Излагаются теоретические основы и алгоритмические особенности классического раздела вычислительной математики — методов решения систем обыкновенных дифференциальных уравнений (ОДУ), включая две методологически связанные, но достаточно самостоятельные части, посвященные задачам Коши и краевым задачам. Приводятся основные определения, классификация и примеры изучаемых задач, а также вопросы существования, единственности, гладкости, формы представления и устойчивости решений. Специально рассматриваются актуальные жесткие, сингулярно-возмущенные и гамильтоновы системы ОДУ, а также задачи с периодическими свойствами, предельными циклами, странными аттракторами и бифуркациями.

Для задач Коши исследуются все основные вычислительные схемы одношаговых m-стадийных и многошаговых алгоритмов. Для явных и неявных методов различной точности описываются условия порядка и контрактивности, области устойчивости и различные определения сходимости, а также подходы к выбору счетных параметров, в том числе для жестких систем. Излагаются современные алгоритмы решения гамильтоновых, сингулярно-возмущенных и дифференциально-алгебраических систем уравнений. Для решения краевых задач изучаются сеточные методы конечных разностей, конечных объемов и конечных элементов с аппроксимациями разных порядков. Рассматриваются структурные и спектральные свойства матриц получаемых систем линейных алгебраических уравнений с разреженными матрицами, а также экономичные алгоритмы их решения. В конце каждой главы при-

водятся задачи по соответствующей тематике. Книга предназначена для студентов, аспирантов и научных работников, связанных с вычислительной и прикладной математикой.

ВВЕДЕНИЕ

Обыкновенные дифференциальные уравнения (ОДУ), в отличие от уравнений в частных производных, – это соотношения между функциями и их производными, зависящими только от одной независимой переменной.В этом смысле они представляют собой частный раздел уравнений математической физики (или математической химии, биологии и т. д.). Однако выделение их изучения в отдельную дисциплину совершенно оправдано. Во-первых, здесь удается исследовать достаточно глубоко свойства решений этих уравнений и на основе этого построить наиболее совершенные численные методы. Во-вторых, для многих актуальных приложений система ОДУ — самодостаточный и чрезвычайно важный математический объект, адекватно описывающий сложные процессы и/или явления. Достаточно вспомнить задачи астрономии, привлекающие внимание человечества еще с доисторических времен, а также химической кинетики, электротехники, механики и т. д. В-третьих, во многих нестационарных многомерных задачах после аппроксимации дифференциальных уравнений по пространственным переменным мы приходим к системам ОДУ большого порядка, и здесь приходят на помощь соответствующие эффективные алгоритмы.

В общетеоретическом плане ОДУ давно и достаточно хорошо исследованы, особенно для линейных случаев. Первые дифференциальные уравнения были сформулированы и ре-

шены изобретателями дифференциального и интегрального исчисления Исааком Ньютоном (1642—1727 гг.) и Готфридом Лейбницем (1646—1716 гг.). Начала теории обыкновенных дифференциальных уравнений порождены еще братьями Бернулли Якобом (1654—1705 гг.) и Иоганном (1667—1748 гг.), Л. Эйлером (1707—1783 гг.) — автором метода ломаных, первого численного алгоритма решения ОДУ, Ж.Лагранжем (1736—1813 гг.), за которым Л. Эйлер признал открытие вариационного исчисления, К.Гауссом (1777—1855 гг.), О.Коши (1836 г.— первое доказательство теоремы существования решения системы ОДУ) и основоположником теории устойчивости А. М.Ляпуновым (1857—1918 гг.). Первоочередные вопросы здесь — это существование и единственность решения, его корректность и оценки устойчивости к возмущениям исходных данных, правых частей или параметров уравнений, асимптотические свойства и сингулярности в различных характерных ситуациях. Однако затем главное внимание стали привлекать более сложные проблемы, обусловленные факторами нелинейности и параметризации решений: явления бифуркации, предельные циклы и странные аттракторы, которые привели к появлению теории динамических систем, хаоса и катастроф, заложенной в работах Анри Пуанкаре (1854—1912 гг.).

Наиболее глубоко удается изучить линейные системы ОДУ, а в особенности — самые простые уравнения с постоянными коэффициентами. Это вполне естественно, и основным орудием исследования здесь является аппарат линейной алгебры. Зачастую полученные для простейших случаев ре-

зультаты являются ориентиром при разработке алгоритмов решения более сложных задач — или нелинейных, или даже линейных, но имеющих переменные коэффициенты уравнений.

Однако так бывает далеко не всегда, и яркой тому иллюстрацией являются так называемые жесткие системы ОДУ, к решению которых приковано внимание многих математиков, начиная со второй половины XX в. С практической, или физической, точки зрения такие задачи характерны наличием как гладко меняющихся со временем, так и быстро растущих (или убывающих) компонент решения. В математическом формализме это соответствует сильному разбросу значений коэффициентов, а также очень плохой обусловленности матриц в системах ОДУ. Появился также и специальный термин — нелинейная неустойчивость. В результате возникшие насущные проблемы инициировали новое поколение алгоритмов и поток публикаций с оригинальными теоретическими подходами: одностороннее условие Липшица, понятия контрактивности, логарифмической нормы, B-сходимости и т. д.

Среди направлений, активно развиваемых в последние десятилетия, следует выделить численные методы решения гамильтоновых систем уравнений, основанных на лагранжевых вариационных принципах и обладающих замечательными свойствами сохранения энергии и других инвариантов, которые важно наследовать при построении алгоритмов интегрирования, особенно при моделировании нелинейных волновых процессов на длительных временных интервалах.

Математические проблемы, возникающие при решении

обыкновенных дифференциальных уравнений, методологически и исторически делятся на две достаточно самостоятельные части. Первая относится к решению задач Коши с заданными начальными данными для искомых функций, в которых типичная независимая переменная — это время. Изложению соответствующих вопросов посвящены первые четыре главы книги.

Первая глава содержит описание основных свойств исследуемых объектов: определения, классификация и примеры задач Коши, фазовые траектории и портреты систем ОДУ, вопросы существования, единственности, гладкости и способы представления решений. Здесь же изучаются фундаментальные понятия устойчивости и контрактивности, специальные понятия, характерные для жестких систем ОДУ и для динамических задач с такими особенностями, как периодичность, предельные циклы, странные аттракторы и т.,д., без понимания которых трудно ожидать появления эффективных и робастных (безотказных) вычислительных методов.

Во второй главе даются общие представления, терминология и спецификации численных алгоритмов, такие как локальная и глобальная погрешности, устойчивость и сходимость. Приводятся основные сведения из линейной алгебры, необходимые для исследования свойств рассматриваемых методов, а также предлагаются минимально требуемые сведения по итерационному решению систем нелинейных алгебраических уравнений (СНАУ) при использовании неявных подходов к численному интегрированию нелинейных ОДУ.

Третья и четвертая главы содержат исследования двух

главных типов алгоритмов для решения задач Коши — одношаговых m-стадийных и многошаговых методов. К первому классу относятся явные и неявные методы Рунге—Кутты (МРК), методы типа Розенброка (МтР), а также их разновидности и обобщения — диагонально неявные, разделяющиеся и коллокационные МРК, *abc*-схемы С. С.Филиппова.Для них систематически анализируются условия порядка и функции устойчивости, а также условия A-устойчивости и L-устойчивости. Отдельно рассматриваются симплектические интеграторы для гамильтоновых систем обыкновенных дифференциальных уравнений, а также специальные методы для решения систем ОДУ второго порядка и сингулярно-возмущенных задач, имеющих обширные практические приложения. Исследуются вопросы существования и единственности неявных схем Рунге—Кутты. В методах решения жестких задач определяются и исследуются свойства, базирующиеся на односторонних условиях Липшица: B-устойчивость, контрактивность, AN-устойчивость, B-согласованность и B-сходимость. Описываются важные практические подходы к автоматическому контролю точности и устойчивости вычислительных схем.

Среди многошаговых методов изучаются все основные семейства: явные и неявные алгоритмы Адамса, методы Нюстрема и Милна, формулы дифференцирования назад (ФДН) различных порядков. Рассматриваются локальные погрешности и условия устойчивости многошаговых методов, наивысшие достижимые порядки устойчивых алгоритмов, области абсолютной устойчивости, кривые локуса корней харак-

теристических многочленов, барьеры Далквиста и понятие $A(\alpha)$-устойчивости. Описываются такие обобщения, как многошаговые методы Рунге—Кутты, в том числе одноопорные многошаговые МРК. Излагаются оригинальные явные алгоритмы с переменными шагами В. И.Лебедева для устойчивого решения жестких задач.Исследуются также свойства классических многошаговых методов в случае неравномерных сеток.

Отдельный параграф посвящен новому для решения задач Коши направлению — разрывным методам Галеркина (РМГ), изначально предложенным для интегрирования многомерных уравнений. В данном случае основой является вариационная постановка, когда приближенное решение ищется в пространстве так называемых тестовых функций, по условию ортогональности невязки системы ОДУ к пространству пробных функций. При этом удается, в частности, построить алгоритмы с разными шагами для различных компонент искомого вектора.Такие подходы очень важны, например, в задачах химической кинетики, когда различные реагирующие вещества имеют отличающиеся на порядки скорости реакций.

Последняя глава является самой обширной и фактически представляет собой автономную часть книги. Хотя исторически методы решения краевых задач для ОДУ и алгоритмы численного интегрирования задач Коши описывались в отдельных книгах специалистами в разных направлениях вычислительной математики и на различных методических основах, мы хотим подчеркнуть в данном случае единство методологических принципов, основанных на сеточных аппроксимациях, устойчивости, сходимости и оптимизации алгоритмов

по условиям обеспечения требуемой точности при минимальных вычислительных затратах.

С физической точки зрения в пятой главе рассматриваются стационарные или гармонические по времени одномерные смешанные краевые задачи диффузии, теплопроводности, электродинамики и других многочисленных приложений. Среди различных способов дискретизации исходных непрерывных задач в книге с единых позиций излагаются три основных сеточных подхода. Первый — методы конечных разностей (МКР), основанные на непосредственной замене производных конечно-разностными выражениями, наиболее близок к описанной в первых главах методологии решения задач Коши для систем ОДУ. Второй подход базируется на аппроксимациях интегральных законов сохранения, являющихся следствиями (а возможно, и наоборот) исходных дифференциальных уравнений. В 1960-е г. получающиеся сеточные схемы назывались балансными, или консервативными, но в последующие годы за ними закрепилось название методов конечных объемов (МКО). Наконец, самые распространенные в последние десятилетия вычислительные инструменты — это методы конечных элементов (МКЭ), заключающиеся в построении приближенных обобщенных решений соответствующих вариационных постановок, эквивалентных в определенном смысле классическим дифференциальным формулировкам. Серьезное преимущество МКЭ заключается не только в наличии важных теоретических достижений на базе аппарата соболевских пространств, обеспечивающих разработку и обоснование эффективных методов высоких порядков точно-

сти для самого широкого класса задач математического моделирования, но и в создании уникальных компьютерных технологий, основанных на определении локальных сеточных поэлементных матриц и сборке (ассемблировании) глобальных матриц. Последние качества позволяют обеспечивать как высокий уровень автоматизации построения алгоритмов и программирования, так и возможности масштабируемого распараллеливания расчетов на современных многопроцессорных и многоядерных суперкомпьютерах.

Помимо индивидуальных аппроксимационных свойств МКР, МКО и МКЭ, значительное внимание уделяется во многом общим для этих методов матричным особенностям получаемых систем линейных алгебраических уравнений (СЛАУ), включая оценки возмущений их приближенных решений и вопросы сходимости к точному решению, а также специальным методам их решения и анализу их численной устойчивости.

Специальный параграф посвящен относительно новому направлению — разрывным методам Галеркина, активно завоевывающим широкую популярность при решении различных типов краевых задач. Данный подход в определенной степени является обобщение конечно-объемных алгоритмов, а с другой стороны — формально представляет собой смешанный метод конечных элементов со специальным выбором тестовых и пробных функций. Мы особое внимание уделяем так называемым иммерсионным алгоритмам, позволяющим учитывать особенности решений подсеточного масштаба, что актуально при моделировании процессов в многослойных средах с контрастными свойствами.

Представленный материал основан на курсе лекций, читаемых автором в течение ряда лет на механико-математическом факультете Новосибирского государственного университета. В определенном смысле он может рассматриваться как продолжение предыдущей книги автора — "Численный анализ. Часть I" [30], посвященной классическим и современным приближениям функций, производных и функционалов. Все главы данной книги снабжаются задачами, которые могут использоваться на семинарских занятиях и для самостоятельной работы студентов.

Следует сказать, что основная монографическая и учебная литература по свойствам обыкновенных дифференциальных и методам их решения относится к XX-му столетию. И если теоретические достижения ОДУ обязаны в значительной степени российским ученым и освещены в книгах В. И.Арнольда [2], [4], А. П.Карташева и Б. Л. Рождественского [35], И. Г. Петровского [50], Л. С. Понтрягина [51], М. В. Федорюка [62], то авторами основополагающих результатов по численным методам решения задач Коши, ставшими классическими, являются такие известные зарубежные специалисты, как Дж.Батчер, Дж.Далквист , Э.Хайрер , Х.Штеттер . и другие, см. монографии [10], [19], [20], [64]–[68], [72], [73].

Тем не менее в настоящей книге нашли отражение алгоритмические разработки отечественных, в том числе сибирских ученых, опубликованные в книгах и некоторых цитируемых по тексту препринтах и статьях С. С Артемьева [5], Ю. Е. Бояринцева[13], [14], Г. В. Демидова и Е. А. Новикова [48], А. Ю. Захарова [25],В. И. Лебедева [42] ,Ю. И. Кузнецова

[40], Ю. В. Ракитского [52], В. В. Смелова [58],С. С Филиппова и других.

Необходимо заметить, что рассматриваемые нами вопросы традиционно публикуются в учебниках по методам вычислений, которые также включены в приведенный список литературы. Здесь указаны как являющиеся уже давно настольными у многих математиков книги , И. С. Березина и Н. П.Жидкова [12], К. И.Бабенко [7],Н. С.Бахвалова с коллегами [11], В. И. Крылова с соавторами [39], Г. И. Марчука [45], А. А.Самарского [56], Хэмминга[67], а так и более поздние издания [1], 21], [22], [61], [68].

В данной книге изучаются только детерминистские постановки задач и методы их решения, хотя последние десятилетия получили развитие вероятностные алгоритмы для стохастических дифференциальных уравнений. Методы Монте–Карло и статистического моделирования для решения ОДУ несомненно актуальны во многих приложениях и занимают свою самостоятельную нишу в вычислительной математике, а их изучение — это интересный предмет для отдельной книги.

Важно также отметить, что основная масса классических алгоритмов решения ОДУ, а также и более современные методы, реализованы и имеются в многочисленных программных библиотеках, как свободно доступных в Интернете, так и коммерческих, см., например, [6], [26], [34], [36], [70]. Более того, среди специалистов по численному решению задач Коши различных характерных классов сложился устоявшийся “бенчмарк” (benchmark), заключающийся в накоплении,

как в традиционных публикациях, так и в электронных вариантах, результатов сравнительного тестирования методов на представительном наборе методических и практических задач. Огромная информация по этим вопросам имеется в Интернете, например, на сайте [73], а также в Википедии (см. раздел Software for ODE Solving). Такие имеющиеся обширные материалы, несомненно, очень полезны для использования не только на учебных семинарах, но и на студенческих компьютерных практикумах.

Глава 1

СВОЙСТВА РЕШЕНИЙ ЗАДАЧ КОШИ ДЛЯ СИСТЕМЫ ОДУ

Цель данной главы — представить основные понятия о задачах Коши для ОДУ, провести их классификацию, рассмотреть основные вопросы теории обыкновенных дифференциальных уравнений, а также привести свойства их решений для некоторых характерных задач. Изложение этого материала должно дать предварительное понимание той математической проблемы, для решения которой исследуются далее численные методы ее решения.

Для более глубокого изучения данных вопросов могут быть рекомендованы монографии В. И.,Арнольда [2], [4], Ю. Е.,Бояринцева [13], [14], С. К. Годунова [18],В.Д.Лисейкина [44],[45],И. Г.Петровского [50], Л. С.Понтрягина [53] , С.И.Фаддеева и В.В.Когая [64] и М. Ф.Федорюка [63].

§ 1.1. Определения, классификация и примеры задач Коши

1.1.1. Исходные обозначения и понятия. *Обыкновенным дифференциальным уравнением* (ОДУ) называется соотношение вида

$$F(t, y(t),\ y'(t), ..., y^{(n)}(t)) = 0, \tag{1.1}$$

где F — известная функция, t — независимая переменная, а $y(t)$ — неизвестная функция. Максимальный порядок n производной функции $y(t)$, входящий в (1.1), называется *порядком уравнения*.

Для обозначения производных мы также будем использовать записи $\dot{y} \equiv \dfrac{d\,y}{d\,t},\ \ddot{y} \equiv \dfrac{d^2y}{d\,t^2}$. Простейшим примером дифференциального уравнения является ОДУ первого порядка

$$y'(t) = f(t), \tag{1.2}$$

где функция $f(t)$ предполагается непрерывной на некотором отрезке вещественной оси $\Omega = [a, b]$. Уравнение (1.2) имеет бесконечно много решений, представимых в виде неопределенного интеграла $\int f(t)d\,t$, т.е. совокупности всех первообразных $F(x) + C$, где $F'(x) = f(x)$, а C – некоторая постоянная. Чтобы выделить среди этого семейства единственное решение, достаточно задать значение $y_0 = y(t_0)$ в какой-то точке $t_0 \in \Omega$. В этом случае получаем *задачу Коши* для уравнения (1.2), а ее решение при $t \in \Omega$ дается формулой

$$y(t) = y_0 + \int\limits_{t_0}^{t} f(t)d\,t. \tag{1.3}$$

Естественное обобщение рассматриваемой задачи — это переход к системе ОДУ относительно N функций $y_1(t), ..., y_N(t)$,

общая форма которой имеет вид

$$F_i(t, y_1, y_1', ..., y_1^{(n_1)}, ..., y_N,\ y_N', ..., y_N^{(n_N)}) = 0,\quad i = 1, ..., N. \tag{1.4}$$

Очевидно, что эту систему можно свести к системе из $\bar{N} = n_1 + ... + n_N$ уравнений 1-го порядка при введении соответствующих обозначений, если производные искомых функций определить как новые неизвестные:

$$y_1 = \bar{y}_1,\ y_i^{(j)} = \bar{y}_{N_i+j},\ i = 1, ..., N;\ j = 0, 1, ..., n_i;$$
$$N_0 = n_0 = 0,\ N_i = N_{i-1} + n_{i-1} + 1.$$

Если систему (1.4) представить в разрешенном относительно первых производных виде (будем считать, что она допускает эту возможность, т.е. выполняются достаточные условия теоремы о неявной функции) и добавить начальные данные для всех неизвестных функций, то получим следующую задачу Коши:

$$\begin{aligned} &y(t) = (y_1(t), \ldots, y_{\bar{N}}(t))^T, \\ &f(t, y) = (f_1(t, y), \ldots, f_{\bar{N}}(t, y))^T, \\ &\dot{y} \equiv \frac{d\,y}{d\,t} = f(t, y),\ \ y(t_0) = y_0,\ \ t_0 \leqslant t \leqslant t_e < \infty. \end{aligned} \tag{1.5}$$

Здесь $\bar{N}$ — *размерность системы* ОДУ, $\Omega = [t_0, t_e]$ — *конечный отрезок интегрирования* (далее зачастую без ограничения общности будем полагать $t_0 = 0$), $y_0 = \{y_1(t_0), ..., y_{\bar{N}}(t_0)\}^T$ — вектор начальных данных, t — независимая переменная (аргумент), $f(t, y)$ — заданная вектор-функция (правая часть) с

областью определения $D = \{t, y\} \subset \mathbb{R}^{\bar{N}+1}$, y — искомое векторное решение, а верхний знак "T" означает транспонирование вектора.

В дальнейшем мы будем рассматривать главным образом именно системы ОДУ первого порядка в виде (1.5), разрешенном относительно производных, а размерность системы будем для простоты обозначать через N.

Можно провести следующую классификацию систем уравнений вида (1.5):

- *автономные* ОДУ : $f = f(y)$ (функция f явно не зависит от t), неавтономная система, т. е. не являющаяся автономной, всегда может быть сведена к автономной путем добавления уравнения $\dot{t} = 1$;
- *линейные*: $\dot{y} = A(t)y + f(t), \quad A(t) \in \mathbb{R}^{N,N}, \ f(t) \in \mathbb{R}^N$ – матрица и свободный член системы, элементы которых не зависят от неизвестного решения; если $f(t) = 0$, система называется однородной;
- ОДУ *с постоянными коэффициентами*: $\dot{y} = Ay + b$, где элементы матрицы A не зависят от t;
- $N = 1$ — *скалярные уравнения*; когда это безразлично, мы будем употреблять термин "уравнение", не оговаривая, является ли оно векторным или скалярным.

Решением системы ОДУ называется определенная и непрерывно дифференцируемая на отрезке $[t_0, t_e]$ вектор-функция $y(t)$, которая удовлетворяет системе (1.5) в области определения $D \in \mathbb{R}^{N+1}$ ее правой части $f(t, y(t))$. График кривой $y = y(t), t = t$ при $t \in \Omega$ в $(N + 1)$ – мерном

пространстве $\mathbb{R}^{N+1}$ с координатами $(t, y_1, ..., y_N)$ называется *интегральной кривой* системы ОДУ. *Общее решение* системы — это вектор-функция $y = \varphi(t, C_1, ..., C_N)$, зависящая от постоянных $C_1, ..., C_N$, которая удовлетворяет ОДУ при любых допустимых константах, причем их можно подобрать по условию удовлетворения заданным начальным условиям $y|_{t=t_0} = y_0$. Конкретизация общего решения при каком-то наборе постоянных $C_1, ..., C_N$ называется *частным решением*.

Геометрически общее решение системы ОДУ в области D представляет собой семейство непересекающихся интегральных кривых. *Решение задачи Коши* (1.5) — это интегральная кривая, проходящая через заданную точку $(t_0, y_0) \in \mathbb{R}^{N+1}$. Пример изображения одномерных интегральных кривых приведен на рис. 1.1. Для вектор-функций $y(t)$, соответственно, могут быть определены интегральные непересекающиеся кривые $y_i(t)$ для каждой из компонент решения, представляемые соответствующими графиками.

Рис. 1.1. Пример изображения одномерных интегральных кривых

Если вектор-функция $f(t, y)$ определена в области $D \subset \mathbb{R}^{N+1}$, то совокупность векторов $(f_1(t, y), ..., f_N(t, y), 1)$ образует *векторное поле*, которому принадлежат интегральные кривые, т. е. касаются векторов этого поля в каждой точке $(t, y) \in D$. Таким образом, проблема решения задачи Коши для дифференциального уравнения может быть сформулирована в геометрическом смысле: найти такую интегральную

кривую, проходящую через заданную точку (t_0, y_0), касательная к которой в каждой точке имеет направление, совпадающее с заданным векторным полем.

Нетрудно проверить, что линейные задачи Коши обладают *принципом суперпозиции,* который в простейшем смысле можно определить следующим образом. Пусть $\hat{y}(t)$ и $\check{y}(t)$ суть решения задачи (1.5) (которая и является главным объектом наших исследований) с линейными правыми частями $\hat{f}(t)$ и $\check{f}(t)$ соответственно, удовлетворяющие начальным данным $\hat{y}^0$ и $\check{y}^0$. Если при этом $f(t) = \hat{f}(t) + \check{f}(t)$ и $y_0 = \hat{y}_0 + \check{y}_0$, то имеет место равенство $y(t) = \hat{y}(t) + \check{y}(t)$ для всех $t \in \Omega$, причем $y(t)$ — решение задачи Коши (1.5).

1.1.2. Некоторые приемы построения аналитических решений. Хотя нашей главной целью является построение и изучение численных методов решения ОДУ, необходимо отметить, что в некоторых специальных случаях удается применить приемы для нахождения аналитических решений. Мы укажем на несколько таких подходов в применении к скалярному дифференциальному уравнению.

Задача нахождения или анализа интегральных кривых иногда решается введением *изоклин* — геометрического места точек, в которых касательные к искомым интегральным кривым имеют одинаковое направление. Семейство изоклин дифференциального уравнения определяется соотношением

$$f(t, y) = k,$$

где k – числовой параметр. С помощью построения сети изоклин при разных параметрах можно исследовать свойства ин-

тегральных кривых. Например, нулевая изоклина $f(t,y)=0$ дает уравнение линии, на которых могут находиться точки максимума и минимума интегральных кривых. А приравнивая нулю вторую производную

$$y''=\frac{\partial f}{\partial t}+f(t,y)\frac{\partial f}{\partial y}=0,$$

можно найти геометрическое место точек перегиба.

Уравнения вида

$$\varphi_1(t)\psi_1(y)dt=\varphi_2(t)\psi_2(y)dy,$$

в которых коэффициенты при дифференциалах распадаются на множители, зависящие только от t или только от y, называются *уравнениями с разделяющимися переменными*. Общий интеграл таких уравнений можно записать в форме

$$\int\frac{\varphi_1(t)}{\varphi_2(t)}dt-\int\frac{\psi_2(y)}{\psi_1(y)}dy=C,$$

но здесь надо рассмотреть дополнительно частные решения, соответствующие равенству $\psi_1(y)\varphi_2(t)=0$.

Некоторые дифференциальные уравнения с помощью простых преобразований сводятся к уравнениям с разделяющимися переменными. Например, для уравнения вида

$$y'=f(at+by+c)$$

с постоянными a,b,c это можно сделать заменой переменных $z=at+by+c$, если $b\neq 0$.

Дифференциальное уравнение вида

$$M(t,y)dt + N(t,y)dy = 0$$

называется *уравнением в полных дифференциалах*, если его левая часть представляет полный дифференциал некоторой функции $u(t,y)$, т. е.

$$M\,dt + N\,dy = du = \frac{\partial u}{\partial t}dt + \frac{\partial u}{\partial y}dy.$$

Для возможности такого представления необходимо выполнения условия

$$\partial M/\partial y = \partial N/\partial t,$$

а область определения уравнения должна быть односвязной.

Общий интеграл уравнения при этом имеет вид $u(t,y) = C$ или

$$\int\limits_{t_0}^{t} M(t,y_0)dt + \int\limits_{y_0}^{y} N(t_0,y)dy = C.$$

Для некоторых дифференциальных уравнений можно так сгруппировать члены, что получаются легко интегрируемые комбинации. Например, уравнение

$$(t^3 + ty^2)dt + (t^2y + y^3)dy = 0$$

можно переписать в форме

$$t^3dt + t\,y(y\,dt + t\,dy) + y^3dy = 0,$$

откуда получаем общий интеграл

$$t^4 + 2(ty)^2 + y^4 = C.$$

В некоторых случаях удается подобрать *интегрирующий множитель* — функцию $\mu(t, y)$, после умножения на которую получаем полный дифференциал

$$d\,u = \mu M d\,t + \mu N d\,y.$$

Из условия

$$\partial(\mu\,M)/\partial\,y = \partial(\mu\,N)/\partial\,t$$

для нахождения $\mu(t, y)$ получается уравнение в частных производных, которое в отдельных конкретных случаях удается легко решить. Следует сказать, что сам поиск интегрирующего множителя в общем случае — это отнюдь не простая математическая задача. Рассматриваемые в этом пункте аналитические приемы нахождения решений ОДУ являются только иллюстрациями возможных подходов, хотя в принципе можно поставить проблему об автоматизации построения соответствующего класса алгоритмов, что значительно расширило бы рамки вычислительных технологий для данного класса уравнений.

Если решение линейного однородного дифференциального уравнения каким-то образом получено, то решение соответствующего неоднородного уравнения зачастую можно найти с помощью *метода вариации постоянных*. Проиллюстрируем его на примере уравнения

$$y' + p(t)y = q(t),$$

где $p(t)$ и $q(t)$ — заданные функции от аргумента t, непрерывные в интервале $[t_0, t_e]$. Если $q(t) \equiv 0$, то соответствующее

уравнение называется однородным и является уравнением с разделяющимися переменными, общее решение которого имеет вид

$$y = C\exp(-\int p(t)d\,t).$$

Общее решение неоднородного уравнения ищется методом вариации постоянной в следующей форме:

$$y(t) = C(t)\exp(-\int p(t)d\,t).$$

Здесь $C(t)$ есть новая неизвестная функция, которая находится после подстановки данной формулы в исходное уравнение.

В качестве примера рассмотрим уравнение

$$y' + 2\,t\,y = 2\,te^{-t^2},$$

общее решение которого при нулевой правой части имеет вид

$$y = Ce^{-t^2}.$$

Считая далее, что C есть функция от t, для нее из исходного неоднородного уравнения получаем $C'' = 2\,t$, откуда $C = t^2 + C_1$. Таким образом, искомый результат есть

$$y(t) = (t + C_1)e^{-t^2},$$

где постоянная C_1 при начальном значении $y|_{t=0} = y_0$ определяется как $C_1 = y_0 e^{t_0^2} - t_0$.

1.1.3. Примеры систем ОДУ. Мы приведем только несколько иллюстраций ОДУ, главным образом тех, которые

сыграли историческую роль в развитии не только дифференциальных уравнений, но и обширнейшей теории задач уравнений математической физики.

а. *Уравнения движения* материальной точки в поле сил, используемые еще И. Ньютоном для определения движения планет:

$$m\ddot{x} = f_x, \;\; m\ddot{y} = f_y, \;\; m\ddot{z} = f_z. \tag{1.6}$$

Здесь m — масса частицы, f_x, f_y, f_z — проекции сил, которые определяются законами или гравитации, или электромагнетизма ,или имеют какую-либо другую природу, точки означают дифференцирование по времени t, а x, y, z — координаты частицы. С равным успехом уравнения Ньютона (1.6) описывают систему взаимодействующих N частиц, и тогда все левые части уравнений снабжаются индексами $i = 1, ..., N$.

Рассмотрим для простоты одномерный случай, когда частица движется в потенциальном поле $U(x,t)$, т. е.

$$m\ddot{x} = f_x(x,t) = -\frac{\partial U(x,t)}{\partial x}. \tag{1.7}$$

Если потенциальная энергия не зависит от времени ($U = U(x)$), то при движении в таком поле сохраняется полная энергия

$$E = T(\dot{x}) + U(x) = T(\dot{x}_0) + U(x_0) = \text{const},$$

где $T(\dot{x}) = \dfrac{1}{2}m\dot{x}^2$ есть кинетическая энергия, а x_0 и $\dot{x}_0$ — начальные координата и скорость, однозначно определяющие решение задачи.

б. *Уравнения колебаний* струны были рассмотрены впервые Б. Тейлором (1715 г.) и И. Бернулли (1727 г.), причем на дискретном уровне. Струна или другая упругая среда представляется совокупностью взаимодействующих точек, причем взаимодействуют только соседние точки, а силы, действующие на i-ю точку, пропорциональны разностям смещений $y_{i-1} - y_i$ и $y_i - y_{i+1}$. В итоге мы получаем, по аналогии с (1.7), систему ОДУ

$$\begin{aligned} &y_1'' = c^2(-2\,y_1 + y_2),\\ &\cdots\cdots\cdots\cdots\cdots\\ &y_i'' = c^2(y_{i-1} - 2\,y_i + y_{i+1}),\\ &\cdots\cdots\cdots\cdots\cdots\\ &y_N'' = c^2(y_{N-1} - 2\,y_N), \end{aligned} \tag{1.8}$$

в которой предполагается, что концы струны закреплены, т.е. для крайних точек смещения относительно положения равновесия равны нулю: $y_0 = y_{N+1} = 0$. Еще Д'Аламбером в 1747 г. было указано, что в пределе при $N \to \infty$ из (1.8) следует дифференциальное волновое уравнение $\dfrac{\partial^2 u}{\partial\, t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2}$. Легко видеть, что правые части (1.8) являются конечно-разностными аппроксимациями второй производной по x и $c = a/h$, где h — расстояние между соседними взаимодействующими точками. Отметим ту характерную общность для уравнений (1.7) и (1.8), что они являются дифференциальными уравнениями второго порядка, а это является типичным для задач, связанных с движением.

в. *Дискретные уравнения теплопроводности* впервые бы-

ли описаны Био (1804 г.) и Фурье (1807 г.). Последний представил стержень совокупностью "молекул", каждая из которых имеет температуру y_i, и предположил, что энергия, передаваемая от каждой частицы к соседним, пропорциональна разности температур $y_{i-1} - y_i$ и $y_i - y_{i+1}$. Отсюда он заключил, что цепочка индивидуальных энергетических балансов частиц описывается системой ОДУ первого порядка

$$y_i' = b^2(y_{i-1} - 2\,y_i + y_{i+1}), \;\; i = 1, ..., N, \tag{1.9}$$

в которой штрих означает дифференцирование по времени, а температуры крайних точек y_0 и y_{N+1}, например, являются заданными. Именно из (1.9) Фурье предельным переходом вывел свое знаменитое уравнение теплопроводности $\dfrac{\partial\, u}{\partial\, t} = d^2 \dfrac{\partial^2 u}{\partial\, x^2}$, где $d^2 \mathrm{cong} b^2 h^2$ есть коэффициент теплопроводности, а h — расстояние между "молекулами", которое предполагается одинаковым.

г. *Уравнения лагранжевой механики* имеют происхождение от динамики Д'Аламбера и "принципа наименьшего действия" Лейбница — Мопертью. Они были опубликованы в монументальном труде Ж. Л. Лагранжа "Аналитическая механика" (1788 г.) и положили начало вариационному исчислению.

Уравнения движения Ньютона (1.6), (1.7) в потенциальном поле $U(x, t)$ можно переписать в форме *уравнений Лагранжа*, если ввести функционал, равный разности кинетической и потенциальной энергий и называемый *функцией Лагранжа*, или *лагранжианом*:

$$L(x,\dot{x},t) = T(\dot{x}) - U(x,t),$$
$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x},\ \ T(\dot{x}) = \frac{1}{2}m\dot{x}^2. \tag{1.10}$$

Полученное уравнение замечательно тем, что оно не изменяет свой вид при переходе от декартовых координат к каким-либо другим координатам, т.е. при замене $x = x(q,t)$ имеем

$$\frac{d}{dt}\frac{\partial \tilde{L}}{\partial \dot{q}} = \frac{\partial \tilde{L}}{\partial q},\ \ q = q(x,t),$$
$$\tilde{L}(q,\dot{q},t) = L\Big(x(q,t),\ \frac{dx(q,t)}{dt}, t\Big). \tag{1.11}$$

Данный результат переносится на произвольную механическую систему с $2N$ степенями свободы, функция Лагранжа которой зависит от вектора *обобщенных координат* $y = (y_1, ..., y_N)^T$, вектора *обобщенных скоростей* $\dot{y} = (\dot{y}_1, ..., \dot{y}_N)$ и времени t. Уравнения Лагранжа, представляющие собой переформулировку уравнений движения Ньютона, в данном случае записываются как

$$\frac{d}{dt}\frac{\partial L(y,\dot{y},t)}{\partial \dot{y}_i} = \frac{\partial L(y,\dot{y},t)}{\partial y_i},\ \ i = 1, ..., N. \tag{1.12}$$

На основе лагранжиана определяется функционал S, называемый *действием* и определяемый следующим интегралом:

$$S = \int_{t_1}^{t_2} L(y_1(t), ..., y_N(t), \dot{y}_1(t), ..., \dot{y}_N(t), t)dt. \tag{1.13}$$

Л.Эйлером еще до Лагранжа было показано, что необходимым условием экстремума интеграла (1.13) является выполнение *уравнения Эйлера*, заключающееся в равенстве нулю *вариационной (функциональной) производной*

$$\frac{\delta S}{\delta y(t)} = \frac{\partial L}{\partial q_i} - \frac{d}{d\,t}\frac{\partial\, L}{\partial \dot{q}_i} = 0, \;\; i = 1, ..., N. \tag{1.14}$$

Здесь мы поменяли обозначения аргументов y_i и $\dot{y}_i$ на q_i и $\dot{q}_i$, которые наиболее широко приняты в литературе для записи обобщенных координат и обобщенных скоростей. Отметим, что в этих переменных полная энергия системы частиц связывается с лагранжианом следующим соотношением:

$$E(t) = T(t) + U(t) = \sum_{i=1}^{N} \frac{\partial\, L}{\partial \dot{q}_i}\, \dot{q}_i - L. \tag{1.15}$$

Используемое в (1.14) понятие вариационной производной определяется следующим образом. Пусть задан функционал $F(y)$, означающий правило, по которому каждой реализации $y(t)$ из некоторой совокупности функций, зависящих от аргумента t и составляющих область определения функционала, ставится в соответствие вещественное число F, называемое значением функционала на этой функции. Отметим, что могут быть функциональные объекты вида $F(y, t)$, которые наряду с зависимостью от y являются также функцией от аргумента t. Примером может служить величина

$$F(y, t) = \int G(t, t')y(t')d\,t',$$

где $G(t, t')$ — заданная функция двух аргументов.

Рассмотрим теперь значения функционала F, определенные для двух функций $y(t)$ и $\delta y(t)$, полагая "функцию возмущения" $\delta y(t)$ равной нулю всюду, кроме некоторой окрестности $\Delta\, t$ точки t_0. Вариационной производной функционала $F(y)$ в точке t_0 называется предел

$$\frac{\delta F(y)}{\delta y(t_0)} = \lim_{\substack{|\Delta t|\to 0 \\ max|\delta y|\to 0}} \frac{F(y+\delta y) - F(y)}{\int\limits_{\Delta t} d\,t \delta y(t)},$$

при условии, что он существует и не зависит ни от вида $\delta y(t)$, ни от способа стягивания к нулю окрестности $|\Delta t|$, ни от закона стремления к нулю максимума модуля функции $\delta y(t)$. Если в качестве $\delta y(t)$ взять δ-функцию $\lambda\delta(t-t_0)$, где λ — амплитудный параметр, то можно получить более удобную формулу для вариационной производной:

$$\frac{\delta F(y)}{\delta y(t_0)} = \lim_{\lambda\to 0} \frac{F[y(t)+\lambda\delta(t-t_0)] - F(y(t))}{\lambda} = \Big(\frac{\partial}{\partial\lambda} F[y(t)+\lambda\delta(t-t_0)]\Big)\Big|_{\lambda=0},$$

которая сводит операцию вычисления функциональной производной к обычному дифференцированию.

Как видно из фактического совпадения (1.12) и (1.14), уравнения Эйлера вариационной задачи для функционала (1.13) есть ни что иное, как уравнения Лагранжа для механической системы. Хотя механический смысл этого вариационного принципа — *принципа наименьшего действия* при переходе системы из одного состояния в другое — получил в литературе название *принципа Гамильтона*.

д. *Гамильтоновы системы*. Сыгравшие революционную роль в механике работы Гамильтона, опубликованные в 1834 и 1835 гг., внешне выглядят как переход от одной системы уравнений к другой путем несложных преобразований переменных.

