предполагая, что $l \in R_+^n$. Для этого переформулируем задачу (2.1)–(2.4) в виде $\inf\{f(X^0) + g(X^0)\}$, где

$$f(X^0) = \begin{cases} -F_0(X^0), & \text{если} \quad X^0 \in R_+^m, \\ +\infty, & \text{если} \quad X^0 \notin R_+^m, \end{cases}$$

$$g(X^0) = \begin{cases} 0, & \text{если} \quad \exists X^1, \ldots, X^m, l^1, \ldots, l^n, \text{ удовлетворяющие } (2.2)\text{–}(2.4), \\ +\infty & \text{в противном случае.} \end{cases}$$

Вычислим сопряженные функции:

$$f^*(-p) = \sup\{F_0(X^0) - pX^0 \,|\, X^0 \geq 0\} = \begin{cases} 0, & \text{если } p \geq 0, \quad q_0(p) \geq 1, \\ +\infty & \text{в противном случае,} \end{cases}$$

$$g^*(p) = \sup\{pX^0 - g(X^0)\} = \begin{cases} \inf\{sl \,|\, s \geq 0,\ q_j(s, p) \geq p_j,\ j = 1, \ldots, m\}, & \text{если} \quad p \in R_+^m, \\ +\infty, & \text{если} \quad p \notin R_+^m. \end{cases}$$

По теореме Фенхеля (см. [6, с. 47]) имеем, что

$$F^A(l) = \inf\{f^*(-p) + g^*(p)\} = \inf\{sl \,|\, q_0(p) \geq 1,\ q_j(p, s) \geq p_j,\ j = 1, \ldots, m,\ p \geq 0,\ s \geq 0\}. \tag{3.3}$$

В силу положительной однородности первой степени функций $q_0(p)$, $q_j(s, p)$, $j = 1, \ldots, m$, из (3.1) и (3.3) следует (3.2). Теорема 1 доказана.

Определение 1. Будем называть задачу (3.2) двойственной по Янгу к задаче (2.1)–(2.4).

Предложение 2. *Если множители Лагранжа* $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m) \geq 0$, $\hat{s} \geq 0$ *к задаче* (1)–(4) *удовлетворяют условиям*

$$(\hat{X}^j, \hat{l}^j) \in \operatorname{Arg\,max}\{\hat{p}_j F_j(X^j, l^j) - \hat{p}X^j - \hat{s}l^j \,|\, X^j \geq 0, l^j \geq 0\}, \quad j = 1, \ldots, m, \tag{3.4}$$

$$\hat{p}_j \left(F_j(\hat{X}^j, \hat{l}^j) - \hat{X}_j^0 - \sum_{i=1}^m \hat{X}_j^i \right) = 0, \quad j = 1, \ldots, m, \tag{3.5}$$

$$\hat{s}_k \left(l_k - \sum_{j=1}^m \hat{l}_k^j \right) = 0, \quad k = 1, \ldots, n, \tag{3.6}$$

$$\hat{X}^0 \in \operatorname{Arg\,max}\{F_0(X^0) - \hat{p}X^0 \,|\, X^0 \geq 0\}, \tag{3.7}$$

то является решением задачи (3.2) *для* $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m) \geq 0$, $\hat{s} \geq 0$.

Доказательство. В силу положительной однородности и вогнутости функции $F_j(X^j, l^j)$ получаем из (3.4), что $\hat{p}_j \le q_j(\hat{p}, \hat{s})$, $j = 1, \ldots, m$, $\hat{p}_j F_j(\hat{X}^j, \hat{l}^j) = \hat{p}\hat{X}^j + \hat{s}\hat{l}^j$. Из (3.7) следует, что $q_0(\hat{p}) = 1$, $F_0(\hat{X}^0) = \hat{p}\hat{X}^0$. Из (3.5) имеем, что

$$\sum_{j=1}^{m}(\hat{p}\hat{X}^j + \hat{s}\hat{l}^j) = \sum_{j=1}^{m}\hat{p}_j F_j(\hat{X}^j, \hat{l}^j) = \sum_{j=1}^{m}\hat{p}_j\hat{X}_j^0 + \sum_{j=1}^{m}\hat{p}_j\sum_{i=1}^{m}\hat{X}_i^j = \sum_{j=1}^{m}\hat{p}_j\hat{X}_j^0 + \sum_{i=1}^{m}\hat{p}\hat{X}^i,$$

$$\sum_{j=1}^{m}\hat{s}\hat{l}^j = \sum_{j=1}^{m}\hat{p}_j\hat{X}_j^0.$$

Откуда следует с учетом (3.6) и $\hat{s} \in \partial F^A(l)$, что

$$q_A(\hat{s}) = \frac{\hat{s}l}{F^A(l)} = \frac{\hat{s}l}{F_0(\hat{X}^0)} = q_0(\hat{p}).$$

Предложение 2 доказано.

4. ОБОБЩЕННАЯ КОНВОЛЮЦИЯ

В качестве примера рассмотрим две технологии $F_1(l^1)$ и $F_2(l^2)$, использующие одни и те же производственные факторы. Будем считать, что по этим технологиям выпускается частично взаимозаменяемая продукция и что потребители этой продукции оценивают ее с помощью функции полезности $F_0(X_1^0, X_2^0)$, которая является положительно-однородной, вогнутой, непрерывной на множестве R_+^2 функцией, принимающей положительные значения на множестве $\operatorname{int} R_+^2$.

Рассмотрим задачу распределения ресурсов между технологиями:

$$F_0(F_1(l^1), F_2(l^2)) \to \max, \tag{4.1}$$

$$l^1 + l^2 \le l, \tag{4.2}$$

$$l^1 \in R_+^n, \quad l^2 \in R_+^n. \tag{4.3}$$

Функция $\widehat{F}^A(l)$, задающая зависимость оптимального значения функционала задачи (4.1)–(4.3) от вектора l суммарных затрат производственных факторов, описывает агрегированную технологию. Заметим, что операция построения функции $\widehat{F}^A(l)$ по функциям $F_1(l^1)$ и $F_2(l^2)$ может рассматриваться как обобщение операции конволюции в выпуклом анализе.

Следствие 1. Пусть

$$q_j(\mathbf{s}) = \inf_{\{\mathbf{X}\ge 0 | F_j(\mathbf{X})>0\}} \frac{\mathbf{sX}}{F_j(\mathbf{X})} \quad (j = 1,2), \quad q_0(\alpha_1, \alpha_2) = \inf_{\{Y_1\ge 0, Y_2\ge 0 | F_0(Y_1,Y_2)>0\}} \frac{\alpha_1 Y_1 + \alpha_2 Y_2}{F_0(Y_1, Y_2)}.$$

Функция $q_0(q_1(\mathbf{s}), q_2(\mathbf{s}))$ является функцией себестоимости для производственной функции $\widehat{F}^A(l)$.