Более конкретно, вместо обобщенных скоростей $\dot{q}_i$ вводится новая переменная — *обобщенный импульс* p_i:

$$p_i = \frac{\partial L}{\partial \dot{q}_i}. \tag{1.16}$$

При этом из уравнения Лагранжа (1.14) следует соотношение

$$\dot{p}_i = \frac{\partial L}{\partial q_i}. \tag{1.17}$$

Выражения (1.16), (1.17) позволяют перейти от дифференциального уравнения Лагранжа второго порядка к системе ОДУ 1-го порядка в переменных q_i, p_i. Этот переход называется *преобразованием Лежандра*, а получаемая система $2N$-го порядка имеет вид

$$\dot{q}_i = \frac{\partial H(p,q,t)}{\partial p_i}, \;\; \dot{p}_i = -\frac{\partial H(p,q,t)}{\partial q_i}, \;\; i = 1, ..., N, \tag{1.18}$$

где функционал $H(p,q,t)$ называется *гамильтонианом* и определяется как полная энергия системы E:

$$H(p,q,t) = \sum_{i=1}^{N} p_i \dot{q}_i - L = T + U. \tag{1.19}$$

Уравнения (1.18) называются *гамильтоновой системой* и обладают следующим важным следствием:

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} + \sum_{i=1}^{N} \Big(\frac{\partial H}{\partial q_i}\dot{q}_i + \frac{\partial H}{\partial p_i}\dot{p}_i \Big) = \frac{\partial H}{\partial t}.$$

Отсюда следует, что если функционал H явно не зависит от времени, то гамильтониан системы на решении (1.18) не меняется со временем, т. е. является *инвариантом*.

Уравнения Гамильтона обладают тем свойством, что они сохраняют свою форму (1.18) при так называемых *канонических преобразованиях*, в которых новые переменные P_i, Q_i связаны со старыми условиями

$$p_i = \frac{\partial F(q,Q,t)}{\partial q_i}, \;\; P_i = \frac{\partial F(q,Q,t)}{\partial Q_i}, \;\; i = 1, ..., N, \qquad (1.20)$$

где F — *производящая функция* преобразования, а время t является при этом просто параметром преобразования. Переменные (p, q, P, Q), в которых записываются гамильтониан $H(p, q, t)$ и система (1.18), называются *каноническими переменными*. Если ввести векторы $z = (q, p)^T$ и $\nabla_z = \left(\frac{\partial}{\partial p}, \frac{\partial}{\partial q}\right)^T$, гамильтонова система (1.18) может быть записана в векторно-матричном виде

$$J\dot{z} = \nabla_z H, \qquad (1.21)$$

где $J = \begin{bmatrix} 0 & -I_N \\ I_N & 0 \end{bmatrix}$ — *симплектическая матрица*, которая является кососимметричной и порядка $2N$ (I_N — единичная матрица N-го порядка).

Классическим примером гамильтоновой системы являются *уравнения Кеплера*

$$\dot{q} = p, \quad \dot{p} = f(q) = -q/||q||^3, \;\; ||q||^2 = q^T q,$$

где $q = (q_1, q_2)^T, p = (p_1, p_2)^T$ — координаты положения и угловые моменты движения планеты относительно Солнца, а $||\cdot||$ означает евклидову норму вектора. Гамильтониан системы Кеплера равен

$$H(p, q) = p^T p/2 + 1/||q||.$$

Вопросам исследования свойств гамильтоновых систем дифференциальных уравнений и их многочисленных актуальных приложений посвящена обширная специальная

литература, см., например, [3], [63], [65].

е. *Сингулярно-возмущенные и дифференциально-алгебраические уравнения (ДАУ).* Специфику данного класса задач, очень важного для практики, можно увидеть из простого уравнения второго порядка

$$y'' + \alpha\, y' + y = 0,$$

описывающего нелинейные осцилляции, например, в электрических цепях. При $\alpha > 0$ его решения являются затухающими, а при $\alpha < 0$ — неустойчиво растущими. Если здесь положить

$$\alpha = \mu(y^2 - 1),\ \mu > 0,$$

то мы получаем знаменитое *уравнение Ван-дер-Поля*

$$y'' + \mu(y^2 - 1)y' + y = 0,$$

изученное им в 1920–1926 гг., которое можно переписать в виде системы двух уравнений первого порядка:

$$y_1' = y_2,\ y_2' = \mu(1 - y_1^2)y_2 - y_1.$$

Несложный анализ показывает, что в этой системе малые колебания усиливаются, а большие затухают, причем существует устойчивое периодическое решение (предельный цикл), к которому сходятся все остальные решения.

Приведем теперь уравнение к несколько иной форме, исследованной А.А.Дородницыным в 1947 г.:

$$\varepsilon\, z'' + (z^2 - 1)z' + z = 0,$$

в которой коэффициент при старшей производной будем считать очень малым, т. е. $\varepsilon \ll 1$, и запишем ее в виде системы

$$\begin{aligned} y' &= -z = f(y,z), \\ \varepsilon z' &= y - \left(\tfrac{z^3}{3} - z\right) = g(y,z). \end{aligned}$$

В плоскости (y,z) решения этих уравнений при малых ε имеют быстрые осцилляции, сходящиеся к некоторому многообразию M, определяемому равенством $y = z^3/3 - z$, где они становятся гладкими. Чтобы аппроксимировать решение при очень малых ε, положим $\varepsilon = 0$, в результате чего получаем дифференциально-алгебраическую систему

$$y' = -z = f(y,z),\ \ 0 = y - (y^3/3 - y) = g(y,z),$$

представляющую собой комбинацию из дифференциального и алгебраического уравнений, которая называется *приведенной системой* и легко решается:

$$y' = -z = (z^2 - 1)z',\ \ l\,n|z| - \frac{z^2}{2} = t + c.$$

Последняя задача имеет смысл, если только начальные данные согласованы, т. е. принадлежат многообразию M.

Рассмотрим теперь некоторые свойства систем ОДУ с малым параметром : $\varepsilon \to 0$ при производной, которые называются сингулярно-возмущенными. Их главная особенность заключается в том, что в предельном случае $\varepsilon = 0$ меняется порядок дифференциального уравнения и, как следствие, — характер поведения решения.

Пусть имеется задача

$$y' = f(y,z),\ \ \varepsilon z' = g(y,z),\ \ y(t_0) = y_0,\ \ z(t_0) = z_0,$$

где y и z — векторы, а f и g — достаточное число раз дифференцируемые векторные функции той же размерности, что y и z соответственно. Такой приведенной системой будет ДАУ

$$y' = f(y,z),\ \ 0 = g(y,z),$$

для которого начальные значения согласованы, если $0 = g(y_0, z_0)$. Будем предполагать, что *матрица Якоби* $g_z(y,z)$ обратима (такое дифференциально-алгебраическое уравнение называется уравнением индекса 1), что по теореме о неявной функции обеспечивает существование локально единственного решения алгебраического уравнения $z = G(y)$. После его подстановки в дифференциальное уравнение мы получаем так называемое ОДУ в пространстве состояний

$$y' = f(y, G(y)).$$

Если якобиан $g_z(y,z)$ является вырожденным, то мы приходим к дифференциально-алгебраическим уравнениям высших индексов. Их особенности рассмотрим на простейших линейных ДАУ с постоянными матричными коэффициентами

$$B\,u' + A\,u = s(t).$$

Рассматривая решения вида $e^{\lambda t}u_0$ (когда $s(t) \equiv 0$), мы приходим к рассмотрению пучка матриц $A + \lambda\,B$. Если матрица $A + \lambda\,B$ вырождена при всех значениях λ, то ДАУ при заданных начальных условиях либо не имеет решения, либо имеет бесконечно много решений. Будем предполагать, что

пучок $A + \lambda B$ является регулярным, т. е. многочлен не обращается тождественно в нуль. Для исследования рассматриваемого ДАУ используем следующий результат из теории матриц (Вейерштрасс, 1868):

если $A+\lambda B$ есть регулярный пучок матриц размерности N, то существуют неособенные матрицы P и Q такие, что

$$PAQ = \begin{bmatrix} J_{N-m} & 0 \\ 0 & I_m \end{bmatrix}, \quad PBQ = \begin{bmatrix} I_{N-m} & 0 \\ 0 & J_m \end{bmatrix},$$

где I_m есть единичная матрица порядка m, $J_m = blockdiag(J_1, ..., J_k)$, а каждый блок J_i размерности $m_i, m = \sum_{i=1}^{k} m_i$, имеет жордановый вид

$$J_i = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & & 1 \\ 0 & & & 0 \end{bmatrix}.$$

Умножая наше линейное ДАУ слева на матрицу P и используя преобразования

$$u = Q \begin{bmatrix} y \\ z \end{bmatrix}, \quad P\,s(t) = \begin{bmatrix} f(t) \\ g(t) \end{bmatrix}, \quad z = (z_{m_1}, ..., z_{m_k})^T,$$

мы получаем распадающуюся дифференциально-алгебраическую систему

$$y' + J_{N-m} y = f(t), \;\; J_m z' + z = g(t).$$

Таким образом, для y мы получили обыкновенное дифференциальное уравнение, а система для z, в свою очередь, распадается на k подсистем, каждая из которых записывается в форме

$$\bar{z}_2' + \bar{z}_1 = \bar{g}_1, ..., \bar{z}_{m_i}' + \bar{z}_{m_{i-1}} = \bar{y}_{m_{i-1}}, \bar{z}_{m_i} = \bar{g}_{m_i},$$

где у подвекторов $z_{m_i} = (\bar{z}_1, ..., \bar{z}_{m_i})$, $g_{m_i} = (\bar{g}_1, ..., \bar{g}_{m_i})$ для простоты использована локальная нумерация их компонент. Здесь последнее уравнение определяет $\bar{z}_{m_i}$, а остальные компоненты вычисляются рекурсивно повторным дифференцированием, и в итоге $\bar{z}_1$ зависит от $(m_i - 1)$-й производной $\bar{g}_{m_i}$. Поскольку численное дифференцирование является неустойчивой процедурой, то величина $\bar{m} = \max_i(m_i)$ представляет меру возможного накопления ошибок при реализации алгоритма. Значение $\bar{m}$ называется *индексом нильпотентности пучка матриц* $A + \lambda B$, а в применении к ДАУ с постоянными матрицами — *индексом дифференциально-алгебраического уравнения*. На обобщении и исследовании этого понятия для более сложных задач мы останавливаться не будем, см. [63].

§ 1.2. Фазовые траектории и портрет системы ОДУ

Если $y_i = y_i(t)$ и $y_k = y_k(t)$ — некоторые две компоненты решения, то они определяют кривую в плоскости y_i, y_k. Эта кривая называется *фазовой траекторией*, сама плоскость — *фазовой плоскостью*, а картина, образуемая различными траекториями, — *фазовым портретом* решения системы ОДУ. Рассмотрение фазовых траекторий и портретов систем ОДУ в

различных плоскостях зачастую является удобным средством качественного анализа решений.

Для линейной однородной системы ОДУ (имеющей нулевой свободный член) одна из фазовых траекторий легко находится: такая система имеет нулевое решение $y_i(t) = 0, i = 1, ..., N$, и фазовая траектория — точка $(0, ..., 0)$. Эта точка называется *точкой покоя*, или *положением равновесия*.

Дадим еще одно понятие, которое в определенном смысле привносит геометрический подход к решению ОДУ: *поток φ_t дифференциального уравнения (1.5) по времени t* (далее — просто поток) — это отображение, которое для любой точки y_0 фазового пространства сопоставляет значение $y(t)$ решения обыкновенного дифференциального уравнения с начальным значением $y(0) = y_0$, т. е.

$$\varphi_t : \ \varphi_t(y_0) = y(t), \quad y(0) = y_0. \tag{1.22}$$

Приведем в качестве примера линейную однородную систему ОДУ с постоянными вещественными коэффициентами

$$\begin{aligned} y_1' &= a_{1,1}y_1 + a_{1,2}y_2, \\ y_2' &= a_{2,1}y_1 + a_{2,2}y_2. \end{aligned} \tag{1.23}$$

Пусть λ_1, λ_2 — собственные числа, а $z^{(k)} = (z_1^{(k)}, z_2^{(k)}), k = 1, 2$, — соответствующие собственные векторы матрицы $A = \{a_{i,k}\}$ системы (1.23).

Если оба корня ее характеристического уравнения вещественны и различны, то вектор вещественного решения уравнений (1.23) записывается в виде

$$y(t) = c_1 e^{\lambda_1 t} z^{(1)} + c_2 e^{\lambda_2 t} z^{(2)}, \tag{1.24}$$

где c_1, c_2 — некоторые постоянные. Координаты ξ_1, ξ_2 вектора $y(t)$ из собственных векторов в базисе определяются как

$$\xi_1 = c_1 e^{\lambda_1 t},\ \xi_2 = c_2 e^{\lambda_2 t}.$$

Предположим сначала, что числа λ_1 и λ_2 одного знака. При этом в отдельности надо рассмотреть два случая.

а. $\lambda_1 < 0, \lambda_2 < 0$: если $c_1 > 0, c_2 = 0$, то фазовая траектория — ось ξ_1, если $c_1 = 0, c_2 > 0$ — ось ξ_2; фазовый портрет системы для этого случая приведен на рис. 1.2, где стрелки показывают направление, в котором движется точка фазовой траектории при : $t \to \infty$; поскольку $y(t)$ стремится к точке покоя $(0, 0)$, получаемая картина называется *устойчивым узлом*.

б. $\lambda_1 > 0, \lambda_2 > 0$: фазовый портрет такой системы изображается тем же рис. 1.2, только стрелки на осях ξ_1, ξ_2 в данном случае направлены от начала координат; соответствующая картина называется *неустойчивым узлом*.

Обратимся теперь к случаю разных знаков собственных чисел, считая для определенности $\lambda_1 > 0, \lambda_2 < 0$. Если один из коэффициентов c_1, c_2 равен нулю, то фазовые траектории проходят по одной из осей ξ_1, ξ_2. В противном случае траектории имеют вид «гипербол». Получаемый фазовый портрет называется *седлом* и приведен на рис. 1.3.

Перейдем далее к случаю комплексных собственных чисел матрицы A в системе (1.23). Обозначая $\lambda_1 = \lambda$, будем иметь $\lambda_2 = \bar{\lambda}$, где черта означает комплексное сопряжение. Если z есть собственный вектор матрицы ($A\,z = \lambda\,z$), то получаем

также $A\bar{z} = \bar{\lambda}\bar{z}$, т. е. $\bar{z}$ есть собственный вектор, отвечающий $\bar{\lambda}$. При этом всякое решение системы (1.23) по-прежнему имеет вид (1.24), а каждое вещественное решение представимо в форме

$$y(t) = c\, e^{\lambda t} z + \bar{c} e^{\bar{\lambda} t} \bar{z},$$

где c — произвольная комплексная постоянная. Положим

$$\lambda = \alpha + i\beta, \;\; z = g_1 - i\, g_2, \;\; c = a + i\, b,$$

где i — мнимая единица, а числа α, β, a, b и векторы g_1, g_2 вещественны. Тогда

$$\begin{aligned} y(t) &= \xi_1 g_1 + \xi_2 g_2, \\ \xi_1 &= 2\, e^{\alpha t}(a \cos \beta t - b \sin\, b\, t), \\ \xi_2 &= 2\, e^{\alpha t}(b \cos \beta t + a \sin\, b\, t). \end{aligned}$$

Рис. 1.2. Фазовый портрет для системы (1.23) при $\lambda_1 < 0, \lambda_2 < 0$

Рис. 1.3. Фазовый портрет для системы (1.23) при $\lambda_1 > 0, \lambda_2 < 0$

Отсюда получаем следующие типы фазовых портретов:

а1) *центр*: $\alpha = 0$, т. е. оба корня $\pm i\beta$ чисто мнимые; в этом случае имеем

$$\xi_1 = \rho\cos(\beta\, t + \gamma),\ \ \xi_2 = \rho\sin(\beta t + \gamma),$$
$$\rho = 2\sqrt{a^2 + b^2},\ \ tan\gamma = b/a,$$

а фазовые траектории — эллипсы, изображенные на рис. 1.4; направление их обхода зависит от знака β (рисунок соответствует $\beta > 0$);

Рис. 1.4. Фазовый портрет для мнимых собственных чисел матрицы A в (1.23)

б1) *фокус*: $\alpha \neq 0$; здесь, в свою очередь, различаются случаи $\alpha < 0$ (*устойчивый фокус*) и $\alpha > 0$ (*неустойчивый фокус*); для обоих знаков фазовые траектории являются спиралями, изображенными на рис. 1.5; для отрицательных α спирали закручиваются в начало координат при $t \to \infty$, а для положительных с ростом t точка траектории уходит на бесконечность.

Рис. 1.5. Фазовые траектории для комплексных собственных чисел при $\alpha < 0$
(устойчивый фокус)

Освободимся теперь от требования различия собственных чисел матрицы A системы (1.23). При наличии их кратности отличие наблюдается только в том случае, когда A не является симметричной или нормальной матрицей. При этом существует невырожденная матрица T, с помощью которой *преобразованием подобия* A сводится к *жордановой форме*:

$$T\,A\,T^{-1} = J = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}.$$

В силу этого систему ОДУ (1.23) можно преобразовать (путем умножения слева на T и справа на T^{-1}) к виду

$$y' = J\,y. \tag{1.25}$$

Решение этой системы определяется формулами, в которых, как и выше, e_1 и e_2 обозначают вектор-орты:

$$y_1 = e^{\alpha\,t}e_1,\ \ y_2 = e^{\alpha\,t}(t\,e_1 + e_2).$$

Из рассмотренного простого примера линейной однородной системы наглядно видно многообразие решений ОДУ. Приведенные их представления достаточно легко обобщаются и на системы более высоких порядков, но эти вопросы уже относятся к качественной теории обыкновенных дифференциальных уравнений, см., например, [2], [4].

§ 1.3. Вопросы существования, единственности и гладкости решений

В этом параграфе мы приведем (без доказательства, как в основном и на протяжении всей книги) несколько основных утверждений.

Теорема 1.1 (существования и единственности). *Пусть D — область в пространстве $\mathbb{R}^{N+1}$ с координатами $t, y_1, ..., y_N$, а вектор-функция $f(t, y)$ и ее производные $\dfrac{\partial\, f_i(t,y)}{\partial\, y_j}, 1 \leqslant i, j \leqslant N$, определены и непрерывны для всех*

$(t, y) \in D$. *Тогда решение задачи Коши* (1.5) *при* $(t_0, y_0) \in D$ *существует и единственно на некотором интервале* $(t_0 - \delta, t_0 + \delta), \delta > 0$.

Геометрическая интерпретация этой основной теоремы такова: при сформулированных условиях через каждую точку области D проходит интегральная кривая, и притом только одна.

Свойства гладкости решений определяются следующим утверждением.

Теорема 1.2 (о гладкости решений ОДУ). *Пусть вектор-функция* $f(t, y)$ *обладает непрерывными производными по всем переменным до порядка* $p \geqslant 1$ *включительно. Тогда всякое решение системы ОДУ имеет непрерывные производные до порядка* $p + 1$.

Это означает, что чем более гладкая правая часть, тем глаже само решение.

Для линейной задачи Коши

$$\dot{y} = A(t)\, y + f(t), \;\; y(t_0) = y_0, \;\; A = \{a_{i,j}(t)\} \tag{1.26}$$

существование и единственность решения устанавливается не локально, а глобально, т. е. на всем отрезке $\Omega = (t_0, t_e)$.

Теорема 1.3 (существование и единственность решений линейных ОДУ). *Пусть вектор-функция* $f(t)$ *и матрица* $A(t)$ *непрерывны на интервале* $\Omega = (t_0, t_e)$. *Тогда решение задачи Коши* (1.26) *существует и единственно на всем интервале* Ω.

Данную теорему можно переформулировать как существование однозначной продолжимости решения на весь интервал.

Ниже для линейных систем мы будем рассматривать только решения, продолженные на весь интервал.

Замечание 1.1. Условия теорем 1.1, 1.3 можно ослабить в том смысле, что для существования и единственности решения системы ОДУ (1.5) достаточно выполнения *условия Липшица*

$$|f_i(t,\hat{y}_1,...,\hat{y}_N) - f_i(t,\check{y}_1,...,\check{y}_N)| \leqslant L\sum_{i=1}^{N}|\hat{y}_i - \check{y}_i|, \;\; i=1,...,N, \tag{1.27}$$

где $L>0$ — *постоянная Липшица*.

Сделаем еще одно замечание: скалярное дифференциальное уравнение n-го порядка

$$y^{(n)} + a_1(t)y^{(n-1)} + ... + a_n(t)y = g(t)$$

сводится к системе ОДУ типа (1.26) для векторов с компонентами

$$y_i = y^{(i-1)}, \quad f_n = g(t), \quad f_i(t) = 0, \quad i = 1,...,n-1,$$

а матрица $A = \{a_{i,j}\} \in R^{n,n}$ при этом имеет следующий вид $(a_{i,i+1}=1,\; i=1,...,n-1)$:

$$A = \begin{bmatrix} 0 & 1 & & & & \\ 0 & 0 & 1 & & & 0 \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & \cdot & 0 & 1 \\ -a_n & -a_{n-1} & \dots & \dots & \dots & -a_1 \end{bmatrix}.$$

Таким образом, условия теоремы 1.3 в данном случае — это непрерывность функций $g(t)$ и $a_1(t), ..., a_n(t)$.

§ 1.4. Представления решений ОДУ

Получить выражения для решений и изучить их свойства — это один из главных вопросов теории систем ОДУ, и в достаточно полной форме такое удается сделать только для линейных уравнений. Однако этот первый шаг необходим для исследования более общих случаев.

Простейшая линейная однородная задача Коши

$$\dot{y} = A\,y,\ y(t_0) = y_0,\ \ A = \{a_{i,j}; i, j = 1, ..., N\}$$

с постоянными (не зависящими от времени t) элементами матрицы A имеет единственное решение

$$y(t) = e^{A(t-t_0)}y_0,$$

выражаемое с помощью экспоненциального оператора, на определении и свойствах которого мы остановимся ниже.

1.4.1. Однородные системы ОДУ. Рассмотрим линейную однородную систему

$$\dot{y} = A(t)y, \quad y \in \mathbb{N}, \quad A \in \mathbb{R}^{N,N}, \tag{1.28}$$

в которой матрица-функция $A(t)$ непрерывна на отрезке Ω и $t \in \Omega$.

Определение 1.1. *Фундаментальной системой решений уравнения* (1.28) *называется набор из N векторных линейно независимых на интервале Ω решений $y^1(t), ..., y^N(t)$, т. е. таких, для которых равенство*

$$c_1 y^1(t) + c_2 y^2(t) + ... + c_N y^N(t) \equiv 0, \;\; t \in \Omega,$$

выполняется только при постоянных $c_1, ..., c_N$, *равных нулю одновременно.*

Совокупность решений линейных систем ОДУ образует линейное пространство решений, в котором можно рассматривать различные базисы.

Теорема 1.4. *Фундаментальные системы решений уравнения* (1.28) *существуют, причем всякое решение* (1.28) *может быть представлено в виде*

$$y(t) = c_1 y^1(t) + ... + c_N y^N(t), \tag{1.29}$$

где $c_1, ..., c_N$ — *произвольные постоянные.*

Линейная зависимость или независимость системы N векторных функций $y^i(t)$ на интервале $\Omega = (t_0, t_e)$, каждая из которых имеет N компонент $y^i_j(t), j = 1, ..., N$, определяется введенной в 1810 г. Г. Вронским матрицей

$$W(t) = \{y^i_j(t); i, j = 1, ..., N\} =$$
$$= W(y^1, ..., y^N) = \begin{vmatrix} y^1_1(t) & y^1_2(t) & ... & y^1_N(t) \\ y^2_1(t) & y^2_2(t) & ... & y^2_N(t) \\ \cdots & \cdots & \cdots & \cdots \\ y^N_1(t) & y^N_2(t) & ... & y^N_N(t) \end{vmatrix}, \tag{1.30}$$

носящей его имя (определитель ее называется *вронскианом*).

Теорема 1.5 (о линейной зависимости или независимости системы векторных функций). *Если вронскиан системы вектор-функций* $y^1(t), ..., y^N(t)$ *отличен от нуля хотя бы в одной точке* $t \in \Omega$, *то эти векторные функции*

линейно независимы на Ω. *Если же данные функции линейно зависимы, то их вронскиан тождественно равен нулю на* Ω.

Подчеркнем, что теорема 1.5 относится к системе функций, никак не связанной с системой ОДУ. В противном случае имеет место более сильное утверждение.

Теорема 1.6. *Пусть вектор-функции* $y^1(t), ..., y^N(t)$ — *решения однородной системы* (1.28). *Если их вронскиан* $W(t)$ *обращается в нуль хотя бы в одной точке* $t \in \Omega$, *то эти функции линейно зависимы и их вронскиан тождественно равен нулю на* Ω.

Очевидно, что матрица Вронского фундаментальной системы решений ОДУ удовлетворяет уравнению

$$W'(t) = A(t)W(t), \tag{1.31}$$

а любое решение вида (1.29) может быть записано в форме

$$y(t) = W(t) \cdot c, \quad c = (c_1, ..., c_N)^T. \tag{1.32}$$

Если это решение удовлетворяет начальным данным $y(t_0) = y_0 = (y_0^1, ..., y_0^N)$, то мы имеем $c = W^{-1}(t_0)y_0$, откуда после подстановки c в (1.32) следует формула

$$y(t) = R(t, t_0)y_0, \quad R(t, t_0) = W(t)W^{-1}(t_0), \tag{1.33}$$

где матрица $R(t, t_0) \in \mathbb{R}^{N,N}$ называется *резольвентой*.

Замечание 1.2. Фундаментальная система решений ОДУ, а следовательно, и матрица Вронского $W(t)$ системы (1.28) определены неоднозначно. Однако эта матрица всегда невырождена и определение резольвенты в (1.33) является

корректным. Иногда $W(t)$ называется также *фундаментальной матрицей системы ОДУ*.

Поскольку сама резольвента, как функция от t, является решением задачи Коши

$$R'(t,t_0) = A(t)R(t,t_0), \quad t \in \Omega, \quad R(t_0,t_0) = I,$$

где $I \in \mathbb{R}^{N,N}$ есть единичная матрица, то $R(t,t_0)$ называют также *матрицей Грина системы ОДУ*. Важно отметить, что в силу теоремы 1.3, резольвента для каждой задачи Коши определена однозначно (в отличие от матрицы Вронского).

Для определителя матрицы Вронского $|W(t)| \equiv \det(W(t))$ имеется следующий замечательный результат алгебраического характера.

Теорема 1.7 (Лиувилля). *Для значений вронскиана системы ОДУ справедлива формула Лиувилля*

$$|W(t)| = |W(t_0)| exp\Big(\int\limits_{t_0}^{t} t\, r(A(t'))d\, t'\Big),$$

где $tr(A(t')) = a_{1,1}(t') + ... + a_{N,N}(t')$ *есть след матрицы* $A(t')$.

Как указывалось выше, скалярное обыкновенное дифференциальное уравнение N-го порядка

$$y^{(N)} + a_1(t)y^{(N-1)} + ... + a_N(t)y = 0 \tag{1.34}$$

может быть записано в виде системы ОДУ при введении векторной функции $\bar{y}(t) = (y_1(t), ..., y_N(t))^T$ с компонентами

$$y_1(t) = y(t),\ y_2(t) = y'(t), ..., y_N(t) = y^{(N-1)}(t).$$

Система из N решений $\bar{y}^1(t), ..., \bar{y}^N(t)$ однородного уравнения (1.34), линейно независимых на отрезке $[t_0, t_e]$, называется фундаментальной. А вронскианом системы N решений уравнения (1.34), в соответствии с введенными векторными обозначениями, называется определитель

$$W(t) = \begin{vmatrix} y_1(t) & ... & y_N(t) \\ y_1'(t) & ... & y_N'(t) \\ \cdots & \cdots & \cdots \\ y_1^{(N-1)}(t) & ... & y_N^{(N-1)}(t) \end{vmatrix}.$$

Такое определение вронскиана с использованием производных фундаментальных решений (1.34), очевидно, не противоречит (1.30).

Рассмотрим теперь систему функций $y_1(t), ..., y_N(t)$, заданных на отрезке $[t_0, t_e]$, для которых определены скалярные произведения

$$(y_i, y_j) = \int_{t_0}^{t_e} y_i(t) y_j(t) d\, t, \;\; i, j = 1, ..., N.$$

Теорема 1.8. *Для того чтобы система функций $y_1(t), ..., y_N(t)$ была линейно независимой на $[t_0, t_e]$, необходимо и достаточно, чтобы ее определитель Грама*

$$\Gamma(y_1, ..., y_N = \begin{vmatrix} (y_1, y_1) & (y_1, y_2) & ... & (y_1, y_N) \\ (y_2, y_1) & (y_2, y_2) & ... & (y_2, y_N) \\ \cdots & \cdots & \cdots & \cdots \\ (y_N, y_1) & (y_N, y_2) & ... & (y_N, y_N) \end{vmatrix}$$

равнялся нулю.

Отметим такое существенное отличие определителей Вронского и Грама: если первый есть функция аргумента t, то второй — это просто число.

1.4.2. Неоднородные системы. Перейдем теперь к рассмотрению линейных систем ОДУ с ненулевой правой частью $f(t)$.

Теорема 1.9. *Линейная неоднородная система ОДУ* (1.26) *имеет частное решение*

$$y(t) = W(t) \int_{t_0}^{t} W^{-1}(t') f(t') dt' = \int_{t_0}^{t} R(t,t') f(t') dt',$$

где $W(t) = (y^1(t), y^2(t), ..., y^N(t))^T$ *есть фундаментальная матрица однородной системы* (1.28), *а* $R(t,t')$ — *ее резольвента.*

Приведенная формула частного решения получается методом вариации постоянных. Имеет место следующее утверждение, фактически являющееся фундаментальной теоремой, но тривиально проверяемой: всякое решение линейной неоднородной системы ОДУ есть сумма ее частного решения и общего решения однородной системы. Отсюда решение задачи Коши (1.26) дается формулой

$$y(t) = R(t,t_0) y_0 + \int_{t_0}^{t} R(t,t') f(t') dt'. \qquad (1.35)$$

Отметим также, что решение задачи Коши (1.5) удовлетворяет системе интегральных уравнений, являющихся обобщением (1.3):

$$y_i(t) = y_i(t_0) + \int\limits_{t_0}^{t} f_i(t, y^1(t), ..., y^N(t))\,dt, \quad i = 1, ..., N. \tag{1.36}$$

Непосредственным дифференцированием легко убедиться в справедливости и обратного утверждения: решение интегрального уравнения (1.36) удовлетворяет системе ОДУ (1.5).

На основе представления (1.36) формулируется метод последовательных приближений, рассматривавшийся еще в работах Лиувилля (1838 г.), Коши и Пеано (1888 г.), но наиболее полно исследованный Пикаром в 1891–1896 гг. и носящий его имя.

Теорема 1.10 (Пикара). *Если правая часть системы ОДУ непрерывна по t и удовлетворяет условию Липшица* (1.27) *по y для всех $t \in [t_0, t_e]$, то метод последовательных приближений Пикара*

$$y^n(t) = y(t_0) + \int\limits_{t_0}^{t} f(t', y^{n-1}(t'))\,dt' \tag{1.37}$$

сходится равномерно в области определения ОДУ ($y^n(t) \to y(t)$ при $n \to \infty$), а предел $y(t)$ является единственным решением задачи Коши (1.5).

Формула (1.37) может рассматриваться как алгоритм приближенного решения задачи Коши, однако он является слишком трудоемким из-за необходимости многократного вычисления интеграла. В силу этого данный результат имеет скорее теоретическое значение.

Для исследования решений и построения численных методов важное место занимает представление решения ОДУ с

помощью рядов Тейлора, которое для скалярного уравнения при $y(t) \in C^{p+1}[t_0, t_e]$ имеет вид

$$y(t_0 + h) = y(t_0) + y'(t_0)h + ... + y^{(p)}(t_0)\frac{h^p}{p!} + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\xi), \quad \xi \in [t_0, t_0 + h]. \tag{1.38}$$

Здесь величины $y^{(k)}(t_0)$ могут быть заменены, согласно исходному уравнению, на производные правой части $\frac{d^{k-1}f(t_0)}{d\,t^{k-1}}$. В частности, если ограничимся первыми членами ряда

$$y(t_0 + h) = y_0 + f_0 h + O(h^2),$$

то получим знаменитый метод ломаных Эйлера.

Однако непосредственное использование ряда (1.38) для практического расчета малоцелесообразно, поскольку вычисление полных производных старших порядков представляет собой довольно трудоемкий процесс. Например, из последовательного дифференцирования с учетом самого ОДУ мы имеем

$$\begin{aligned} &y' = f(t, y), \quad y'' = f_t + f_y y' = f_t + f_y f, \\ &y''' = f_{t\,t} + 2\, f_{t\,y} f + f_{y\,y} f^2 + f_y(f_t + f_y f). \end{aligned} \tag{1.39}$$

Для нахождения производных более высоких степеней приходится уже применять или специальную операторную технику, или другие типы рядов, или же автоматизацию аналитических выкладок на компьютере, см. [63], [64].

1.4.3. Линейные системы с постоянной матрицей. Для линейных систем ОДУ вида (1.26), имеющих постоянные коэффициенты (элементы $a_{i,j}(t)$ не зависят от t), резольвента, или матрица Грина, имеет вид

$$R(t, t_0) = e^{A(t-t_0)}, \tag{1.40}$$

где *матричная экспонента* определяется как сумма сходящегося степенного ряда

$$e^{A\,t} = \sum_{k=0}^{\infty} \frac{1}{k!}(A\,t)^k. \tag{1.41}$$

Поскольку матрица есть представление линейного оператора $A\colon \mathbb{R}^N \to \mathbb{R}^N$ в конкретной координатной системе, то равенство (1.41) является также определением экспоненты линейного оператора. Как видно, оно фактически основывается на понятии предела последовательности линейных операторов, а сам оператор e^A (множитель t опускаем как несущественный) может быть определен еще следующим эквивалентным способом:

$$e^A = \lim_{k\to\infty} (I + \frac{1}{k}A)^k, \tag{1.41}$$

где I — единичный оператор или матрица. Как видно из (1.41), (1.41a), данные соотношения являются непосредственными обобщениями числового ряда

$$e^z = 1 + z + \frac{z^2}{2!} + ...,\ z \in C,$$

сходящегося абсолютно и равномерно в каждом круге $|z| \leqslant a$, и *формулы Эйлера* для экспоненты комплексного числа $z = u + iv$:

$$e^z = e^u(\cos\, v + i \sin\, v) = \lim_{n\to\infty} (1 + \frac{z}{n})^n.$$

Если матрица A с постоянными коэффициентами является *нормальной*, т. е. имеет полный базис из собственных векторов z_k с соответствующими собственными числами λ_k, то она

диагонализируется преобразованием подобия $A = Q\Lambda Q^{-1}$, где $\Lambda = \mathrm{diag}\{\lambda_k\}$ — диагональная матрица из собственных значений, а квадратная матрица $Q = \{z_{k,i}\}$ своими столбцами имеет собственные векторы z_k матрицы A. При этом однородная система ОДУ распадается на N независимых уравнений вида

$$\frac{d\,z_k}{d\,t} = \lambda_k z_k, \quad k = 1, ..., N, \tag{1.42}$$

а вектор искомого решения выражается как $y(t) = Q\,z(t)$, $z = (z_1, ..., z_N)^T$. Данный факт легко проверяется из того, что проблема собственных значений для матрицы A записывается в матричном виде как $A\,Q = Q\Lambda$, а система $\dot{y} = A\,y$ сводится к (1.42) с помощью прямой подстановки и умножения слева на Q^{-1}. Отметим, что данная матрица является вещественной, если таковыми являются собственные числа λ_k. Это имеет место, например, для симметричных матриц $A = \{a_{i,j} = a_{j,i}\}$, являющихся частным случаем нормальных.

Если же вектор начальных данных y_0 в задаче Коши (1.27) разложить в ряд по собственным векторам, т.е. представить его в виде $y_0 = Q\,z_0$, где компоненты вектора $z_0 = \{z_i^{(0)}\}$ определяются из равенства $z_0 = Q^{-1}y_0$, то для векторов $z_k(t)$ получаем задачи Коши

$$\dot{z}_k = \lambda_k z_k,\ z_k(t_0) = z_k^{(0)},$$

имеющими своими решениями $z_k(t) = \exp(\lambda_k(t - t_0))z_k^{(0)}$.

Из соотношений (1.41), (1.41а) получаем следующую формулу для диагональной матричной экспоненты:

$$\exp(\Lambda t) = \mathrm{diag}\{\exp(\lambda_k t)\}.$$

Напомним, что аналогичный факт имеет место и для любой функции $F(A)$ от диагональной матрицы A, т. е. результат есть диагональная матрица с элементами $F(\lambda_k)$.

Если же нормальная матрица не является диагональной, то результат действия матричной функции

$$u = F(A)v$$

определяется с помощью следующих преобразований:

$$Q^{-1}u = Q^{-1}F(A)Q\,w = F(\Lambda)w,$$

где $w = Q^{-1}v$ и $F(\Lambda) = \mathrm{diag}\{F(\lambda_k)\}$. Отсюда получаем соотношение

$$u = Q\,F(\Lambda)Q^{-1}v,$$

которое дает правило для вычисления матричной функции, в том числе экспоненты —

$$F(A) = Q\,F(\Lambda)Q^{-1}.$$

Отметим, что если матрица Q является ортогональной, т. е. $Q^TQ = Q\,Q^T = I$, что может иметь место для симметричных матриц A, то $Q^{-1} = Q^T$ и соответствующую подстановку можно сделать во всех предшествующих формулах.

Для комплексного собственного числа $\lambda = \alpha + i\omega$ решение можно записать в форме

$$y(t_{n+1}) = e^{\alpha\,h}(\cos\omega h + i\sin\omega h)y(t_n), \tag{1.43}$$

где $t_n = n\,h$, $n = 0, 1, ..., N$, а $h > 0$ — некоторая вещественная постоянная. Модуль комплекснозначной функции

$$r(\lambda\, h) = e^{\alpha\, h}(\cos\omega\, h + i \sin\omega\, h) \tag{1.44}$$

не зависит от ω и обладает следующими свойствами:

$$|r(\lambda\, h)| = e^{\alpha\, h} < 1 \quad \text{при} \quad \alpha < 0,$$
$$|r(\lambda\, h)| = 1 \quad \text{при} \quad \alpha = 0,$$
$$|r(\lambda\, h)| \to 0 \quad \text{при} \quad \alpha\, h \to -\infty.$$

§ 1.5. Устойчивость систем ОДУ

На качественном уровне устойчивость ОДУ означает относительно малое изменение его решения при малом возмущении исходных данных или параметров системы уравнений. Поэтому естественно сначала рассмотреть зависимость решений от параметров и начальных условий.

1.5.1. Зависимость решений от параметров и начальных данных. Рассмотрим следующую задачу Коши для системы ОДУ:

$$\dot{y} = f(t, y, \mu), \quad y(t_0, \mu) = y^0(\mu), \tag{1.45}$$

где $\mu = (\mu_1, ..., \mu_m)^T$ есть вектор параметров.