Двойственный к функции полезности $F_0(Y_1, Y_2)$ индекс цены выражается с помощью преобразования Янга:

$$q_0(\alpha_1, \alpha_2) = \inf_{\{Y_1\ge 0, Y_2\ge 0 | F_0(Y_1,Y_2)>0\}} \frac{\alpha_1 Y_1 + \alpha_2 Y_2}{F_0(Y_1, Y_2)}.$$

Предположим, что в силу преобразования Янга технологиям $F_1(l^1)$ и $F_2(l^2)$ соответствуют функции себестоимости $q_1(s_1x_1, s_2x_2)$ и $q_1(s_1y_1, s_2y_2)$.

Рассмотрим задачу об агрегировании технологий: для каких функций $q_0(\beta_1,\beta_2)$, $q_1(\beta_1,\beta_2)$, $q_2(\beta_1,\beta_2)$ существуют положительные числа z_1, z_2 такие, что для любых $s_1 \ge 0$, $s_2 \ge 0$ справедливо, что

$$q_0\left(q_1(s_1x_1,s_2x_2),\, q_1(s_1y_1,s_2y_2)\right) = q_2(s_1z_1,s_2z_2).$$

Предложение 3. *Пусть функции* $q_0(\beta_1,\beta_2)$, $q_1(\beta_1,\beta_2)$, $q_2(\beta_1,\beta_2)$ *положительно однородны, вогнуты, непрерывны на* R^2_+, *принимают положительные значения на множестве* $\operatorname{int} R^2_+$ *и удовлетворяют для любых* $\beta_1 \ge 0$, $\beta_2 \ge 0$ *условиям* $q_j(\beta_1,\beta_2) = q_j(\beta_2,\beta_1)$, $q_j(0,\beta_2) = \beta_2$, $j = 0, 1, 2$. *Тогда имеем*

$$q_0(\beta_1,\beta_2) = q_1(\beta_1,\beta_2) = q_2(\beta_1,\beta_2) = \left((\beta_1)^{-\sigma} + (\beta_2)^{-\sigma}\right)^{-1/\sigma},$$

где $-1 \le \sigma < 0$.

Доказательство. Полагая $s_2 = 0$, получаем, что

$$q_0\left(q_1(s_1x_1,0), q_1(s_1x_2,0)\right) = q_0(s_1x_1,s_1x_2) = s_1q_0(x_1,x_2) = q_2(s_1z_1,0) = s_1z_1,$$

откуда следует, что $z_1 = q_0(x_1,x_2)$. Полагая $s_1 = 0$, аналогично получаем, что $z_2 = q_0(y_1,y_2)$. Таким образом, задача сводится к вопросу о существовании функции $q_2(\beta_1,\beta_2)$ такой, что для любых $\beta_1 \ge 0$, $\beta_2 \ge 0$ $q_2(\beta_1,\beta_2) = q_2(\beta_2,\beta_1)$, $q_2(0,\beta_2) = \beta_2$ и для любых $x_1 \ge 0$, $x_2 \ge 0$, $y_1 \ge 0$, $y_2 \ge 0$ справедливо равенство

$$q_0\left(q_1(x_1,x_2), q_1(y_1,y_2)\right) = q_2\left(q_0(x_1,y_1), q_0(x_2,y_2)\right).$$

Полагая в этом соотношении $y_1 = y_2 = 0$, получаем, что

$$q_0\left(q_1(x_1,x_2), q_1(0,0)\right) = q_0\left(q_1(x_1,x_2),0\right) = q_1(x_1,x_2),$$

$$q_2\left(q_1(x_1,0), q_1(x_2,0)\right) = q_2(x_1,x_2).$$

Откуда следует, что $q_1(x_1,x_2) = q_2(x_1,x_2)$. Таким образом, для любых $x_1 \ge 0$, $x_2 \ge 0$, $y_1 \ge 0$, $y_2 \ge 0$ справедливо равенство

$$q_0\left(q_1(x_1,x_2), q_1(y_1,y_2)\right) = q_1\left(q_0(x_1,y_1), q_0(x_2,y_2)\right).$$

Полагая в этом соотношении $x_2 = y_1 = 0$, получаем, что

$$q_0\left(q_1(x_1,0), q_1(0,y_2)\right) = q_0(x_1,y_2),$$

$$q_1\left(q_0(x_1,0), q_0(0,y_2)\right) = q_1(x_1,y_2).$$

Из этого следует, что $q_1(x_1,y_2) = q_0(x_1,y_2)$. Таким образом, для любых $x_1 \ge 0$, $x_2 \ge 0$, $y_1 \ge 0$, $y_2 \ge 0$ справедливо равенство

$$q_0\left(q_0(x_1,x_2), q_0(y_1,y_2)\right) = q_0\left(q_0(x_1,y_1), q_0(x_2,y_2)\right) = q_0\left(q_0(x_1,y_2), q_0(x_2,y_1)\right).$$

По теореме Дебре–Гормана–Кукушкина (см. [7, теорема 1, с. 29]) существуют непрерывные, строго монотонные, обращающиеся в нуле в нуль функции $\lambda(\beta)$, $\mu(\beta)$ такие, что для любых $x_1 \ge 0$, $x_2 \ge 0$, $y_1 \ge 0$, $y_2 \ge 0$ справедливо равенство

$$q_0\left(q_0(x_1,x_2), q_0(y_1,y_2)\right) = \lambda\left(\mu(x_1) + \mu(x_2) + \mu(y_1) + \mu(y_2)\right).$$

Полагая $y_1 = y_2 = 0$, получаем, что для любых $x_1 \ge 0$, $x_2 \ge 0$ справедливо равенство

$$q_0(x_1,x_2) = q_0\left(q_0(x_1,x_2), q_0(0,0)\right) = \lambda\left(\mu(x_1) + \mu(x_2) + \mu(0) + \mu(0)\right) = \lambda\left(\mu(x_1) + \mu(x_2)\right).$$

Заметим, что для любого $x_1 \ge 0$ справедливо равенство

$$x_1 = q_0(x_1,0) = \lambda\left(\mu(x_1) + \mu(0)\right) = \lambda\left(\mu(x_1)\right),$$

т.е. функция $\lambda(\beta)$ является обратной функцией к $\mu(\beta)$. Таким образом, для любых $x_1 \ge 0, x_2 \ge 0$ справедливо равенство

$$q_0(x_1,x_2) = \mu^{-1}\left(\mu(x_1), \mu(x_2)\right).$$

Без ограничения общности, умножая, если нужно, функцию $\mu(x)$ на положительное число, будем считать, что $\mu(1) = 1$. Из положительной однородности функции $q_0(x_1, x_2)$ следует, что если числа $x_1 \geq 0$, $x_2 \geq 0$ таковы, что $\mu(x_1) + \mu(x_2) = 1$, то для любого $t \geq 0$ справедливо равенство

$$\mu(tx_1) + \mu(tx_2) = \mu(t),$$

откуда следует, что $\mu(\beta) = \beta^{-\sigma}$, где $\sigma < 0$. Из этого следует, что

$$q_0(\beta_1, \beta_2) = q_1(\beta_1, \beta_2) = q_2(\beta_1, \beta_2) = \left((\beta_1)^{-\sigma} + (\beta_2)^{-\sigma}\right)^{-1/\sigma},$$

где $-1 \leq \sigma < 0$ в силу предположения о вогнутости функции. Предложение 3 доказано.