Теорема 1.11 (о дифференцируемости решений по параметру). *Пусть в некоторой области D пространства (t, y, μ) вектор-функция f имеет непрерывные производные до порядка $p \geqslant 1$ включительно по переменным t, y, μ. Тогда решение задачи Коши* (1.45) *$y(t, \mu)$ имеет p непрерывных производных по переменным t, μ.*

Зависимость решения от начальных данных проанализируем для простоты на примере скалярной задачи Коши

$$\dot{y} = f(t, y), \quad y(t_0) = y^0.$$

Здесь величины t_0, y^0 можно рассматривать как параметры. С помощью замены переменных

$$\tilde{t} = t - t_0, \quad \tilde{y} = y - y^0$$

получаем задачу Коши

$$\dot{\tilde{y}} = \tilde{f}(\tilde{t}, \tilde{y}), \quad \tilde{y}(0) = 0,$$

в которой правая часть $\tilde{f}(\tilde{t}, \tilde{y}) = f(t_0 + \tilde{t}, y^0 + \tilde{y})$ зависит от начальных данных как от параметров. Отсюда следует, что в условиях теоремы 1.9 (в данном случае — наличие у $f(t, y)$ непрерывных производных до порядка $p \geqslant 1$ включительно) решение $y(t, t_0, y^0)$ имеет p непрерывных производных по переменным t, t_0, y^0.

Выражения же для первых производных могут быть выписаны явным образом.

Теорема 1.12. *Если частная производная $f(t, y)$ по y существует и непрерывна, то справедливы следующие формулы для производных решения $y(t, t_0, y^0)$:*

$$\frac{\partial y(t, t_0, y^0)}{\partial y^0} = R(t, t_0), \quad \frac{\partial y}{\partial t_0} = -R(t, t_0) f(t_0, y^0). \qquad (1.46)$$

В более общем виде возмущение решения при вариации исходных данных определяется следующим утверждением.

Теорема 1.13 (Грёбнера). *Пусть $y(t)$ и $z(t)$ являются решениями исходной и возмущенной задач Коши*

$$\begin{aligned} &\dot{y} = f(t,y), \quad y(t_0) = y^0, \\ &\dot{z} = f(t,z) + g(t,z), \quad z(t_0) = z^0 \end{aligned} \tag{1.47}$$

соответственно, а производная $\partial f/\partial y$ *существует и непрерывна. Тогда возмущенное решение связано с исходным соотношением*

$$\begin{aligned} z(t) = y(t) + \int\limits_{t_0}^{t} \frac{\partial y}{\partial z}(t,t',z(t'))g(t,z(t'))dt' + \\ + \int\limits_{t_0}^{t} \frac{\partial y}{\partial z}(t,t_0,y^0 + s(z^0 - y^0))(z^0 - y^0)ds. \end{aligned} \tag{1.48}$$

Рассмотрим, наконец, возмущение решения задачи (1.45), обусловленное малой вариацией параметра. Данный анализ проведем на основе введения *матрицы Якоби*, или *якобиана*:

$$\frac{\partial f}{\partial y} = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \dots & \frac{\partial f_1}{\partial y_N} \\ \dots & \dots & \dots \\ \frac{\partial f_N}{\partial y_1} & \dots & \frac{\partial f_N}{\partial y_N} \end{bmatrix} = J \in \mathbb{R}^{N,N}, \tag{1.49}$$

а также вектора производных по параметру $\dfrac{\partial f(t,y(t),\mu)}{\partial \mu}$. Наряду с (1.45) введем «возмущенную» задачу Коши

$$\dot{z} = f(t,z,\nu), \quad z(t_0) = y^0, \tag{1.50}$$

в которой значение ν считается достаточно близким к μ из (1.45), а также сформулируем вспомогательную систему с якобианом

$$w'(t) = J(t)w(t), \quad J = \left\{ \frac{\partial f_i}{\partial z_j} \right\}, \tag{1.51}$$

резольвенту которой обозначим через $G(t, t_0)$, чтобы отличать ее от $R(t, t_0)$ в (1.46).

Теорема 1.14. *Пусть $y(t)$ является решением задачи* (1.45), *тогда решение возмущенной задачи* (1.50) *задается формулой*

$$z(t) = y(t) + (\nu - \mu) \int_{t_0}^{t} G(t, t') \frac{\partial f}{\partial \mu}(t', y(t'), \mu) d\, t' + O(|\nu - \mu|). \tag{1.52}$$

Отсюда видно, что чувствительность решения к изменениям параметра зависит не только от величины $\partial f / \partial \mu$, но и от устойчивости дифференциального уравнения (1.51) с матрицей Якоби.

Рассмотрим теперь класс задач Коши, возмущенных по отношению к исходной задаче (1.5):

$$\tilde{y}' = f(t, \tilde{y}) + \Delta(t), \;\; \tilde{y}(t_0) = y^0 + \Delta_0, \;\; t \in [t_0, t_e],$$

где Δ_0 и $\Delta(t)$ — возмущение, а $\tilde{y}(t)$ — возмущенное решение.

Определение 1.2. *Если при всех $t \in [t_0, t_e]$ для любого $\varepsilon > 0$ существует величина $\delta > 0$ такая, что при выполнении условий*

$$||\Delta(t)|| \leqslant \delta, \;\; ||\Delta_0|| \leqslant \delta$$

справедливо неравенство $||y(t) - \tilde{y}(t)|| \leqslant \varepsilon$, то задача Коши называется абсолютно устойчивой.

Абсолютная устойчивость — вполне естественное понятие, не накладывающее слишком больших требований к системе ОДУ. Для ее получения достаточно выполнение условия Липшица для функции правой части $f(t, y)$, которое мы считаем выполненным на протяжении всей книги. В то же время интуитивно ясно, что если задача Коши не является абсолютно устойчивой, то построение для нее приемлемого численного решения представляет большую проблему.

1.5.2. Теория устойчивости А. М. Ляпунова. К вопросам устойчивости динамических систем было приковано пристальное внимание многих европейских математиков XIX в. Первый серьезный шаг в этом направлении сделал Раус, работа которого получила премию Адамса в 1877 г. Свое завершение в определенном плане теория устойчивости получила в знаменитой работе А. М. Ляпунова 1892 г.

Определение 1.3. *Решение $\tilde{y}(t)$ задачи Коши для системы ОДУ (1.5) называется устойчивым по Ляпунову, если для любого $\varepsilon > 0$ существует $\delta > 0$ такое, что все решения $y(t)$, удовлетворяющие условию $|y(t_0) - \tilde{y}(t_0)| < \delta$, определены для всех $t_0 \leqslant t \leqslant t_e$, и при этих значениях аргументов выполняется неравенство*

$$|y(t) - \tilde{y}(t)| < \varepsilon. \tag{1.53}$$

Решение $\tilde{y}(t)$ называется асимптотически устойчивым, если оно устойчиво и, кроме того,

$$\lim_{t \to \infty} |y(t) - \tilde{y}(t)| = 0. \tag{1.54}$$

Слово “определены” выше можно понимать как “могут быть продолжены на $[t_0, t_e]$, и для любого продолжения верно

(1.53)". Если решение системы ОДУ является асимптотически устойчивым при заданных начальных данных, то будем также говорить, что *асимптотически устойчива соответствующая задача Коши для системы ОДУ*.

Очевидно, мы всегда можем свести исследование устойчивости к случаю

$$\tilde{y} = \{\tilde{y}_i,\ i = 1, ..., N\} \equiv 0,$$

взяв вместо $y(t)$ новую неизвестную вектор-функцию

$$y(t) - \tilde{y}(t) = \{y_i(t) - \tilde{y}_i(t)\},$$

которая имеет нулевое начальное значение при $t = t_0$.

Следует отметить, что задача Коши специфицируется как функциями f_i, так и начальной координатой t_0. Поэтому система ОДУ (1.5) с одной и той же правой частью в разных начальных точках может быть как устойчивой, так и неустойчивой. Простой физический пример — движение шарика по волнистой поверхности. Он может иметь положения равновесия с нулевыми скоростями или на дне впадины, или на верху возвышения. Но если в первом случае при небольшом отклонении шарик возвращается назад на дно, то во втором — скатывается с вершины дальше. Таким образом, мы имеем устойчивое или неустойчивое положение равновесия.

Важным инструментом анализа устойчивости является *функция Ляпунова*, которая определяется как дифференцируемый неотрицательный функционал $(y_1, ..., y_N)$, т. е. $V(y_1, ..., y_N) \geqslant 0$, причем $V(y_1, ..., y_N) = 0$ тогда и только тогда, когда $y_1 = ... = y_N = 0$, обладающий на решениях

задачи Коши свойством $\dfrac{d\,V(y_1,...,y_N)}{d\,t} \leqslant 0$.

Существование такой функции Ляпунова является достаточным условием устойчивости задачи Коши с начальной точкой $t_0 = 0$. Однако общего метода построения функций Ляпунова нет.

В качестве примера рассмотрим систему

$$\dot{y}_1 = y_2,\ \dot{y}_2 = -y_1,$$

для которой можно выбрать неотрицательную функцию Ляпунова $V = y_1^2 + y_2^2$. Поскольку для данной системы

$$d\,V/d\,t = 2\,y_1\dot{y}_1 + 2\,y_2\dot{y}_2 = 2\,y_1y_2 - 2\,y_1y_2 = 0,$$

то точка покоя $t_0 = y_1(t_0) = y_2(t_0) = 0$ является устойчивой. Однако приведенная система не обладает асимптотической устойчивостью, поскольку она имеет траектории в форме окружностей, не стремящиеся к точке $(0,0)$ при $t \to \infty$.

Для линейной системы с постоянными коэффициентами устойчивость задачи Коши

$$y' = A\,y, \quad y(t_0) = y^0 \tag{1.55}$$

однозначно определяется следующим утверждением, причем результат в данном случае справедлив для любого t_0.

Теорема 1.15 (об устойчивости системы с постоянными коэффициентами). *Решение задачи* (1.55) *устойчиво по Ляпунову тогда и только тогда, когда все собственные значения матрицы A удовлетворяют условию $Re(\lambda) \leqslant 0$, а для кратных корней, порождающих жордановы блоки, выполняется строгое неравенство $Re(\lambda) < 0$.*

Другой критерий устойчивости линейной задачи имеет следующую знаменитую формулировку, породившую огромное количество исследований по устойчивости различного типа динамических систем и уравнений.

Теорема 1.16 (Ляпунова). *Для асимптотической устойчивости модельной системы ОДУ* (1.55) *необходимо, чтобы для любой и достаточно, чтобы для какой-нибудь симметричной положительно определенной матрицы Q решение X матричного уравнения Ляпунова*

$$X\,A + A^T X = Q \tag{1.56}$$

было отрицательно определенной матрицей.

Несмотря на несомненную важность двух последних теорем, они являются мало конструктивными в том смысле, что приведенные условия являются трудно проверяемыми.

Исследование на устойчивость решения системы ОДУ (1.5) можно свести к анализу устойчивости нулевого (тривиального) решения $y_i \equiv 0,\ i = 1, ..., N$ некоторой аналогичной системы

$$y_i' = F_i(t, y_1, ..., y_N),\ i = 1, ..., N, \tag{1.56}$$

где $F_i(t, 0, ..., 0) \equiv 0$ для $i = 1, ..., N$.

Говорят, что точка $y_1 = ... = y_N = 0$ есть *точка покоя* системы (1.56a). Применительно к ней определения устойчивости или неустойчивости могут быть сформулированы следующим образом. Точка покоя $y_i = 0, i = 1, ..., N$ устойчива

по Ляпунову, если для любого $\varepsilon > 0$ существует $\delta > 0$ такое, что для любого решения $y_i(t), i = 1, ..., N$, с начальными данными $y_{i,0} = y_i(t_0)$, удовлетворяющими условию

$$|y_{i,0}| < \delta, \;\; i = 1, 2, ..., N,$$

для всех $t \geqslant t_0$ выполняются неравенства

$$|y_i(t)| < \varepsilon, \;\; i = 1, 2, ..., N.$$

Если при этом дополнительно выполняется условие

$$\lim_{t \to +\infty} |y_i(t)| = 0, \;\; i = 1, 2, ..., N,$$

то точка покоя называется асимптотически устойчивой. Но она является неустойчивой, если при сколь угодно малом $\delta > 0$ хотя бы для одного решения $y_i(t)$ условие $|y_i(t)| < \varepsilon, \; t \geqslant t_0$, не выполняется при любом $\varepsilon > 0$.

Очевидно, что понятие “точка покоя” означает, другими словами, положение равновесия тривиального решения.

Примеры анализа точек покоя для однородной линейной системы ОДУ с постоянной матрицей 2-го порядка фактически приведены в § 1.2, где имеют место следующие случаи:

- устойчивый узел (рис. 1.2) и устойчивый фокус (рис. 1.5) — асимптотически устойчивые положения равновесия (или точки покоя);
- центр (рис. 1.4) — устойчивое по Ляпунову, но не асимптотически устойчивое положение равновесия;
- седло (рис. 1.3), неустойчивый узел, неустойчивый фокус — неустойчивые положения равновесия.

Отметим, что утверждения теорем 1.13 и 1.14 не имеют места для матриц с переменными элементами $A(t)$, даже если знаки собственных чисел $\lambda(t)$ удовлетворяют приведенным условиям для всех $t \geqslant t_0$.

Для нелинейных систем относительно простого вида

$$y' = A\,y + g(t, y) \tag{1.57}$$

с постоянными матрицами A при помощи функций Ляпунова доказывается следующий результат.

Теорема 1.17. *Пусть матрица A нелинейной системы* (1.57) *имеет не зависящие от t элементы, а ее собственные числа отвечают условию $Re(\lambda) \leqslant -\rho < 0$. Предположим далее, что для любого $\varepsilon > 0$ существует такое $\delta > 0$, что*

$$||g(t, y)|| \leqslant \varepsilon ||y|| \text{ при } ||y|| < \delta, t \geqslant t_0.$$

Тогда начало координат $t_0 = 0$ является устойчивым положением равновесия для задачи (1.57).

Фактически данная теорема обосновывает *метод линеаризации*, в котором в окрестности начальной точки $t_0 = 0$ исходная система ОДУ вида (1.5) аппроксимируется линеаризованной системой вида (1.55), где $A = \left\{a_{i,j} = \dfrac{\partial f_i}{\partial y_j}(0)\right\}$ представляется матрицей Якоби в начале координат, а для последней уже применяется теорема 1.13.

Здесь и далее через $\|\cdot\|$ обозначается *евклидова*, или *сферическая, норма вектора*, определяемая через скалярное произведение

$$\|y\| \equiv \|y\|_2 = (y,y)^{1/2} = \Big(\sum_{i=1}^{N} |y_i|^2\Big)^{1/2}.$$

Важно отметить, что матрица $A \in \mathbb{R}^{N,N}$ — это представление линейного оператора $A\colon \mathbb{R}^N \to \mathbb{R}^N$ в конкретной координатной системе, от выбора которой само операторное преобразование не зависит. Множество L всех линейных операторов $A\colon \mathbb{R}^N \to \mathbb{R}^N$ с метрикой (расстоянием между двумя операторами) $\rho(A,B) = ||A-B||$ само является полным нормированным линейным пространством над полем вещественных чисел.

Поскольку элементы L можно складывать и умножать на числа, а последовательности Коши в L имеют пределы, теория функциональных операторных (матричных) рядов буквально повторяет теорию числовых рядов и переносится на функции со значениями из L.

Однако существенным отличием операторных рядов является возможность их обрыва.

Определение 1.4. *Оператор (матрица) называется нильпотентным, если некоторая его (ее) степень равна нулю.*

Простейшим примером нильпотентной матрицы является

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Вообще, любая строго треугольная матрица $A = \{a_{i,j}\}$ с ненулевыми элементами только при $i > j$ или при $i < j$ является нильпотентной.

Определение 1.5. *Оператор* $\Lambda\colon \mathbb{R}^N \to \mathbb{R}^N$ *называется диагональным, если его матрица в каком-нибудь базисе диагональна. Такой базис называется собственным.*

Отметим имеющееся отличие с терминологией в теории матриц: $A = \{a_{i,j}\} \in \mathbb{R}^{N,N}$ называется диагональной, если $a_{i,j} = 0$ при $i \neq j$. А если матрица каким-то преобразованием подобия сводится к диагональной матрице $\tilde{A} = Q\,A\,Q^{-1}$, то она называется *диагонализируемой*.

Определение 1.6. *Определителем оператора A, обозначаемым detA, называется определитель матрицы оператора в каком-нибудь базисе.*

Определитель матрицы $A = \{A_i\}$ — это ориентированный объем параллелепипеда с ребрами $A_1, ..., A_N \in \mathbb{R}^N$, задаваемыми столбцами матрицы. Определитель есть подмножество $\mathbb{R}^N$, состоящее из всех точек вида

$$a_1 A_1 + ... + a_N A_N,\ 0 \leqslant a_i \leqslant 1.$$

Например, при $N = 2$ (рис. 1.6) определитель

$$\begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}$$

есть площадь параллелограмма, натянутого на векторы $\vec{A}_1, \vec{A}_2$ с координатами (x_1, y_1) и (x_2, y_2), взятая со знаком плюс, если упорядоченная пара векторов $(\vec{A}_1, \vec{A}_2)$ задает ту же ориентацию $\mathbb{R}^2$, что и базисная пара векторов $(\vec{e}_1, \vec{e}_2)$ (как на рис. 1.6), и со знаком минус в противном случае.

Столбец A_i матрицы оператора A в базисе с векторами-ортами $(\vec{e}_1, \vec{e}_2)$ ($\vec{e}_i$ имеет единственную ненулевую i-ю компо-

ненту, равную единице) составлен из координат образа вектора $A\vec{e}_i$. Поэтому определитель оператора A — это ориентированный объем образа единичного куба (параллелепипеда с ребрами $\vec{e}_1, ..., \vec{e}_N$ при отображении A).

То же самое относится и к любой фигуре: ее ориентированный объем после применения A меняется в $\det A$ раз, и данный факт не зависит от системы координат. Геометрически это вовсе не очевидно, поскольку форма фигуры при линейном преобразовании сильно меняется.

Рис. 1.6. Иллюстрация к определению ориентированного объема параллелограмма

Определение 1.7. *Следом матрицы* $A = \{a_{i,j}\}$ *называется сумма ее диагональных элементов:*

$$trA = a_{1,1} + ... + a_{N,N}.$$

Следом оператора A *называется след его матрицы в каком-нибудь базисе.*

След матрицы, как и ее собственные числа λ_k, не зависит от выбора базиса, поскольку применение формулы Виета к характеристическому многочлену $\det(A - \lambda I)$ показывает, что след равен сумме собственных чисел. Напомним также, что определитель матрицы равен произведению ее собственных чисел.

Между понятиями определителя и следа имеется следующая связь: при $\varepsilon \to 0$ выполняется соотношение

$$\det(I + \varepsilon A) = 1 + \varepsilon \mathrm{tr}(A) + o(\varepsilon^2).$$

§ 1.6. Одностороннее условие Липшица, контрактивность и устойчивость жестких систем

Результаты из п. 1.5.1 можно дополнить следующей теоремой, имеющей фундаментальный характер в том смысле, что она дает оценки чувствительности решения к возмущениям исходных данных при достаточно общих предположениях о задаче Коши.

Теорема 1.18. *Пусть $\tilde{y}(t)$ — приближенное решение системы ОДУ $y' = f(t, y)$ при $y(t_0) = y^0$, которое удовлетворяет условиям*

$$\begin{aligned} &||\tilde{y}(t_0) - y(t_0)|| \leqslant \rho, \\ &||\tilde{y}'(t) - f(t, y(t))|| \leqslant \varepsilon, \\ &||f(t, \tilde{y}) - f(t, y)|| \leqslant L||\tilde{y} - y||, \end{aligned} \tag{1.58}$$

где L есть постоянная Липшица. Тогда при $t \geqslant t_0$ справедлива оценка погрешности

$$||y(t) - \tilde{y}(t)|| \leqslant \rho e^{L(t-t_0)} + \frac{\varepsilon}{L}(e^{L(t-t_0)} - 1). \tag{1.59}$$

Очевидный “недостаток” последнего неравенства заключается в том, что данная оценка возмущения решения имеет экспоненциальный рост, даже если само решение затухает. И чем больше константа Липшица, тем определяемая соотношением (1.59) неустойчивость ОДУ оказывается сильнее.

Такие ситуации возникают в так называемых *жестких задачах*, которые характеризуются наличием в решении как гладких, медленно меняющихся со временем t компонент, так и присутствием быстро меняющихся, или жестких, компонент, описывающих *переходные процессы*.

Хорошим примером является скалярное уравнение

$$y'(t) = \lambda\, y(t) + F'(t) - \lambda\, F(t), \quad t \geqslant 0, \quad y(0) = y^0, \qquad (1.60)$$

где $\lambda \ll 0$, а F — медленно меняющаяся функция только от t. Решение (1.60) имеет вид

$$y(t) = F(t) + e^{\lambda\, t}(y^0 - F(0)),$$

в котором зависящий от начальных значений второй член быстро убывает и при $\lambda t \ll 0$ в $y(t)$ фактически не присутствует, т. е. при достаточно больших значениях аргумента t преобладающей частью будет "гладкая" функция $F(t)$. Однако при малых t как раз наличие второй (жесткой) компоненты может привести к резкому экспоненциальному изменению $y(t)$.

Жесткие системы ОДУ требуют использования специальных методов численного интегрирования. Понятие жесткости является относительным в том смысле (как и устойчивость), что нельзя по количественному значению какого-то параметра строго провести грань, разделяющую нежесткие и жесткие ОДУ.

Определение 1.8. *Линейная система ОДУ вида* $\dot{y}(t) = A\, y(t) + g(t)$ *с постоянными коэффициентами называется жесткой, если*

а) *все собственные числа* λ_i *матрицы* A *имеют неположительные вещественные части:* $Re\lambda_i \leqslant 0, \quad i = 1, ..., N$;

б) *ненулевые вещественные части* λ_i *имеют большой разброс по величине:*

$$\max_{1 \leqslant i \leqslant N} \{Re(-\lambda_i)\} / \min_{\substack{1 \leqslant i \leqslant N \\ Re\lambda_i \neq 0}} \{R(-\lambda_i)\} \gg 1;$$

в) *интервал интегрирования имеет относительно большую длину:*

$$\max_{1 \leqslant i \leqslant N} \{Re(-\lambda_i)\} \cdot (t_{N_e} - t_0) \gg 1.$$

Как видно отсюда, данные свойства определяются только спектром матрицы A. Иногда жестким системам "разрешают" иметь собственные числа с "небольшими" положительными вещественными частями $\mathrm{Re}\lambda_i$.

Определение 1.9. *Нелинейную систему ОДУ вида* (1.5) *называют жесткой при следующих характерных признаках:*

а) *большой разброс по величине модулей* $|\dot{y}_i|, i = 1, ..., N$, *первых производных решения;*

б) *наличие больших по модулю отрицательных вещественных частей собственных чисел матрицы Якоби от правой части;*

в) *большой разброс по величине элементов матрицы Якоби исходных уравнений;*

г) *относительно длинный интервал интегрирования* (t_0, t_e).

Можно привести также определение жесткости “по Ламберту”, который ввел это понятие в 1973 г.: задача Коши для системы (1.5) называется жесткой на отрезке $I = [t_0, t_e]$, если для всех $t \in I$ имеем

$$Re(-\lambda_i) < 0,\ i = 1, ..., N,$$
$$S(t) = \max_i Re(-\lambda_i) / \min_i Re(-\lambda_i) \gg 1,$$

где λ_i – собственные числа матрицы $\partial f / \partial y$, в которой подставлено решение $y(t)$ в точке t.

Отношение $S(t)$ можно назвать локальным коэффициентом жесткости. Отметим, что строго говоря, понятие жесткости следует применять не к самим уравнениям, а к задаче Коши для системы ОДУ , поскольку при разных начальных данных свойства жесткости могут или быть, или отсутствовать.

Примером жесткой задачи может быть система ОДУ, которая описывает поведение химических взаимодействий веществ, содержащих ряд быстрых и медленных реакций.

Другими словами, для жестких систем характерно наличие “быстрых” и “медленных” компонент решения. Простейшей иллюстрацией может быть автономная линейная система ОДУ с постоянной матрицей второго порядка

$$\dot{y} = Ay,\ \ A = \begin{bmatrix} -1000 & 999 \\ 1 & -2 \end{bmatrix},$$

собственные числа которой равны $\lambda_1 = -1001$ и $\lambda_2 = -1$.

Решение этой системы имеет вид

$$y_1(t) = 0.999[y_1(0) - y_2(0)]e^{\lambda_1 t} + [0.001y_1(0) + 0.999y_2(0)]e^{\lambda_2 t},$$
$$y_2(t) = -0.001[y_1(0) - y_2(0)]e^{\lambda_1 t} + [0.001y_1(0) + 0.999y_2(0)]e^{\lambda_2 t},$$

а его простой анализ показывает наличие некоторого *пограничного слоя* $[0, \tau]$, величина которого определяется экспонентой с показателем λ_1, причем компонента y_1 изменяется намного быстрее, чем y_2. После прохождения пограничного слоя, т. е. при $t > \tau$, величины производных решения невелики и определяются экспонентой с показателем λ_2.

Наличие погранслоев, характерных для жестких задач, зачастую связано с наличием малых коэффициентов при производных, а такие уравнения называются *сингулярно возмущенными*. Рассмотрим дифференциальное уравнение первого порядка

$$\mu \dot{y} = f(t, y),\ \mu > 0,\ \ f(t, y) \in C^{0,1}_{t,y},\ y \in \mathbb{R}^1.$$

Будем считать, что μ есть малый параметр, причем при $\mu \to 0$ получаемое вырожденное уравнение

$$f(t, \bar{y}) = 0$$

имеет единственное достаточно гладкое решение $\bar{y}(t) = G(t)$, а в окрестности производная $\partial f / \partial y$ отрицательна (последнее условие является достаточным для устойчивости решения $\bar{y}(t) = g(t)$). Характер поведения решения данного уравнения приведен на рис. 1.7, где штриховыми линиями показано поле направлений, касательных к интегральным кривым. Для достаточно малого параметра сингулярности μ эти касательные даже при небольшом отклонении от функции $G(t)$ почти перпендикулярны оси t. И чем меньше величина μ, тем быстрее

осуществляется сближение интегральных кривых и решения вырожденного уравнения.

У любой интегральной кривой выделяются два участка с существенно различным поведением решения, причем длительность первого намного меньше второго. Первый участок отражает быстрое стремление $y(t)$ к $\bar{y}(t)$, а на втором производные решения значительно меньше и все интегральные кривые практически совпадают с графиком $g(t)$. Отметим, что даже при небольшом отклонении начальной точки (t_0, y_0) от точки графика $(t_0, G(t_0))$ производная решения dy/dt в пограничном слое намного больше величины dg/dt по модулю. Приведенный пример свидетельствует, что понятие жесткости может относиться не только к системе ОДУ, но даже и к одному уравнению.

Рис. 1.7. Иллюстрация пограничного слоя в решении жесткой задачи

Мы уже отмечали,что понятие жесткости относится не только к самому уравнению, но и к решаемой задаче, которая характеризуется также начальными данными и длиной интервала интегрирования $[t_0, t_e]$. Например, система двух линейных ОДУ с постоянной матрицей A, имеющей достаточно близкие собственные значения $\lambda_1 = -1, \lambda_2 = -2$, характеризуется своими частными решениями в виде соответствующих экспонент. Такая система не является жесткой на интервале $[0, 1]$, но является таковой на отрезке $[0, 100]$, поскольку в последнем случае проявляется резкое различие между частными решениями и тривиальным (нулевым).

Другой пример: задача с характерным погранслоем $[0, \tau]$ при интегрировании на интервале $[0, \tau/2]$ не должна рассматриваться как жесткая, если здесь не будет сильного изменения характера решения.

Важно отметить, что для систем ОДУ с переменными коэффициентами, даже линейных, одних локальных спектральных матричных свойств недостаточно для характеристики жесткости или нежесткости решения.

Хорошим подтверждением этому служит оригинальное *уравнение Деккера—Вервера* [20], иллюстрирующее возможность получения произвольной жесткости задачи, при сохранении типичного поведения решения:

$$\dot{y}(t) = A(t)y(t), \;\; A(t) = E(t)DE^{-1}(t), D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}, \; E(t) = \begin{bmatrix} \cos(\omega t - \\ \sin(\omega t - \end{bmatrix}$$

где d_1, d_2, ω и θ — константы. Сделав замену переменной

$$z = E^{-1}y,$$

получаем новое уравнение

$$\dot{z}(t) = C\, z(t)$$

с не зависящей от t матрицей

$$C = D - E^{-1}(t)\dot{E}(t) = \begin{bmatrix} d_1 + \omega \mathrm{ctan}\, \theta & -\omega \sin^{-1}(\theta) \\ \omega \sin^{-1}(\theta) & d_2 - \omega \mathrm{ctan}\, (\theta) \end{bmatrix}.$$

Собственные значения $\lambda(C)$, определяющие экспоненциальное поведение решения $y(t)$, удовлетворяют характеристическому уравнению

$$\lambda^2 - (d_1 + d_2)\lambda + d_1 d_2 + (d_2 - d_1)\omega \operatorname{ctan}\theta + \omega^2 = 0.$$

В случае $d_1 = -1, d_2 = -10, \omega = 6$, $\operatorname{ctan}\theta = 4/3$ приходим к известному примеру Винограда (1952 г.), для которого собственные числа $\lambda(C) = 2, -13$, а искомое решение имеет следующий вид ($\bar{t} = 6\,t$):

$$y(t) = C_1 e^{2t} \begin{bmatrix} \cos\bar{t} + 2\sin\bar{t} \\ 2\cos\bar{t} - \sin\bar{t} \end{bmatrix} + C_2 e^{-13t} \begin{bmatrix} \sin\bar{t} - 2\cos\bar{t} \\ 2\sin\bar{t} + cos\bar{t} \end{bmatrix},$$

где C_1 и C_2 — произвольные постоянные. При этом собственные значения исходной переменной матрицы

$$A(t) = \begin{bmatrix} -1 - 9\cos^2\bar{t} + 6\sin 2\bar{t} & 12\cos^2\bar{t} + 4.5\sin 2\bar{t} \\ -12\sin^2\bar{t} + 4.5\sin 2\bar{t} & -1 - 9\sin^2\bar{t} - 6\sin 2\bar{t} \end{bmatrix}$$

не зависят от времени t и равны $\lambda_1 = -1, \lambda_2 = -10$. Видно, что экспоненты $e^{\lambda_1 t}$ и $e^{\lambda_2 t}$ отсутствуют в решении и даже $||y(t)|| \to \infty$ при $C_1 \neq 0$. Таким образом, спектр матрицы $A(t)$ нельзя использовать для получения надежной информации о локальном поведении решения.

Жесткие дифференциальные уравнения из-за наличия больших собственных значений всегда обладают большой константой Липшица L. Это, в свою очередь, приводит к слишком грубым оценкам возмущения решения. Такую ситуацию удается исправить с помощью нового понятия — *одностороннего условия Липшица*.

Определение 1.10. *Пусть $f(t, y(t))$ — вещественная векторная функция с областью определения $D_t \subset \mathbb{R}^N$ для каждого t, а $l(t) : \Omega = [t_0, t_e] \to \mathbb{R}$ — кусочно-непрерывная*

функция такая, что при всех $t \in \Omega$ *и* $\hat{y}, \check{y} \in D_t$ *справедливо неравенство*

$$\begin{aligned}&((f(t,\hat{y}(t)) - f(t,\check{y}(t)), (\hat{y}(t) - \check{y}(t)))) \leqslant \\ &\leqslant l(t)||\hat{y}(t) - \check{y}(t)||_2^2.\end{aligned} \tag{1.61}$$

Тогда $f(t, y(t))$ *называется функцией, удовлетворяющей при* $t \in \Omega$ *одностороннему условию Липшица с постоянной* $l(t)$*, определенной для каждого значения параметра* t*.*

Если функция $f(t,\cdot)$ удовлетворяет классическому условию Липшица с константой $L(t)$, то эта $L(t)$ является также односторонней постоянной Липшица $t(t)$ для $f(t,\cdot)$. Это следует из неравенства Шварца

$$\begin{aligned}&((f(t,\hat{y}) - f(t,\check{y})), (\hat{y} - \check{y}) \leqslant ||f(t,\hat{y}) - f(t,\check{y}))||_2||\hat{y} - \check{y}||_2 \leqslant \\ &\leqslant L(t)||\hat{y} - \check{y}||_q^2.\end{aligned}$$

Однако обратное утверждение неверно. Например, если $f\colon \mathbb{R} \to \mathbb{R}$ есть монотонная невозрастающая функция при растущих $y \in \mathbb{R}$, то при $\hat{y}, \check{y} \in \mathbb{R}$ имеем

$$(f(\hat{y}) - f(\check{y})), (\hat{y} - \check{y}) \leqslant 0,$$

так что неравенство (1.61) выполняется при $l(t) = 0$. Предположим теперь, что для некоторого $y \in \mathbb{R}$ функция f имеет бесконечную производную. Тогда последнее неравенство имеет смысл, но в окрестности этого y функция f не удовлетворяет классическому условию Липшица. Отметим, более того, что для жестких задач может быть $l(t) \leqslant 0$ одновременно с $L \gg 0$.

Пусть теперь $y(t)$ — точное решение задачи Коши

$$y' = f(t, y(t)), \quad y(t_0) = y^0, \tag{1.62}$$

а $\tilde{y}(t)$ — возмущенное решение при искаженных начальных данных $\tilde{y}(t_0) = \tilde{y}^0$. Тогда для квадрата нормы погрешности

$$\delta(t) = ||\tilde{y}(t) - y(t)||^2$$

справедливо уравнение

$$\delta'(t) = 2((\tilde{y}'(t) - y'(t)), (\tilde{y}(t) - y(t))),$$

из которого при выполнении (1.61) следует дифференциальное неравенство

$$\delta'(t) \leqslant 2l(t)\delta(t), \quad t \in [t_0, t_e], \quad \delta(0) = ||\tilde{y}(0) - y(0)||^2.$$

После умножения обеих его частей на величину

$$\eta(t) = \exp(-2\int_{t_0}^{t} l(\tau)d\tau)$$

мы получаем соотношение

$$\frac{d}{d\,t}(\delta(t)\eta(t)) \leqslant 0,$$

означающее монотонное уменьшение произведения $\delta(t)\eta(t)$ для $t \in [t_0, t_e]$. Отсюда следует важное утверждение.

Теорема 1.19. *Пусть правая часть задачи Коши* (1.62) *удовлетворяет одностороннему условию Липшица* (1.61) *при* $t \in [t_0, t_e]$, *а* $\tilde{y}(t)$ *есть возмущенное решение той же системы ОДУ с начальным условием* $\tilde{y}(0) = \tilde{y}^0$. *Тогда для любых* t_1, t_2, *удовлетворяющих неравенствам* $t_0 \leqslant t_1 \leqslant t_2 \leqslant t_e$, *справедлива оценка*

$$||\tilde{y}(t_2) - y(t_2)|| \leqslant \exp(\int_{t_1}^{t_2} l(t) d\, t)\ ||\tilde{y}(t_1) - y(t_1)||. \qquad (1.63)$$

Последнее соотношение позволяет установить свойство *контрактивности* решений ОДУ: если $l(t) \leqslant 0$ для $t_0 \leqslant t_1 \leqslant t \leqslant t_2 \leqslant t_e$, то выполняется неравенство

$$||\tilde{y}(t_2) - y(t_2)|| \leqslant ||\tilde{y}(t_1) - y(t_1)||, \qquad (1.64)$$

означающее "притягивание" двух различных решений друг к другу. В литературе системы ОДУ, обладающие этим свойством, и соответствующие функции $f(t, \cdot)$ называются также *диссипативными*.

Рассмотрим теперь подробнее, что дает использование одностороннего условия Липшица для линейной однородной системы ОДУ с постоянной матрицей A:

$$\dot{y}(t) = f(t, y) \equiv A\, y, \quad A \in \mathbb{R}^{N,N}. \qquad (1.65)$$

Так как решение этого уравнения имеет вид

$$y(t) = \exp(A\, t) y^0,$$

то отсюда вследствие (1.61) имеем $||\exp(A\, t) y^0||_2 \leqslant \exp(lt)||y^0||_2$. Таким образом, с помощью односторонней постоянной Липшица получаем оценку матричной экспоненты

$$||\exp(A\, t)||_2 \leqslant \exp(lt). \qquad (1.66)$$

С точки зрения получения наилучшей оценки вида (1.63) для модельной задачи (1.65), т. е. при $f(t, y) = A\, y$, возникает

вопрос о нахождении наименьшего значения l, удовлетворяющего неравенству (1.66). Эта величина называется *логарифмической нормой матрицы* и определяется выражением

$$\mu(A) = \lim_{\delta \to 0^+} \frac{||I + \delta A||_2 - 1}{\delta}, \tag{1.67}$$

где предел понимается как односторонний.

В этом определении можно, в принципе, использовать различные матричные нормы. Если же используемую в (1.67) сферическую норму расписать как подчиненную векторной евклидовой норме, то получим

$$\mu(A) = \max_{y \neq 0} \frac{(A\,y, y)}{(y, y)}. \tag{1.68}$$

При этом наименьшая возможная односторонняя константа Липшица для матрицы есть

$$\mu(A) = \lambda_{max}(A + A^T)/2. \tag{1.69}$$

Можно показать, что в (1.67) предел существует для любой из гельдеровских норм, и мы можем записать

$$\mu_q(A) = \lim_{\delta \to 0^+} \frac{\|I + \delta A\|_q - 1}{\delta} = \lim_{\delta \to 0^+} \frac{\ln \|e^{\delta A}\|_q}{\delta}. \tag{1.70}$$

Величину $\mu_q(A)$ можно интерпретировать как одностороннюю производную отображения $\|\cdot\|_q \colon \mathbb{R}^{N,N} \to \mathbb{R}_+$ в точке $I \in \mathbb{R}^{N,N}$ по направлению, задаваемому матрицей $A \in \mathbb{R}^{N,N}$.

Логарифмические нормы, подчиненные гельдеровским нормам с $q = 1$ и $q = \infty$, в соответствии со своим определением (1.70), характеризуются как

$$\mu_1(A) = \max_j\{a_{j,j} + \sum_{i\neq j}|a_{i,j}|\},$$
$$\mu_\infty(A) = \max_i\{a_{i,i} + \sum_{j\neq i}|a_{i,j}|\}. \tag{1.71}$$

Хотя величина $\mu_q(A)$ может быть отрицательна, логарифмическая норма обладает рядом свойств, которые напоминают свойства нормы:

$$-\|A\| \leqslant -\mu(-A) \leqslant \mathrm{Re}(\lambda_i(A)) \leqslant \mu(A) \leqslant \|A\|, \quad i = 1, ..., N,$$
$$\mu(a\,A + b\,I) = a\mu(A) + b \ \text{ для } \ a \geqslant 0 \ \text{ и } \ b \in \mathbb{R},$$
$$-\mu(-A) + \mu(B) \leqslant \mu(A + B) \leqslant \mu(A) + \mu(B).$$

С использованием логарифмической нормы матрицы Якоби $\{f'(t, y)\}$ нелинейной системы ОДУ теорема 1.17 может быть усилена следующим образом.

Теорема 1.20 (Далквиста). *Пусть* $l\colon [t_0, t_e] \to \mathbb{R}$ *есть кусочно-непрерывная функция, удовлетворяющая соотношению*

$$\mu(f'(t, y)) \leqslant l(t), \quad t \in [t_0, t_e],$$

для всех y *из области определения решений задачи Коши* (1.62). *Тогда для любых двух ее решений* $\tilde{y}(t), y(t)$ *при произвольных* q *и* t_1, t_2, *удовлетворяющих неравенствам* $t_0 \leqslant t_1 \leqslant t_2 \leqslant t_e$, *выполняется соотношение*

$$||\tilde{y}(t_2) - y(t_2)||_q \leqslant exp\Big(\int_{t_1}^{t_2} l(\tau)\,d\,\tau\Big)||\tilde{y}(t_1) - y(t_1)||_q.$$

Данный результат позволяет оценивать устойчивость и возмущения решений ОДУ в различных нормах, которые

оказываются наиболее естественными для конкретной задачи.

§ 1.7. Периодические решения, предельные циклы, странные аттракторы

При изучении динамических систем, описываемых системами ОДУ, зачастую важно знать характер поведения решений для больших значений времени. Такими примерами являются периодические или почти периодические астрономические задачи, процессы химической кинетики с характерными циклами, колебания в электротехнических цепях и т. д. Построение эффективных численных методов для решения многих сложных случаев невозможно без учета таких тонких особенностей, поскольку не существует одного универсального метода, подходящего на все случаи жизни. А наличие примеров уравнений с типовыми фазовыми портретами, исследованными теоретически, является очень важным для тестирования алгоритмов и сравнительного анализа их практической эффективности.

Одно из математических явлений, обнаруженное еще А. Пуанкаре в 1882 г., — это предельные циклы в решениях систем ОДУ. Известная иллюстрация к этому — *уравнение Ван-дер-Поля* (1926 г.)

$$y'' + \alpha y' + y = 0, \tag{1.72}$$

исследовавшееся автором при моделировании нелинейных осцилляций. Решения уравнения (1.72) являются затухающими при $\alpha > 0$ и неустойчивыми при $\alpha < 0$. В качестве простей-

шего выражения можно взять $\alpha = \mu(y^2 - 1)$ при $\mu > 0$. Переписывая для наглядности уравнение как систему

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= \mu(1 - y_1^2) - y_1, \end{aligned}$$

нетрудно обнаружить, что здесь с ростом времени t малые колебания усиливаются, а большие — затухают. Поэтому можно ожидать, что существует устойчивое периодическое решение, называемое *предельным циклом*, к которому сходятся все остальные решения. Характерный пример изображен на рис. 1.8.

Отметим, что предельный цикл уравнения Ван-дер-Поля асимптотически устойчив по Ляпунову, а вещественные части собственных значений матрицы Якоби принимают вдоль него как отрицательные, так и положительные значения.

Для автономной системы ОДУ 2-го порядка исчерпывающий результат устанавливается следующим утверждением.

Теорема 1.21 (Пуанкаре). *Любое ограниченное решение системы*

$$y_1' = f_1(y_1, y_2), \quad y_2' = f_2(y_1, y_2), \tag{1.73}$$

когда f_1 и f_2 суть многочлены, либо стремится к критической точке $f_1 = f_2 = 0$ для бесконечного множества точек $t \to \infty$, либо является периодическим, либо стремится к предельному циклу.

Рис. 1.8. Характерный предельный цикл для уравнения Ван-дер-Поля

Данная теорема для систем уравнений большего порядка уже не верна. Контрпримером может служить знаменитая задача (рассмотренная впервые Ж. Лагранжем в 1788 г.) о механическом движении сферического маятника без трения. В этом случае уравнения движения в сферических координатах принимают вид

$$\begin{aligned} \varphi'' &= -2\frac{\cos\psi}{\sin\psi}\varphi'\psi', \\ \psi'' &= \sin\psi\cos\psi(\varphi')^2 - \sin\psi, \end{aligned} \tag{1.74}$$

а его решение может не стремиться ни к периодическому, ни к стационарному решению.

Предельный цикл является частным случаем более общего явления — *аттрактора*, которым называется ограниченное притягивающее предельное множество траекторий решений системы ОДУ. Другими словами, все решения, оказывающиеся в окрестности аттрактора, будут стремиться к нему. В частности, предельный цикл, изображаемый в фазовом пространстве изолированной замкнутой кривой, называется *регулярным аттрактором*.

В нелинейных системах ОДУ размерности $N \geqslant 3$ при некоторых значениях параметров может возникнуть *странный аттрактор*, который характеризуется режимом установившихся непериодических колебаний. Его открытие сделано Э. Лоренцом в 1979 г. при построении искусственной математической модели для тестирования численных методов прогноза погоды:

$$\begin{aligned} y_1' &= -\sigma y_1 + \sigma y_2, \\ y_2' &= -y_1 y_3 + r\, y_1 - y_2, \\ y_3' &= y_1 y_2 - b\, y_3, \end{aligned} \tag{1.75}$$

где σ, r и b — положительные константы.

Системы ОДУ со странными аттракторами являются неустойчивыми по Ляпунову, и для них при $f(0) = 0$ характерной особенностью является наличие положительного собственного значения у матрицы Якоби $\left\{\dfrac{\partial f}{\partial y}(0)\right\}$. Важно отметить, что странный аттрактор — это не частный патологический случай, а один из характерных видов решения ОДУ.

Уравнения (1.75) дают пример наличия параметризованной задачи. В параграфе 1.5 мы рассматривали вопросы зависимости решения от входных параметров, которая в "хороших" случаях имеет достаточно гладкий вид. Однако так бывает далеко не всегда, и переход из одного режима движения в качественно другой, или *фазовый переход*, при непрерывном изменении параметров системы ОДУ называется *бифуркацией*. Соответствующие величины параметров, при которых происходят качественные фазовые переходы, называются *бифуркационными значениями параметров*, или *точками бифуркации*. Изучение изменения качественной картины решений при изменении параметров — это область глубокой теории бифуркаций, основы которых заложены в работах А.Пуанкаре, А. А.Андронова и В. И.Арнольда, см. [4].

Типичная формальная причина появления бифуркации — изменение знака вещественных частей корней характеристического уравнения матрицы Якоби для исходной системы ОДУ. Известная иллюстрация в данном случае — это так на-

зываемая *бифуркация Андронова–Хопфа*, обнаруживаемая в простейшей модели химической реакции — "брюсселятора", (см. подробнее [64]), которая описывается системой 2-го порядка

$$\begin{aligned} y_1' &= A + y_1^2 y_2 - (B+1)y_1, \\ y_2' &= B\, y_1 - y_1^2 y_2, \end{aligned} \tag{1.76}$$

где A и B — положительные постоянные. Уравнения (1.76) имеют особую точку $y_1' = y_2' = 0$ при $y_1 = A,\ y_2 = B/A$, а при $B > A^2 + 1$ — единственный предельный цикл, изображенный на рис. 1.9.

Когда B приближается к $A^2 + 1$, предельный цикл становится все меньше и меньше и, наконец, исчезает в критическую точку. Если же данную модель усложнить: считать B переменной величиной и положить в (1.76) $B = y_3, A = 1$, получаем систему ОДУ 3-го порядка

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - (y_3+1)y_1, \\ y_2' &= y_1 y_3 - y_1^2 y_2, \\ y_3' &= -y_1 y_3 + \alpha, \end{aligned} \tag{1.77}$$

в котором новое добавленное уравнение содержит параметр α.

Рис. 1.9. Фазовый портрет уравнений "брюсселятора" (1.76) при $A = 1, B = 3$

Последняя система имеет особую точку $y_1 = 1, y_2 = y_3 = \alpha$ с матрицей Якоби

$$\frac{\partial f}{\partial y} = \begin{bmatrix} \alpha - 1 & 1 & -1 \\ -\alpha & -1 & 1 \\ -\alpha & 0 & 1 \end{bmatrix}.$$

Корни ее характеристического многочлена λ удовлетворяют уравнению

$$p_3(\lambda) = \lambda^3 + (3-\alpha)\lambda^2 + (3-2\alpha)\lambda + 1 = 0.$$

При нарушении условия устойчивости, т.е.для значений α, немного превышающих бифуркационное значение $\alpha_0 = (9 - \sqrt{17})/4 \approx 1.219$, в решении системы (1.77) обнаруживается предельный цикл. Однако если α продолжает расти, то около $\alpha \approx 1.5$ предельный цикл "взрывается" и $y_1 \to 0$ при одновременном $y_2, y_3 \to \infty$. Таким образом, поведение системы (1.77) кардинально отличается от упрощенной модели (1.76).

Приведем еще один классический пример, представляющий задачу с периодическим решением. Это так называемая модель Лотки—Вольтерра эволюции биологических популяций типа "хищник — жертва", описываемая системой двух уравнений (см. подробнее [63])

$$\begin{aligned} \dot{y}_1 &= y_1(y_2 - 2), \\ \dot{y}_2 &= y_2(y_1 - 1). \end{aligned}$$

Решение данной системы обнаруживает три повторяющиеся стадии: рост популяции жертв, увеличение популяции хищников за счет поедания жертв, уменьшение популяции жертв за счет поедания хищниками.

После деления приведенных уравнений одного на другое и разделения переменных мы получаем соотношение

$$\frac{d\,H}{d\,t} = \frac{1-y_1}{y_1}\dot{y}_1 + \frac{y_2-2}{y_2}\dot{y}_2 = 0,$$

где величина

$$H(y_1, y_2) = \ln y_1 - y_1 + 2\ln y_2 - y_2,$$

как видно, не меняется со временем, т. е. является инвариантом решения данных ОДУ. В фазовой плоскости $y = (y_1, y_2)$ каждое из решений представляет собой замкнутую кривую, являющуюся линией уровня функционала $H(y_1, y_2) = \text{const}$.

Отметим также, что *уравнения Лотки—Вольтерра* могут быть записаны в форме гамильтоновой системы

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{bmatrix} 0 & y_1 y_2 \\ -y_1 y_2 & 0 \end{bmatrix} \bigtriangledown H(y_1, y_2).$$

§ 1.8. Задачи к главе 1

1.8.1. Вычислить матрицы e^{At}, если матрица A имеет один из четырех вариантов:

$$a.\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},\ b.\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},\ c.\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},\ d.\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

1.8.2. Доказать, что если оператор A в евклидовом пространстве кососимметрический, то e^A — ортогональный оператор.

1.8.3. Последовательность чисел Фибоначчи $0, 1, 1, 2, 3, ...$ определяется при $a_0 = 0, a_1 = 1$ рекуррентной формулой $a_n = a_{n-1} + a_{n-2}$, или в матричном виде

$$z_n = A\, z_{n-1},\ A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix},\ z_n = \begin{bmatrix} a_n \\ a_{n-1} \end{bmatrix}.$$

Показать, что a_n растет, как геометрическая прогрессия, и найти предел $\alpha = \lim\limits_{n\to\infty} (a_n/n)$.

Ответ: $\alpha = \ln((\sqrt{5}+1)/2),\ a_n = (\lambda_1^n - \lambda_2^n)/\sqrt{5}$, где $\lambda_{1,2} = (1 \pm \sqrt{5})/2$ — собственные числа матрицы A.

1.8.4. Доказать, что если матрицы $A, B \in \mathbb{R}^N$, или линейные операторы $A, B\colon \mathbb{R}^N \to R^N$, коммутируют ($AB = B\,A$), то $e^A e^B = e^B e^A = e^{A+B}$.

Указание: сравните формальные (абсолютно сходящиеся) ряды для матричных экспонент.

1.8.5. Пусть $\{a_{i,j}\}$ — матрица оператора A в ортонормированном базисе. Показать справедливость неравенств

$$\max_j \sum_{i=1}^{N} a_{i,j}^2 \leqslant ||A||_2^2 \leqslant \sum_{i,j}^{N} a_{i,j}^2.$$

1.8.6. Доказать, что собственные числа диагонального оператора вещественны.

1.8.7. Показать, что если все N собственных чисел оператора $A : \mathbb{R}^N \to \mathbb{R}^N$ вещественны и различны, то он диагонален.

1.8.8. Какие из следующих отображений прямой (с обычной метрикой) сжаты?

а) $y = \sin\, t$; б) $y = \sqrt{1+t^2}$; в) $y = \operatorname{arc\,tg}\, t$.

1.8.9. Удовлетворяют ли условию Липшица следующие отображения?

а) $y = t^2,\ t \in \mathbb{R}$; б) $y = \sqrt{t},\ t > 0$; в) $y = t^2,\ |t| \leqslant 1$.

1.8.10. Исследовать, являются ли следующие функции линейно независимыми в их области определения:

а) $1, x$; б) $x, 2\,x, x^2$; в) $x, x\,e^x, x^2 e^x$; г) $1, \sin x, \cos 2\,x$.

1.8.11. Найти определитель Вронского для указанных систем функций:

а) $1, x$; б) x, x^{-1}; в) $1, 2, x^2$; г) $x, \ln x$.

1.8.12. Составить линейные однородные ОДУ, зная их характеристические уравнения:

а) $\lambda^3 = 0$; б) $(\lambda^2+1)^2 = 0$; в) $9\lambda^2 - 6\lambda + 1 = 0$; г) $(\lambda - 1)^2 = 0$.

1.8.13. Составить линейные однородные ОДУ, если известны корни характеристических уравнений:

а) $\lambda_1 = 1, \lambda_2 = 2$; б) $\lambda_1 = \lambda_2 = 1$; в) $\lambda_1 = \lambda_2 = \lambda_8 = 1$.

1.8.14. Составить линейные однородные ОДУ, если заданы их фундаментальные системы решений:

а) e^{-t}, e^{t}; б) $1, e^{t}$; в) $1, x$; г) $\sin x, \cos 2\,x$.

1.8.15. Решить задачу Коши

$$(1 + e^{t})yy' = e^{t},\ \ y|_{t=0} = 1,$$

методом разделения переменных.

1.8.16. Проинтегрировать следующие уравнения с помощью разделения переменных:

а) $(1 + y^2)d\,t = t\,d\,y$; б) $e^{-y}(1 + y') = 1$; в) $y' = \sin(x - y)$.

1.8.17. Решить методом вариации постоянных следующие неоднородные ОДУ:

а) $y' + 2\,y = e^{-t}$; б) $(2\,t - y^2)y = 2\,t$; в) $ty' - 2\,y = t^3 \cos t$.

1.8.18. Решить следующие задачи Коши:

а) $t^2 + ty' = y,\ \ y|_{t=1} = 0$; б) $y' + y\cos t = \cos t,\ \ y|_{t=0} = 1$.

1.8.19. Доказать, что $\det e^{A} \neq 0$. **1.8.20.** Доказать, что для матричного многочлена $P(A_1, ..., A_n), A_1, ..., A_n \in \mathbb{R}^{N,N}$, с неотрицательными коэффициентами справедливо неравенство

$$||P(A_1, ..., A_n)|| \leqslant P(||A_1||, ..., ||A_n||).$$

1.8.21. Исходя из определения устойчивости по Ляпунову, исследовать на устойчивость решения следующих задач Коши:

а) $y' = y + t, y_0 = 1$; б) $y' = 2 + t, y_0 = 1$; в) $y' = t^2 - y, y(1) = 1$.

1.8.22. Определить характер точек покоя для следующих систем ОДУ:

а) $y' = 3\,y_1 + y_2,\ y_2' = -2\,y_1 + y_2$;

б) $y_1' = 3\,y_1,\ \ y_2' = 3\,y_2$

1.8.23. При каких значениях α точка покоя (0,0) устойчива для системы

$$y_1' = -3\,y_1 + \alpha y_2, y_2' = -2\,y_1 + y_2.$$

1.8.24. Исследовать на устойчивость тривиальные решения следующих систем:

а) $\dot{y}_1 = -y_1, \dot{y}_2 = -y_2$; б) $\dot{y}_1 = -2\,y_1 - 3\,y_2, \dot{y}_2 = y_1 - y_2$.

1.8.25. Исследовать на устойчивость систему ОДУ

$$\dot{y}_1 = y_2 - \dot{y}_1^2,\ \ \dot{y}_2 = -y_1 - 3\,y_2^2.$$

Указание: использовать функцию Ляпунова $V = y_1^2 + y_2^2$.

Глава 2

ОСНОВНЫЕ ПОНЯТИЯ И КЛАССИФИКАЦИЯ ЧИСЛЕННЫХ МЕТОДОВ

Целью данной главы является ознакомление с терминологией, общей классификацией и основными понятиями численных методов, которые будут подробнее изучаться в последующих разделах книги. В нижеследующих параграфах содержится большой объем описательного материала: основные подходы к дискретизации задач,различия многостадийных и многошаговых методов, определения погрешности аппроксимации, ошибки и сходимости численного решения, многочисленные типы устойчивости и соответствующие теоремы (как правило, без доказательств, что придает излагаемому материалу справочный и обзорный характер). Два последних раздела являются вспомогательными и содержат некоторые сведения из линейной алгебры, на которых в значительной степени базируется аппарат исследования алгоритмов, а также методы решения нелинейных уравнений, которые неизбежно приходится использовать при реализации неявных схем.

§ 2.1. Терминология численных решений и методов

Если в рассматриваемых нами задачах Коши

$$\dot{y} = f(t, y(t)), \quad y(t_0) = y_0, \quad t \in [t_0, t_e], \tag{2.1}$$

искомое решение y есть функция непрерывного аргумента t, то численное решение — это набор значений $y_n, n = 1, 2, ...$, соответствующий совокупности дискретных значений независимой переменной $t_0 \leqslant t_n \leqslant t_e,\ n = 0, 1, ..., N_e$, которая называется *сетка*. Отображение t_n на интервале интегрирования $[t_0, t_e]$ называется узлом сетки, а N_e есть количество узлов, или размер сетки. Важно отметить, что в абсолютном большинстве случаев численное решение является приближенным $y_n \approx y(t_n)$, во-первых, из-за погрешности метода и, во-вторых, из-за ошибок арифметических операций (машинных округлений) при вычислениях с конечным числом значащих цифр.

Сетка называется *равномерной*, когда все ее *шаги* $h_n = t_n - t_{n-1}$ одинаковы, и *неравномерной* — в противном случае. Будем обозначать через $h = \max\limits_n\{h_n\}$ максимальный шаг неравномерной сетки и через $h_{min} = \min\limits_n\{h_n\}$ — минимальный шаг (очевидно, что для равномерной сетки $h = h_{min} = h_n$ при всех n). Исследования численных методов, или алгоритмов (будем считать эти термины синонимами), как правило, проводятся на *последовательности сгущающихся сеток* $\Omega_h^{(k)} = \{t_n^{(k)}, n = 0, 1, ..., N_e^{(k)}\}$, для которой выполняются правила $h_{min}^{(k)} \leqslant h_{min}^{(k-1)}, h^{(k)} \leqslant h^{(k-1)}, N_e^{(k)} \geqslant N_e^{k-1}$. Последовательность сеток будем называть *сходящейся* при $h^{(k)} \to 0$ для $k \to \infty$, а также *квазиравномерной*, или *регулярной*, ес-

ли $h^{(k)}/h^{(k)}_{min} \leqslant \Delta_k < \infty$ для всех k, где Δ_k характеризуем как *коэффициент неравномерности* сетки $\Omega^{(k)}$. Естественно, что при этом полагается $h^{(k)}_{min} > 0$, т. е. исключаются из рассмотрения сетки с кратными узлами. Сетки называются *вложенными*, если $\Omega^{(k-1)}_h \subset \Omega^{(k)}_h$, т. е. узлы более редкой сетки являются также узлами густой сетки. Когда наличие последовательности сеток явно не акцентируется, то индекс "k" в обозначениях будем опускать и писать $\Omega_h, h_n, h, h_{min}$. В большинстве случаев неявно полагается, что $N_e \gg 1$, а величины шагов h_n достаточно малы, и при этом величины $O(h)$ и $o(h)$ понимаются в обычном функциональном смысле.

2.1.1. Инструменты построения численных алгоритмов и простейшие алгоритмы Эйлера. Если в задаче (2.1) система ОДУ имеет порядок N, то для каждого n-го момента времени $y(t_n) = \{y_l(t_n), l = 1, ..., N\} \in \mathbb{R}^N$ есть вектор той же размерности. Можно также рассматривать и "глобальный" вектор решения $\bar{y} = \{y_l(t_n), n = 0, 1, ..., N_e,\ \ l = 1, ..., N\} \in \mathbb{R}^{N(N_e+1)}$, определенный на сетке Ω_h.

Формулы нахождения новых значений можно представить в виде

$$y_{n+1} = \varphi_n(y_n) = \bar{\varphi}_h(0),$$

причем отображение $\bar{\varphi}_h$ может интерпретироваться как сеточная аппроксимация потока φ_t для ОДУ, определенного в (1.22).

Построение различных методов решения ОДУ осуществляется главным образом на основе двух математических инструментов. Первый из них — это использование интегрального представления решения

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(t,y)dt \tag{2.2}$$

с применением квадратурных формул каких-то порядков для приближенного вычисления интеграла в (2.2) и оценки получаемой при этом погрешности, см. [12], [30].

Замечание 2.1. Вообще говоря, вместо интегрального соотношения (2.2) можно было бы использовать любое из равенств вида

$$y_{n+k_1} = y_{n-k_2} + \int_{t_{n-k_2}}^{t_{n+k_1}} f(t,y)dt.$$

Однако такие кардинальные подходы распространения не получили, за исключением случая $k_1 = k_2 = 1$, рассмотренного в главе 4.

Второй подход основан на *формуле Тейлора*

$$y(t) = P_p(t) + r_p(t), \tag{2.3}$$

представляющей решение (если оно достаточно гладкое в окрестности точки t_n) в виде *многочлена Тейлора* p-го порядка

$$P_p(t) = \sum_{k=0}^{p} \frac{y^{(k)}(t_0)}{k!}(t-t_n)^k \tag{2.4}$$

и остаточного члена (погрешности) $r_p(t)$. Если функция $y(t)$ непрерывно дифференцируема $p+1$ раз, т. е. принадлежит классу $C^{p+1}[t_n, t_{n+1}]$ в некоторой окрестности точки t_n, то ошибка может быть записана в форме Лагранжа

$$r_p(t) = \frac{y^{(p+1)}(\xi)}{(p+1)!}(t-t_n)^{p+1}, \quad \xi \in [t_n, t_n+h]. \tag{2.5}$$

Если же $(p+1)$-я производная решения $y(t)$ интегрируема на отрезке $[t_n, t_n+h]$, то справедлива также интегральная форма погрешности

$$r_p(t) = \frac{1}{n!}\int\limits_{t_n}^{t} y^{(p+1)}(\xi)(t-\xi)^p d\xi. \tag{2.6}$$

Простейший алгоритм решения ОДУ — это *явный метод Эйлера* (1678 г.), заключающийся в рекуррентных вычислениях по соотношениям

$$y_{n+1} = y_n + h_n f(t_n, y_n), \tag{2.7}$$

которые следуют из применения квадратурной формулы левых прямоугольников в (2.2). Если же здесь применить формулу правых прямоугольников, то получаем *неявный метод Эйлера*

$$y_{n+1} = y_n + h_n f(t_{n+1}, y_{n+1}). \tag{2.8}$$

Другая интерпретация формул (2.7) и (2.8) заключается в том, что производная $\dot{y}$ в уравнении (2.1) аппроксимируется односторонней конечной разностью $(y_{n+1}-y_n)/h_n$, которая в первом случае представляет собой правую разность относительно точки t_n, а во втором случае — левую конечную разность, или формулу дифференцирования назад, относительно точки t_{n+1}.

Сравнение формул (2.7) и (2.8) показывает их принципиальную разницу с точки зрения трудоемкости реализации: в

отличие от явного метода, где осуществляются простые рекуррентные вычисления, неявный алгоритм требует для нахождения неизвестного y_{n+1} решения нелинейной (в общем случае) системы уравнений N-го порядка на каждом n-м шаге, что на практике чаще всего осуществляется с помощью какого-либо итерационного алгоритма. Более того, требует исследования и вопрос о существовании и единственности самого решения этой системы.

В силу легко проверяемых соотношений

$$y(t_{n+1}) = y(t_n) + \int\limits_{t_n}^{t_{n+1}} f(t,y)dt = y(t_n) + h_n f_n + \frac{h_n^2}{2} f'(\xi_1) =$$

$$= y(t_n) + h_n f_{n+1} - \frac{h_n^2}{2} f'(\xi_2); \;\; \xi_1, \xi_2 \in [t_n, t_{n+1}],$$

в предположении $y(t_n) = y_n$ мы можем записать

$$\psi_h^{(l)} \equiv y(t_{n+1}) - y_{n+1}^{(l)} = (-1)^{l+1} \frac{h_n^2}{2} f'(\xi_l), \qquad (2.9)$$

где $l = 1, 2$ для явного и неявного методов Эйлера соответственно. Величина $\psi_h^{(l)}$ называется *локальной ошибкой* метода, или *погрешностью аппроксимации*. Как видно из (2.9), она для обоих методов Эйлера есть $O(h^2)$, если функция $f(t) \in C^1[t_0, t_e]$.

Рассмотренные алгоритмы относятся к классу *одношаговых методов*, требующих для вычисления произвольного y_{n+1} только одного предыдущего значения y_n. В отличие от них, в *многошаговых методах* определение некоторого y_{n+k} требует знания величин $y_{n+k-1}, ..., y_n$. Общего вида *k-шаговый метод* записывается в форме сеточного уравнения

$$F_h(y_{n+k}, y_{n+k-1}, ..., y_n, f_{n+k}, ..., f_n) = 0, \tag{2.10}$$

где F_h — некоторая сеточная нелинейная функция от указанных в скобках величин.

Сеточное уравнение (2.10) иногда удобно представлять также в разрешенной относительной неизвестной y_{n+k} форме:

$$y_{n+k} = \varphi_h(y_{n+k}, ..., y_n, f_{n+k}, ..., f_n). \tag{2.11}$$

2.1.2. Погрешности аппроксимации и ошибки приближенных решений. Данные понятия являются первичными в теории любых численных методов. А при изучении алгоритмов для систем ОДУ они требуют формального уточнения, поскольку в литературе по данным вопросам имеются некоторые терминологические расхождения.

Определение 2.1. *Погрешностью аппроксимации (локальной ошибкой) k-шагового метода называется сеточная функция*

$$\psi_n^h = F_h(y(t_{n+k}), y(t_{n+k-1}), ..., y(t_n), f(y(t_{n+k}), ..., f(y(t_n))), \quad n = 0, 1, ..., N \tag{2.12}$$

где $y(t_i)$ — вектор значения точного решения в точке t_i.

Слово “локальная"в данном определении имеет тот смысл, что фактически величина ψ^h характеризует ошибку метода при реализации только одного шага, т. е. разность $y(t_{n+k}) - y_{n+k}$, в предположении точного знания решения на предыдущих шагах ($y_{n+j-1} = y(t_{n+j-1})$ для $j = 1, 2, ..., k-1$).

Определение 2.2. *Порядок погрешности равен γ (метод согласован с порядком γ), если $||\psi^h||_\infty = \max\limits_n \{|\psi_n^h|\} = O(h^{\gamma+1})$ при $h \to 0$.*

Если погрешность ψ^h допускает представление

$$\psi^h = h^{\gamma+1}\psi(t_n, y(t_n)) + O(h^{\gamma+2}), \tag{2.13}$$

где ψ — некоторая не зависящая от h ограниченная функция, то сеточная функция $h^{\gamma+1}\psi(t_n, y(t_n))$ называется главным членом асимптотического разложения локальной ошибки метода в узле сетки t_n.

Для неявного метода Эйлера, например, имеем

$$F_h(y(t_{n+1}), y(t_n), f(t_{n+1}, y(t_{n+1})) = y(t_{n+1}) - y(t_n) - h_n f(t_{n+1}, y(t_{n+1})) =$$

Согласно введенным определениям, погрешности аппроксимации явного и неявного методов Эйлера имеют первый порядок (согласованы с первым порядком).

Очевидно, что величина и порядок локальной ошибки γ зависят от той нормировки, которая выбирается в уравнении (2.10). Чтобы устранить имеющийся формальный произвол, мы здесь и далее естественно предполагаем, что F_h записывается таким образом, при котором искомая величина y_{n+k} входит с коэффициентом единица (или порядка единицы). Сделаем еще одно замечание: порядок погрешности аппроксимации γ предполагает определенную гладкость искомого решения и правой части (как правило, $f \in C^{\gamma}[t_0, t_e]$, $y \in C^{\gamma+1}[t_0, t_e]$).

Погрешность аппроксимации ψ_n^h можно также определить на основе формулы (2.11), разрешенной относительно неизвестной y_{n+k}:

$$\psi_n^h = y(t_{n+k}) - \varphi_h(y(t_{n+k}), ..., y(t_n), f(y(t_{n+k})), ..., f(y(t_n)). \tag{2.14}$$

В силу выбранной нами нормировки сеточной функции F_h, определения локальной ошибки ψ_n^h по формулам (2.12) и (2.14) совпадают. Обоим определениям погрешности аппроксимации можно дать следующую наглядную интерпретацию: ψ_n^h есть невязка формул (2.12) или (2.14) численного метода, т. е. величина, получаемая при замене приближенных решений $y_n, ..., y_{n+k}$ на их точные значения $y(t_n), ..., y(t_{n+k})$.

Определение 2.3. *Ошибкой, или глобальной погрешностью, численного решения называется сеточная функция*

$$z^h = (y)^h - y^h = \{y(t_n) - y_n\}, \tag{2.15}$$

где $(y)^h$ *есть вектор, компоненты которого суть значения точного решения в узлах сетки.*

Иногда вместо понятия глобальной погрешности используется термин "полная погрешность". Отметим еще, что далее зачастую мы будем употреблять термин *порядок точности метода*, подразумевая под этим порядок глобальной погрешности.

Определение 2.4. *Метод сходится на задаче Коши с порядком* $\gamma > 0$ *(порядок глобальной ошибки равен* γ*), если*

$$\begin{aligned} &||z^h||_\infty = O(h^\gamma) \quad \text{при} \quad h \to 0, \\ &||z^h||_\infty = \max_{l,n}\{|z^h_{n,l}|,\ l = 1, ..., p, \quad n = 1, ..., N\}. \end{aligned} \tag{2.16}$$

Если в контексте величина порядка γ не имеет значения, то просто говорят, что метод сходится на задаче Коши, а более строго — приближенное решение сходится к точному при $h \to 0$.

2.1.3. Формы представления многошаговых и многостадийных методов. Чтобы яснее ориентироваться в имеющемся многообразии алгоритмов, их можно заранее классифицировать. И что особенно важно — установить на дальнейшее по возможности удобные обозначения.

В сеточных уравнениях вида (2.11) функция f может или зависеть от искомой величины y_{n+k}, или не зависеть. В первом случае для ее нахождения необходимо решать систему алгебраических уравнений, и соответствующие методы называются *неявными*. В противном случае значение y_{n+k} определяется из более простых рекуррентных соотношений, и такой метод называется *явным*. Реализация отдельного n-го шага в неявных методах более трудоемка, чем в явных. Отличие особенно значительно для нелинейных систем ОДУ, когда решение вспомогательных уравнений может потребовать проведения итераций.

Среди многошаговых алгоритмов наибольшее распространение имеют *линейные многошаговые методы* (ЛММ), записываемые в виде

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h_{n+k} \sum_{i=0}^{k} \beta_i f_{n+i}, \quad h_{n+k} = t_{n+k} - t_{n+k-1}, \qquad (2.17)$$

где α_i, β_i — некоторые числовые параметры, $f_{n+i} = f(t_{n+i}, y_{n+i})$ и дополнительно предполагается

$$\alpha_k \neq 0, \quad |\alpha_0| + |\beta_0| > 0.$$

Первое допущение означает, что уравнение (2.17) (даже если оно неявное) однозначно разрешимо относительно y_{n+k},

по крайней мере при достаточно малых h. Это как раз соответствует ранее выбранной нормировке общей формулы (2.10) k-шагового метода, которая для (2.17) принимает вид

$$F_h = \sum_{i=0}^{k} (\alpha_i y_{n+i} - h_{n+k}(\beta_i f_{n+i})) = 0.$$

Второе условие здесь просто указывает, что метод является k-шаговы и его всегда можно удовлетворить, уменьшив при необходимости индекс k. За счет выбора значений α_i и β_i можно обеспечить тот или иной порядок погрешности. Чем больше свободных параметров многошагового метода, тем выше достижимый порядок. Многошаговые методы (слово "линейные", означающее, что в (2.17) величины y_{n+i} и f_{n+i} входят линейным образом, мы зачастую будем опускать) чаще всего применяются с постоянным шагом, и поэтому далее вместо h_{n+k} используем h. Отметим также, что если функция f зависит от решения y, то в правой части (2.17) для явных методов сумма берется не до k, а до $k-1$, т. е. $\beta_k = 0$.

Если метод (2.17) согласован, т. е. порядок аппроксимации по крайней мере $\gamma \geqslant 1$, то решение должно обладать какой-то гладкостью, например, $y(t) \in C^1[t_0, t_e]$ как минимум откуда на коэффициенты α_i, β_i накладываются условия

$$\sum_{i=0}^{k} \alpha_i = 0, \quad \sum_{i=0}^{k} i\,\alpha_i = \sum_{i=0}^{k} \beta_i.$$

Чтобы в этом убедиться, достаточно подставить в (2.17) два различных решения: $y = 1$ и $y = t$ при соответствующих правых частях $f = 0$ и $f = 1$.

Замечание 2.2. Данные соотношения можно интерпретировать и обобщить следующим образом. Если порядок аппроксимации равен $\gamma > 1$, то уравнения (2.17) должны выполняться на многочленах $f = (q+1)t^q$ и $y = t^{q+1}$, удовлетворяющих уравнению $y' = f$. Поэтому вместо (2.13) для ЛММ мы можем записать *условия коллокации*

$$\sum_{i=0}^{k}(\alpha_i i^{q+1} - \beta_i i^q) = 0, \;\; q = 0, 1, ..., \gamma. \tag{2.18}$$

Очевидно, что эти соотношения являются необходимыми условиями согласованности с порядком γ. Можно также сформулировать и такое определение: условия коллокации (2.18) характеризуют алгебраическую согласованность метода, т. е. максимальный порядок многочленов, на которых его погрешность аппроксимации равна нулю.

Замечание 2.3. Формула (2.17) k-шагового метода содержит всего $2k+1$ свободных параметров, если он неявный, и $2\,k$ параметров, если он явный (один параметр уходит на нормировку). С другой стороны, как мы видели только что выше, если порядок погрешности равен γ, то коэффициенты α_i, β_i должны удовлетворять $\gamma + 1$ условию согласованности. Таким образом, наивысший порядок, который мы можем надеяться получить, равен $2k$ для неявного многошагового метода и $2\,k - 1$ — для явного. Такие алгоритмы будем называть *оптимальными по порядку*.

Замечание 2.4. Поскольку формула (2.17) для k-шагового метода предполагает знание уже вычисленных нескольких предыдущих шагов, то ее использование требует

нахождение стартовых значений с помощью каких-то других процедур вида

$$y_r = s_r(h), \;\; 0 \leqslant r = k-1.$$

Например, это могут быть рассмотренные ниже одношаговые методы. Для этих стартовых процедур аналогично определяются локальные ошибки:

$$\psi_r^h = y(t_r) - s_r(h), \;\; r = 0, ..., k-1.$$

Для одношаговых методов повышение точности требует введения дополнительных параметров. Это осуществляется путем использования промежуточных точек на интервале $[t_n, t_{n+1}]$. Получаемые одношаговые алгоритмы называются *методами Рунге—Кутты* (МРК) и в достаточно общем виде определяются формулами

$$\begin{aligned} y_{n+1} &= y_n + \sum_{j=1}^{m} b_j k_j, \\ k_j &= h_n f(t_n + c_j h_n, y_n + \sum_{l=1}^{d} a_{j,l} k_l), \;\; j = 1, ..., m, \end{aligned} \tag{2.19}$$

где $a_{j,l}, b_j$ и c_j — подбираемые по условиям аппроксимации числовые параметры, а целое $m > 1$ называется *числом стадий* (этапов) метода. При $d = m$ эти алгоритмы называются неявными (НМРК), при $d = j - 1$ — явными (ЯМРК), а в промежуточных случаях $d = j$ — полунеявными, или диагонально неявными (ДНМРК). Их существенное различие заключается в способах реализации: если N есть порядок системы ОДУ (и вектора y_n), то в НМРК на каждом n-м шаге

необходимо решать алгебраическую систему порядка $m\,N$, в НМРК все вычисления осуществляются простым рекуррентным образом, а ДНМРК представляет собой промежуточный вариант, т. е. решение одной системы $m\,N$-го порядка сводится к решению m систем порядка N, что является гораздо более легкой задачей.

К многостадийным алгоритмам относится также семейство *явных одношаговых методов типа Розенброка*(МтР), описываемых соотношениями

$$
\begin{aligned}
&y_{n+1} = y_n + \sum_{j=1}^{m} b_j k_j,\\
&k_j = h_n[I - a\,h_n J(t_n, y_n)]^{-1} f(t_n + c_j h_n, y_n + \sum_{l=1}^{j-1} a_{j,l} k_l),\\
&J = \frac{\partial\, f}{\partial\, y},\ j = 1, ..., m.
\end{aligned}
\tag{2.20}
$$

Здесь $a_{j,l}, b_j$ и c_j — числовые параметры, как и в МРК, I — единичная матрица N-го порядка, а J — матрица Якоби решаемой системы ОДУ. Дополнительный параметр a полагается всегда положительным, а при $a = 0$ формулы (2.20) определяют явные методы Рунге—Кутты.

Принцип построения МтР можно пояснить следующим образом. Запишем для автономного ОДУ m-стадийный полунеявный метод Рунге—Кутты в следующей форме:

$$
\begin{aligned}
&y_{n+1} = y_n + \sum_{j=1}^{m} b_j k_j,\\
&k_j = h\, f(y_n + \sum_{l=1}^{j-1} a_{j,l} k_l + a_{j,j} k_j),\ j = 1, ..., m.
\end{aligned}
$$

Как видно, последнее соотношение при заданных коэффициентах $a_{j,l}$ представляет собой для каждого j нелинейное уравнение относительно одного неизвестного k_j (или систему уравнений порядка N, если исходная система ОДУ имеет соответствующий порядок). Общий подход к решению таких нелинейных систем заключается в проведении итераций, чаще всего — по методу Ньютона, для чего на каждой итерации надо вычислять матрицу Якоби J. Идея МтР заключается в упрощении вычислений — мы ограничиваемся выполнением только одной ньютоновской итерации. Другими словами, мы проводим линеаризацию уравнения для k_j и получаем

$$k_j = h\, f(g_j) + h\, f'(g_j) a_{j,j} k_j,$$
$$g_j = y_n + \sum_{p=1}^{j-1} a_{j,l} k_j,$$

откуда и следует (2.20) после введения формально обобщающего параметра a, оптимизация которого может повысить точность или устойчивость алгоритма.

2.1.4. Другие методы: блочные, гибридные, предиктор-корректор, экстраполяционные. В рамках главного разделения методов на многошаговые и многостадийные возможны различные модификации и обобщения.