5. АГРЕГИРОВАНИЕ МЕЖОТРАСЛЕВОГО БАЛАНСА

Рассмотрим вопрос об агрегировании межотраслевого баланса (2.1)–(2.4). Предположим, что множество номеров отраслей и выпускаемых ими продуктов $\{1, \ldots, m\}$ разбито на непересекающиеся подмножества $\{I_\alpha \mid \alpha = 1, \ldots, \nu\}$. Обозначим через $Z^\alpha = (X_j^0 \mid j \in I_\alpha)$, где $\alpha = 1, \ldots, \nu$. Предположим, что функция полезности внешних потребителей имеет структуру $F_0(X^0) = G_0(G_1(Z^1), \ldots, G_\nu(Z^\nu))$, где функции G_α, $\alpha = 1, \ldots, \nu$, являются положительно-однородными первой степени, вогнутыми, монотонно неубывающими, непрерывными функциями.

Лемма 1 (см. [8]). *Пусть $G_1(Z^1), \ldots, G_\nu(Z^\nu)$ – положительно-однородные, вогнутые, непрерывные функции, принимающие положительные значения при положительных значениях аргументов, а функция $G_0(Y_1, \ldots, Y_\nu)$ является положительно-однородной, вогнутой, непрерывной на множестве R_+^ν функцией, принимающей положительные значения на множестве $\operatorname{int} R_+^\nu$. Пусть функции*

$$h_\alpha(p^\alpha) = \inf_{\{Z^\alpha \geq 0 \mid G_\alpha(Z^\alpha) > 0\}} \frac{p^\alpha Z^\alpha}{G_\alpha(Z^\alpha)}, \quad \alpha = 1, \ldots, \nu,$$

$$h_0(\beta_1, \ldots, \beta_\nu) = \inf_{\{Y_1 \geq 0, \ldots, Y_\nu \geq 0 \mid G_0(Y_1, \ldots, Y_\nu) > 0\}} \frac{\beta_1 Y_1 + \ldots + \beta_\nu Y_\nu}{G_0(Y_1, \ldots, Y_\nu)}$$

являются их преобразованиями Янга. Тогда получим

$$\inf_{\{Z^1 \geq 0, \ldots, Z^\nu \geq 0 \mid G_0(G_1(Z^1), \ldots, G_\nu(Z^\nu)) > 0\}} \frac{p^1 Z^1 + \ldots + p^\nu Z^\nu}{G_0(G_1(Z^1), \ldots, G_\nu(Z^\nu))} = h_0(h_1(p^1), \ldots, h_\nu(p^\nu)).$$

Здесь $Z^\alpha = (X_j^0 \mid j \in I_\alpha)$, $p^\alpha = (p_j \mid j \in I_\alpha)$, $\alpha = 1, \ldots, \nu$.

По лемме 1 функция $h_0(h_1(p^1), \ldots, h_\nu(p^\nu))$ в силу преобразования Янга будет двойственной к функции $F_0(X^0) = G_0(G_1(Z^1), \ldots, G_\nu(Z^\nu))$.

Рассмотрим вспомогательную задачу

$$H^\alpha(p^1, \ldots, p^{\alpha-1}, p^{\alpha+1}, \ldots, p^\nu, s) = \sup\left\{h_\alpha(p^\alpha) \,\middle|\, p^\alpha \geq 0,\ q_j(p, s) \geq p_j,\ j \in I_\alpha\right\}. \tag{5.1}$$

Теорема 2. *Пусть*

$$H^\alpha(p^1, \ldots, p^{\alpha-1}, p^{\alpha+1}, \ldots, p^\nu, s) = r^\alpha(h_1(p^1), \ldots, h_{\alpha-1}(p^{\alpha-1}), h_{\alpha+1}(p^{\alpha+1}), \ldots, h_\nu(p^\nu), s), \quad \alpha = 1, \ldots, \nu, \tag{5.2}$$

где $r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s)$ – положительно-однородные, вогнутые, непрерывные функции, принимающие положительные значения при положительных значениях аргументов. Тогда

$$q_A(s) = \sup\left\{h_0(h_1, \ldots, h_\nu) \,\middle|\, r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s) \geq h_\alpha \geq 0,\ \alpha = 1, \ldots, \nu\right\}. \tag{5.3}$$

Кроме того, если $\hat{p} = (\hat{p}^\alpha \mid \alpha = 1, \ldots, \nu)$ – решение задачи (3.2), *то $\left\{\hat{h}_\alpha = h^\alpha(\hat{p}^\alpha) \,\middle|\, \alpha = 1, \ldots, \nu\right\}$ – решение задачи* (5.3).

Доказательство. Пусть $\hat{p} = (\hat{p}^\alpha \,|\, \alpha = 1, \ldots, \nu)$ – решение задачи (3.2). Положим $\hat{h}_\alpha = h^\alpha(\hat{p}^\alpha)$, $\alpha = 1, \ldots, \nu$. Из (5.1) следует, что $\hat{h}_\alpha \le r^\alpha(\hat{h}_1, \ldots, \hat{h}_{\alpha-1}, \hat{h}_{\alpha+1}, \ldots, \hat{h}_\nu, s)$, $\alpha = 1, \ldots, \nu$. Поскольку $q_0(\bar{p}) = h_0(h_1(\bar{p}^1), \ldots, h_\nu(\bar{p}^\nu))$, получаем, что

$$q_A(s) \le \sup\left\{h_0(h_1, \ldots, h_\nu) \,\middle|\, r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s) \ge h_\alpha \ge 0,\ \alpha = 1, \ldots, \nu\right\}.$$

В обратную сторону, пусть $(\hat{h}_1, \ldots, \hat{h}_\nu)$ – решение задачи

$$\sup\left\{h_0(h_1, \ldots, h_\nu) \,\middle|\, r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s) \ge h_\alpha \ge 0,\ \alpha = 1, \ldots, \nu\right\}. \tag{5.4}$$

Используя свойства функций $h_\alpha(p^\alpha)$, выберем $\hat{p} = (\hat{p}^\alpha \,|\, \alpha = 1, \ldots, \nu) \ge 0$ так, чтобы

$$\hat{h}_\alpha = h_\alpha(\hat{p}^\alpha), \quad \alpha = 1, \ldots, \nu, \quad q_j(\hat{p}, s) \ge \hat{p}_j, \quad j = 1, \ldots, m.$$

Поскольку $q_0(\bar{p}) = h_0(h_1(\bar{p}^1), \ldots, h_\nu(\bar{p}^\nu))$, получаем, что

$$q_A(s) \ge \sup\left\{h_0(h_1, \ldots, h_\nu) \,\middle|\, r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s) \ge h_\alpha \ge 0,\ \alpha = 1, \ldots, \nu\right\}.$$

Теорема 2 доказана.

Теорема 3. *Пусть*

$$R^\alpha(Y_1, \ldots, Y_{\alpha-1}, Y_{\alpha+1}, \ldots, Y_\nu, l) = \inf_{\left\{h_1 \ge 0, \ldots, h_{\alpha-1} \ge 0, \ldots h_\nu \ge 0 \,\middle|\, r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s)\right\}} \frac{h_1 Y_1 + \ldots + h_{\alpha-1} Y_{\alpha-1} + h_{\alpha+1} Y_{\alpha+1} + \ldots + h_\nu Y_\nu + sl}{r^\alpha(h_1, \ldots, h_{\alpha-1}, h_{\alpha+1}, \ldots, h_\nu, s)},$$

и $\{\hat{X}^0, \ldots, \hat{X}^m, \hat{l}^1, \ldots, \hat{l}^m\}$ *– решение задачи* (2.1)–(2.4). *Положим*

$$\hat{L}^\alpha = \sum_{j \in I^\alpha} \hat{l}^j, \quad \hat{Y}^0_\beta = G_\beta\left(\hat{X}^0_i \,\middle|\, i \in I^\beta\right), \quad \beta = 1, \ldots, \nu,$$

$$\hat{Y}^\alpha_\beta = G_\beta\left(\sum_{j \in I^\alpha} \hat{X}^j_i \,\middle|\, i \in I^\beta\right), \quad \alpha \ne \beta, \quad \alpha = 1, \ldots, \nu, \quad \beta = 1, \ldots, \nu.$$

Тогда $\left\{\hat{Y}^0_\beta, \hat{Y}^\alpha_\beta, \hat{L}^\alpha \,\middle|\, \alpha = 1, \ldots, \nu;\ \beta = 1, \ldots, \nu\right\}$ *является решением задачи*

$$G_0(Y^0_1, \ldots, Y^0_\nu) \to \max, \tag{5.5}$$

$$R^\alpha(Y^\alpha_1, \ldots, Y^\alpha_{\alpha-1}, Y^\alpha_{\alpha+1}, \ldots, Y^\alpha_\nu, L^\alpha) \ge Y^0_\alpha + \sum_{\beta \ne \alpha} Y^\beta_\alpha, \quad \alpha = 1, \ldots, \nu, \tag{5.6}$$

$$\sum_{\alpha=1}^{\nu} L^\alpha \le l, \tag{5.7}$$

$$Y^0_\beta \ge 0, \quad Y^\alpha_\beta \ge 0, \quad \beta = 1, \ldots, \nu, \quad \beta \ne \alpha, \quad L^\alpha \ge 0, \quad \alpha = 1, \ldots, \nu. \tag{5.8}$$

Доказательство. По теореме 1 задача (5.5)–(5.8) является двойственной по Янгу к задаче (5.4). Из теоремы 2 следует, что оптимальные значения функционалов в задачах (3.2) и (5.4) равны, а значит, в силу двойственности по Янгу, равны и оптимальные значения функционалов в задачах (2.1)–(2.4) и (5.5)–(5.8). Из (2.3), (2.4) по построению следует, что $\left\{\hat{Y}^0_\beta, \hat{Y}^\alpha_\beta, \hat{L}^\alpha \,\middle|\, \alpha = 1, \ldots, \nu;\ \beta = 1, \ldots, \nu\right\}$ удовлетворяют (5.7) и (5.8). Таким образом, для доказательства теоремы достаточно проверить, что $\left\{\hat{Y}^0_\beta, \hat{Y}^\alpha_\beta, \hat{L}^\alpha \,\middle|\, \alpha = 1, \ldots, \nu;\ \beta = 1, \ldots, \nu\right\}$ удовлетворяют (5.6).

В силу предложения 1 существуют $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_m) \ge 0$, $\tilde{s} = (\tilde{s}_1, \ldots, \tilde{s}_n) \ge 0$ такие, что выполняются условия

$$(\hat{X}^j, \hat{l}^j) \in \operatorname{Arg\,max}\left\{\tilde{p}_j F_j(X^j, l^j) - \tilde{p} X^j - \tilde{s} l^j \,\middle|\, X^j \ge 0, l^j \ge 0\right\}, \quad j = 1, \ldots, m, \tag{5.9}$$

$$\tilde{p}_j\left(F_j(\hat{X}^j,\hat{l}^j)-\hat{X}^0_j-\sum_{i=1}^m\hat{X}^i_j\right)=0,\quad j=1,\ldots,m, \tag{5.10}$$

$$\tilde{s}_k\left(l_k-\sum_{j=1}^m\hat{l}^j_k\right)=0,\quad k=1,\ldots,n, \tag{5.11}$$

$$\hat{X}^0\in\operatorname{Arg\,max}\left\{F_0(X^0)-\tilde{p}X^0\,\middle|\,X^0\geq 0\right\}. \tag{5.12}$$

В силу положительной однородности и вогнутости функции $F_j(X^j,l^j)$ получаем из (5.9), что $\tilde{p}_j=q_j(\tilde{p},\tilde{s})$ и $\tilde{p}_jF_j(\hat{X}^j,\hat{l}^j)=\tilde{p}\hat{X}^j+\tilde{s}\hat{l}^j$ (здесь $j=1,\ldots,m$). Откуда следует, что

$$\sum_{j=1}^m\tilde{p}_jF_j(\hat{X}^j,\hat{l}^j)=\sum_{j=1}^m(\tilde{p}\hat{X}^j+\tilde{s}\hat{l}^j).$$

С учетом (5.10), (5.12) получаем

$$\sum_{j=1}^m(\tilde{p}\hat{X}^j+\tilde{s}\hat{l}^j)=\sum_{j=1}^m\tilde{p}_jF_j(\hat{X}^j,\hat{l}^j)=\sum_{j=1}^m\tilde{p}_j\hat{X}^0_j+\sum_{j=1}^m\tilde{p}_j\sum_{i=1}^m\hat{X}^j_i=\sum_{j=1}^m\tilde{p}_j\hat{X}^0_j+\sum_{i=1}^m\tilde{p}\hat{X}^i,$$

$$\sum_{j=1}^m\tilde{s}\hat{l}^j=\sum_{j=1}^m\tilde{p}_j\hat{X}^0_j.$$

Из (5.12) следует, что $q_0(\tilde{p})=h_0(h_1(\tilde{p}^1),\ldots,h_\nu(\tilde{p}^\nu))=1$, $F_0(\hat{X}^0)=\tilde{p}\hat{X}^0$, т.е.

$$h_0(h_1(\tilde{p}^1),\ldots,h_\nu(\tilde{p}^\nu))G_0\left(G_1\left(\hat{X}^0_j\,\middle|\,j\in I^1\right),\ldots,G_\nu\left(\hat{X}^0_j\,\middle|\,j\in I^\nu\right)\right)=\sum_{\alpha=1}^\nu h_\alpha(\tilde{p}^\alpha)G_\alpha\left(\hat{X}^0_j\,\middle|\,j\in I^\alpha\right),$$

$$h_\alpha(\tilde{p}^\alpha)G_\alpha\left(\hat{X}^0_j\,\middle|\,j\in I^\alpha\right)=\sum_{j\in I^\alpha}\tilde{p}_j\hat{X}^0_j,\quad \alpha=1,\ldots,\nu.$$