Один из подходов к обобщению многошаговых схем — это построение блочных, или сверхнеявных, алгоритмов. Суть их заключается в том, что на n-ом шаге ищется не только величина y_{n+1}, а группа из $m > 1$ неизвестных $y_{n+1}, ..., y_{n+m}$. Если порядок системы ОДУ равен N, то в этом случае на каждом шаге требуется решать алгебраическую систему по-

рядка mN. Например, при $m = 2$ мы имеем блочный метод Клиппинджера—Димсдейла:

$$y_{n+1} - \frac{1}{2}y_{n+2} = \frac{1}{2}y_n + \frac{h}{4}(f_n - f_{n+2}),$$
$$y_{n+2} = y_n + \tfrac{h}{3}(f_n + 4f_{n+1} + f_{n+2}).$$

Если блочный вектор $(y_{n+1}, y_{n+2})^T$ известен, то по этим формулам определяется следующий блок $(y_{n+2}, y_{n+3})^T$. В связи с этим данную схему можно было бы назвать блочным неявным одношаговым методом и записать в соответствующей векторно-матричной форме:

$$\begin{bmatrix} 1 & -0.5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{n+1} \\ y_{n+2} \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_n \\ y_{n+1} \end{bmatrix} + \begin{bmatrix} 0.25h(f_n - f_{n+2}) \\ h(f_n + 4\,f_{n+1} + f_{n+2})/3 \end{bmatrix}$$

Понятно, что численное решение систем ОДУ можно проводить с помощью гибридных методов, использующих методологию как многостадийных, так и многошаговых алгоритмов. Такие схемы под названиями *общие линейные методы, или многошаговых методов Рунге—Кутты*, будут рассмотрены в четвертой главе. Кроме того, применяются иногда также неоднородные вычислительные схемы, включающие различные типы формул на разных временных шагах. Однако, как правило, мы будем рассматривать только однородные вычислительные алгоритмы, использующие одинаковые формулы на всех шагах $n = 1, 2, ...$

Укажем еще на такой достаточно общий подход к "улучшению" алгоритмов решения ОДУ, как методы *предиктор — корректор*, или *прогноза и коррекции*. Здесь суть заключается в том, что искомая величина сначала определяется приближенно по достаточно грубой формуле, а затем каким-либо

образом “исправляется”. Более того, процедура уточнения может производиться несколько раз, и фактически при этом осуществляется какой-то итерационный процесс.

В качестве примера рассмотрим предложенный Г. В.Демидовым и Е. А.Новиковым оригинальный метод второго порядка точности, требующий для реализации каждого шага только однократного вычисления правой части, см. ДемидовГ. В., Новиков Е. А. Экономичный алгоритм интегрирования нежестких систем обыкновенных дифференциальных уравнений. “Численные методы математической физики”.Новосибирск : ВЦ СО АН СССР, 1979, 69—83.

Данный алгоритм основан на введении вспомогательных величин $\hat{y}_n$, при условии $\hat{y}_0 = y_0$, и на проведении расчетов по формулам

$$\hat{y}_{n+1} = y_n + h_n f(t_n, \hat{y}_n),$$
$$y_{n+1} = y_n + 0.5h_n[f(t_n, \hat{y}_n) + f(t_{n+1}, \hat{y}_{n+1})].$$

Первое из этих соотношений представляет собой явный метод Эйлера, применяемый для прогноза, а второе — процедуру коррекции, осуществляемую с помощью квадратурной формулы трапеций. Строго говоря, в нем для аппроксимации интеграла в (2.2) надо использовать величины y_n, y_{n+1} вместо $\hat{y}_n, \hat{y}_{n+1}$, но здесь учитывается тот факт, что такая замена не изменяет порядка локальной ошибки.

Возможные вариации методов предиктор — корректор продемонстрируем на двухшаговой схеме Адамса (подробнее это семейство алгоритмов рассматривается в главе 4), для которого формулы прогноза и коррекции имеют вид

$$\hat{y}_{n+1} = y_n + 0.5\,h(3\,f_n - f_{n-1}),$$
$$y_{n+1} = y_n + 0.5\,h(f_n + \hat{f}_{n+1}),\quad \hat{f}_{n+1} = f(\hat{y}_{n+1}).$$

Наглядное описание этапов вычислений мы дадим с помощью принятой англоязычной абревиатуры: P (predictor) означает однократное применение предсказывающей формулы, C (corrector) — однократное применение исправляющей формулы и E (evaluation) — вычисление функции f.

В этих обозначениях вычислительный процесс с использованием одной итерации на каждом шаге можно записать как PECE и представить в следующей форме (верхний индекс означает номер итерации):

$$\begin{aligned}
&P\colon y_{n+1}^{(0)} = y_n^{(1)} + 0.5\,h(3\,f_n^{(1)} - f_{n-1}^{(1)}),\\
&E\colon f_{n+1}^{(0)} = f(t_{n+1}, y_{n+1}^{(0)}),\\
&C\colon y_{n+1}^{(1)} = y_n^{(1)} + 0.5\,h(f_n^{(1)} + f_{n+1}^{(0)}),\\
&E\colon f_{n+1}^{(1)} = f(t_{n+1}, y_{n+1}^{(1)}).
\end{aligned}$$

Как отсюда видно, этот алгоритм требует на каждом шаге двукратного вычисления правой части. Поскольку этап коррекции основан на квадратурной формуле трапеций, легко показать, что локальная погрешность данного метода имеет второй порядок.

Данную схему можно упростить, убрав одно вычисление правой части, а получаемый алгоритм можно обозначить как PEC и записать следующим образом:

$$\begin{aligned}
&P\colon y_{n+1}^{(0)} = y_n^{(1)} + 0.5\,h(3\,f_n^{(0)} - f_{n-1}^{(0)}),\\
&E\colon f_{n+1}^{(0)} = f(t_{n+1}, y_{n+1}^{(0)}),\\
&C\colon y_{n+1}^{(1)} = y_n^{(1)} + 0.5\,h(f_n^{(0)} + f_{n+1}^{(0)}).
\end{aligned}$$

С другой стороны, приведенные методы можно уточнить, если операции evaluation — correction повторять несколько раз, в результате чего получим схемы вида $P(EC)^r E$ или $P(EC)^r$, $r > 1$. Очевидно, что при этом порядок локальной ошибки не повысится, однако за счет дополнительных вычислений можно уменьшить количественно погрешность аппроксимации и повысить устойчивость численного метода.

Общую вычислительную схему методов прогноза и коррекции типа PECE можно сформулировать следующим образом, обозначая через $\hat{\alpha}_i, \hat{\beta}_i, \hat{y}_{n+i}, \hat{f}_{n+i} = f(\hat{y}_{n+i})$ и $\hat{k}$ величины, относящиеся к этапу предиктора:

$$\sum_{i=0}^{\hat{k}} (\hat{\alpha}_i \hat{y}_{n+k} - h\hat{\beta}_i \hat{f}_{n+i}) = 0,$$
$$\alpha_k y_{n+k} - h\,\beta_k f_{n+k} - \textstyle\sum_{i=0}^{k-1} (\alpha_i \hat{y}_{n+i} - h\beta_i \hat{f}_{n+i}) = 0.$$

Одним из универсальных подходов к повышению точности алгоритмов решения ОДУ является *экстраполяция* численных результатов, полученных на разных сетках. Исходная идея основана на разложении решения в ряд Тейлора и связывается с именами Рунге и Ричардсона. Ее применение проиллюстрируем на формуле трапеций

$$y_{n+1} = y_n + 0.5\,h(f(t_n, y_n) + f(t_{n+1}, y_{n+1})).$$

Можно показать, что если $y(t) \in C^{2\,M+1}[t_0, t_e]$, то при $y_0 = y(t_0)$ для данного численного решения имеет место представление

$$y_n = y(t_n) + \sum_{i=1}^{M} c_i(t_n) h^{2\,i} + O(h^{2\,M+1}),$$

где $c_i(t) \in C^{2M+1-2i}[t_0, t_e]$, а $y(t_n)$ — значение точного решения ОДУ при $t = t_n$.

Отсюда несложно получить, например, что если по рассматриваемому методу трапеций найдены решения $y(t,h)$ и $y(t,h/2)$ с шагами h и $h/2$ соответственно, то для их линейной комбинации

$$\tilde{y}(t,h) = y(t,h/2) + (y(t,h/2) - y(t,h))/3$$

справедливо в каждой точке $t \in \{t_n\}$ разложение

$$\tilde{y}(t,h) = y(t) + \sum_{i=2}^{M} d_i(t)h^{2i} + O(h^{2M+1})$$

для некоторых вектор-функций $d_i(t) \in C^{2M+1-2}[t_0, t_e]$. Таким образом, если глобальная ошибка для формулы трапеций есть $O(h^2)$, то для экстраполированного численного решения — $O(h^4)$. Более того, на основе последнего разложения несложно построить линейные комбинации $\tilde{y}(t,h)$ для исключения членов с коэффициентами $d_2, d_3, ...$ и получить экстраполяционные формулы более высоких порядков. Данный общий прием называется *экстраполяцией Ричардсона*.

Другим важным моментом использования асимптотических тейлоровских разложений, кроме уточнения численных решений, является возможность апостериорной оценки ошибки. Особенно это эффективно можно осуществить, если удается построить *двусторонние приближения* к искомому решению.

Например, явный и неявный методы Эйлера (2.7) и (2.8), как следует из (2.9), для скалярного ОДУ имеют локальные погрешности разных знаков, если первая производная f'

не меняет знака на интервале $[t_n, t_{n+1}]$. Таким образом, если $y_n = y(t_n)$, то модуль ошибки на $(n+1)$-м шаге не превосходит модуля разности двух численных решений, полученных разными методами:

$$\max_l |y(t_{n+1}) - y_{n+1}^{(l)}| \leqslant |y_{n+1}^{(0)} - y_{n+1}^{(1)}|,$$

где $l = 0, 1$ для явного и неявного методов Эйлера соответственно.

Даже если двусторонние приближения обосновать не удается, апостериорную оценку ошибки можно получить с помощью следующего *правила Рунге*.

Пусть с помощью какого-то метода порядка γ на сетках с шагами h и $h/2$ получены два численных решения, для которых справедливы представления

$$y_{n+1}^h = y(t_{n+1}) + h^\gamma \psi + o(h^\gamma),\ y_{n+1}^{h/2} = y(t_{n+1}) + \left(\frac{h}{2}\right)^\gamma \psi + o(h^\gamma).$$

Тогда отсюда следует соотношение

$$y_{n+1}^h - y_{n+1}^{h/2} = (1 - 2^{-\gamma}) h^\gamma \psi + o(h^\gamma),$$

которое после исключения ψ позволяет получить апостериорную оценку

$$\left\| y_{n+1}^{h/2} - y(t_{n+1}) \right\| = (2^{-\gamma} - 1)^{-1} \left\| y_{n+1}^h - y_{n+1}^{h/2} \right\| + o(h^\gamma).$$

Данная методика позволяет также вместо получения апостериорной оценки повысить порядок численного результата за счет использования линейной комбинации двух сеточных решений y_{n+1}^h и $y_{n+1}^{h/2}$:

$$\tilde{y}^h_{n+1} = (1 - 2^{-\gamma})^{-1}(y^{h/2}_{n+1} - 2^{-\gamma} y^h_{n+1}) = o(h^\gamma).$$

Более того, если порядок γ заранее неизвестен, то его можно приближенно вычислить апостериори за счет использования решения на третьей сетке с шагом $\dfrac{h}{4}$. В силу представления

$$y^{h/4}_{n+1} = y(t_{n+1}) + \left(\frac{h}{4}\right)^\gamma \psi + o(h^\gamma),$$

из которого следует соотношение

$$y^{h/2}_{n+1} - y^{h/4}_{n+1} = (1 - 2^{-\gamma}) 2^{-\gamma} h^\gamma \psi + o(h^\gamma),$$

мы получаем

$$\gamma \approx \log \frac{||y^{h/2}_{n+1} - y^{h/4}_{n+1}||}{||y^h_{n+1} - y^{h/2}_{n+1}||} / \log 2.$$

§ 2.2. Устойчивость и сходимость методов

Существует достаточно большое разнообразие понятий и характеристик устойчивости одношаговых и многошаговых методов решения ОДУ. Зачастую они связаны с исследованиями частного вида систем уравнений. Простейшим примером является однородное линейное скалярное уравнение

$$\dot{y} = \lambda y, \quad \lambda \in \mathbb{C}, \quad t \geqslant 0, \quad y(0) = y_0, \tag{2.21}$$

называемое модельным уравнением Далквиста, который начал его активно использовать в исследованиях с 1959 г. Далее обозначим через $\mathbb{C}_-$ левую полуплоскость комплексной плоскости, включая мнимую ось.

Более содержательная модельная задача — это однородная автомодельная система ОДУ

$$y' = A\,y,\ \ A = \{a_{i,j}\} \in \mathbb{R}^{N,N},\ \ y \in \mathbb{R}^N,$$

с постоянной матрицей A. Здесь наиболее простой случай имеем, если A принадлежит классу диагонализируемых матриц, т. е. существует невырожденная матрица B, которая реализует преобразование подобия

$$B^{-1} A\, B = \Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_N),$$

где λ_i суть собственные числа A. Отметим, что хотя мы рассматриваем только вещественные системы ОДУ, несимметричная матрица A может иметь комплексные собственные значения, что объясняет использование комплексных λ в (2.21) (отметим, что если рассматриваются скалярные вещественные ОДУ, т. е. $N = 1$, то достаточно ограничиваться вещественными λ).

С помощью рассмотренного преобразования вышеприведенное матричное уравнение распадается на N независимых скалярных уравнений вида (2.21):

$$\hat{y}' = \{\hat{y}'_i\} = B^{-1} y' = B^{-1} A\, B \cdot B^{-1} y = \Lambda \hat{y} = \{\lambda_i \hat{y}_i\},$$
$$\hat{y}'_i = \lambda \hat{y}_i,\ \ i = 1, ..., p.$$

Таким образом, подобные задачи тоже могут быть исследованы с помощью уравнения Далквиста (2.21).

И наконец, дальнейшим шагом к усложнению модельной задачи является рассмотрение линеаризованной автономной системы

$$y' = J\,y,$$

где матрица Якоби $J = \dfrac{\partial f}{\partial y}$ может иметь зависящие от аргумента t элементы. Исследование в данном случае усложняет-

ся, поскольку преобразующая матрица B, да и сами собственные числа, будут переменными во времени.

2.2.1. Понятие устойчивости по норме и оценки погрешности одношаговых методов. Исследования устойчивости методов решения систем ОДУ в значительной степени используют такой естественный принцип, как изучение отдельных частных ситуаций, в силу чего возникает много различных понятий и терминов. Однако это никак не отменяет общего понятия устойчивости, лежащего в основе классической теоремы эквивалентности Лакса.

Традиционно анализ устойчивости проводится с помощью различных подходов для одношаговых и многошаговых методов. Общую процедуру численного интегрирования системы ОДУ с помощью одношаговых алгоритмов можно записать в виде

$$\begin{aligned} &y_0 = s(h_0), \\ &y_{n+1} = B_n y_n + h_n g_n(t_n, y_n, h_n), \end{aligned} \tag{2.22}$$

где s — некоторая стартовая процедура, B_n — квадратная *матрица перехода*, или шага интегрирования, а g_n — функция приращения решения на n-м шаге.

С помощью разложения в ряд Тейлора несложно проверить, что если каждая функция g_n удовлетворяет условию Липшица по переменной y_n, то из (2.22) для вектора ошибки численного решения $z_{n+1} = y(t_{n+1}) - y_{n+1}$ следует рекуррентное соотношение

$$z_{n+1} = B_n z_n + h_n \psi_{n+1}, \quad \|\psi_{n+1}\| \leqslant ch^\gamma, \quad n = 0, 1, ... \tag{2.23}$$

где ψ_{n+1} есть локальная погрешность аппроксимации, а c —

некоторая постоянная.

Определение 2.5. *Одношаговый метод* (2.22) *называется устойчивым по норме, если в соотношении* (2.23) *для матричной нормы* $\|B_n\|$*, подчиненной произвольной векторной норме* $\|z_n\|$*, выполняется неравенство*

$$\|B_n\| \leqslant 1 + hb, \quad 0 < b < \infty, \tag{2.24}$$

где b *— независящая от* n *константа.*

Из (2.23) при условии (2.24) следует цепочка неравенств

$$\|z_{n+1}\| \leqslant (1 + h\,b)\|z_n\| + h\|\psi_{n+1}\| \leqslant$$
$$\leqslant (1 + h\,b)^{n+1}\|z_0\| + h[\|\psi_{n+1}\| + (1 + h\,b)\|\psi_n\| + ... + (1 + hb)^n\|\psi_0\|].$$

Отсюда, в силу соотношения

$$(1 + h\,b)^{n+1} \leqslant \exp\{b(t_e - t_0)\},$$

с учетом (2.23) получаем оценку для глобальной ошибки метода (2.22):

$$\|z_{n+1}\| \leqslant \exp\{b(t_e - t_0)\}(\|z_0\| + ch^\gamma/b). \tag{2.25}$$

Таким образом, мы фактически получаем следующее важное утверждение, являющееся усилением знаменитой теоремы эквивалентности Лакса: из аппроксимации и устойчивости метода следует сходимость приближенного решения.

Теорема 2.1. *Если для одношагового метода* (2.22) *выполняются соотношения* (2.23) *и условие устойчивости*

(2.24), *то при* $z_0 = O(h^\gamma)$ *приближенное решение сходится с порядком* γ.

Сформулированный результат можно усилить в том плане, если рассмотреть вместо (2.22) возмущенное численное решение

$$\tilde{y}_0 = \tilde{s}(h_0),$$
$$\tilde{y}_{n+1} = \tilde{y}_n + h_n[g_n(t_n, \tilde{y}_n, h_n) + r_{n+1}],$$

где вектор возмущения r_{n+1} может быть обусловлен или погрешностью арифметических операций, или неточностью "недоитерирования" в неявных методах, или приближенным способом вычисления функции правой части f.

Теперь вместо (2.23) будем предполагать справедливость следующих неравенств для возмущенной ошибки $\tilde{z}_{n+1} = y(t_{n+1}) - \tilde{y}_{n+1}$:

$$\tilde{z}_{n+1} = B_n\tilde{z}_n + h_n(\psi_{n+1} + z_{n+1}),$$
$$\|\psi_{n+1} + r_{n+1}\| \leqslant ch^\gamma + d, \quad n = 0, 1, ...,$$

с независящими от h постоянными c, d. Отсюда при выполнении того же условия устойчивости (2.24) вместо (2.25) приходим к оценке

$$\|\tilde{z}_{n+1}\| \leqslant \exp\{b(t_e - t_0)\}[\|z_0\| + (ch^\gamma + d)/b], \tag{2.26}$$

из которой можно определить вклады в глобальную ошибку погрешностей начальных данных, локальной аппроксимации и возмущения алгоритма.

Отметим, что за определение устойчивости вместо (2.24) в случае постоянных матриц $B_n = B$ можно было бы принять

равномерную по n ограниченность матриц B^n, поскольку отсюда (при выполнении условия Липшица для функций g_n) следует неравенство (2.24).

Теорема 2.1 основана на неравенстве (2.25) и ее справедливость имеет асимптотический характер, т.е. при $h \to 0$. Однако на конкретной сетке с шагами h_n при росте n величина ошибки может сильно возрастать, особенно при больших b. Отсюда возникает потребность как-то ужесточить понятие устойчивости.

2.2.2. Абсолютная устойчивость, D-устойчивость и A-устойчивость методов. Данные понятия имеют основополагающую роль в теории численных схем, хотя они получены из анализа самого тривиального модельного уравнения Далквиста.

Определение 2.6. *Метод называется абсолютно устойчивым с шагом h, если при его применении с этим шагом к модельному уравнению* (2.21) *при $Re\lambda < 0$ численное решение $y_n \to 0$ для $n \to \infty$. Термин "абсолютно" здесь означает при любой величине $|\lambda|$.*

Определение 2.7. *Множество $D \subset \mathbb{C}_-$ называется областью абсолютной устойчивости метода, если метод является абсолютно устойчивым с любым шагом $h > 0$, для которого $\mu = \lambda h \in \mathbb{D}$.*

Введенные понятия формально носят универсальный характер, но наиболее актуальны они для многошаговых методов вида (2.17), к рассмотрению устойчивости которых мы сейчас переходим.

Определение 2.8. *Полином*

$$\rho(\theta) = \sum_{i=0}^{k} \alpha_i \theta^i$$

первым характеристическим полиномом, или производящим многочленом, линейного k-шагового метода (2.17).

Определение 2.9. *Линейный многошаговый метод называется устойчивым (нуль-устойчивым, или D-устойчивым в честь Далквиста), если его первый характеристический (производящий) полином удовлетворяет корневому условию:*

- *все корни θ_j многочлена $\rho(\theta)$ лежат внутри единичного круга или на единичной окружности,*
- *корни $|\theta_j| = 1$, принадлежащие единичной окружности, являются простыми.*

Мотивировку данных определений можно объяснить следующим образом. Равенство $\rho(\theta) = 0$ является характеристическим для однородного разностного уравнения

$$\alpha_k y_{n+k} + ... + \alpha_0 y_n = 0, \tag{2.27}$$

которое можно рассматривать как формулу аппроксимации тривиального ОДУ $y' = 0$, имеющего своим решением константу. Подставляя в данное сеточное уравнение решение в виде $y_j = \theta^j$ и разделив на θ^n, мы получим

$$\rho(\theta) = \alpha_k \theta^k + \alpha_{k-1} \theta^{k-1} + ... + \alpha_0 = 0. \tag{2.28}$$

Пусть это уравнение имеет корни $\theta_1, ..., \theta_l$ кратности $m_1, ..., m_l$ соответственно, $m_j \geqslant 1, m_1 + ... + m_l = k$. Тогда, как показано Лагранжем (1792 г.), общее решение однородного уравнения (2.27) представляется суперпозицией его k линейно независимых решений вида

$$\binom{n}{i}\theta_j^{n-l}, \;\; 1 \leqslant i \leqslant m_j - 1, \;\; 1 \leqslant j \leqslant l \leqslant n.$$

Так как устойчивость метода предполагает ограниченность y_n при $n \to \infty$, то отсюда и следует необходимость выполнения корневых условий.

Отметим, что в силу условия согласованности сумма всех α_k равна нулю и характеристическое уравнение (2.28) обязательно имеет корень $\theta_j = 1$.

Отсюда, в частности, следует важное утверждение: одношаговый метод вида (2.17) всегда является D-устойчивым (поскольку его характеристический полином имеет первый порядок с единственным корнем $\theta_1 = 1$).

Из общих соображений устойчивости понятно, что чем меньше корней характеристического многочлена со свойством $|\theta_j| = 1$, тем лучше. В связи с этим дадим следующее понятие.

Определение 2.10. *Говорят, что многошаговый метод* (2.17) *удовлетворяет сильному корневому условию, если его характеристический полином* $\rho(\theta)$ *имеет простой корень* $+1$, *а все остальные корни лежат строго внутри единичного круга.*

Теорема 2.2 (теорема эквивалентности Лакса). *Если линейный многошаговый метод устойчив и согласован с*

порядком γ, то он с тем же порядком γ.

Выполнение корневого условия гарантирует сходимость согласованных линейных многошаговых методов на равномерной сетке, когда коэффициенты $\alpha_i,\ i = 0, ..., k$ являются постоянными величинами. На неравномерной сетке при частом изменении шага у многошаговых методов может возникнуть неустойчивость. Применение численных методов подразумевает возможность использования достаточно больших размеров шага в процессе счета, чтобы численное решение сохраняло хотя бы некоторые свойства точного решения. Анализ поведения решения при больших размерах шага обычно проводится для модельного скалярного уравнения.

Определение 2.11. *Пересечение области абсолютной устойчивости с вещественной осью называется интервалом абсолютной устойчивости.*

Данное понятие вводится для простого сравнительного анализа свойств устойчивости различных методов, поскольку их области абсолютной устойчивости могут иметь сложную конфигурацию.

При решении линейных систем ОДУ желательно, чтобы величины $\lambda_j(t)h,\ j = 1, ..., p$, где $\lambda_j(t)$ — собственные значения матрицы Якоби от правой части системы ОДУ на ее решении, находились в области абсолютной устойчивости метода. Нарушение требования абсолютной устойчивости метода дает о себе знать ростом модуля численного решения. Так как собственные значения $\lambda_j(t)$ в общем случае комплексные, то выход $\lambda_j(t)h$ из области абсолютной устойчивости метода может происходить как за счет $Re\lambda_j(t)h$, так и $Im\lambda_j(t)h$. При реше-

нии нелинейных систем ОДУ явление неустойчивости может наблюдаться даже при интегрировании со сверхмалым шагом.

Определение 2.12. *Метод называется A-устойчивым, если он абсолютно устойчивый с любым шагом h, т.е. его область абсолютной устойчивости $\mathbb{D} = \mathbb{C}_-$ есть вся левая полуплоскость комплексной плоскости, а интервал абсолютной устойчивости — это вещественная полуось $(-\infty, 0]$.*

При использовании A-устойчивых методов на практике размер шага интегрирования определяется только требованием задаваемой точности вычисления решения ОДУ на участках его быстрого и медленного изменения.

2.2.3. Функция устойчивости и L-устойчивость методов. При применении одношаговых методов к модельному уравнению $y' = \lambda y$ получается выражение

$$y_{n+1} = R(\mu) y_n, \;\; \mu = \lambda\, h. \tag{2.29}$$

Определение 2.13. *Функция $R(\mu)$ называется функцией устойчивости метода, а функция $R(\mu)e^{-\mu}$ — функцией относительной устойчивости.*

Область абсолютной устойчивости одношагового метода определяется неравенством

$$|R(\mu)| < 1, \tag{2.30}$$

а *область относительной устойчивости* — неравенством

$$|R(\mu)e^{-\mu}| < 1. \tag{2.31}$$

Иногда удобнее вместо (2.31) рассматривать неравенство $|R(\mu)| < e^{Re\mu}$. Функция относительной устойчивости связывает два понятия: точность и устойчивость метода. В области

относительной устойчивости одношагового метода не допускаются большие уклонения $|R(\mu)|$ от $e^{Re\mu}$.

Если неравенство (2.30) выполняется для всех $\mu \in \mathbb{C}_-$, то функция устойчивости метода называется *A—допустимой*. Очевидно, что одношаговый метод A-устойчив, если и только если его функция устойчивости A—допустима.

Мы дадим еще одну разновидность свойства устойчивости, которое важно для алгоритмов, ориентированных на численное решение жестких задач Коши с длительными интервалами интегрирования.

Определение 2.14. *Одношаговый метод называется L-устойчивым, если он —устойчивый и* $|R(\mu)| \to 0$ *при* $Re\mu \to -\infty$.

Мотивировкой данного определения может послужить следующая иллюстрация. Пусть имеется задача Коши

$$\dot{y} = -a(y - \cos t),\ y(0) = 0,\ 0 \leqslant t \leqslant 1.5,$$

где коэффициент a — достаточно большое положительное число, например, $a = 2000$. Решение $y(t)$ сначала в окрестности нуля быстро растет (короткий переходной период, или пограничный слой), а затем начинает убывать примерно по формуле $y = \cos(t)$. И именно на этом достаточно гладком периоде установления "хороший" алгоритм должен иметь функцию чувствительности $R(z)$, обеспечивающей демпфирование численного решения, т. е. должно выполняться $R(z) \to 0$ при $z \to \infty$. Для рассматриваемой задачи методы, у которых величина $|R(z)| < 1$, но с ростом z остается близкой к единице, будут давать катастрофически плохие результаты.

Пример 2.1. Для явного метода Эйлера (2.7) имеем $\mathbb{R}(\mu) = 1 + \mu$,и его область абсолютной устойчивости в комплексной плоскости $\mu = x + iy$ определяется как $(1+x)^2 + y^2 < 1$, т.е. представляет собой внутренность круга единичного радиуса с центром в точке $(-1, 0)$, а область относительной устойчивости характеризуется неравенством

$$(1 + x)^2 + y^2 < e^{2\,x}.$$

Очевидно, что явный метод Эйлера не является ни A-устойчивым, ни тем более L-устойчивым, а его интервал абсолютной устойчивости — отрезок $(-1, 0]$.

Пример 2.2. Неявный метод Эйлера (2.8) имеет функцию устойчивости $\mathbb{R}(\mu) = (1 - \mu)^{-1}$, при этом неравенство $R(\mu) < 1$ определяет внешность круга единичного радиуса с центром $(+1, 0)$. Соответствующая область относительной устойчивости определяется неравенством

$$(1 - x)^2 + y^2 > e^{-2\,x},$$

а интервал относительной устойчивости — неравенством

$$e^{-x} < \frac{1}{1 - x} < e^{x}.$$

Легко проверить, что неявный метод Эйлера является L-устойчивым и тем более A-устойчивым, а его интервал абсолютной устойчивости есть объединение двух отрезков $(-\infty, 0] \cup [1, \infty)$.

Отметим, что если функция устойчивости $R(\mu)$ не имеет полюсов в $\mathbb{C}_-$, то неравенство (2.31) в силу принципа максимума модуля аналитической функции можно заменить на требование

$$|R(i\omega)| \leqslant 1 \tag{2.32}$$

для всех ω.

Если одношаговый метод применить к однородной линейной системе ОДУ с постоянной матрицей A, то получится выражение

$$y_{n+1} = R(A\,h)y_n. \tag{2.33}$$

Матрицу $R(Ah)$ называют *матрицей перехода* одношагового метода. Если матрица A является устойчивой (т.е. ее собственные значения $\lambda_i,\ i = 1, ..., N$ лежат в $\mathbb{C}_-$), а одношаговый метод A—устойчивый, то для любого $h > 0$ выполняются неравенства $|R(\lambda_i h)| < 1,\ i = 1, ..., N$, где $R(\mu)$ — функция устойчивости метода. Но тогда спектральный радиус матрицы $R(Ah)$ удовлетворяет неравенству $\rho[R(A\,h)] < 1$ для любого $h > 0$ и представляет собой $|y_n| \to 0 \quad n \to \infty$. Это означает, что свойство A-устойчивости одношаговых численных методов автоматически переносится на асимптотически устойчивые линейные системы ОДУ с постоянными коэффициентами.

Иногда может быть полезна следующая

Теорема 2.3. *Для абсолютной устойчивости с шагом $h > 0$ одношагового численного метода относительно асимптотически устойчивой линейной системы ОДУ необходимо, чтобы для любой и достаточно, чтобы для какой-нибудь симметричной отрицательно определенной матрицы Q, решение X дискретного уравнения Ляпунова*

$$R^T(A\,h)X\,R(A\,h) - X = Q \tag{2.34}$$

было положительно определенной матрицей.

Справедливости ради следует сказать, что хотя данный результат в теоретическом плане исчерпывающе закрывает проблему устойчивости, его практическое применение в конкретных задачах далеко не просто.

2.2.4. Характеристические полиномы многошаговых методов. Приведенное выше понятие D-устойчивости использует свойства только первого характеристического многочлена, и вполне естественно рассмотреть влияние остальных коэффициентов многошаговой схемы. Мы это сделаем на примере *уравнения Далквиста* $y' = \lambda y$.

Определение 2.15. *Полином*

$$\sigma(\theta) = \sum_{i=0}^{k} \beta_i \theta^i \tag{2.35}$$

— называется вторым характеристическим многочленом линейного k-шагового метода, а полином

$$\pi_\mu(\theta) = \rho(\theta) - \mu\sigma(\theta) \tag{2.36}$$

называется характеристическим полиномом линейного k-шагового метода (иногда также и многочленом устойчивости).

Так как корни полинома являются непрерывными функциями его коэффициентов, то корни $\theta_j(\mu)$ многочлена

$$\pi_\mu(\theta) = (\alpha_k - \mu\beta_k)\theta^k + ... + \alpha_0 - \mu\beta_0$$

стремятся при $\mu \to 0$ к корням первого характеристического полинома $\rho(\theta)$. Для более детального исследования поведения глобальной ошибки метода (2.17) остановимся на модельном

уравнении Далквиста $y' = \lambda\, y$, точное решение которого есть $y(t_n) = y_0 e^{\lambda(t_n - t_0)}$. Если метод согласован с порядком $\gamma > 0$, то мы имеем равенство

$$\pi_\mu(y(t_n)) = \sum_{i=0}^{k} (\alpha_i - \mu\beta_i) e^{i\mu} y_0 e^{\lambda(t_n - t_0)} = \psi = O(h^{\gamma+1}), \;\; \mu = \lambda\, h.$$

Поскольку величины y_0 и $\lambda(t_n - t_0) = \lambda\, h\, n$ ограничены, то мы можем также записать

$$\pi_\mu(e^{\lambda\, h}) = O(h^{\gamma+1}).$$

Предполагая теперь, что определяемый многочленом π_μ метод является D-устойчивым, мы можем утверждать, что θ_1 — единственный корень, который может стремиться к значению $+1$ при $h \to 0$, откуда следует представление

$$\pi_\mu(e^{\lambda\, h}) = (e^{\lambda\, h} - \theta_1)\varkappa_\mu(\lambda\, h), \;\; |\varkappa_\mu(\lambda\, h)| \underset{h\to 0}{\to} |\varkappa| = O(h^0),$$

означающее справедливость соотношения

$$\theta_1(\mu) = e^{\lambda\, h} + O(h^{p+1}) \underset{h\to 0}{\to} +1.$$

Таким образом, если $\mathrm{Re}\lambda\, h \leqslant 0$, то мы имеем $|\theta_1| \leqslant 1$, а при $\mathrm{Re}\lambda h > 0$ величина $\mu = \lambda\, h$ находится вне области устойчивости $R(\mu)$.

Однако может быть ситуация, когда $\theta_j(\mu) \underset{h\to 0}{\to} \theta_j(0)$ и $|\theta_j(0)| = 1$, но сами значения $|\theta_j(\mu)| > 1$ при любых малых значениях $\mu = \lambda\, h$. В качестве примера рассмотрим оптимальную по порядку *двухшаговую схему Симпсона* (см. подробнее главу 4)

$$y_{n+2} = y_n + \frac{h}{3}(f_{n+2} + 4\, f_{n+1} + f_n),$$

для которой характеристический многочлен имеет вид

$$\pi_\mu(\theta) = (1 - \lambda\, h/3)\theta^2 - \frac{4}{3}\lambda\, h\theta - (1 + \lambda\, h/3).$$

Корни этого полинома равны

$$\theta_{1,2} = \frac{2\mu/3 \pm \sqrt{1 + \mu^2/3}}{1 - \mu/3},$$

откуда при $|\lambda\, h| \ll 1$ получаем

$$\theta_1 = 1 + \lambda\, h + O(h^2),\ \ \theta_2 = -1 + \lambda\, h/3 + O(h^2).$$

Следовательно, если величина $R\, e\lambda\, h$ мала и положительна, то $|\theta_1| > 1$, а при малых $\mathrm{Re}\lambda\, h < 0$ имеем $|\theta_2| > 1$. Таким образом, для метода Симпсона не существует интервала абсолютной устойчивости, по крайней мере, в окрестности начала координат. Более детальный анализ показывает, что граница области устойчивости $\partial\, R(\mu)$ представляет собой отрезок мнимой оси от $-\sqrt{3}i$ до $+\sqrt{3}i$. Для тех $\lambda\, h$, которые принадлежат этому отрезку, корни $\theta_j(\mu)$ лежат на единичной окружности, а для всех других $\lambda\, h$ полином $\pi_\mu(\theta)$ имеет корни, находящиеся вне единичного круга.

Из приведенного примера ясно, что если предельные значения $\theta_j(\mu \to 0)$, представляющие собой корни первого характеристического полинома $\rho(\theta)$, обладают свойством $|\theta_j(0)| = 1$, то возникает опасность, что область $R(\mu)$ окажется пустой. Однако, что эта опасность не фатальна, как подтверждает следующий пример Штеттера (1965 г.). Для двухшагового метода

$$y_{n+2} = y_n + h(f_{n+1} + 3\, f_n)/2$$

характеристический полином имеет вид

$$\pi_\mu(\theta) = \theta^2 - \lambda\, h\theta/2 - (1 + 3\lambda\, h/2),$$

а его корни при $\lambda\, h \ll 1$ равны

$$\theta_1 = 1 + \lambda\, h + O(h^2),\ \ \theta_2 = -1 - \lambda\, h/2 + O(h^2).$$

Отсюда для малых положительных $\mathrm{Re}\lambda\, h$ имеем $|\theta_1| > 1,\ |\theta_2| > 1$, в то время как для малых отрицательных $\mathrm{Re}\lambda\, h$ получаем $|\theta_1| < 1,\ |\theta_2| < 1$ и, очевидно, область $R(\mu)$ не пуста.

Теорема 2.4. *Пусть $\theta_i(\mu),\ i = 1, ..., k$ — корни характеристического полинома $\pi_\mu(\theta)$ линейного k-шагового метода. Тогда областью абсолютной устойчивости метода является множество $\mathbb{D} = \{\mu \in \mathbb{C}_- : |\theta_i(\mu)| < 1,\ i = 1, ..., k\}$.*

Для исследования многошагового метода может быть полезно представить его формально в виде одношагового алгоритма, но в векторном пространстве большей размерности.

Перепишем сначала схему (2.17), которую для простоты будем рассматривать на равномерной сетке, в форме

$$y_{n+k} = -\sum_{i=0}^{k-1} \alpha'_i y_{n+i} + h \sum_{i=0}^{k} \beta'_i f_{n+i}, \tag{2.37}$$

где $\alpha'_i = \alpha_i/\alpha_k, \beta'_i = \beta_i/\alpha_k$. Введя теперь векторную величину $\varphi_n = \varphi(t_n, y_n, ..., y_{n+k-1}) \in \mathbb{R}^N$, неявно определяемую уравнением

$$\varphi = \beta'_k f(t_n + kh, h\varphi - \sum_{i=0}^{k-1} \alpha'_i y_{n+i}) + \sum_{i=0}^{k-1} \beta'_i f(t_n + ih, y_{n+i}),$$

приведем формулу (2.37) к виду

$$y_{n+k} = -\sum_{i=0}^{k-1} \alpha'_i y_{n+i} + h\varphi_n. \tag{2.38}$$

Далее для k-шагового метода определим "большой" вектор размерности kN, где N — порядок системы ОДУ:

$$Y_n = (y_{n+k-1}, ..., y_{n+1}, y_n)^T \in \mathbb{R}^{kN}. \tag{2.39}$$

С его помощью уравнение (2.38) может быть представлено в форме одношаговой схемы

$$Y_{n+1} = B\, Y_n + h\Phi_n, \;\; n \geqslant 0, \tag{2.40}$$

где вектор $\Phi_n \in \mathbb{R}^{kN}$ и квадратная матрица перехода $B \in \mathbb{R}^{k\,N, k\,N}$ записываются с помощью тензорного произведения Кронекера:

$$\Phi_n = (e_1 \otimes I)\varphi_n, \;\; B = A \otimes I. \tag{2.41}$$

Здесь $e_1 = (1, 0, ..., 0) \in -R^k$ есть вектор-орт, I — единичная матрица N-го порядка, а $A \in -R^{N,N}$ — квадратная матрица

$$A = \begin{bmatrix} -\alpha'_{k-1} & -\alpha'_{k-2} & \dots & -\alpha'_0 \\ 1 & 0 & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix}. \tag{2.42}$$

Напомним, что если $A = \{a_{i,j}\} \in \mathbb{R}^{k,k}$, $I \in \mathbb{R}^{N,N}$, то $B = A \otimes I = \{B_{i,j}\} \in \mathbb{R}^{k\,N, k\,N}$ есть блочная матрица с блоками $B_{i,j} = a_{i,j} I \in \mathbb{R}^{N,N}$.

Определение 2.16. *Метод* (2.40) *называется сильно устойчивым, если единица является простым собственным числом B и матрица B^n равномерно ограничена при $n \geq 0$.*

В связи с понятием сильной устойчивости полезно вспомнить определение D-устойчивости, связанное с корневым условием первого характеристического многочлена $\rho(\theta)$. В силу условий согласованности (2.18) данный многочлен всегда имеет единицу своим собственным числом. Естественно, что устойчивость многошагового метода в общем случае будет лучше, если полином $\rho(\theta)$ не имеет других корней на единичной окружности. Поэтому очевидно, что понятие сильной устойчивости связано со свойством сильного корневого условия, представленного ранее в определении 2.10.

2.2.5. Уравнение Протеро—Робинсона. Согласно теореме Лакса, из аппроксимации и устойчивости следует сходимость многошаговых методов. Рассмотрим ее характер на примере задачи Коши для скалярного *уравнения Протеро—Робинсона*

$$y' = \lambda y + g(t), \;\; y(t_0) = y_0, \tag{2.43}$$

где λ — комплексное число, а функция $g(t)$ такая, что решение $y(t)$ имеет достаточное число производных, ограниченных величинами, не зависящими от значения λ, которое мы будем называть параметром жесткости ОДУ (2.43).

Применение многошагового метода (2.17) к уравнению (2.43) дает формулу

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h\lambda \sum_{i=0}^{k} \beta_i y_{n+i} + h \sum_{i=0}^{k} \beta_i g(t_{n+i}).$$

Легко увидеть, что глобальная ошибка $z_n = y(t_n) - y_n$ удовлетворяет разностному уравнению

$$\sum_{i=0}^{k}(\alpha_i - h\lambda\beta_i)z_{n+i} = -\delta(t_n) = \\ = -\sum_{i=0}^{k}[\alpha_i y(t_n + ih) - h\beta_i y'(t_n + ih)], \tag{2.44}$$

правая часть которого не зависит от λ. Кроме того, если порядок локальной погрешности есть γ, то $\delta(t) = O(h^{\gamma+1})$.

Вводя вектор ошибки

$$Z_n = (z_{n+k-1}, ..., z_{n+1}, z_n)^T \in \mathbb{R}^k,$$

а также соответствующую матрицу

$$B(\mu) = \begin{bmatrix} c_{k-1}(\mu) & \dots & c_1(\mu) & c_0(\mu) \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{bmatrix} \in \mathcal{R}^{k,k}, \tag{2.45}$$

$$c_i(\mu) = -(\alpha_i - \mu\beta_i)/(\alpha_k - \mu\beta_k), \;\; \mu = \lambda\, h,$$

и вектор правой части $\Delta_n = (\delta(t_n)/(\alpha_k - \mu\beta_k), 0, ..., 0)^T$, получаем рекуррентное соотношение

$$Z_{n+1} = B(\mu)E_n + \Delta_n,$$

отсюда следует выражение для глобальной ошибки

$$Z_{n+1} = B(\mu)^{n+1}Z_0 + \sum_{i=0}^{n} B(\mu)^{n-i}\Delta_i. \tag{2.46}$$

Таким образом, для получения оценки ошибки многошагового метода необходимо, чтобы степени матрицы $B(\mu)$

были равномерно ограничены по $\mu = \lambda h$.

§ 2.3. Некоторые сведения из линейной алгебры

Исследование алгоритмов решения систем ОДУ, как и других методов вычислительной математики, в значительной степени базируется на аппарате линейной алгебры, в силу чего в данном параграфе мы представим некоторые необходимые понятия.

2.3.1. Векторные и матричные нормы. Для вещественных векторов $v \in \mathcal{R}^N$ определяются следующие *нормы Гельдера*:

$$||v||_q = (\sum_{i=1}^{N} |v_i|^q)^{1/q},\ q = 1, 2, ..., \infty. \tag{2.47}$$

Среди них наиболее употребительны три нормы:

$$||v||_1 = \sum_{i=1}^{N} |v_i|,\ \ ||v||_2^2 = (v, v) = \sum_{i=1}^{N} v_i^2,\ \ ||v||_\infty = \max_i\{|v_i|\},$$

которые называются *октаэдрической*, *сферической (евклидовой)* и *кубической (равномерной)* соответственно.

Множество N-мерных вещественных векторов, снабженное скалярным произведением $(u, v) = \sum_{i=1}^{N} u_i v_i$, образует гильбертово конечномерное пространство. Матрица A называется *положительно определенной* (п. о.),если для всех $v \in \mathbb{R}^N$ $(Av, v) \geqslant \delta(v, v),\ \delta > 0$. Если же при этом $A = A^T$, то она называется *симметричной положительно определенной* (с.п.о.). Каждая с.п.о. матрица определяет A-скалярное произведение и соответствующую норму

$$(u,v)_A = (Au,v),\ \ ||v||_A^2 = (Av,v).$$

Для всех векторных норм определяются *подчиненные матричные* нормы (индекс “q для краткости опускаем):

$$||A|| = \max_{v\neq 0}\frac{||A\,v||}{||v||}.$$

В частности, при $q = \infty, 1, 2$ имеем

$$\begin{aligned} &||A||_\infty = \max_i\{\sum_{j=1}^N |a_{i,j}|\},\\ &||A||_1 = \max_j\{\sum_{i=1}^N |a_{i,j}|\},\ ||A||_1 = ||A^T||_\infty,\\ &||A||_2 = \max_q\{\nu_q\},\ A\,A^T z_q = \nu_q^2 z_q, \end{aligned} \tag{2.48}$$

где $\nu_q \geqslant 0$ — *сингулярные числа* матрицы A, которые для симметричных матриц равны собственным числам λ_q, определяемым равенствами

$$Az_q = \lambda_q z_q,$$

где z_q — соответствующие собственные векторы.

Максимальный модуль собственных чисел

$$\rho(A) = \max_q\{|\lambda_q|\}$$

называется *спектральным радиусом*, который для симметричных матриц совпадает с евклидовой нормой.

Рассмотрим еще для вещественной матрицы A норму вида

$$||A||_s = \Big(\sum_{i=1}^N\sum_{j=1}^N a_{i,j}^2\Big)^{1/2},$$

называемую *нормой Шура*. Отметим, что хотя для всех векторов $v \in \mathbb{R}^N$ справедливо неравенство

$$||Av||_2 \leqslant ||A||_s ||v||_2,$$

величина $||A||_s$ не является подчиненной матричной нормой для сферической нормы вектора $||v||_2$, поскольку в последнем соотношении $||A||_s$ не является точной верхней гранью ($||A||_s \leqslant \nu_{max} = ||A||_2$).

2.3.2. Числа обусловленности и оценки возмущения решения систем линейных алгебраических уравнений (СЛАУ). Для каждой из норм определяется число обусловленности невырожденной матрицы

$$\mathrm{cond}(A) = ||A|| \cdot ||A^{-1}||.$$

В частности, при $q = 2$ имеем *спектральное число обусловленности*

$$\mathrm{cond}_2(A) = \max_q |\nu_q| / \min_q |\nu_q|. \tag{2.49}$$

Очевидно, что для любой из норм $\mathrm{cond}(A) \geqslant 1$ и равенство достигается, в частности, на единичной матрице $A = I$. Матрицы с большим числом обусловленности, а также соответствующие СЛАУ ($\mathrm{cond}(A) \gg 1$) называются *плохо обусловленными*.

Важно отметить, что с помощью числа обусловленности определяется оценка возмущения решения уравнения. Пусть мы имеем невозмущенное и возмущенное СЛАУ

$$A\,v = f, \;\; A(v + \delta\, v) = f + \delta\, f,$$

откуда следует, что возмущение решения $\delta\, v$ связано с возмущением правой части $\delta\, f$ уравнением

$$A\delta\, v = \delta\, f.$$

Отсюда в предположении $f \neq 0, v \neq 0$ легко получаем следующую цепочку неравенств:

$$\begin{aligned} &||\delta\, v|| = ||A^{-1}\delta\, f|| \leqslant ||A^{-1}|| \cdot ||\delta\, f||, \;\; ||v|| \leqslant ||A|| \cdot ||f||, \\ &\frac{||\delta\, v||}{||v||} \leqslant ||A|| \cdot ||A^{-1}|| \frac{||\delta\, f||}{||f||} = \Delta_f. \end{aligned} \tag{2.50}$$

Итоговое соотношение можно интерпретировать следующим образом: относительная величина возмущения правой части и обусловленная им относительная ошибка решения СЛАУ связаны неравенством, коэффициент в котором есть число обусловленности матрицы A. Важно отметить, что все неравенства в (2.50) являются неулучшаемыми в том смысле, что для любой матрицы найдутся f и $\delta\, f$ такие, при которых все неравенства переходят в равенства.

Если же правая часть известна точно, но вычисления ведутся с возмущенной (невырожденной) матрицей $A = A + \delta\, A$, то приближенное решение $\hat{v}$ удовлетворяет системе $(A + \delta\, A)\hat{v} = f$ и его погрешность оценивается неравенством

$$\frac{\|v - \hat{v}\|}{\|v\|} \leqslant \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta\, A\|}{\|f\|} = \Delta_A.$$

Наконец, когда приближенно даны элементы и матрицы, и правой части, то для решения системы $(A + \delta\, A)(v + \delta\, v) = f + \delta\, f$ получаем следующий результат:

$$\frac{\|\delta\, v\|}{\|v\|} \leqslant \Delta_f + \Delta_A + \Delta_f.\Delta_A.$$

Это неравенство также является неулучшаемым, или точным, в том смысле, что найдутся такие A, f, δA и δf, для которых неравенство перейдет в равенство.

Полученные неравенства показывают, что при фиксированных ошибках исходных данных $||\delta f||, ||\delta A||$ оценки погрешности Δ_f и Δ_A прямо пропорциональны $\mathrm{cond}(A)$. С одной стороны, это оправдывает название "плохих" матриц, а с другой — характеризует относительность понятия устойчивости: можно говорить, что одна матрица хуже обусловлена, чем другая, но нельзя провести строгую границу между "хорошо" и "плохо" обусловленными СЛАУ.

2.3.3. Треугольные разложения матриц. Аппарат исследования алгебраических систем уравнений и методы их решения в значительной степени базируются на различных разложениях, или факторизациях, матриц. Естественно по этим вопросам дать предварительно общие сведения из теории матриц, что мы сейчас и сделаем.

Если у матрицы A все главные миноры не равны нулю (здесь можно использовать не вполне устоявшееся определение — *строго, или вполне, невырожденная матрица)*, то она представима в виде произведения нижней и верхней треугольных матриц: $A = L\,U$. Такое *разложение на треугольные множители*, или факторизация, не единственно, но если у матрицы L или U зафиксировать диагональные элементы, то оно становится единственным. Обычно под термином "$L\,U$-разложение" понимают представление, в котором диагональные элементы матрицы L равны единице. Диагональные элементы матрицы U при этом равны отношению соответству-

ющих главных миноров $|A_i|$ матрицы A:

$$u_{i,i} = |A_i|/|A_{i-1}|, \qquad i = 1, 2, \ldots, N.$$

На такой именно факторизации основан классический *метод исключения Гаусса*, сводящий задачу решения системы $A\,u = f$ к последовательному решению вспомогательных систем с нижней и верхней треугольными матрицами $L\,y = f$, $U\,u = y$.

Не вдаваясь в алгоритмические особенности метода исключения Гаусса, укажем, что количество арифметических действий Q и объем оперативной памяти P, необходимые для его выполнения, оцениваются при больших N величинами

$$Q = \frac{2}{3}N^3 + O(N^2), \qquad P = N^2 + O(N).$$

Реализация метода Гаусса существенно упрощается, если матрица A является ленточной с шириной полуполосы m, т. е. ее элементы $a_{i,j}$ равны нулю при $|i - j| > m$. В этом случае матрицы L, U оказываются также ленточными и затраты вычислительных ресурсов уменьшаются до величин

$$Q = \frac{2}{3}m^2 N + O(m\,N), \qquad P = 2\,m\,N + O(N).$$

Очевидно, что для строго невырожденной матрицы существуют и разложения вида

$$A = L\,D\,U, \tag{2.51}$$

где $D = \{d_{i,i}\}$ — диагональная матрица. Для одной из возможных факторизаций элементы матриц-множителей определяются формулами

$$d_{i,j} = 1/(|A_{i-1}||A_i|), \qquad l_{i,j} = |A_j(i,j)|, \qquad u_{i,j} = |A_i(i,j)|.$$

Здесь $|A_i(i,j)|$ — определитель окаймленной подматрицы i-го порядка, имеющей вид

$$A_i(i,j) = \begin{bmatrix} A_{i-1} & \bar{A}_{i,j} \\ \\ \bar{A}_i & a_{i,j} \end{bmatrix},$$

где A_{i-1} — главная подматрица $(i-1)$-го порядка, а $\bar{A}_i$, $\bar{A}_{i,j}$ — вектор-строка и вектор-столбец $(i-1)$-го порядка, состоящие из первых элементов строки и столбца матрицы A с соответствующими номерами.

Если матрица A симметрична, то в разложении (2.51) матрицы L и U – транспонированные по отношению друг к другу, и мы можем записать $A = L\,D^{1/2} \cdot D^{1/2}L^T = \bar{L}\bar{L}^T, \quad \bar{L} = L\,D^{1/2}$. Такая факторизация симметричной матрицы единственна, и она называется *разложением Холесского*.

2.3.4. Общие решения сеточных СЛАУ. Различные методы решения ОДУ порождают системы разностных, или сеточных, систем уравнений, для представления решений которых существует общая теория, аналогичная теории решения обыкновенных дифференциальных уравнений. Естественно, что и для тех и для других серьезные достижения ограничиваются линейными задачами, а для нелинейных уравнений результаты носят только “штучный” характер.

Универсальная формулировка для линейных уравнений заключается в следующем: общее решение неоднородного уравнения есть сумма общего решения однородного уравнения и частного решения неоднородного уравнения. В свою

очередь, общее решение однородного уравнения есть линейная комбинация линейно-независимых решений, коэффициенты которых определяются из каких-либо дополнительных условий (начальных или краевых).

Будем рассматривать в данном пункте сеточные (разностные) линейные неоднородные уравнения k-го порядка в следующем виде:

$$y(x+k) + p_1(x)y(x+k-1) + ... + p_k(x)y(x) = q(x), \quad (2.52)$$

где $p_1(x), ..., p_k(x)$ и $q(x)$ — заданные функции от аргумента x, который считаем принимающим целые значения $0, 1, 2, ...$ Все функции предполагаем имеющими конечные и определенные значения на этом множестве, а $p_k(x)$ на нем не равен тождественно нулю. Если $q(x) \equiv 0$, то уравнение (2.52) называется однородным. Функция $y(x)$ является искомым решением уравнения (2.52).

Следующее утверждение можно назвать принципом суперпозиции.

Теорема 2.5. *Если $y_1(x), y_2(x), ..., y_k(x)$ — решения однородного уравнения* (2.53), *причем определитель k-го порядка*

$$D[y_1(0), ..., y_k(0)] = \begin{vmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,k} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,k} \\ \cdots & \cdots & \cdots & \\ y_{k,1} & y_{k,2} & \cdots & y_{k,k} \end{vmatrix} \quad (2.53)$$

отличен от нуля, то общее решение линейного однородного уравнения имеет вид

$$y(x) = C_1 y_1(x) + ... + C_k y_k(x), \tag{2.54}$$

где $C_1, ..., C_k$ *— произвольные постоянные.*

В формуле (2.53) используются обозначения

$$y_{i,j} = y_i(j-1),\ i, j = 1, ..., k.$$

Легко показать, что все решения (2.52) при $q(x) \equiv 0$ содержатся в совокупности функций (2.54).

Теорема 2.6. *Общее решение линейного неоднородного уравнения* (2.52) *представляется в виде суммы его частного решения* $y_0(x)$ *и общего решения линейного однородного уравнения, т. е.*

$$y(x) = y_0(x) + C_1 y_1(x) + ... + C_k y_k(x), \tag{2.55}$$

где $y_1(x), ..., y_k(x)$ *— частные решения однородного уравнения, удовлетворяющие условию* $D[y_1(0), ..., y_k(0)] \neq 0$.

Определение 2.17. *Функции* $y_1(x), ..., y_k(x)$ *при* $x = 0, 1, 2, ...$ *называются линейно зависимыми, если для данных значений аргументов имеет место соотношение*

$$C_1 y_1(x) + ... + C_k y_k(x) = 0, \tag{2.56}$$

где $C_1, ..., C_k$ *— постоянные, не равные нулю одновременно.*

Теорема 2.7. *Если функции* $y_1(x), ..., y_k(x)$ *линейно зависимы, то определитель*

$$D[y_1(x), ..., y_k(x)] = \begin{vmatrix} y_1(x) & \cdots & y_k(x) \\ y_1(x+1) & \cdots & y_k(x+1) \\ y_1(x+k-1) & \cdots & y_k(x+k-1) \end{vmatrix} \tag{2.57}$$

равен нулю при всех значениях x. *Обратно, если определитель* (2.57) *равен нулю при* $x = 0, 1, ...,$ *а* $D[y_2(x)...y_k(x)] \neq 0$ *при* $x = 0, 1, ...,$ *то функции* $y_1(x), ..., y_k(x)$ *линейно зависимы и* $y_1(x)$ *входит в соотношение* (2.56) *с коэффициентом* $C_1 \neq 0$.

Рассмотрим теперь некоторые свойства частных решений линейного однородного уравнения.

Определение 2.18. *Функции* $y_1(x), ..., y_k(x)$ *называются линейно независимыми решениями однородного уравнения* (2.52), *если они при всех целых* $x \geqslant 0$ *принимают конечные значения и удовлетворяют этому уравнению, а соотношение* (2.56) *при любых постоянных* $C_1, ..., C_k$, *одновременно не равных нулю, хотя бы для одного* $x \geqslant 0$ *не выполняется.*

Нетрудно показать, что определитель k линейно независимых решений однородного уравнения (2.52) k-го порядка не может быть тождественно равен нулю, т. е. равенство $D[y_1(x), ..., y_k(x)]$ при всех $x = 0, 1, ...$ невозможно.

Теперь мы можем дать другую формулировку теоремы 2.5: если $y_1(x), ..., y_k(x)$ — линейно независимые решения однородного уравнения (2.52), то всякое другое решение $y(x)$ этого уравнения, принимающее конечные и определенные значения при $x \geqslant 0$, может быть представлено в виде (2.55), где $C_1, ..., C_k$ суть произвольные постоянные.

Перейдем далее к рассмотрению важного частного случая — линейного однородного уравнения с постоянными коэффициентами, которое запишем в форме

$$y(x + k) + a_1 y(x + k - 1) + ... + a_k y(x) = 0. \qquad (2.58)$$

Будем искать решение (2.58) в виде $y(x) = \lambda^x$, где число $\lambda \neq 0$ подлежит определению. После подстановки этого представления в исходное уравнение и сокращения на λ^x получаем характеристическое уравнение

$$\lambda^k + a_1\lambda^{k-1} + ... + a_k = 0. \tag{2.59}$$

Если уравнение (2.59) имеет k различных корней $\lambda_1, ..., \lambda_k$, то однородное уравнение (2.52), очевидно, имеет k различных решений

$$y_i(x) = \lambda_i^x, \;\; i = 1, 2, ..., k. \tag{2.60}$$

Составленный из этих решений определитель равен

$$D[\lambda_1^x, ..., \lambda_k^x] = (\lambda_1\lambda_2 \cdots \lambda_k)^x V_k, \tag{2.61}$$

где V_k есть определитель Вандермонда:

$$V_k = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_1^{k-1} & \lambda_2^{k-1} & \cdots & \lambda_k^{k-1} \end{vmatrix} = \prod_{i>j}(\lambda_i - \lambda_j), \tag{2.62}$$

который в наших предположениях не равен нулю. Очевидно, что свободный член в уравнении (2.59) равен

$$a_k = (-1)^k \lambda_1\lambda_2 \cdots \lambda_k$$

и не равен нулю, поскольку в противном случае исходное уравнение (2.52) имело бы порядок $k-1$ (для этого достаточно было бы заменить в нем аргумент x на $x-1$). Следовательно, определитель (2.61) не равен нулю, а решения (2.60) являются линейно независимыми.

Отметим, что вещественное характеристическое уравнение (2.59) может иметь простые корни как вещественные, так и комплексные. Если в последнем случае мы хотим определить все же действительные решения однородного уравнения (2.52), то необходимо воспользоваться тем фактом, что комплексные корни встречаются только сопряженными парами. Пусть такая пара имеет вид

$$\lambda_p = \rho(\cos\omega + i\sin\omega),\ \lambda_q = \rho(\cos\omega - i\sin\omega),$$

где ρ и ω — вещественные модуль и аргумент данной пары. Считая соответствующие постоянные C_p и C_q в представлении общего решения (2.55) комплексными, сумму $C_p\lambda_p^x + C_q\lambda_q^x$ мы всегда можем преобразовать к виду

$$\bar{C}_p\rho^x\cos\omega x + \bar{C}_q\rho^x\sin\omega x,$$

где $\bar{C}_p$ и $\bar{C}_q$ будут вещественными константами. Таким образом, в случае простых корней общее решение представляется линейной комбинацией из выражений вида $\lambda^x, \rho^x\cos\omega x, \rho^x\sin\omega x$.

Если среди корней характеристического уравнения есть кратные, то из выражений (2.60) нельзя составить k линейно независимых решений. Пусть корень λ_s имеет кратность m_s. Можно показать, что такому корню соответствует m_p линейно независимых решений

$$\lambda_s^x, x\lambda_s^x, ..., x^{m_{p-1}}\lambda_s^x.$$

Предположим теперь, что характеристическое уравнение (2.59) имеет корни $\lambda_1, ..., \lambda_l$, где $l \leqslant k$, а каждый из корней

имеет кратность $m_s, 1 \leqslant s \leqslant l, m_1 + ... + m_l = k$. Тогда k линейно независимых решений представляются набором функций

$$\lambda_1^x, x\lambda_1^x, ..., x^{m_1-1}\lambda_1^x, ..., \lambda_l^x, x\lambda_l^x, ..., x^{m_l-1}\lambda_l^x. \tag{2.63}$$

О построении решений однородных уравнений с переменными коэффициентами следует сказать, что здесь общего вида аналитических представлений нет, за исключением частных случаев, и вопрос остается открытым, т.е. фактически необходимо прибегать только к численным методам.

Относительно аналитических решений неоднородных разностных уравнений вида (2.52) мы ограничимся только замечанием, что их построение производится методом вариации постоянных, см. [16], [47], поскольку их изучение не входит в цели данной книги. Общая идея этого подхода та же, что и в решениях неоднородных дифференциальных уравнений, рассмотренных в 1-й главе: в формуле для общего решения (2.54) величины $C_1, ..., C_k$ полагаются зависящими от аргумента x функциями, и для их нахождения решаются разностные уравнения, получаемые после подстановки (2.54) в исходное уравнение (2.52).

§ 2.4. Итерационное решение нелинейных уравнений в неявных методах

Методы решения нелинейных уравнений — обширный раздел вычислительной математики, и целью данного параграфа является только краткое изложение наиболее характерных итерационных алгоритмов, применяемых в рассматриваемых

нами приложениях. Более полное изложение теоретических и практических вопросов можно найти в [49], а также в различных книгах по методам вычислений, например, [11], [12], [61].

Применение безытерационных алгоритмов в нелинейных уравнениях — это сугубо исключительные случаи, и такие "штучные" результаты мы не рассматриваем.

2.4.1. Особенности нелинейных задач при решении ОДУ. Как уже отмечалось в предыдущих параграфах, применение любого из неявных численных методов решения задачи Коши для ОДУ требует решения системы нелинейных уравнений (СНУ) на каждом временном шаге. Например, в неявном методе Эйлера (2.8) возникает необходимость для каждого n решать систему

$$g_n(y_{n+1}) \equiv y_{n+1} - h_n f(t_{n+1}, y_{n+1}) - y_n = 0, \tag{2.64}$$

а для общего k-шагового метода вида (2.10) такая СНУ записывается формально как

$$g_n(y_{n+k}) \equiv F_h(y_{n+k}, ..., y_n,\ f_{n+k}, ..., f_n) = 0, \tag{2.65}$$

где f_{n+k}, в свою очередь, зависит от неизвестного y_{n+k}.

В применении к линейному многошаговому методу (2.17) соответствующая система N-го порядка конкретизуется к виду

$$g_n(y_{n+k}) \equiv \sum_{i=0}^{k} (\alpha_i y_{n+i} - h_{n+k}\beta_i f_{n+i}) = 0. \tag{2.66}$$

И наконец, для одношаговых многостадийных неявных методов Рунге—Кутты в форме (2.19) при $j \leqslant d \leqslant m$ нелинейные системы возникают на каждой j-й стадии при вычислении коэффициентов $k_j,\ j = 1, ..., m$:

$$g_{n,j} \equiv k_j - h_n f(t_n + c_j h_n, y_n + \sum_{l=1}^{d} \alpha_{j,l} k_l) = 0, \qquad (2.67)$$

где у искомых величин k_j индекс n опущен ради краткости.

Все приведенные уравнения мы будем рассматривать в единообразной форме

$$g(y) = \{g_i(y),\ i = 1, ..., N\} = 0,\ \ y \in \mathbb{R}^N, \qquad (2.68)$$

не использующей индексов номера временного шага и номера стадии.

Проблемы численного решения СНУ, в отличие от невырожденных систем линейных алгебраических уравнений, включают вопросы выбора начального приближения y^0 (от которого может зависеть сходимость или расходимость итераций), а также возможности неединственности решений.

Существенной особенностью наших задач, значительно облегчающей ситуацию, является то, что при решении СНУ на n-м временном шаге мы имеем уже решение с предыдущего шага, которое ненамного отличается от искомого и является хорошим начальным приближением. При этом вполне естественно допустить, что в его окрестности неизвестное решение является единственным. Более того, поскольку шаг h_n численного интегрирования должен быть относительно малым, то и оператор решаемой нелинейной системы должен быть в определенном смысле близок к единичному.

В силу этого количество “нелинейных” итераций предполагается небольшим, хотя вопрос их сокращения даже в несколько раз является актуальным с точки зрения ресурсоемкости решения всей системы ОДУ.

Вопрос о критериях окончания итераций при решении СНУ тоже является существенным, поскольку "недоитерированность" означает внесение дополнительного возмущения в неявную схему, что может сказаться как на погрешности аппроксимации, так и на устойчивости численного решения. Однако при решении задач Коши данная проблема зачастую упрощается и решается "волевым образом", как это делается в алгоритмах предиктор — корректор и в методах типа Розенброка.

В заключение данного пункта сделаем еще следующее замечание: в силу указанной выше специфики рассматриваемых СНУ свойства описываемых ниже алгоритмов мы исследуем только в локальном смысле, т. е. в окрестности начального приближения y^0 предполагается существование единственного решения y системы нелинейных уравнений вида (2.68), и вопросы сходимости (или расходимости) итерационных приближений y^n изучаются исключительно в аспекте $y^n \to y$.

Для одного нелинейного уравнения $g(y) = 0$, т. е. $N = 1$ в (2.68), в данном случае имеется заманчивый по своей прозрачности *метод бисекции*, заключающийся в следующем.

Пусть в окрестности искомого (единственного) корня y имеется два приближения $\hat{y}^k, \check{y}^k$, для которых непрерывная функция $g(y)$ имеет разные знаки, т. е. $g(\hat{y}^k)g(\check{y}^k) < 0$. Разделим отрезок $[\hat{y}^k, \check{y}^k]$ пополам и вычислим среднюю точку $y^{k+1} = (\hat{y}^k + \check{y}^k)/2$.

Далее определим

$$\hat{y}^{k+1} = y^{k+1}, \check{y}^{k+1} = \check{y}^k \text{ или } \hat{y}^{k+1} = \hat{y}^k, \check{y}^{k+1} = y^{k+1}$$

таким образом, чтобы выполнялись двусторонние приближения

$$g(\hat{y}^{k+1})g(\check{y}^{k+1}) < 0.$$

Очевидно, что такие последовательные приближения будут сходиться к искомому корню

$$y \in [\hat{y}^{k+1}, \check{y}^{k+1}],$$

причем длины "доверительных" интервалов $\rho(\hat{y}^{k+1}, \check{y}^{k+1})$, которым гарантировано принадлежит неизвестное решение, убывают по геометрической прогрессии:

$$\rho(\hat{y}^{k+1}, \check{y}^{k+1}) \leqslant 2^{-1}\rho(\hat{y}^{k}, \check{y}^{k}) \leqslant 2^{-(k+1)}\rho(\hat{y}^{0}, \check{y}^{0}).$$

Таким образом, данный алгоритм обеспечивает не только гарантированную достаточно быструю скорость сходимости итераций, при минимальных требованиях к свойствам функции $g(y)$, но и апостериорную оценку погрешности приближений:

$$\rho(y, y^{k+1}) \leqslant \rho(\hat{y}^{k}, \check{y}^{k}).$$

К сожалению, для СНУ высокого порядка N эффективность этого метода резко падает и он практически не применяется.

2.4.2. Метод простой итерации. Перепишем систему из N уравнений (2.68) в виде

$$y = F(y), \;\; y \in \mathbb{R}^N, \tag{2.69}$$

где в качестве векторной функции F можно взять, например, $F(y) = y - g(y)$.

Определим в $\mathbb{R}^N$ полное метрическое пространство M с расстоянием $\rho(y_1, y_2) \geqslant 0,\ y_1, y_2 \in M$. Отображение $F\colon M \to M$ называется *сжимающим*, если для некоторого $0 \leqslant q < 1$

$$\rho(F(y_1), F(y_2)) \leqslant q\rho(y_1, y_2);\ y_1, y_2 \in M. \tag{2.70}$$

Решение y уравнения $y = F(y)$ называется *неподвижной точкой* отображения F.

Выбирая некоторый начальный вектор y^0, определим итерационный процесс

$$y^{k+1} = F(y^k),\ \ k = 0, 1, ..., \tag{2.71}$$

который называется *методом простой итерации*.

Теорема 2.8. *Пусть отображение F в* (2.69) *является сжимающим. Тогда уравнение* (2.69) *имеет единственное решение, и для любого начального приближения $y^0 \in M$ метод простой итерации сходится со скоростью геометрической прогрессии, т. е. для погрешности итерационного решения выполняется неравенство*

$$\rho(y^k, y) \leqslant \frac{q^k}{1-q}\rho(y^1, y^0). \tag{2.72}$$

Если $y = (y_1, ..., y_N)^T$ и отображение $F(y) = (F_1(y), ..., F_N(y))^T$ является непрерывно дифференцируемым в окрестности единственной неподвижной точки $y = F(y)$, то при выборе начального приближения y^0 из данной окрестности достаточным условием сходимости метода простой итерации будет условие $s(F'(y)) \leqslant q < 1$, где s — спектральный радиус матрицы-якобиана

$$F'(y) = \left\{ \frac{\partial F_i(y)}{\partial y_j}, \;\; i,j = 1,...,N \right\},$$

и при этом будет выполняться неравенство (2.72).

Тривиальным обобщением метода простой итерации является выбор $F(y) = y - \alpha g(y)$, где α — некоторая постоянная, за счет оптимизации которой иногда удается уменьшить спектральный радиус q и таким образом ускорить итерационный процесс, записываемый в данном случае как

$$y^{k+1} = y^k - \alpha g(y^k). \tag{2.73}$$

2.4.3. Итерационный метод Ньютона—Канторовича. В качестве естественного обобщения алгоритма (2.73) напрашивается в этой формуле взять переменные значения итерационных параметров α_k вместо постоянного α.

Наводящие соображения к такой оптимизации легко привести для случая одного нелинейного уравнения $g(y) = 0$. Очевидно, что если y^* — корень этого уравнения, то он будет также корнем уравнения

$$y = F(y) \equiv y - \varphi(y)g(y),$$

где $\varphi(y)$ — непрерывная в окрестности $y = y^*$ функция. На основе последнего равенства можно рассматривать итерационный процесс вида

$$y^{k+1} = y^k - \varphi(y^k)g(y^k), \tag{2.74}$$

который обладает тем свойством, что если последовательность y^k сходится к y^∞, то при $\varphi(y^*) \neq 0$ этот предел является искомым решением y^*.

Предположим, что y^k — достаточно хорошее приближение к искомому решению y и что $g'(y^k) \neq 0$. Тогда при ограниченности $g''(\xi^k)$ для $\xi_k \in [y, y^k]$ мы можем использовать линейный интерполяционный многочлен Эрмита (другими словами — отрезок ряда Тейлора)

$$\begin{aligned} H(y) &= g(y^k) + g'(y^k)(y - y^k), \\ g(y) &= H(y) + \psi_k, \end{aligned} \tag{2.75}$$

где погрешность данного приближения определяется как

$$\psi_k = \psi_k(y) = 0.5\, g''(\xi^k)(y - y^k)^2, \;\; \xi_k \in [y, y^k]. \tag{2.76}$$

Поскольку $g(y) = 0$, в предположении малости величины ψ_k можно записать

$$y \approx y^k - g(y^k)/g'(y^k),$$

откуда следует, что за следующее приближение естественно взять единственный корень полинома Эрмита из (2.75), т. е. полагаем $H(y^{k+1}) = 0$, что дает

$$y^{k+1} = y^k - g(y^k)/g'(y^k), \tag{2.77}$$

т. е. данные итерации соответствуют $\varphi(y^k) = (g'(y^k))^{-1}$ в (2.74).

Наиболее распространенное название этого итерационного процесса — *метод Ньютона*, но зачастую к нему добавляют фамилию Л. В.Канторовича, который провел его исследование в функциональных пространствах.

Данный итерационный процесс имеет наглядную геометрическую интерпретацию, если $y = y^*$ является действительным корнем уравнения $g(y) = 0$: значение y^{k+1} есть абсцисса точки пересечения касательной, проведенной к кривой $x = g(y)$ в точке y^k, с осью y, см. рис. 2.1.

Из левого графика на этом рисунке видно, что начальное при-

Рис. 2.1. *Геометрическая интерпретация метода Ньютона—Канторовича*

ближение целесообразно выбирать удовлетворяющим условию

$$g(y^0)g''(y^0) > 0.$$

Если при этом на интервале $[y^0, y^*]$ производные $g'(y)$ и $g''(y)$ не меняют своих знаков, то последовательность y^k монотонно сходится к искомому корню.

В случае $g(y^0)g''(y^0) < 0$, как видно из примера на рис. 2.1 (справа), итерации y^k могут расходиться, если начальное приближение не очень хорошее.

Для рассматриваемого одномерного случая ($N = 1$) скорость сходимости этого алгоритма выводится очень просто.Из соотношений (2.75) получаем следующие два равенства:

$$\begin{aligned} g(y) - H(y) &= g''(\xi^k)(y - y^k)^2, \\ H(y^{k+1}) - H(y) &= g'(y^k)(y^{k+1} - y). \end{aligned} \tag{2.78}$$

А отсюда, в силу равенств $g(y) = H(y^{k+1}) = 0$, для ошибок итерационного приближения $z_k = y - y^k$ имеем равенство

$$z_{k+1} = -\frac{g''(\xi_k)}{2\,g'(y^k)} z_k^2, \tag{2.79}$$

которое означает *квадратичную сходимость* итерационного метода Ньютона—Канторовича, если функция $g(y)$ имеет в окрестности своего единственного корня ограниченную вторую производную.

По определению, последовательность z^k сходится к z с порядком r, если

$$\limsup_{k\,\to\,\infty} \left| \frac{z_{k+1}}{z_k^r} \right| \leqslant c < +\infty.$$

При $r = 1$ сходимость называется линейной, а при $r > 1$ — сверхлинейной.

В приведенных примерах, помимо требований $g'(y^k) \neq 0$, естественно наложить условие $g'(y) \neq 0$, означающее, что y является простым корнем решаемого уравнения. В общем случае величина y называется корнем кратности m, если $g^{(j)}(y) = 0$ при $0 \leqslant j \leqslant m-1$ и $g^{(m)}(y) \neq 0$.

Метод Ньютона может сходиться и для кратного корня, но сходимость при этом не обязана быть квадратичной. Например, для $g(y) = y^2$ имеем $z_{k+1} = z_k/2$, т. е. сходимость линейная.

Данный итерационный процесс естественным образом переносится на системы нелинейных уравнений p-го порядка. Обозначая через

$$G'(y) = \left\{ \frac{\partial\, g_i(y)}{\partial\, y_j};\ \ i, j = 1, ..., N \right\} \tag{2.80}$$

матрицу-якобиан, метод Ньютона—Канторовича можно записать в аналогичной (2.76) форме:

$$y^{k+1} = y^k - (G'(y^k))^{-1} g(y^k), \;\; k = 0, 1, ... \tag{2.81}$$

Свойства сходимости этого алгоритма характеризуются следующим утверждением.

Теорема 2.9. *Пусть y есть единственное решение СНУ $g(y)$ в замкнутом шаре $\Omega = \{x \colon ||x - y||_\infty \leqslant \delta\}$. Пусть также в этом шаре существует невырожденный якобиан отображения $g : \mathbb{R}^n \to \mathbb{R}^n$, удовлетворяющий условию Липшица*

$$||G'(x) - G'(y)||_\infty \leqslant c||x - y||_\infty, \; c > 0,$$

при любых $x, y \in \Omega$. Тогда для любого начального приближения

$$y^0 \in \{x \colon ||x - y||_\infty \leqslant \varepsilon\}$$

метод Ньютона—Канторовича (2.79) *сходится и для погрешностей $z^k = y - y^k$ выполняются неравенства*

$$||z^{k+1}||_\infty \leqslant \gamma ||z^k||_\infty^2, \;\; ||z^k||_\infty \leqslant \gamma^{-1} (\gamma ||z^0||_\infty)^{2^k}, \tag{2.82}$$

где использованы обозначения

$$\gamma = c \max_{||x-y||_\infty \leqslant \delta} ||(G'(x))^{-1}||_\infty, \; 0 < \varepsilon < \min(\delta, \gamma^{-1}).$$

2.4.4. Другие итерационные алгоритмы. Рассмотренные выше методы являются наиболее распространенными в неявных схемах решения ОДУ. “Для полноты картины” мы дадим также краткое описание еще нескольких итерационных процессов для решения СНУ.

Приведенный ранее для одного уравнения вывод метода Ньютона из эрмитовой интерполяции подсказывает общий метод построения итерационных алгоритмов. Имея вычисленные значения функции $g(y^k), ..., g(y^{k-m})$ в соответствующих точках $y^k, ..., y^{k-m}$ (мы опять для простоты возвращаемся к случаю одного уравнения, т. е. $N = 1$), можно построить интерполяционный многочлен Лагранжа $L_m(y)$ степени m и взять в качестве y_{k+1} его соответствующий корень. Таким образом можно получить итерационный процесс со скоростью сходимости $r = m + 1$. Если же использовать точки $y_n, ..., y_{n-m}$, все или некоторые, как кратные узлы (в методе Ньютона линейный многочлен Эрмита строится по одному двукратному узлу), т. е. использовать еще и значения производных функции $g(y)$, то в принципе можно повысить и порядок интерполяционного полинома, и скорость сходимости итераций. Одно из главных условий успеха такой идеи — наличие хорошего способа для вычисления "подходящего" корня интерполяционного многочлена.