Рассмотрим вспомогательную задачу

$$G_\alpha\left(X^0_j\,\middle|\,j\in I^\alpha\right)\to\max, \tag{5.13}$$

$$F_j(X^j,l^j)\geq X^0_j+\sum_{i\in I^\alpha}X^i_j+\sum_{\beta\neq\alpha}\sum_{i\in I^\beta}\hat{X}^i_j,\quad j\in I^\alpha, \tag{5.14}$$

$$\sum_{j\in I^\alpha}X^j_i\leq\sum_{j\in I^\alpha}\hat{X}^j_i,\quad i\notin I^\alpha, \tag{5.15}$$

$$\sum_{j\in I^\alpha}l^j\leq L^\alpha, \tag{5.16}$$

$$X^0_j\geq 0,X^j\geq 0,\quad l^j\geq 0,\quad j\in I^\alpha. \tag{5.17}$$

Набор переменных $\left\{\hat{X}^0_j,\hat{X}^j,\hat{l}^j\,\middle|\,j\in I^\alpha\right\}$ является решением задачи (5.13)–(5.17). Множители Лагранжа p, s удовлетворяют

$$(\hat{X}^j,\hat{l}^j)\in\operatorname{Arg\,max}\left\{\tilde{p}_jF_j(X^j,l^j)-\tilde{p}X^j-\tilde{s}l^j\,\middle|\,X^j\geq 0,l^j\geq 0\right\},\quad j\in I^\alpha,$$

$$\tilde{p}_j\left(F_j(\hat{X}^j,\hat{l}^j)-\hat{X}^0_j-\sum_{i=1}^m\hat{X}^i_j\right)=0,\quad j\in I^\alpha,$$

$$h_\alpha(\tilde{p}^\alpha)G_\alpha\left(\hat{X}^0_j\,\middle|\,j\in I^\alpha\right)=\sum_{j\in I^\alpha}\tilde{p}_j\hat{X}^0_j.$$

Следовательно, набор $\left\{h_\alpha(\tilde p^\alpha), \tilde p, \tilde s\right\}$ является множителями Лагранжа для задачи (5.13)–(5.17).

Задача (5.1) является двойственной по Янгу к задаче (5.13)–(5.17). В силу предложения 2 вектор $\tilde p^\alpha$ является решением задачи (5.1), т.е.

$$H^\alpha(\tilde p^1, \ldots, \tilde p^{\alpha-1}, \tilde p^{\alpha+1}, \ldots, \tilde p^\nu, \tilde s) = h_\alpha(\tilde p^\alpha),$$

и, значит,

$$\left(\sum_{j\in I^\alpha} \hat X_i^j, L^\alpha \,\middle|\, i \in I^\beta, \beta \ne \alpha\right) \in \partial H^\alpha(\tilde p^1, \ldots, \tilde p^{\alpha-1}, \tilde p^{\alpha+1}, \ldots, \tilde p^\nu, \tilde s).$$

Кроме того,

$$\begin{aligned} R^\alpha\left(G_\beta\left(\sum_{j\in I^\alpha} \hat X_i^j \,\middle|\, i \in I^\beta\right), L^\alpha \,\middle|\, \beta \ne \alpha\right) &= \inf_{\left\{p^\beta \ge 0, s\ge 0 \,\middle|\, \beta\ne\alpha, H^\alpha(p^\beta, s|\beta\ne\alpha)>0\right\}} \frac{\sum\limits_{\beta\ne\alpha}\sum\limits_{i\in I^\beta}\sum\limits_{j\in I^\alpha} p_i \hat X_i^j + sL^\alpha}{H^\alpha(p^1, \ldots, p^{\alpha-1}, p^{\alpha+1}, \ldots, p^\nu, s)} = \\ &= \frac{\sum\limits_{\beta\ne\alpha}\sum\limits_{i\in I^\beta}\sum\limits_{j\in I^\alpha} \tilde p_i \hat X_i^j + \tilde s L^\alpha}{H^\alpha(\tilde p^1, \ldots, \tilde p^{\alpha-1}, \tilde p^{\alpha+1}, \ldots, \tilde p^\nu, \tilde s)} = \frac{\sum\limits_{\beta\ne\alpha}\sum\limits_{i\in I^\beta}\sum\limits_{j\in I^\alpha} \tilde p_i \hat X_i^j + \tilde s L^\alpha}{h_\alpha(\tilde p^\alpha)}. \end{aligned} \tag{5.18}$$

Заметим, что

$$\begin{aligned} 0 &= \sum_{j\in I^\alpha} \tilde p_j \left(F_j(\hat X^j, \hat l^j) - \hat X_j^0 - \sum_{i=1}^m \hat X_j^i\right) + \tilde s\left(L^\alpha - \sum_{j\in I^\alpha} \hat l^j\right) = \\ &= \sum_{j\in I^\alpha} (\tilde p_j F_j(\hat X^j, \hat l^j) - \tilde p \tilde X^j - \tilde s \tilde l^j) - \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_j \hat X_j^i + \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_i \hat X_i^j + \tilde s L^\alpha = \\ &= \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_i \hat X_i^j + \tilde s L^\alpha - \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_j \hat X_j^i. \end{aligned}$$

Отсюда следует, что

$$\sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_i \hat X_i^j + \tilde s L^\alpha = \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_j \hat X_j^i. \tag{5.19}$$

В силу определения преобразования Янга справедливо неравенство

$$\sum_{j\in I^\alpha}\sum_{i\in I^\beta} \tilde p_j \hat X_j^i \ge h_\alpha(\tilde p^\alpha) G_\alpha\left(\sum_{i\in I^\beta} \hat X_j^i \,\middle|\, j \in I^\alpha\right).$$

Из (5.18) и (5.19) получаем, что

$$\begin{aligned} h_\alpha(\tilde p^\alpha) R^\alpha\left(G_\beta\left(\sum_{j\in I^\alpha} \hat X_i^j \,\middle|\, i \in I^\beta\right), L^\alpha \,\middle|\, \beta \ne \alpha\right) &= \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_i \hat X_i^j + \tilde s L^\alpha = \sum_{\beta\ne\alpha}\sum_{i\in I^\beta}\sum_{j\in I^\alpha} \tilde p_j \hat X_j^i \ge \\ &\ge \sum_{\beta\ne\alpha} h_\alpha(\tilde p^\alpha) G_\alpha\left(\sum_{i\in I^\beta} \hat X_j^i \,\middle|\, j \in I^\alpha\right) = \sum_{\beta\ne\alpha} h_\alpha(\tilde p^\alpha) Y_\alpha^\beta. \end{aligned}$$

Теорема 3 доказана.

Замечание. Агрегированные балансы (5.6) не зависят от конкретного вида функции G_0, описывающей спрос внешнего потребителя. Поэтому конструкцию построения балансов (5.6) можно рассматривать как агрегирование межотраслевых балансов (2.2). Будем называть условия (5.2) из теоремы 2 условиями агрегирования балансов. Анализ условия (5.2) сводится к задаче о слабой отделимости.

Предложение 4. *Пусть* $H(p^1, p^2) \in \Phi_{m+n} \cap C^2(R_+^{m+n})$, $h(p^1) \in \Phi_m \cap C^2(R_+^m)$. *Для того чтобы* $H(p^1, p^2) = r(h(p^1), p^2)$, *где* $r(h, p^2) \in \Phi_{n+1} \cap C^2(R_+^{n+1})$, $p^1 = (p_1^1, \ldots, p_m^1) \in R_+^m$, $p^2 = (p_1^2, \ldots, p_n^2) \in R_+^n$, *необходимо и достаточно, чтобы существовала функция* $\psi(p^1, p^2) \in C^1(R_+^{m+n})$ *такая, что*

$$\frac{\partial H(p^1, p^2)}{\partial p_j^1} = \psi(p^1, p^2)\frac{\partial h(p^1)}{\partial p_j^1}, \quad j = 1, \ldots, m,$$

$$\sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 H(p^1, p^2)}{\partial p_i^1 \partial p_j^1} u_i u_j = \psi(p^1, p^2) \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 h(p^1)}{\partial p_i^1 \partial p_j^1} u_i u_j$$

для любых $u = (u_1, \ldots, u_m)$ *таких, что*

$$\sum_{j=1}^m \frac{\partial h(p^1)}{\partial p_j^1} u_j = 0.$$

Доказательство. *Необходимость.* Поскольку $H(p^1, p^2) = r(h(p^1), p^2)$, где $r(h, p^2) \in \Phi_{n+1} \cap C^2(R_+^{n+1})$, то

$$\frac{\partial H(p^1, p^2)}{\partial p_j^1} = \left.\frac{\partial r(h, p^2)}{\partial h}\right|_{h=h(p^1)} \frac{\partial h(p^1)}{\partial p_j^1}, \quad j = 1, \ldots, m.$$

Положим

$$\psi(p^1, p^2) = \left.\frac{\partial r(h, p^2)}{\partial h}\right|_{h=h(p^1)}.$$

При сделанных предположениях $\psi(p^1, p^2) \in C^1(R_+^{m+n})$.

Если

$$\sum_{j=1}^m \frac{\partial h(p^1)}{\partial p_j^1} u_j = 0,$$

то

$$\sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 H(p^1, p^2)}{\partial p_i^1 \partial p_j^1} u_i u_j = \left.\frac{\partial r(h, p^2)}{\partial h}\right|_{h=h(p^1)} \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 h(p^1)}{\partial p_i^1 \partial p_j^1} u_i u_j +$$
$$+ \left.\frac{\partial^2 r(h, p^2)}{\partial h^2}\right|_{h=h(p^1)} \left(\sum_{j=1}^m \frac{\partial h(p^1)}{\partial p_j^1} u_j\right)\left(\sum_{i=1}^m \frac{\partial h(p^1)}{\partial p_i^1} u_i\right) =$$
$$= \left.\frac{\partial r(h, p^2)}{\partial h}\right|_{h=h(p^1)} \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 h(p^1)}{\partial p_i^1 \partial p_j^1} u_i u_j = \psi(p^1, p^2) \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 h(p^1)}{\partial p_i^1 \partial p_j^1} u_i u_j.$$

Достаточность. В силу тождества Эйлера

$$\sum_{j=1}^m \frac{\partial h(p^1)}{\partial p_j^1} p_j^1 = h(p^1) > 0 \quad \text{при} \quad p^1 > 0.$$

Пусть для определенности $\dfrac{\partial h(p^1)}{\partial p_k^1} > 0$. По теореме о неявной функции существует дважды непрерывно дифференцируемая функция $\widehat{p}_k^1(h, p_1^1, \ldots, p_{k-1}^1, p_{k+1}^1, \ldots, p_m^1)$ такая, что $h(p_1^1, \ldots, p_{k-1}^1, \widehat{p}_k^1(\widehat{h}, p_1^1, \ldots, p_{k-1}^1, p_{k+1}^1, \ldots, p_m^1), p_{k+1}^1, \ldots, p_m^1) = \widehat{h}$.

Причем для $j \neq k$ имеем

$$\frac{\partial h}{\partial p_k^1}\frac{\partial \widehat{p}_k^1}{\partial p_j^1} + \frac{\partial h}{\partial p_j^1} = 0.$$

Сделаем в функции $H(p^1, p^2)$ замену переменных. Перейдем от переменных (p^1, p^2) к переменным $(p_1^1, \ldots, p_{k-1}^1, h, p_{k+1}^1, \ldots, p_m^1, p^2)$, полагая $p_k^1 = \widehat{p}_k^1(h, p_1^1, \ldots, p_{k-1}^1, p_{k+1}^1, \ldots, p_m^1)$. Заметим, что для $j \neq k$ справедливо равенство

$$\begin{aligned}
&\frac{\partial H(p_1^1, \ldots, p_{k-1}^1, \widehat{p}_k^1(h, p_1^1, \ldots, p_{k-1}^1, p_{k+1}^1, \ldots, p_m^1), p_{k+1}^1, \ldots, p_m^1, p^2)}{\partial p_j^1} = \\
&= \frac{\partial H(p^1, p^2)}{\partial p_k^1}\frac{\partial \widehat{p}_k^1(h, p_1^1, \ldots, p_{k-1}^1, p_{k+1}^1, \ldots, p_m^1)}{\partial p_j^1} + \frac{\partial H(p^1, p^2)}{\partial p_j^1} = \\
&= \psi(p^1, p^2)\left(\frac{\partial h(p^1)}{\partial p_k^1}\frac{\partial \widehat{p}_k^1(h, p_1^1, \ldots, p_{k-1}^1, p_{k+1}^1, \ldots, p_m^1)}{\partial p_j^1} + \frac{\partial h(p^1)}{\partial p_j^1}\right) = 0.
\end{aligned}$$

Откуда следует, что $H(p^1, p^2) = r(h(p^1), p^2)$, где $r(h, p^2) \in C^2(R_+^{n+1})$ и $\psi(p^1, p^2) = \left.\dfrac{\partial r(h, p^2)}{\partial h}\right|_{h=h(p^1)}$.