Вместо описанной только что *прямой интерполяции* для конструирования итераций можно применять *обратную интерполяцию*, что приводит к близким, но отличающимся алгоритмам. Для этого достаточно построить полином $P_m(g)$ степени m, интерполирующий по точкам $g^k, ..., g^{k-m}$ и значениям $y^k, ..., y^{k-m}$ обратную функцию $g^{-1}(y)$, а после этого в качестве следующего приближения взять $y^{k+1} = P_m(0)$.

И прямая, и обратная линейные интерполяции Лагранжа приводят к так называемому *методу секущих*. В этом случае по точкам y_k, y_{k-1} строится полином Лагранжа 1-го порядка

$$L(y) = g(y^{k-1})\frac{y - y^k}{y^{k-1} - y^k} + g(y^k)\frac{y - y^{k-1}}{y_k - y^{k-1}},$$

единственный корень которого определяется формулой

$$y^{k+1} = y^k - g(y^k)\frac{y^k - y^{k-1}}{g(y^k) - g(y^{k-1})}, \tag{2.83}$$

которую можно рассматривать как частный случай (2.74) при $\varphi(y^k) = (y^k - y^{k-1})/(g(y^k) - g(y^{k-1}))$. В итерациях (2.83) нетрудно видеть аналогию с методом Ньютона, поскольку последняя дробь представляет собой конечно-разностную аппроксимацию выражения $(g'(y^k))^{-1}$ из (2.76).

Взаимосвязь данных алгоритмов наглядно иллюстрирует рис. 2.2, на котором y_n^{k+1} и y_c^{k+1} (геометрически означающих точки пересечения указанных касательной и хорды с осью y) обозначают новые приближения по методу Ньютона и методу секущих, полученные по известным точкам $(y^k, g(y^k))$ и $(y^{k-1}, g(y^{k-1}))$.

Отсюда можно построить и обобщение метода секущих на многомерный случай $p > 1$: отталкиваясь от формулы метода Ньютона (2.81), в элементах якобиана $G'(y_k)$ все частные производные заменим на соответствующие конечные разности, а полученную матрицу $G'(y_k)$ будем считать аппроксимацией якобиана. В итоге получим очень похожий на (2.81) многомерный метод секущих —

$$y^{k+1} = y^k - (G'_h(y^k))^{-1} g(y^k),\ k = 0, 1, ...$$

Если же применить обратную линейную интерполяцию, то аппроксимирующий многочлен имеет вид

$$P(g) = y^{k-1}\frac{g - g(y^k)}{g(y^{k-1}) - g(y^k)} + y^k\frac{g - g(y^{k+1})}{g(y^k) - g(y^{k-1})}$$

и для него легко проверяется свойство $P(0) = y^{k+1}$, см. (2.83).

Рис. 2.2. Иллюстрация к методу Ньютона и методу секущих

Если y — искомый единственный корень и $g \in C^2$, то для метода секущих получаем соотношения

$$g(y) - L(y) = g''(\xi_k)(y - y^k)(y - y^{k-1}),$$
$$L(y^{k+1}) - L(y) = \tfrac{g(y^k)-g(y^{k-1})}{y^k-y^{k-1}}(y^{k+1} - y) = g'(\eta_k)(y^{k+1} - y),$$

откуда для погрешностей $z^k = y - y^k$ следует равенство

$$z^{k+1} = -0.5\, g''(\xi_k) z_k z_{k-1}/g'(\eta_k), \tag{2.84}$$

имеющее очевидную схожесть с (2.79). Использованные в (2.84) неизвестные величины ξ_k, η_k — это некоторые точки из минимального отрезка, включающего координаты y^k, y^{k+1} и y.

Если использовать обозначение γ из (2.82), то из равенства (2.84) получаем неравенство

$$|z^{k+1}| \leqslant \gamma |z^k| \cdot |z^{k-1}|. \tag{2.85}$$

Скорость сходимости данных итераций элегантно выражается через *числа Фибоначчи*, определяемые следующей последовательностью:

$$\begin{aligned} &\gamma_0 = \gamma_1 = 1, \\ &\gamma_k = \gamma_{k-1} + \gamma_{k-2}, \;\; k = 2, 3, ... \end{aligned} \tag{2.86}$$

Для них справедливы соотношения

$$\gamma_k = \frac{1}{\sqrt{5}} \Big(\Big(\frac{1+\sqrt{5}}{2} \Big)^{k+1} - \Big(\frac{1-\sqrt{5}}{2} \Big)^{k+1} \Big),$$

из которых при больших k получаем

$$\gamma_k = O(\beta^k), \; \beta = \frac{1+\sqrt{5}}{2} \approx 1.618. \tag{2.87}$$

Для анализа скорости сходимости метода секущих введем величины $d_k = \gamma |z^k|$ и предположим, что

$$d_0 < d < 1, \;\; d_1 < d < 1.$$

Тогда из неравенств (2.85) получаем

$$d_2 \leqslant d_1 d_0 \leqslant d^2, \;\; d_3 \leqslant d_2 d_1 \leqslant d^5, \;\; d_k \leqslant d^{\gamma_k} \approx d^{\beta^k}.$$

А поскольку $\beta < 2$, то скорость сходимости оказывается ниже квадратичной, которая имеется в методе Ньютона (как видно из (2.82), для него поведение ошибки имеет характер d^{2^k}).

Однако сравнение формул реализации этих двух алгоритмов показывает значительный выигрыш в экономичности каждой итерации метода секущих, поскольку он требует вычисления только функции $g(y_k)$ и не требует нахождения якобиана $G'(y_k)$.

Существуют различные компромисные модификации методов типа Ньютона, в которых пытаются сократить объем вычислений без существенного уменьшения скорости сходимости итераций. Такие приемы основываются, например, на

расчетах якобианов не на каждой итерации, но на этих вопросах мы останавливаться не будем.

Интересный теоретически и важный практически вопрос — как и при каких условиях можно построить итерационные процессы высших порядков? В 1838 г. П. Л.Чебышев предложил метод отыскания действительных корней уравнения $g(y) = 0$, основанный на представлении функции, обратной к функции $g(y)$, с помощью отрезка ряда Тейлора. В некотором смысле обобщением данного подхода является итерационный *метод Кенига*, основанный на следующем утверждении.

Теорема 2.10 (Кенига). *Если $g(y)$ и $\varphi(y)$ — аналитические функции в области $|y| < R < \infty$, содержащей единственный корень $y = y^*$ уравнения $g(y) = 0$ и $\varphi(y^*) \neq 0$, то*

$$y^* = \lim_{k\to\infty} (c_k/c_{k+1}),$$

где c_k — коэффициент при y^k в разложении дроби $\varphi(y)/g(y)$ по степеням y: $\varphi(y)/g(y) = \sum\limits_{k=0}^{\infty} c_k y^k$.

Поскольку коэффициенты данного степенного ряда определяются как

$$c_k(z) = \frac{1}{k!}\Big(\frac{\varphi(y)}{g(y)}\Big)^{(k)}_{y=z},$$

то на основе теоремы Кенига можно показать, что итерационные приближения

$$y^{k+1} = y^k + (r+1)\Big(\frac{\varphi(y)}{g(y)}\Big)^{(r)}_{y=y^k} \Big/ \Big(\frac{\varphi(y)}{g(y)}\Big)^{(r+1)}_{y=y^k} \tag{2.88}$$

сходятся к y^*, если начальное приближение y^0 достаточно близко к y^*, и при этом порядок итераций (2.88) не ниже $r+2$.

Рассмотренные методы решения СНУ относятся к так называемому классу алгоритмов *одновременных смещений* (якобиевого типа), в которых на каждой итерации все неизвестные компоненты y_i перевычисляются "параллельно" и независимо друг от друга. При этом получаемые новые приближения y_i^{k+1} не зависят от способа нумерации компонент.

Альтернативой этому служат методы *последовательных смещений* (зейделевого типа), в которых при выбранной упорядоченности неизвестных вычисление y_i^{k+1} производится с использованием уже найденных на данной итерации y_j^{k+1} при $j < i$:

$$\begin{aligned} &g_i(y_1^{k+1}, ..., y_{i-1}^{k+1}, \hat{y}_i^{k+1}, y_{i+1}^k, ..., y_p^k) = 0, \\ &y_i^{k+1} = \omega \hat{y}_i^{k+1} + (1-\omega) y^k, \;\; i = 1, ..., p. \end{aligned} \tag{2.89}$$

Здесь для общности включен *релаксационный параметр* ω, значение которого выбирается вещественным в интервале от 0 до 2. При $1 < \omega < 2$ итерационный процесс называется верхней релаксацией, а при $0 < \omega < 1$ — нижней. В большинстве практических задач алгоритм (2.89) используется с $\omega > 1$ и последовательным циклическим перебором индексов $i = 1, .., p$, а сложившееся наименование для такого подхода — *последовательная верхняя релаксация* (ПВР). Для симметричных положительно определенных СЛАУ существует теория Янга—Франкела по оптимизации выбора параметра ω в ПВР, но для нелинейных систем такая законченная теория отсутствует.

Как видно из (2.89), для каждого i мы должны при вычислении y_i^{k+1} решать одно нелинейное уравнение, что может

быть сделано эффективно с помощью того же метода Ньютона. Получаемый при этом итерационный процесс называется *методом ПВР—Ньютона*, и зачастую он оказывается чрезвычайно экономичным.

Естественное обобщение "последовательного" подхода — это *блочный метод ПВР* (в том числе, возможно, с ускорением Ньютона). Такой алгоритм также описывается формулой (2.89), но при этом под $y_i,\ i = 1, ..., I$, подразумевается блок из компонент размерности p_i, причем $p_1 + ... + p_I = p$. Блочные варианты итерационных процессов, как правило, ускоряют их сходимость, но одновременно увеличивают вычислительную сложность каждой итерации, так что вопросы их оптимизации — это "штучные" темы исследований в различных практических задачах. Очевидно, что методы зейделевого типа видоизменяются при перемене нумерации компонент и соответствующей последовательности их вычислений. Более того, эти последовательности могут динамически меняться от итерации к итерации, что открывает широкие просторы для фантазии энтузиастов — разработчиков новых алгоритмов. В вычислительной линейной алгебре успешно применяются многочисленные итерационные *методы переменных направлений*, аналоги которых могут создаваться и для решения СНУ.

Блочные подходы, в том числе ньютоновского типа, разумеется, не возбраняется применять и в алгоритмах одновременных смещений. Например, вместо (2.89) можно рассмотреть *блочный метод Якоби*

$$\begin{aligned} &g_i(y_1^k, ..., y_{i-1}^k, y_i^{k+1}, y_{i+1}^k, ..., y_I^k) = 0, \\ &i = 1, ..., I,\ \ k = 0, 1, ..., \end{aligned} \tag{2.90}$$

где y_i — это подвекторы размерности p_i, так что $p_1+...+p_I = p$. На каждой k-й итерации метода (2.90) надо решать СНУ p_i-го порядка, для чего можно с успехом применять ньютоновское ускорение. Получаемый при этом итерационный процесс можно назвать *блочным методом Якоби–Ньютона*.

"Блочные" направления алгоритмов, очевидно, наиболее актуальны для решения систем высокого порядка, т. е. со значениями p в сотни, тысячи и более, где, например, уже большое значение имеют вопросы распараллеливания вычислений на многопроцессорных компьютерах.

В заключение данного пункта отметим, что методы решения СНУ можно конструировать из принципа минимизации функционалов, что дает различные обобщения градиентных и оптимизационных алгоритмов, применяемых для решения СЛАУ. Данные вопросы составляют большой самостоятельный раздел вычислительной алгебры, не нашедший (по крайней мере пока) широкого применения в неявных схемах решения ОДУ.

§ 2.5. Задачи к главе 2

2.5.1. Пусть $F \in C^1[z-\delta, z+\delta]$, где z — единственная неподвижная точка для F. Может ли метод простой итерации сходиться к z, если $|F'(z)| = 1$? Может ли он расходиться в этом случае?

2.5.2. Выяснить сходимость метода простой итерации при различных начальных приближениях для следующих уравнений:

$$x = e^{2x} - 1;\ x + l\,nx = \frac{1}{2};\ x = t\,g\,x.$$

2.5.3. Доказать, что если A, B — перестановочные с.п.о. матрицы, то матрица $B\,A$ также положительно определена.

2.5.4. Функция $f \in C^{p+1}$ имеет изолированный нуль z кратности p. Рассмотрите итерационный процесс

$$x_{k+1} = x_k - p\frac{f(x_k)}{f'(x_k)}$$

и докажите, что если он сходится к z, то сходимость квадратичная и для погрешностей $e_k \equiv z - x_k$ справедливо предельное соотношение

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^2} = \frac{f^{(p+1)}(z)}{p(p+1)f^{(p)}(z)}.$$

2.5.5. Используя метод Ньютона для решения уравнения $y^2 - a = 0$, построить итерацию второго порядка для вычисления $\sqrt{a}$.

2.5.6. Показать, что функция

$$f(y) = y\frac{(m-1)y^m + (m+1)a}{(m+1)y^m + (m-1)a}$$

определяет итерации $y^{k+1} = f(y^k)$ третьего порядка для вычисления величины $\sqrt[m]{a}$ (воспользоваться теоремой Кенига с $g(y) = y^m - a$ и $\varphi(y) \equiv 1$).

2.5.7. Показать, что для нормальной матрицы A справедливо равенство $||A||_2 = \rho(A)$, где ρ — спектральный радиус.

2.5.8. Доказать неулучшаемость оценки (2.50) в норме $||\cdot||_2$.

2.5.9. Показать, что МРК инвариантны относительно линейных преобразований $y = B\,z$, т.е. применение методов к уравнениям $y' = f(t,y)$ и $z' = B^{-1}(t, B\,z)$ с начальными значениями $y_0 = B\,z_0$ дает соотношения $y_1 = B\,z_1$.

2.5.10. Показать, что для неотрицательно определенной симметричной матрицы $A = \{a_{i,j}\}$ справедливо неравенство

$$|a_{i,j}| \leqslant \sqrt{a_{i,i}a_{j,j}}.$$

2.5.11. Доказать, что симметричная матрица $A = \{a_{i,j}\} \in \mathbb{R}^{N,N}$ является неотрицательно определенной, если и только если

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}(u_i, u_j \geqslant 0) \text{ для всех } u_i \in \mathbb{R}^N.$$

2.5.12. Показать, что для невырожденной матрицы $A \in \mathbb{R}^{N,N}$ в евклидовой норме выполняется неравенство

$$||A^{-1}|| \leqslant ||A||^{N-1}/|\det A|.$$

Указание: используйте сингулярное разложение матарицы $A = \bigcup^T \Lambda\, V$, где U и V — ортогональные матрицы, а $\Lambda = \mathrm{diag}(\sigma_1, ..., \sigma_N), \sigma_1 \geqslant ... \geqslant \sigma_N > 0$.

2.5.13. Доказать, что если матрица A вещественна и кососимметрична, то матрица $(I + A)^{-1}(I - A)$ ортогональна.

2.5.14. Показать, что для произвольных матриц $A, B \in \mathbb{R}^{N,N}$ спектры $A\,B$ и $B\,A$ совпадают.

2.5.15. Доказать, что если A есть с.п.о. матрица, а B — симметричная матрица, то все собственные числа матрицы $A\,B$ вещественные.

2.5.16. Показать, что для экстремальных собственных значений симметричной матрицы $A = \{a_{i,j}\}$ справедливы оценки

$$\lambda_{max} \geqslant \max_i\{a_{i,i}\},\ \lambda_{min} \leqslant \min_i\{a_{i,i}\}.$$

2.5.17. Показать, что в с.п.о. матрице максимальный по модулю элемент стоит на главной диагонали.

2.5.18. Показать, что оценка погрешности СЛАУ (см. 2.50)

$$||\delta v||/||v|| \leqslant ||A|| \cdot ||A^{-1}|| \cdot ||\delta f||/||f||$$

является неулучшаемой.

Глава 3

МЕТОДЫ РУНГЕ–КУТТЫ (МРК)

В данной главе мы рассмотрим одношаговые многостадийные методы, носящие имена первооткрывателей Рунге и Кутты (Carl Runge, 1856—1927, Wilhelm Kutta, 1867—1944 гг.), основополагающие работы которых относятся к 1895 и 1901 гг. соответственно. Последовательно будут исследованы вопросы аппроксимации, устойчивости и сходимости численных решений для явных и неявных алгоритмов, а также примыкающих к ним методов типа Розенброка. Для жестких и нелинейных задач исследования базируются на свойствах уравнений с односторонним условием Липшица, а также на понятиях контрактивности и B-устойчивости. Помимо общих аспектов МРК, мы рассмотрим особенности их применения для актуальных ОДУ специального вида, в том числе гамильтоновых систем, дифференциальных уравнений 2-го порядка и сингулярно возмущенных задач, а также подходы к контролю точности, устойчивости и выбору шагов интегрирования.

§ 3.1. Общая схема и классификация m-стадийных МРК

Методы Рунге—Кутты — это m-стадийные одношаговые

схемы вида

$$y_{n+1} = y_n + \sum_{j=1}^{m} b_j k_{j,n},$$
$$k_{j,n} = h_n f(t_n + c_j h_n, y_n + \sum_{l=1}^{d} a_{j,l} k_{l,n}),\ j = 1, ..., m, \tag{3.1}$$

где m фактически означает, сколько промежуточных точек используется при реализации одного шага численного интегрирования.

Величины $b_j, c_j, a_{j,l}$ суть вещественные параметры, а целое m определяет число стадий метода и является мерой его сложности, так как количество вычислений f на шаге больше или равно m. При $d = j - 1$ методы называются явными, при $d = j$ — полунеявными, или диагонально неявными, при $j < d \leqslant m$ — неявными, см. пояснения к формуле (2.10) в п. 2.1.3. Коэффициенты c_j и b_j называются *узлами и весами МРК*, а для их нахождения осуществляется разложение решения ОДУ в ряд Тейлора в окрестности точки (t_n, y_n):

$$y(t_n + h) = y(t_n) + h\, y'(t_n) + \frac{h^2}{2!} y''(t_n) + ... = y(t_n) + h\, f(t_n, y_n) + \frac{h^2}{2}(f_t +$$
$$+\tfrac{h^3}{6}(f_{t\,t} + 2\, f_{t\,y} f + f_{y\,y} f^2 + f_y f_t + f_y^2 f)(t_n, y_n) + ... \tag{3.2}$$

Для вывода членов с большими степенями h необходимо выражать старшие производные решения $\dfrac{d^k y}{d\, t^k}$ через частные производные правой части, используя условие удовлетворения $y(t)$ исходному дифференциальному уравнению. Получаемые при этом громоздкие формулы можно записать, например, с использованием следующих компактных соотношений:

$$y'' = A_1(f),\ \ y''' = A_1[(A_1 f)] = A_2(f) + A_1(f)\frac{\partial f}{\partial y}, y^{(4)} = A_3(f) + 3\,A_1(f)A_1\Big(\frac{\partial f}{\partial y}\Big) + A_2(f)\frac{\partial f}{\partial y} + A_1(f)\Big(\frac{\partial f}{\partial y}\Big)^2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \tag{3.3}$$

Здесь для удобства обозначений введены операторы

$$A_m(u) = \Big(\frac{\partial}{\partial t} + f\frac{\partial}{\partial y}\Big)^m u = \sum_{k=0}^{m} C_m^k f^k \frac{\partial^m u}{\partial t^{m-k}\partial y^k}.$$

Количество априори неизвестных узлов c_j, весовых множителей b_j и коэффициентов $a_{j,l}$ определяется длиной используемого отрезка ряда Тейлора в (3.2), а их значения ищутся из условий порядка аппроксимации исходных ОДУ и устойчивости получаемых схем.

Члены сумм в правой части соотношения (3.2), получаемые из расписывания выражений для $y^{(k)}|_{h=0}$, называются элементарными дифференциалами порядка k, а их количество равно $(k-1)!$. Уравнения для определения числовых параметров метода выписывают, приравнивая коэффициенты, стоящие при одинаковых элементарных дифференциалах в выражениях для $[y(t_n+h)]^{(k)}|_{h=0}$ и для тейлоровского разложения численного решения $(y_{n+1})^{(k)}|_{h=0}$.

Таким образом, теоретически процесс получения уравнений для коэффициентов методов весьма прост, но на практике с ростом k он скоро приводит к очень громоздким формулам. Для решения этой проблемы разработаны специальные технологии B-рядов и P-рядов с графическим представлением элементарных дифференциалов, см. [64], [65], но на этих деталях ниже мы останавливаться не будем.

Дело в том, что данные вопросы практически решены не только в плане уже имеющихся опубликованных схем самых различных порядков, построенных “вручную”, но и в смысле создания программных систем по автоматическому выводу соответствующих формул (см. Тыглиян А. В., Филиппов С. С. “Элементарные дифференциалы, их графы и коды”. Математическое моделирование, т. 21, N 8, 2009, 37—43).

3.1. 3.2. Явные МРК

Схематическое представление методов Рунге—Кутты принято наглядно изображать с помощью *таблиц Батчера*, позволяющих давать наглядную запись вычислительной схемы через ее числовые параметры. Особенность их для явных МРК заключается в том, что подматрица коэффициентов $a_{j,l}$ является строго нижней треугольной, см. табл. 3.1.

Общее представление явного одношагового метода можно записать как

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n), \tag{3.4}$$

где Φ — некоторая выражаемая через f функция (для явного метода Эйлера имеем $\Phi = f(t_n, y_n)$).

Табл. 3.1. Схема (таблица) Батчера для явного m-стадийного МРК

c_2	$a_{2,1}$				
c_3	$a_{3,1}$	$a_{3,2}$			
...	...	...	...	...	
c_m	$a_{m,1}$	$a_{m,2}$	...	$a_{m,m-1}$	
	b_1	b_2	...	b_{m-1}	b_m

Формулу явного МРК можно записать в виде

$$\begin{aligned} &F_h(y_{n+1}, y_{n,m-1}, ..., y_n) = y_{n+1} - y_n - \sum_{j=1}^{m} b_j k_{j,n} = 0, \\ &y_{n,j} = y_n + \sum_{l=1}^{j-1} a_{j,l} k_{l,n}, \;\; k_{j,n} = h_n f(t_{n,j}, y_{n,j}), \\ &t_{n,j} = t_n + c_j h_n, \;\; j = 1, ..., m, \end{aligned} \tag{3.5}$$

причем здесь $t_{n+1} = t_{n,m}$ и обычно неформально полагается (хотя это и не обязательно)

$$y_{n,j} \approx y(t_{n,j}) = y_n + \int\limits_{t_n}^{t_{n,j}} f(t, y(t)) d\, t. \tag{3.6}$$

Это соответствует тому, что все значения $y_{n,j}$, по которым вычисляются правые части в МРК, являются приближениями к решению в точках $t_{n,j}$.

Погрешность аппроксимации (*локальная погрешность*) МРК определяется как

$$\psi = F_h(y(t_{n+1}), y(t_{n,m-1}), ..., y(t_n)). \tag{3.7}$$

Определение 3.1. *МРК имеет локальную погрешность порядка* γ*, если в (3.7)*

$$||\psi|| \leqslant Ch^{\gamma+1},$$

где константа C *не зависит от* h *(величина* C *характеризуется формулой МРК и свойствами гладкости решения ОДУ).*

Погрешность МРК, или локальную ошибку, можно записать в следующем виде (без потери общности считаем $h_n = h$):

$$\psi_n = y(t_{n+1}) - y(t_n) - h\Phi(t_n, y(t_n)), \;\; \Phi(t_n, y(t_n)) = \sum_{j=1}^{m} b_j h\, f(t_n + c_j h, y_{n,j} \tag{3.8}$$

Раскладывая здесь все величины в ряд Тейлора в окрестности точки t_n, группируя члены с одинаковыми степенями по h и с одинаковыми типами частных производных f, получаем для нахождения неопределенных коэффициентов $a_{j,l}, b_j, c_j$ системы нелинейных алгебраических уравнений (СНАУ), называемыми также условиями согласованности МРК. Эти СНАУ могут быть как переопределенными, так и недоопределенными.

Выбор искомых коэффициентов производится так, чтобы функция $\psi_n(h)$ имела возможно большой порядок при произвольных t и $f(t,y)$, т. е. выполнялись условия

$$\psi_n(0) = \psi_n'(0) = ... = \psi_n^{(\gamma)}(0) = 0, \;\; \psi_n^{(\gamma+1)} \neq 0,$$

где все указанные производные предполагаются существующими и непрерывными. При этом погрешность аппроксимации будет равна остаточному члену отрезка ряда Тейлора

$$\psi_n(h) = \frac{h^{\gamma+1}}{(\gamma+1)!}\psi_n^{(\gamma+1)}(\xi), \;\; 0 \leqslant \xi \leqslant h. \tag{3.9}$$

Для получения разложения численного решения в ряд Тейлора достаточно рассматривать y_{n+1} как дифференцируемую функцию от переменной h с условием $y_{n+1}(h)|_{h=0} = y_n$:

$$y_{n+1}(h) = y_n + h\frac{d\,y_{n+1}}{d\,h}(0) + ... + \frac{h^\gamma}{\gamma!}\,\frac{d^\gamma y_{n+1}}{d\,h}(0) + O(h^{\gamma+1}).$$

Уравнения согласованности МРК получаются достаточно громоздкими, и построению их решений с теми или иными свойствами искомых коэффициентов посвящено огромное количество исследований многочисленных специалистов.

Приведем один характерный пример. Трехстадийный явный метод Рунге—Кутты имеет шесть параметров $a_{2,1}, a_{3,1}, a_{3,2}, b_1, b_2, b_3$ и будет согласован с третьим порядком, если данные параметры удовлетворяют уравнениям

$$\begin{aligned} &b_1 + b_2 + b_3 = 1, \qquad 2(c_2 b_2 + c_3 b_3) = 1, \\ &6\,c_2 a_{3,2} b_3 = 1, \qquad 3(c_2^2 b_2 + c_3^2 b_3) = 1. \end{aligned}$$

Общее решение этой системы можно записать в виде

$$b_3 = \frac{2 - 3\,c_2}{6\,c_3(c_3 - c_2)}, \qquad b_2 = \frac{1 - 2\,c_3 b_3}{c_2},$$
$$b_1 = 1 - b_2 - b_3, \qquad a_{3,2} = (c_2 b_3)^{-1},$$

где параметры c_2 и c_3 — свободные.

Для явных МРК ради простоты нахождения коэффициентов полагается

$$c_j = \sum_{l=1}^{j-1} a_{j,l}, \quad j = 2, ..., m, \tag{3.10}$$

что соответствует условию (3.6). В результате m-стадийный метод имеет $m(m+1)/2$ параметров $a_{j,l}$ и b_j. Для того чтобы метод имел порядок γ, необходимо удовлетворить $r(\gamma)$ уравнениям согласованности, причем их число быстро растет с увеличением γ, см. таблицу 3.2. Чтобы получаемые уравнения были разрешимы, должно выполняться неравенство

$$m(m+1)/2 \geqslant r(\gamma). \tag{3.11}$$

Табл. 3.2. Зависимость числа параметров и числа уравнений согласованности от γ, $m \leqslant 9$

m	$m(m+1)/2$	γ	$r(\gamma)$
1	1	1	1
2	3	2	2
3	6	3	4
4	10	4	8
5	15	5	17
6	21	6	37
7	28	7	85
8	36	8	200
9	45	9	486

Из соотношения (3.11) и из таблицы видно, что для числа стадий $m \leqslant 4$ достижим порядок $\gamma = m$, однако для $m \geqslant 5$ — уже нет. Так, пятый порядок может иметь методы с $m \geqslant 6$, а шестой порядок — с $m \geqslant 9$.

Погрешность аппроксимации ψ_n можно связать с глобальной ошибкой $z_{n+1} = y(t_{n+1}) - y_{n+1}$, если вычесть почленно соотношения (3.8) и (3.4):

$$z_{n+1} = z_n + h[\Phi(t_n, y(t_n)) - \Phi(t_n, y_n)] + \psi_n,$$

а далее к выражению в квадратных скобках применить теорему Лагранжа о среднем —

$$\Phi(t_n, y(t_n)) - \Phi(t_n, y_n) = \frac{\partial \Phi(t_n, \xi_n)}{\partial y} z_n, \ \xi_n \in [y(t_n), y_n].$$

Отсюда получаем

$$z_{n+1} = B_n z_n + \psi_n, \;\; B_n = I + h\frac{\partial \Phi(t_n, \xi_n)}{\partial\, y}, \tag{3.12}$$

где I — единичный оператор. Из последнего равенства следуют неравенства для норм

$$||z_{n+1}|| \leqslant ||B_n||||z_n|| + ||\psi_n||, \tag{3.13}$$

которые при условии (2.24) позволяют установить оценку вида (2.25) из § 2.2. Далее следует такой важный результат, что если правая часть системы ОДУ удовлетворяет условию Липшица, то явные методы Рунге—Кутты являются устойчивыми, в смысле определения 2.5, т. е. норма оператора в (3.13) удовлетворяет неравенству $||B_n|| \leqslant 1 + h\, b$ при всех n, откуда из (2.25) следует при достаточно малых h, что порядки их сходимости совпадают с соответствующими порядками погрешности аппроксимации. Несложно также проверить и более тонкие утверждения: при $||B_n|| > 1$ старые ошибки усиливаются, при $||B_n|| = 1$ — накапливаются линейно, а при $||B_n|| < 1$ — подавляются.

Области устойчивости явных МРК (см. определение 2.13) характеризуются следующим образом. Функции устойчивости для явных МРК являются комплексными многочленами , и, следовательно, неравенство $|R(z)| < 1$ выполняется только в ограниченной области из $\mathbb{C}_-$. Таким образом, никакой явный МРК не может быть A-устойчивым и тем более L-устойчивым. Для m-стадийного явного МРК m-го порядка для $1 \leqslant m \leqslant 4$ область абсолютной устойчивости определяется неравенством

$$|R_m(z)| = \Big| \sum_{i=0}^{m} \frac{z^i}{i!} \Big| < 1.$$

Теорема 3.1. *Если явный МРК имеет порядок γ, то*

$$R(z) = 1 + z + \frac{z^2}{2} + ... + \frac{z^\gamma}{\gamma!} + O(z^{\gamma+1}).$$

Доказательство следует из того, что $e^{\lambda y}$ является точным решением задачи

$$y' = \lambda y, \;\; y_0 = 1,$$

и из равенств

$$y_{n+1} = R(\lambda h) y_n, \;\; e^z - R(z) = O(z^{\gamma+1}).$$

□

В m-стадийном явном МРК выбор параметров можно проводить не только из условия минимальной локальной ошибки, но и из расширения области устойчивости. Например, если требуется для больших значений аргумента t решить ОДУ с достаточно гладким решением, т.е. с малыми значениями производных, то из соображений экономичности метода можно "пожертвовать" порядком аппроксимации за счет возможности использования более крупного шага h_n при обеспечении устойчивости и требуемой точности численного решения.

Приведем некоторые примеры явных МРК двухстадийный метод Рунге 2-го порядка:

$$m = 2, \gamma = 2, \;\; c_2 = a_{2,1} = 1/2, \;\; b_1 = 0, \;\; b_2 = 1,$$

для которого алгоритм имеет вид

$$y_{n+1} = y_n + h\, f(t_n + h/2, y_n + k_1/2), \;\; k_1 = h\, f(t_n, y_n). \tag{3.14}$$

Очевидно, что данный алгоритм можно интерпретировать как результат применения квадратурной формулы центральных прямоугольников к вычислению интеграла в (3.6) при $j = m = 2$:

$$\int\limits_{t_n}^{t_{n+1}} f(t, y(t))\, d\, t = h\, f(t_{n+1/2}, y_{n+1/2}) + O(h^3).$$

Здесь, в свою очередь, $y_{n+1/2}$ обозначает приближение к значению $y(t_{n+1/2})$, т. е.

$$y_{n+1/2} \equiv y_n + hf(t_n, y_n)/2 = y(t_{n+1/2}) + O(h^2).$$

Четырехстадийный метод Рунге 3-го порядка:

$$\begin{gathered} m = 4, \gamma = 3, \;\; c_2 = a_{2,1} = 1/2, \;\; c_3 = a_{3,2} = c_4 = a_{1,1} = 1, \\ a_{3,1} = a_{4,1} = a_{4,2} = 0, \;\; b_1 = 1/6, \;\; b_2 = 2/3, \;\; b_3 = 0, \;\; b_4 = 1/6. \end{gathered} \tag{3.15}$$

Как следует из табл. 3.2, третий порядок можно получить и при $m = 3$ (даже не единственным образом, поскольку в данном случае для нахождения 6 неизвестных коэффициентов требуется удовлетворить только четырем уравнениям). В частности, хорошими свойствами обладает трехстадийный метод Хойна 3-го порядка:

$$m = \gamma = 3, \;\; c_2 = a_{2,1} = 1/3, \;\; c_3 = a_{3,1} = 2/3,$$

$$a_{3,1} = 0,\ \ b_1 = 1/4,\ \ b_2 = 0,\ \ b_3 = 3/4,$$

со схемой реализации

$$\begin{aligned} &y_{n+1} = y_n + (k_1 + 3\,k_3)/4,\ \ k_1 = h\,f(t_n, y_n),\\ &k_2 = h\,f(t_n + h/3, y_n + k_1/3),\ \ k_3 = h\,f(t_n + 2\,h/3, y_n + 2\,k_2/3). \end{aligned} \tag{3.16}$$

Условия согласования 4-стадийного МРК включают следующие 8 уравнений:

$$\begin{aligned} &\sum_i b_i = 1,\ \ \sum_i b_i c_i = 1/2,\ \ \sum_i b_i c_i^2 = 1/3,\\ &b_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = 1/6,\ \ b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = 1/4,\\ &b_3 c_3 a_{32} c_2 + b_4 c_4 (a_{42} c_2 + a_{43} c_3) = 1/8,\\ &b_3 a_{32} c_2^2 + b_4 (a_{42} c_2^2 + a_{43} c_3^2) = 1/12,\ \ b_4 a_{43} a_{32} c_2 = 1/24. \end{aligned} \tag{3.17}$$

Существует бесчисленное множество решений этой системы, т. е. значений b_j и c_j (веса и узлы МРК), а также коэффициентов $a_{j,l}$. Их значения выбираются из условий устойчивости (см. примеры ниже), а также уменьшения множителя $\psi_n^{(\gamma+1)}$ в главном члене погрешности $O(h^5)$.

Приведем схемы Батчера для двух разработанных Куттой методов четвертого порядка, из которых один наиболее популярен (“классический” МРК), а второй отличается повышенной точностью (“правило 3/8”), см. Табл. 3.3 и 3.4.

Табл. 3.3. Схема явного “классического” МРК

Для первого из них формулы имеют вид

1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	2/6	2/6	1/6

$$y_{n+1} = y_n + (k_1 + 2\,k_2 + 2\,k_3 + k_4)/6,$$

$$k_1 = h\,f(t_n, y_n), \;\; k_2 = h\,f(t_n + h/2, y_n + k_1/2),$$

$$k_3 = h\,f(t_n + h/2, y_n + k_3/2), \;\; k_4 = h\,f(t_n + h, y_n + k_3), \tag{3.18}$$

а для второго —

$$y_{n+1} = y_n + (k_1 + 3\,k_2 + 3\,k_3 + k_4)/8,$$

$$k_1 = h\,f(t_n, y_n), \;\; k_2 = h\,f(t_n + h/3, y_n + k_1/3),$$

$$k_3 = h\,f(t_n + 2\,h/3, y_n + k_2 - k_1/3), \;\; k_4 = h\,f(t_n + h, y_n + k_1 - k_2 + k_3). \tag{3.19}$$

Табл. 3.4. Схема метода Кутты 4-го порядка по “правилу 3/8”

Напомним, что при $m = 5$ повысить порядок ψ_h до $O(h^6)$ невозможно, в силу чего пятистадийные МРК не используются на практике. Методы шестого и более высоких

1/3	1/3			
2/3	−1/3	1		
1	1	−1	1	
	1/8	3/8	3/8	1/8

порядков (например, десятого) существуют, но по причине их громоздкости применяются только в исключительных случаях.

§ 3.3. Неявные МРК (НМРК)

Схема Батчера для неявных m-стадийных МРК выражается табл. 3.5, главное отличие которой от табл. 3.1 для явных МРК заключается в том, что элементы $a_{j,l}$ уже не составляют треугольную матрицу.

Соответствующие формулы МРК имеют вид

$$\begin{aligned} k_j &= f(t_n + c_j h, y_n + \sum_{l=1}^{m} a_{j,l} k_l), \quad j = 1, ..., m, \\ y_{n+1} &= y_n + h \sum_{j=1}^{m} b_j k_j. \end{aligned} \tag{3.20}$$

Табл. 3.5. Схема Батчера для неявных МРК

Для нахождения величин $k_j, j = 1, ..., m$, при этом на каж-

c_1	$a_{1,1}$	$a_{1,2}$	$\dots$	$a_{1,m}$
c_2	$a_{2,1}$	$a_{2,2}$	$\dots$	$a_{2,m}$
$\dots$	$\dots$	$\dots$	$\dots$	
c_m	$a_{m,1}$	$a_{m,2}$	$\dots$	$a_{m,m}$
	b_1	b_2	$\dots$	b_m

дом шаге приходится решать нелинейную систему уравнений размерности, в общем случае равной mN. Для полунеявных (диагонально-неявных) МРК значения k_j находятся последовательно по стадиям, т. е. реализация одного шага сводится к решению m нелинейных систем порядка N.

3.3.1. Вопросы существования, единственности и примеры НМРК

Теорема 3.2 (1-я теорема Батчера). *Пусть f — функция, непрерывная и удовлетворяющая условию Липшица с постоянной L в некоторой окрестности точки t_n. Тогда при выполнении условия*

$$h_n < L\Big(\max_j |\sum_l a_{j,l}|\Big)^{-1}$$

система уравнений НМРК имеет единственное решение, которое может быть получено итерированием. Если функция $f(t,y)$ p раз непрерывно дифференцируема, то все k_j, как и функции h, также принадлежат множеству C^p.

Доказательство существования единственного решения следует из сходимости последовательных приближений k_j^s

в итерационном процессе Якоби $k_j^{s+1} = f(t_n + c_j h, y_n + h\sum_{j=l}^{m} a_{j,l} k_j^s)$, которое в условиях теоремы имеет место при достаточно малом h.

Дифференцируемость величин k_j следует из теоремы о неявных функциях.

Порядок погрешности неявного МРК определяется так же, как и для явного.

Теорема 3.3 (2-я теорема Батчера). *Если справедливы равенства*

$$C(\eta): \quad \sum_{l=1}^{m} a_{j,l} c_j^{q-1} = c_j^q/q, \;\; j = 1, ..., m; \;\; q = 1, ..., \eta,$$

$$D(\xi): \quad \sum_{j=1}^{m} b_j c_j^{q-1} a_{j,l} = b_l(1 - c_l^q)/q, \;\; l = 1, ..., m, \;\; q = 1, ..., \xi,$$

$$B(p): \quad \sum_{j=1}^{m} b_j c_j^{q-1} = q^{-1}, \;\; q = 1, ..., p,$$

причем выполняются условия $p \leqslant \xi + \eta + 1$, $p \leqslant 2\eta + 2$, то неявный МРК имеет порядок p.