Кроме того, $\lambda r(h(p^1), p^2) = \lambda H(p^1, p^2) = H(\lambda p^1, \lambda p^2) = r(h(\lambda p^1), \lambda p^2) = r(\lambda h(p^1), \lambda p^2)$ при $\lambda > 0$. Откуда получаем, что если $\lambda > 0$, то справедливо равенство $r(\lambda h, \lambda p^2) = \lambda r(h, p^2)$. Заметим, что в силу положительной однородности функции $h(p^1)$ справедливо равенство

$$\sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 h(\lambda p^1)}{\partial p_i^1 \partial p_j^1} p_i^1 p_j^1 = \frac{d^2 h(\lambda p^1)}{d\lambda^2} = 0 \quad \text{при} \quad \lambda > 0, \quad p^1 > 0.$$

В силу вогнутости функции $H(p^1, p^2)$ для любых $u = (u_1, \ldots, u_m)$, $v = (v_1, \ldots, v_n)$ справедливо неравенство

$$\begin{aligned}
0 \geq &\sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 H(p^1, p^2)}{\partial p_i^1 \partial p_j^1} u_i u_j + \sum_{i=1}^m \sum_{j=1}^n \frac{\partial^2 H(p^1, p^2)}{\partial p_i^1 \partial p_j^2} u_i v_j + \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 H(p^1, p^2)}{\partial p_i^2 \partial p_j^2} v_i v_j = \\
= &\left.\frac{\partial r(h, p^2)}{\partial h}\right|_{h=h(p^1)} \sum_{i=1}^m \sum_{j=1}^m \frac{\partial^2 h(p^1)}{\partial p_i^1 \partial p_j^1} u_i u_j + \left.\frac{\partial^2 r(h, p^2)}{\partial h^2}\right|_{h=h(p^1)} \left(\sum_{j=1}^m \frac{\partial h(p^1)}{\partial p_j^1} u_j\right)\left(\sum_{i=1}^m \frac{\partial h(p^1)}{\partial p_i^1} u_i\right) + \\
&+ \sum_{j=1}^n \left.\frac{\partial^2 r(h, p^2)}{\partial h \partial p_j^2}\right|_{h=h(p^1)} v_j \left(\sum_{i=1}^m \frac{\partial h(p^1)}{\partial p_i^1} u_i\right) + \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 r(h, p^2)}{\partial p_i^2 \partial p_j^2} v_i v_j.
\end{aligned}$$

Полагая $u = p^1$ и выбирая

$$\alpha = \pm \sum_{j=1}^m \frac{\partial h(p^1)}{\partial p_j^1} p_j^1,$$

получаем, что

$$\left.\frac{\partial^2 r(h, p^2)}{\partial h^2}\right|_{h=h(p^1)} \alpha^2 + \sum_{j=1}^n \left.\frac{\partial^2 r(h, p^2)}{\partial h \partial p_j^2}\right|_{h=h(p^1)} v_j \alpha + \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 r(h, p^2)}{\partial p_i^2 \partial p_j^2} v_i v_j \leq 0$$

для произвольных значений α, v. Следовательно, $r(h, p^2) \in \Phi_{n+1} \cap C^2(R_+^{n+1})$. Предложение 4 доказано.

6. АГРЕГИРОВАНИЕ В СЛУЧАЕ ФУНКЦИЙ С ПОСТОЯННОЙ ЭЛАСТИЧНОСТЬЮ ЗАМЕЩЕНИЯ

Обозначим через J_0 множество номеров товаров, которые выпускаются рассматриваемыми отраслями и используются конечными потребителями. Обозначим через J_i множество номеров товаров, которые выпускаются рассматриваемыми отраслями и используются в качестве производственных факторов i-й отраслью, а через I_i – множество номеров первичных ресурсов, которые используются в качестве производственных факторов i-й отраслью. Будем предполагать, что каждая отрасль использует хотя бы один вид первичных ресурсов, т.е. $I_i \neq \varnothing$. Предположим, что функции

$$F_0(X) = \left(\sum_{j\in J_0}\left(\frac{X_j}{\lambda_j}\right)^{-\rho}\right)^{-1/\rho}, \quad F_i(X,l) = \left(\sum_{j\in J_i}\left(\frac{X_j}{w_{ij}}\right)^{-\rho} + \sum_{k\in I_i}\left(\frac{l_k}{v_{ik}}\right)^{-\rho}\right)^{-1/\rho}, \quad i = 1, \ldots, m,$$

где $\rho \in (-1,0)\cup(0,+\infty)$, $\lambda_j > 0$ $(j \in J_0)$, $w_{ij} > 0$ $(j \in J_i, i = 1, \ldots, m)$, $v_{ik} > 0$ $(k \in I_i, i = 1, \ldots, m)$.

Обозначим

$$a_{ij} = \begin{cases} (w_{ij})^{\frac{\rho}{1+\rho}}, & \text{если} \quad j \in J_i, \\ 0, & \text{если} \quad j \notin J_i, \end{cases}$$

$$b_{ik} = \begin{cases} (v_{ik})^{\frac{\rho}{1+\rho}}, & \text{если} \quad k \in I_i, \\ 0, & \text{если} \quad k \notin I_i. \end{cases}$$

Рассмотрим матрицы

$$A = \|a_{ij}\|_{j=1,\ldots,m}^{i=1,\ldots,m}, \quad B = \|b_{ik}\|_{k=1,\ldots,n}^{i=1,\ldots,m},$$

E – единичную матрицу $(m\times m)$. Если матрица A продуктивна, то матрица $C = \|c_{ik}\|_{k=1,\ldots,n}^{i=1,\ldots,m} = (E-A)^{-1}B$ является неотрицательной.

Следствие 2. Пусть матрица A продуктивна. Задача

$$q_0(p) \to \max, \tag{6.1}$$

$$q_i(p,s) \geq p_i \geq 0, \quad i = 1, \ldots, m, \tag{6.2}$$

имеет решение вида

$$p_j = \left(\sum_{k=1}^{n} c_{jk}(s_k)^{\frac{\rho}{1+\rho}}\right)^{\frac{1+\rho}{\rho}}, \quad j = 1, \ldots, m.$$

Агрегированная функция себестоимости и агрегированная производственная функция имеют вид

$$q_A(s) = \left(\sum_{k=1}^{n}(\gamma_k s_k)^{\frac{\rho}{1+\rho}}\right)^{\frac{1+\rho}{\rho}}, \quad F^A(l) = \left(\sum_{k=1}^{n}\left(\frac{l_k}{\gamma_k}\right)^{-\rho}\right)^{-1/\rho},$$

где

$$\gamma_k = \left(\sum_{j=1}^{m} c_{jk}(\lambda_j)^{\frac{\rho}{1+\rho}}\right)^{\frac{1+\rho}{\rho}}, \quad k = 1, \ldots, n.$$

Доказательство. Если $\rho \in (0,+\infty)$, то неравенства $q_i(p,s) \geq p_i \geq 0$, $i = 1, \ldots, m$, эквивалентны неравенствам

$$\sum_{j \in J_i} a_{ij} (p_j)^{\frac{\rho}{1+\rho}} + \sum_{k \in I_i} b_{ik} (s_k)^{\frac{\rho}{1+\rho}} \geq (p_i)^{\frac{\rho}{1+\rho}}, \quad i = 1, \ldots, m.$$

В силу продуктивности матрицы A существует неотрицательная матрица $(E - A)^{-1}$ (см. [9, с. 132]), и эти неравенства эквивалентны неравенствам

$$\sum_{k=1}^{n} c_{ik} (s_k)^{\frac{\rho}{1+\rho}} \geq (p_i)^{\frac{\rho}{1+\rho}}, \quad i = 1, \ldots, m.$$

Если $\rho \in (-1,0)$, то, аналогично, неравенства $q_i(p,s) \geq p_i \geq 0$, $i = 1, \ldots, m$, эквивалентны неравенствам

$$\sum_{k=1}^{n} c_{ik} (s_k)^{\frac{\rho}{1+\rho}} \leq (p_i)^{\frac{\rho}{1+\rho}}, \quad i = 1, \ldots, m.$$

Для любых $\rho \in (-1,0) \cup (0,+\infty)$ имеем систему неравенств

$$\left(\sum_{k=1}^{n} c_{ik} (s_k)^{\frac{\rho}{1+\rho}} \right)^{\frac{1+\rho}{\rho}} \geq p_i \geq 0, \quad i = 1, \ldots, m.$$

Функция

$$q_0(p) = \left(\sum_{j \in J_0} (\lambda_j p_j)^{\frac{\rho}{1+\rho}} \right)^{\frac{1+\rho}{\rho}} \tag{6.3}$$

монотонно не убывает по переменным p, откуда следует, что решение задачи (6.1), (6.2) имеет вид

$$p_j = \left(\sum_{k=1}^{n} c_{jk} (s_k)^{\frac{\rho}{1+\rho}} \right)^{\frac{1+\rho}{\rho}}, \quad j = 1, \ldots, m.$$

Подставляя решение в (6.3), получаем выражение для функции $q_A(s)$. Используя выражение для преобразования Янга CES-функции, получаем выражение для $F^A(l)$. Следствие 2 доказано.

Рассмотрим агрегирование межотраслевого баланса в случае CES-функций. Так же, как в разд. 5, предположим, что множество номеров отраслей и выпускаемых ими продуктов $\{1, \ldots, m\}$ разбито на непересекающиеся подмножества $\{I_\alpha \mid \alpha = 1, \ldots, \nu\}$. Будем считать, что производственные функции отраслей являются CES-функциями:

$$F_i(X,l) = \left(\sum_{j \in J_i} \left(\frac{X_j}{w_{ij}} \right)^{-\rho} + \sum_{k \in I_i} \left(\frac{l_k}{v_{ik}} \right)^{-\rho} \right)^{-1/\rho}, \quad i = 1, \ldots, m,$$

где $\rho \in (-1,0) \cup (0,+\infty)$, $w_{ij} > 0$, $(j \in J_i, i = 1, \ldots, m)$, $v_{ik} > 0$ $(k \in I_i, i = 1, \ldots, m)$.

Будем искать агрегирующие функции также в классе CES-функций:

$$G_\alpha\left(X_j \,\middle|\, j \in I^\alpha\right) = \left(\sum_{j \in I^\alpha} \left(\frac{X_j}{\lambda_{\alpha j}} \right)^{-\rho} \right)^{-1/\rho}, \quad \text{где} \quad \lambda_j^\alpha \geq 0, \quad j \in I^\alpha; \quad \alpha = 1, \ldots, \nu. \tag{6.4}$$

Обозначим через $A_{\alpha\beta} = \|a_{ij}\|_{j\in I^\beta}^{i\in I^\alpha}$, $\alpha,\beta = 1, \ldots, \nu$, подматрицу матрицы A, а через $E_{\alpha\alpha}$ – единичную матрицу, у которой $\left|I^\alpha\right|$ строк. Определим вектор-строку

$$\theta^\alpha = \left((\lambda_{\alpha j})^{\frac{\rho}{1+\rho}} \,\middle|\, j \in I_\alpha \right).$$

Следствие 3. Пусть матрица A продуктивна. Тогда матрицы $A_{\alpha\alpha}$, $\alpha = 1, \ldots, \nu$, продуктивны. Для того чтобы функции (6.4) удовлетворяли условиям агрегирования балансов (5.2), необходимо и достаточно, чтобы для любой упорядоченной пары (α,β), где $\alpha = 1, \ldots, \nu$, $\beta = 1, \ldots, \nu$, $\beta \neq \alpha$, существовало число $\mu_{\beta\alpha} \geq 0$ такое, что выполняется равенство

$$\mu_{\alpha\beta}\theta^\beta = \theta^\alpha (E_{\alpha\alpha} - A_{\alpha\alpha})^{-1} A_{\alpha\beta}. \tag{6.5}$$

Заметим, что матрицы $A_{\alpha\alpha}$, $\alpha = 1, \ldots, \nu$, продуктивны, так как продуктивна матрица A. Доказательство следствия 3 непосредственно вытекает из следствия 2.

Замечание 2. Из (6.5) следует, что вектор θ^α должен быть собственным вектором матрицы $(E_{\alpha\alpha} - A_{\alpha\alpha})^{-1} A_{\alpha\beta} (E_{\beta\beta} - A_{\beta\beta})^{-1} A_{\beta\alpha}$, где $\beta \neq \alpha$. По теореме Фробениуса–Перрона такой неотрицательный собственный вектор существует для каждого $\beta \neq \alpha$. Более того, если $\mu_{\alpha\beta} > 0$, то неотрицательный вектор $\theta^\beta = \dfrac{1}{\mu_{\alpha\beta}} \theta^\alpha (E_{\alpha\alpha} - A_{\alpha\alpha})^{-1} A_{\alpha\beta}$ является собственным вектором матрицы $(E_{\beta\beta} - A_{\beta\beta})^{-1} A_{\beta\alpha} (E_{\alpha\alpha} - A_{\alpha\alpha})^{-1} A_{\alpha\beta}$, при этом выполняются соотношения (6.5) для пар (α,β) и (β,α).

СПИСОК ЛИТЕРАТУРЫ

1. *Барро Р.Д., Сала-и-Мартин Х.* Экономический рост. М.: БИНОМ, Лаборатория знаний, 2017. 824 с.
2. *Асемоглу Д.* Введение в теорию современного экономического роста: в 2 кн. М.: Дело., РАНХ и ГС, 2018. 1624 с.
3. *Acemoglu D., Ozdaglar A., Tahbaz-Salehi A.* The network origins of aggregate fluctuations // Econometrica. 2012. V. 80. № 5. P. 1977–2016.
4. *Agaltsov A.D., Molchanov E.G., Shananin A.A.* Inverse problems in models of resource distribution // J. of Geometric Analysis. 2018. V. 28. № 1. P. 726–765.
5. *Ашманов С.А.* Введение в математическую экономику. М.: Физматлит, 1984. 294 с.
6. *Обен Ж.-П.* Нелинейный анализ и его экономические приложения. М.: Мир, 1988. 264 с.
7. *Kukushkin N.S.* Separable aggregation and existence of Nash equilibrium. Germany, University of Beilefeld, 1995, working paper № 28. 34 p.
8. *Вратенков С.Д., Шананин А.А.* Анализ структуры потребительского спроса с помощью экономических индексов. М.: ВЦ АН СССР, 1991. 62 с.
9. *Никойдо Х.* Выпуклые структуры и математическая экономика. М.: Мир, 1972. 517 с.