Упрощающие условия $B(p), C(\eta), D(\xi)$ — основа алгебраической теории МРК. Соотношения $B(p)$ означают, что квадратурная формула в интегральном соотношении

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t) d\, t$$

точна, если f — полином степени не выше $p-1$. Условия $C(\eta)$ соответствуют тому, что для каждой промежуточной точки $t_{n,j} = t_n + c_j h$ выражение $\sum_{l=1}^{m} a_{j,l} f(t_n + c_l h)$ точно аппрокси-

мирует интеграл $y_{n,j} = \int\limits_{t_n}^{t_{n,j}} f(t)dt$, если f — многочлен степени меньше η. Неявные МРК, удовлетворяющие данным условиям, называются *коллокационными методами*.

Условие $D(\xi)$ для m-стадийного НМРК с ненулевыми весами $b_1, ..., b_m$ и различными узлами $c_1, ..., c_m$, как можно показать, связано следующим образом с условиями $C(\eta)$ и $B(p)$:

- из $C(m)$ и $B(m+\xi)$ следует $D(\xi)$;
- из $D(m)$ и $B(m+\eta)$ следует $C(\eta)$.

Следствием 2-й теоремы Батчера является существование порядка $p = 2m$, который получается при значениях $\eta = \xi = m$ в условиях теоремы. При этом узлы c_j и веса b_j выбираются в соответствии с квадратурными формулами Гаусса, имеющими максимальную алгебраическую точность, см. [30], для интегрального представления решения ОДУ.

Простейшие неявные МРК получаются из применения известных квадратурных формул к соотношению

$$y(t_{n+1}) = y(t_n) + \int\limits_{t_n}^{t_{n+1}} f(t, y(t))dt.$$

Отсюда семейство “весовых” методов первого порядка имеет вид (Коши, 1824)

$$y_{n+1} = y_n + h\, f(t_n + \theta\, h, y_n + \theta(y_{n+1} - y_n)),$$

из которого при $\theta = 0$ и $\theta = 1$ получаем явный и неявный алгоритмы Эйлера.

При использовании формулы центральных прямоугольников, что соответствует $\theta = 1/2$, получаем МРК с погрешностью второго порядка

$$y_{n+1} = y_n + h_n k_1, \;\; k_1 = f(t_n + h_n/2, y_n + h_n k_1/2), \qquad (3.21)$$

который называется *неявным правилом средней точки* и имеет локальную ошибку

$$\psi_n = -\frac{h_n^3}{24} f''(\xi_1), \;\; \xi_1 \in [t_n, t_{n+1}].$$

Для этого одностадийного метода второго порядка ненулевые коэффициенты таблицы Батчера равны

$$a_{1,1} = c_1 = 1/2, \;\; b_1 = 0.$$

Если же аппроксимировать интеграл по *правилу трапеций*, то имеем неявный МРК того же второго порядка

$$y_{n+1} = y_n + h_n[f(t_n, y_n) + f(t_{n+1}, y_{n+1})]/2 \qquad (3.22)$$

с погрешностью $\psi_h = \dfrac{h_n^2}{12} f''(\xi_2)$, $\xi_2 \in [t_n, t_{n+1}]$. Данный алгоритм формально является уже двухстадийным, но на каждом n-м шаге правую часть нужно вычислять только один раз.

Схемы Батчера для методов средней точки и трапеций приведены ниже, слева и справа соответственно:

$1/2$	$1/2$
	1

1	$1/2$	$1/2$
	$1/2$	$1/2$

Сравнивая ошибки двух последних алгоритмов, обнаруживаем, что они дают двусторонние приближения к точному

решению, если у функции f вторая производная не меняет знак на $[t_n, t_{n+1}]$.

Если же к интегралу применить простейшую квадратурную *формулу Радо* (левый узел совпадает с концом интервала, а второй оптимизируется в алгебраическом смысле, см. [30]), то получаем формулу

$$y_{n+1} \approx y_n + h_n[f(t_n, y_n) + 3\, f(t_n + \frac{2}{3}h_n), y(t_n + \frac{2}{3}h_n)]. \tag{3.23}$$

После аппроксимации последнего члена в правой части с помощью квадратичной интерполяции по значениям y_n, $y'_n = f_n$, y_{n+1}

$$y(t_n + \frac{2}{3}h) \approx \frac{5}{9}y_n + \frac{4}{9}y_{n+1} + \frac{2}{9}h_n f(t_n, y_n)$$

приходим к неявному двухстадийному МРК третьего порядка, предложенному Хаммером и Холлингсуортом (1955):

$$\begin{aligned} &y_{n+1} = y_n + \frac{h_n}{4}(k_1 + 3\, k_2), \quad k_1 = f(t_n, y_n), \\ &k_2 = f(t_n + \frac{2}{3}h_n), y_n + \frac{h_n}{3}(k_1 + k_2). \end{aligned} \tag{3.24}$$

Неявные двустадийные МРК порядка 3 в общем случае должны удовлетворять следующим четырем условиям для коэффициентов:

$$\begin{aligned} &b_1 + b_2 = 1, \;\; b_1 c_1 + b_2 c_2 = 1/2, \;\; b_1 c_1^2 + b_2 c_2^2 = 1/3, \\ &b_1(a_{11}c_1 + a_{12}c_2) + b_2(a_{21}c_1 + a_{22}c_2) = 1/6. \end{aligned} \tag{3.25}$$

Здесь удается построить однократно-диагонально-неявный МРК (ОДНМРК, характеризуемый одинаковыми диагональными коэффициентами $a_{1,1} = a_{2,2} = \varkappa$), предложенный

С. Нерсеттом в 1974 г. и описываемый схемой Батчера в табл. 3.6.

Табл. 3.6. Схема двухстадийного ОДНМРК

$\varkappa$	$\varkappa$	0
$1-\varkappa$	$1-2\varkappa$	$\varkappa$
	$1/2$	$1/2$

$\varkappa = (3 \pm \sqrt{3})/6.$

Коэффициенты двухстадийного неявного МРК четвертого порядка должны удовлетворять восьми нелинейным уравнениям, которые имеют единственное решение, представляемое следующей схемой Хаммера, см. табл. 3.7.

При этом узлы и веса соответствуют двухточечной квадратурной формуле Гаусса (максимального алгебраического порядка). Неявные МРК различного числа стадий m можно также получать путем аппроксимации интеграла с помощью *квадратурных формул Лобатто* (1852). В данном случае первый и последний узлы полагаются совпадающими с концами интервала $[t_n, t_{n+1}]$, т. е. $c_1 = 0, c_m = 1$, а остальные выбираются оптимальными в смысле полиномиальной точности.

Табл. 3.7. Схема двухстадийного НМРК

В результате получаются алгоритмы порядка $2m-2$. Батчер (1964 г.) обнаружил, что таким образом можно получить неявные МРК, у которых первая строка и последний столбец таблицы коэффициентов являются нулевыми. При этом первая и последняя стадии становятся явными, число неявных стадий понижается до $m-2$, а порядок погрешности остается

$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$
$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} + \frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

равным $2m-2$. Пример таблицы Батчера для коэффициентов 3-стадийного МРК приводится в табл. 3.8.

Табл. 3.8. Пример схемы 3-стадийного МРК

0	0	0	0
1/2	1/4	1/4	0
1	0	1	0
	1/6	2/3	1/6

3.3.2. Устойчивость неявных МРК. Если неявный метод Эйлера $y_{n+1} = y_n + h\, f(t_{n+1}, y_{n+1})$ применить к модельному (скалярному) уравнению Далквиста $y' = \lambda y$, то получим равенство $y_{n+1} = y_n + h\lambda y_{n+1}$. Отсюда получаем функцию устойчивости

$$y_{n+1} = R(h\lambda) y_n, \;\; R(z) = (1-z)^{-1}. \tag{3.26}$$

Таким образом, область устойчивости ($|R(z)| < 1$) в данном случае — это внешность круга радиуса 1 с центром в $+1$, т.е. алгоритм оказывается A-устойчивым.

Для неявного m-стадийного МРК общего вида

$$\begin{aligned} g_j &= y_n + h \sum_{l=1}^{m} a_{j,l} f(t_n + c_j h, g_j), \quad l = 1, ..., m, \\ y_{n+1} &= y_n + h \sum_{j=1}^{m} b_j f(t_n + c_j h, g_j) \end{aligned} \tag{3.27}$$

в применении к уравнению $y' = \lambda y$ имеем (после исключения величин g_j) $y_{n+1} = R(h\lambda)y_n$, где функция устойчивости определяется как

$$R(z) = 1 + z\, b^T (I - zA)^{-1} e, \tag{3.28}$$

где $e = (1, ..., 1)^T$, $b = (b_1, ..., b_m)^T$, $A = \{a_{l,j}\} \in \mathbb{R}^{m,m}$. Отсюда видно, что условие L-устойчивости неявных методов Рунге—Кутты имеет вид

$$R(\infty) = 1 - b^T A^{-1} e = 0,$$

что дает следующие требования на соотношения коэффициентов:

$$\sum_{j=1}^{m} b_j \sum_{l=1}^{m} \bar{a}_{j,l} = 1,$$

где $\bar{a}_{j,l}$ — элементы матрицы A^{-1}.

Поскольку последнее соотношение является не очень наглядным, можно привести более конструктивное утверждение: неявный МРК с невырожденной матрицей A является L-устойчивым при выполнении одного из следующих условий:

$$a_{m,j} = b_j, \quad j = 1, ..., m,$$

$$a_{i,1} = b_1, \quad i = 1, ..., m.$$

Первое из этих условий (оно называется жесткой точностью и играет большую роль при решении сингулярно-возмущенных задач, см. ниже § 3.12) в матрично-векторной записи означает $A^T e_m = b$, где $e_m = (0, ..., 0, 1)^T$, откуда получаем $b^T = e_m^T A$ и $R(\infty) = 1 - e_m A\, A^{-1} e = 0$.

Второе условие записывается как $A\,e_1 = e\,b_1$, где $e_1 = (1, 0, ..., 0)^T$, и доказывается аналогично: $R(\infty) = 1 - b^T A^{-1} A\,e_1 b_1^{-1} = 0$.

Как показал Х. Штеттер (1973 г.), $R(z)$ является рациональной функцией, которую можно представить в виде

$$R(z) = P(z)/Q(z),\ \ P(z) = |I - z\,A + z\,e\,b^T|,$$
$$Q(z) = |I - z\,A|,\ \ \deg P = k \leqslant m,\ \ \deg Q = s \leqslant m. \tag{3.29}$$

Если неявный МРК имеет порядок γ, то при $z \to 0$

$$e^z - R(z) = O(z^{\gamma+1}).$$

Теорема 3.4. *Неявный МРК A-устойчив тогда и только тогда, когда*

- $|R(iy)| \leqslant 1$ *при любых вещественных* y,
- $R(z)$ — *аналитическая функция при* $Re\,z < 0$.

Первое из условий означает устойчивость на всей мнимой оси (это называется *I-устойчивостью)* и эквивалентно требованию, чтобы многочлен

$$E(y) = |Q(i\,y)|^2 - |P(i\,y)|^2 = Q(i\,y)Q(-iy) - P(i\,y)P(-i\,y) \tag{3.30}$$

для всех $y \in \mathbb{R}$ удовлетворял условию $E(y) \geqslant 0$.

Теорема 3.5. *Рациональная функция устойчивости* $R(z) = P(z)/Q(z)$ *порядка* $\gamma \leqslant 2m-2$ *будет* I*-устойчива, если и только если* $|R(\infty)| \leqslant 1$.

Рациональные функции, которые для заданных степеней числителя и знаменателя имеют наивысший порядок аппроксимации некоторой функции, называются аппроксимациями Паде. Если неявный МРК имеет порядок γ, то его функция чувствительности есть *аппроксимация Паде* для экспоненты:

$$|e^z| - \left|\frac{P(z)}{Q(z)}\right| = O(z^{\gamma+1}).$$

Рациональная функция $R(z)$ называется *диагональной аппроксимацией Паде*, если степени многочленов k и s в числителе и знаменателе одинаковые (ошибка аппроксимации при этом есть $O(z^{k+s+1})$).

Пример 3.1. Для приведенного выше 2-стадийного НМРК 4-го порядка (Хаммер) функция устойчивости имеет вид

$$R(z) = \frac{1 + z/2 + z^2/12}{1 - z/2 + z^2/12} \tag{3.31}$$

и является A-допустимой, но $\lim\limits_{\mathrm{Re}\, z \to -\infty} R(z) = 1$. Данная функция аппроксимирует экспоненту с погрешностью $0(z^5)$, а соответствующий метод не является L-устойчивым.

3.3.3. Диагонально-неявные МРК. Как уже отмечалось, в данных методах на каждом шаге надо решить m

СНАУ порядка N, которые при использовании итераций ньютоновского типа требуют решать СЛАУ с матрицами вида $I - h\,a_{j,j}\frac{\partial f}{\partial y}$. Если $a_{ii} = a$ (однократно диагонально неявные МРК — ОДНМРК, см. табл. 3.9), то эти матрицы одинаковые и для них можно только один раз выполнить трудоемкое $L\,U$-разложение.

Здесь возможно упрощение условий согласования — перенос в правую часть зависящих от a членов; при этом слева останутся укороченные суммы $\sum'$ (без диагональных коэффициентов, как в явных МРК).

В итоге мы получаем следующие уравнения: обычные условия —

$$\sum b_j = 1,\ \sum b_j a_{jk} = 1/2,\ \sum b_j a_{jk} a_{kl} = 1/3, ...$$

и упрощенные условия —

$$\sum b_j = 1,\ \sum{}' b_j a_{jk} = 1/2 - a,\ \sum{}' b_j a_{jk} a_{kl} = 1/3 - a + a^2, ...$$

Функция устойчивости ДНМРК имеет вид $R(z) = \dfrac{P(z)}{(1 - a_{1,1}z)\cdots(1 - a_{m,m}z)}$, где $P(z)$ — многочлен степени не выше m.

Табл. 3.9. Схема Батчера для ОДНМРК

Для ОДНМРК ($a_{1,1} = ...a_{m,m} = a$) вид $R(z)$ упрощается:

$$R(z) = (-1)^m \sum_{j=0}^{m} L_m^{(m-j)}(a^{-1})(a\,z)^j/(1 - a\,z)^m. \qquad (3.32)$$

Здесь $L_m^{(k)}$ — k-я производная полинома Лагерра степени m, который описывается формулой

c_1	a			
c_2	$a_{2,1}$	a		
$\dots$	$\dots$	$\dots$	$\dots$	
c_m	$a_{m,1}$	$a_{m,2}$	$\dots$	a
	b_1	b_2	$\dots$	b_m

$$L_m(x) = \sum_{j=0}^{m} (-1)^j \left(\frac{m}{j}\right) \frac{x^j}{j!}.$$

Области A-устойчивости для порядков $\gamma \leqslant m$ при $m = 1, 2, 3$ представляются следующими интервалами значений a соответственно:

$$1/2 \leqslant a < \infty, \;\; 1/4 \leqslant a < \infty, \;\; 1/3 \leqslant a < 1.068. \tag{3.33}$$

§ 3.4. Методы типа Розенброка (МтР)

В применении к автономному ОДУ $y' = f(y)$ формулы НМРК можно записать в виде

$$\begin{aligned} k_j &= h\, f\Big(y_n + \sum_{l=1}^{j-1} a_{j,l} k_l + a_{j,j} k_j\Big), \;\; j = 1, ..., m, \\ y_{n+1} &= y_n + \sum_{j=1}^{m} b_j k_j. \end{aligned} \tag{3.34}$$

Если здесь линеаризовать выражение для k_j, то получим

$$k_j = h\, f(g_j) + h\, f'(g_j) a_{j,j} k_j,$$
$$g_j = y_n + \sum_{l=1}^{j-1} a_{j,l} k_l.$$

Такие соотношения используются на каждой ньютоновской итерации при решении нелинейных уравнений для k_j последовательно на стадиях $j = 1, ..., m$, полагая при этом $k_j^{(0)} = h\, f(g_j)$. Используя приведенные выражения (вместо итераций) как новый класс алгоритмов, получаем методы типа Розенброка.

Значительное сокращение трудоемкости каждого шага МтР достигается путем замены матрицы Якоби $f'(g_j)$ на $J = f'(y_n)$. При этом существенное увеличение степеней свободы, т. е. возможностей выбора неопределенных коэффициентов, допускает введение добавочных линейных комбинаций членов Jk_j, что дает следующий класс m-стадийных МтР:

$$\begin{aligned} k_j &= h[f(y_n + \sum_{l=1}^{j-1} a_{j,l} k_l) + J \sum_{l=1}^{j} \beta_{j,l} k_l], \\ y_{n+1} &= y_n + \sum_{j=1}^{m} b_j k_j. \end{aligned} \tag{3.35}$$

На каждой стадии этого метода требуется решать СЛАУ N-го порядка с неизвестным вектором $k_j = (k_{j,1}, ..., k_{j,N})^T$ и матрицей $I - h\beta_{j,j} J$. Дальнейшую экономию дают "однократные" МтР с условием $\beta_{1,1} = ... = \beta_{m,m} = \beta$, что позволяет на

каждом шаге делать только одно LU-разложение матрицы $I - h\beta J$.

В случае неавтономных ОДУ $y' = f(t, y)$ МтР можно применить к расширенной "автономизированной" системе. Тогда компоненты, соответствующие переменной t, вычисляются явно, тогда получаем формулы

$$k_j = h[f(t_n + c_j h,\ y_n + \sum_{l=1}^{j-1} a_{j,l} k_l) + \beta_j h \frac{\partial f}{\partial t}(t_n, y_n) + \\ + \frac{\partial f}{\partial y}(x_n, y_n) \sum_{l=1}^{j} \beta_{j,l} k_l],\ \ y_{n+1} = y_n + \sum_{j=1}^{m} b_j k_j, \\ c_j = \sum_{l=1}^{j-1} a_{j,l},\ \ \beta_j = \sum_{l=1}^{j} \beta_{j,l}. \tag{3.36}$$

Если решаются *неявные дифференциальные уравнения*, т. е. не разрешенные относительно производных:

$$M\, y' = f(t, y), \tag{3.37}$$

где $M \in -R^{N,N}$ — невырожденная постоянная матрица, то данную систему ОДУ можно умножить на M^{-1} и применить рассмотренные выше методы (3.34)—(3.36) к уравнению $y' = M^{-1} f(t, y)$. Умножая потом полученные соотношения на матрицу M, приходим к следующему алгоритму:

$$M\,k_j = h[f(t_n + c_j h,\ y_n + \sum_{l=1}^{j-1} a_{j,l}k_l)+$$
$$+\beta_j h\frac{\partial\, f}{\partial\, t}(t_n, y_n) + \frac{\partial\, f}{\partial\, y}(t_n, y_n)\sum_{l=1}^{j}\beta_{j,l}k_l], \qquad (3.38)$$
$$y_{n+1} = y_n + \sum_{j=1}^{m} b_j k_j.$$

Преимущество такой записи в том, что не нужно обращать M, а также в сохранении возможной ленточной формы матриц M и $\dfrac{\partial\, f}{\partial\, y}$. Поскольку трудоемкость одного шага МтР существенно выше, чем у явного МРК, применять их целесообразно только к решению жестких систем, в которых явные методы требуют сверхмалых шагов сетки h.
Уравнения согласованности МтР, как обычно, получаются с использованием разложений в ряд Тейлора точных и численных решений ОДУ.

Приведем несколько простых примеров методов типа Розенброка:

одностадийный МтР второго порядка —

$$y_{n+1} = y_n + h[I - \frac{h}{2}\frac{\partial\, f}{\partial\, y}(t_n, y_n)]^{-1} f(t_n, y_n). \qquad (3.39)$$

Двухстадийный МтР второго порядка для автономной ОДУ —

$$y_{n+1} = y_n + \frac{3}{4}k_1 + \frac{1}{4}k_2,$$

$$k_1 = h[I - \frac{h}{4}\frac{\partial f}{\partial y}(y_n)]^{-1} f(y_n), \tag{3.40}$$

$$k_2 = h[I - \frac{h}{4}\frac{\partial f}{\partial y}(y_n)]^{-1} f(y_n + k_1).$$

Функции устойчивости МтР определяются так же, как и для ДНРК.

В частности, если применить методы (3.39) и (3.40) к уравнению $y' = \lambda y$, то соответственно получим

$$R_1(z) = \frac{1 + z/2}{1 - z/2}, \;\; R_2(z) = \frac{(1 + z/4)^2}{(1 - z/4)^2}. \tag{3.41}$$

Для первого из них область относительной устойчивости в комплексной плоскости $z = x + i\,y$ определяется неравенством

$$\frac{(1 + x/2)^2 + y^2/4}{(1 - x/2)^2 + y^2/4} < e^{2\,x}.$$

Оба данных алгоритма являются A-устойчивыми, но не L-устойчивыми.

В заключение данного параграфа рассмотрим предложенное С. С.Филипповым трехпараметрическое семейство одношаговых методов решения ОДУ, названное автором $a\,b\,c$ — *схемами* (С. С.Филиппов. $A\,B\,C$ — схемы для жестких систем обыкновенных дифференциальных уравнений. ДАН, т.399, N 2, 2004, 170—172). Формально они являются обобщением методов типа Розенброка, поскольку, кроме матрицы Якоби диф-

ференциального уравнения, в них еще используется и квадрат якобиана.

Данные алгоритмы ориентированы на решение автономной системы ОДУ

$$\dot{y} = f(y(t)), \;\; y(t_0) = y_0, \;\; t \in [t_0, t_e]$$

и представляются следующей формулой:

$$(I + a\,h\,f_y + b\,h^2 f_y^2)(y_1 - y_0) = h\,f + c\,h^2 f_y f,$$

где f_y — якобиан, a, b, c — числовые параметры, а I, y и f суть единичная матрица и векторы порядка N. Здесь все значения функций берутся для значений аргумента $t = t_0$, причем первый шаг рассматривается как репрезентативный и все последующие значения вычисляются аналогично, с заменой y_0, y_1 на y_n, y_{n+1} соответственно.

Если функция $f(y)$ достаточное число раз дифференцируема, то с помощью тейлоровских разложений для a, b, c–схем можно показать следующие порядки локальных ошибок аппроксимации:

- при любых вещественных значениях a, b и c - схема имеет порядок не ниже первого;
- при выполнении условия

  $$c = a + 1/2$$

  схема имеет второй порядок, а главный член локальной ошибки имеет вид

$$\psi(h) = \frac{h^3}{3!}[f_{y,y} f\, f + (1 + 3\,a + 6\,b) f_y^2 f];$$

- при дополнительном условии

$$a + 2\,b + 1/3 = 0$$

$a\,b\,c$–схемы образуют однопараметрическое семейство методов третьего порядка для решения линейных автономных ОДУ с главным членом локальной ошибки

$$\psi(h) = -\Big(\frac{h^4}{4!}\Big)(1 + 2\,a) f_y^3 f.$$

Устойчивость $a\,b\,c$–схем можно исследовать на основе анализа их функции устойчивости $R(z = h\lambda)$, получаемой после применения данных алгоритмов к модельному уравнению Далквиста $\dot{y} = \lambda y$:

$$R(\mu) = \frac{1 + (1 + a)\mu + (b + c)\mu^2}{1 + a\,z + b\,z^2},\ \mu = \lambda\,h.$$

Отсюда нетрудно получить следующие утверждения:

- $a\,b\,c$—схемы второго порядка A-устойчивы тогда и только тогда, когда выполняются неравенства

$$a \leqslant -1/2,\ \ b \geqslant -\frac{a}{2} - \frac{1}{4};$$

- для L-устойчивости $a\,b\,c$—схемы второго порядка необходимо и достаточно выполнение условий

$$a + b + 1/2 = 0,\ \ b > 0;$$

- функция устойчивости $a\,b\,c$—схем третьего порядка имеет вид

$$R(\mu) = \frac{1 + (1+a)\mu + (1/3 + a/2)\mu^2}{1 + a\mu - (1/6 + a/2)\mu^2}$$

эти методы A-устойчивы тогда и только тогда, когда $a \leqslant -1/2$; данное семейство содержит единственный L-устойчивый алгоритм

$$\Big(I - \frac{2\,h}{3} f_y + \frac{h^2}{6} f_y^2\Big)(y_1 - y_0) = h\,f - \frac{h^2}{6} f_y f,$$

получаемый при значениях параметров $a = -\frac{2}{3}, b = -c = \frac{1}{6}$ и имеющий функцию устойчивости

$$R(\mu) = \frac{1 + \mu/3}{1 - 2\mu/3 + \mu^2/6};$$

- метод, получаемый при $a = -\frac{1}{2}, b = \frac{1}{12}, c = 0$:

$$\Big(I - \frac{h}{2} f_y + \frac{h^2}{12} f_y^2\Big)(y_1 - y_0) = h\,f,$$

является единственной A-устойчивой $a\,b\,c$–схемой, имеющей четвертый порядок аппроксимации для линейных автономных задач, причем главный член локальной ошибки в этом случае равен

$$\delta(h) = \Big(\frac{h^5}{5!}\Big) f_y^4 f/6,$$

а функция устойчивости имеет вид

$$R(\mu) = \frac{1 + \mu/2 + \mu^2/12}{1 - \mu/2 + \mu^2/12}.$$

Приведем еще один полезный пример: в семействе A-устойчивых $a\,b\,c$–схем второго порядка имеется единственный метод

$$\Big(I - \frac{h}{2} f_y\Big)(y_1 - y_0) = h\, f,$$

не использующий квадрата матрицы Якоби f_y; его функция устойчивости $R(\mu) = (1 + \mu/2)/(1 - \mu/2)$ такая же, как в неявных алгоритмах средней точки и трапеций.

§ 3.5. B-устойчивость НМРК

Напомним введенное в главе 1 определение 1.5: неравенство для евклидовой нормы

$$((f(t,y) - f(t,z)), (y - z)) \leqslant l||y - z||^2 \tag{3.42}$$

называется односторонним условием Липшица для f, число l есть односторонняя константа Липшица. Для комплекснозначных $y, f \in C^N$ это условие следует заменить на

$$\mathrm{Re}((f(t,y) - f(t,z)), (y - z)) \leqslant l||y - z||^2.$$

Введение данного понятия позволяет сформулировать следующее важное утверждение.

Теорема 3.6. *Пусть $f(t,y)$ — непрерывная функция, удовлетворяющая* (3.42). *Тогда для любых двух решений $y(t)$ и $z(t)$ уравнения $y' = f(t,y)$ при $t \geqslant t_0$ выполняется неравенство*

$$||y(t) - z(t)|| \leqslant ||y(t_0) - z(t_0)||e^{l(t-t_0)}. \tag{3.43}$$

Напомним (см. (1.64)), что при $l \leqslant 0$ из (3.43) следует свойство контрактивности решений ОДУ, т. е.

$$||y(t) - z(t)|| \leqslant ||y(t_0) - z(t_0)||. \tag{3.44}$$

Последнее означает, что расстояние между любыми двумя решениями есть невозрастающая функция t.

Рассмотренные свойства решений дифференциальных уравнений важно сохранять и на дискретном уровне при построении численных алгоритмов.

Определение 3.1. *Неявный МРК*

$$y_{n+1} = y_n + h\sum_{j=1}^{m} b_j f(t_n + c_j h, g_j),$$
$$g_j = y_n + h\sum_{l=1}^{m} a_{j,l} f(t_n + c_l h, g_l),$$

называется B-устойчивым, если из условия контрактивности f при любых $h > 0$ следует

$$||y_{n+1} - \hat{y}_{n+1}|| \leqslant ||y_n - \hat{y}_n||, \tag{3.45}$$

где y_n и $\hat{y}_n$ — два численных решения с начальными данными y_0 и $\hat{y}_0$.

Как видно из сравнения (3.45) и (3.44), свойство B-устойчивости метода — это наследование свойства контрактивности решений ОДУ.

Теорема 3.7 (третья теорема Батчера). *Неявный МРК является B-устойчивым, если его коэффициенты удовлетворяют условиям*

а) $b_j \geqslant 0$ *для* $j = 1, ..., m$ *;*

б) матрица $M = \{m_{i,j} = b_i a_{i,j} + b_j a_{j,i} - b_i b_j; i, j = 1, ..., m\}$ *положительно определенная.*

Определение 3.2. *Неявный МРК называется алгебраически устойчивым, если он удовлетворяет условиям а,б.*

Из теоремы 3.7 очевидно следует, что свойство алгебраической устойчивости является достаточным, но не необходимым условием B-устойчивости НМРК.

Теорема 3.8 (Хайрер). *Пусть* $R(z) = P(z)/Q(z)$*, где* $P(0) = Q(0) = 1$*,* $deg\ P \leqslant m$*,* $deg\ Q = m$*, — есть неприводимая A-устойчивая функция такая, что* $R(z) - e^z = 0(z^{\gamma+1})$ *для некоторого* $\gamma \geqslant 1$*. Тогда существует* m*-стадийный B-устойчивый МРК порядка* γ *с* $R(z)$ *в качестве функции устойчивости.*

Из этой теоремы выводятся следующие утверждения о порядковых барьерах НМРК.

а. Порядок алгебраически устойчивого ДНМРК не выше 4.

б. Порядок ОДНМРК со всеми положительными весами b_j не выше 4.

в. Порядок ДНМРК со всеми $b_j > 0$ не выше 6.

§ 3.6. Контрактивность МРК

Сначала установим важное вспомогательное утверждение: поскольку для решения линейной системы ОДУ $y' = Ay$ с постоянной или переменной матрицей имеем

$$\frac{d}{d\,t}||y||^2 = \frac{d}{d\,t}(y,y) = 2\,\mathrm{ReRe}(y,y') = 2\,\mathrm{Re}(y,A\,y),$$

то норма этого решения затухает при

$$\mathrm{Re}(y,A\,y) \leqslant 0, y \in \mathbb{R}^N. \tag{3.46}$$

Отсюда несложно получить следующий алгебраический результат.

Теорема 3.9 (Дж. фон Нейман). *Пусть рациональная функция $R(z)$ ограничена при $\mathrm{Re}\,z \leqslant 0$, а постоянная матрица A удовлетворяет неравенству* (3.46). *Тогда справедливо соотношение*

$$||R(A)|| \leqslant \sup_{\mathrm{Re}\,z \leqslant 0} |R(z)|.$$

Следствие 3.1. Если рациональная функция $R(z)$ A-устойчива ($\max\limits_{R\,e\,z \leqslant 0} |R(z)| \leqslant 1$), то численное решение $y_{n+1} = R(h\,A)y_n$ при условии (3.46) *контрактивно в евклидовой норме* (т. е. $||y_{n+1}|| \leqslant ||y_n||$).

Следствие 3.2. Если при любом $v \in \mathbb{R}^N$ справедливо неравенство $\mathrm{Re}(v,A\,v) \leqslant l||v||^2$, то $||R(A)|| \leqslant \sup\limits_{\mathrm{Re}\,z \leqslant l} |R(z)|$.

Теорема 3.10. *Пусть для произвольной нормы матрицы A при некотором $\lambda \geqslant 0$ справедливо неравенство $||A + \lambda I|| \leqslant \lambda$. Если функция устойчивости метода является абсолютно монотонной, т. е.*

$$R(0) = 1, \;\; R^{(j)}(z) \geqslant 0 \qquad z \in [-\rho, 0] \qquad j = 0, 1, 2, ..,$$

то при $h\lambda \leqslant \rho$ имеет место численная контрактивность, т. е. $||R(h\,A)|| \leqslant 1$.

Определение 3.3. *Наибольшее число ρ, для которого функция чувствительности абсолютно монотонна, называется пороговым коэффициентом $R(z)$.*

Теорема 3.11 (об алгебраической устойчивости). *Пусть МРК имеет все различные c_j и положительные b_j, а также удовлетворяет условиям $B(2\,m-2), C(m-1), D(m-1)$ из теоремы* 3.3 *Батчера. Тогда для его алгебраической устойчивости необходимо и достаточно, чтобы $|R(\infty)| \leqslant 1$, где $R(z)$ — его функция устойчивости.*

§ 3.7. AN - устойчивость

Понятия A-устойчивости и B-устойчивости основаны на исследованиях двух крайних случаев: f есть число или нелинейная функция достаточно общего вида. Рассмотрим промежуточный вариант – скалярное неавтономное уравнение

$$y' = \lambda(t)y, \;\; Re\lambda(t) \leqslant 0,$$

которое является линейным, но содержит переменный коэффициент λ.

Для данной задачи НМРК записывается в виде

$$\begin{aligned} g_j &= y_n + h\sum_{l=1}^{m} a_{j,l}\lambda(t_n + c_j h)g_l,\ j = 1, ..., m, \\ y_{n+1} &= y_n + h\sum_{j=1}^{m}\lambda(t_n + c_j h)b_j g_j, \end{aligned} \tag{3.47}$$

который после применения векторно-матричных обозначений принимает форму

$$(I - A\,Z)g = y_n e,\ \ y_{n+1} = y_n + b^T Z\,g, \tag{3.48}$$

а после исключения вектора $g = y_n(I - A\,Z)^{-1}e$ дает

$$\begin{gathered} y_{n+1} = R(Z)y_n,\ R(Z) = 1 + b^T Z(I - A\,Z)^{-1}e, \\ b = (b_1, ..., b_m)^T,\ e = (1, ..., 1)^T,\ Z = \mathrm{diag}\{z_j\} \in \mathbb{R}^{m,m}, \\ z_j = h\lambda(t_n + c_j h),\ g = (g_1, ..., g_m)^T,\ e = \{1\} \in \mathbb{R}^m, \\ g = e\,y_n + A\,Z\,g. \end{gathered} \tag{3.49}$$

Определение 3.4. *МРК называется AN-устойчивым, если* $|R(Z)| \leqslant 1$ *для всех* $Z = diag\{z_j; Re\,z_j \leqslant 0\}$ *и* $z_j = z_l$ *при* $c_j = c_l (j, l = 1, ..., m)$.

Легко проверить, что для $z_j = z,\ j = 1, ..., m$ имеем $R(Z) = R(z)$.

Очевидно следующее утверждение о соотношениях определений устойчивости: *B-устойчивость* $\Longrightarrow$ *AN-устойчивость* $\Longrightarrow$ *A-устойчивость*..

Пример 3.2. Для правила трапеций $y_{n+1} = y_n + h(f(t_n, y_n)+f(t_{n+1}, y_{n+1}))/2$ имеем $R(Z) = (1+z_1/2)/(1-z_2/2)$. Поскольку, например, из $z_2 = 0, z_1 = -\infty$, следует $R(Z) = -\infty$, то данный метод не является $A\,N$-устойчивым.

Теорема 3.12 (о достаточном условии B-устойчивости). *Если $|R(Z)| \leqslant 1$ при всех $Re\, z_j \leqslant 0$ и $|z_j| \leqslant \varepsilon$ для некоторого $\varepsilon > 0$, то НМРК является алгебраически устойчивым и, следовательно, B-устойчивым.*

§ 3.8. Существование и единственность НМРК для жестких ОДУ

Как уже отмечалось ранее, если шаг сетки h достаточно мал, то нелинейная система уравнений для нахождения величин k_j в неявных МРК вида (3.20) имеет единственное решение. Однако в жестких системах ОДУ с большой константой Липшица L, когда условие $h\,L \ll 1$ не выполняется, вопросы существования и единственности НМРК требуют специального рассмотрения.

Введем в рассмотрение скалярное произведение векторов $(u, v)_D = u^T D\, v$, порождаемое диагональной положительно определенной матрицей $D = \mathrm{diag}\{d_i > 0\} \in \mathbb{R}^{N,N}$, а также определим величину

$$\alpha_D(A^{-1}) = \max\{\alpha\} : (u, A^{-1}u)_D \geqslant \alpha(u, u)_D,\ u \in \mathbb{R}^N. \tag{3.50}$$

Если выполняется неравенство $\alpha > 0$, то это определяет свойство коэрцитивности матрицы A^{-1}. Введем также необхо-

димое для дальнейшего обозначение $\alpha_0(A^{-1}) = \sup\limits_{d_j>0} \alpha_D(A^{-1})$.

В качестве иллюстрации отметим, что для m-стадийного диагонально неявного МРК с матрицей $A = \{a_{i,j}; a_{i,j} = 0 \text{ для } i < j\} \in \mathbb{R}^{m,m}$ имеем $\alpha_0(A^{-1}) = \min\limits_{j=1,\dots,m} \{a_{j,j}^{-1}\}$.

Теорема 3.13 (существования НМРК). *Пусть f — непрерывно дифференцируемая функция, удовлетворяющая одностороннему условию Липшица*

$$((f(t,y) - f(t,z)), (y-z)) \leqslant l||y-z||^2.$$

Если матрица A m-стадийного НМРК обратима и $hl \leqslant \alpha_0(A^{-1})$, то у нелинейной системы

$$g_j = y_n + h\sum_{l=1}^{m} a_{j,l} f(t_n + c_j h, g_l),\ j = 1, ..., m,$$

существует решение $g = (g_1, ..., g_m)^T$.

Отметим, что алгебраически устойчивый НМРК для контрактивной задачи ($l = 0$) может не допускать решения системы для g_j.

Рассмотрим теперь возмущенный НМРК

$$\begin{aligned} \tilde{g}_j &= y_n + h\sum_{j=1}^{m} a_{j,l} f(t_n + c_j h, \tilde{g}_j) + \delta_j, \\ \tilde{y}_{n+1} &= y_n + h\sum_{j=1}^{m} b_j f(t_n + c_j h, \tilde{g}_j), \end{aligned} \tag{3.51}$$

где $\delta_j \in \mathbb{R}^N$ — векторы каких-либо привнесенных возмущений.

Теорема 3.14 (о возмущениях НМРК). *В условиях теоремы существования* 3.13 *справедливы оценки*

$$||\tilde{g}_j - g_j||_D \leqslant \frac{||A^{-1}||_D}{\alpha_D(A^{-1}) - hl}||\delta_j||_D,\ j = 1, ..., m,$$
$$||\tilde{y}_{n+1} - y_{n+1}||_D \leqslant ||b^T A^{-1}||_D\Big(1 + \tfrac{||A^{-1}||_D}{\alpha_D(A^{-1})-hl}\Big)\max_j ||\delta_j||_D. \tag{3.52}$$

Следствием данной теоремы является следующее утверждение.

Теорема 3.15 (о единственности НМРК). *При условиях теоремы существования решение системы для коэффициентов НМРК единственно.*

§ 3.9. *B*-согласованность и *B*-сходимость НМРК

Как мы уже видели, в линейных задачах проблема взаимоотношения локальных и глобальных ошибок полностью закрывается теоремой эквивалентности Лакса 2.2: из аппроксимации и устойчивости следует сходимость метода. В нелинейных системах ОДУ данная теорема не работает и ситуация оказывается более сложной. Для иллюстрации рассмотрим феномен снижения порядка для жестких ОДУ.

Хорошим примером в данном случае является *модельная задача Протеро–Робинсона*

$$y' = \lambda(y - \varphi(t)) + \varphi'(t),\ y(t_0) = y_0,\ \mathrm{Re}\lambda \leqslant 0, \tag{3.53}$$

которая является частным случаем (2.43) при $g(t) = \varphi'(t) - \lambda\varphi(t)$. Как легко проверить, задача (3.53) имеет точное решение

$$y = \varphi(t) + e^{\lambda t}[y_0 - \varphi(t_0)], \tag{3.54}$$

а при $y_0 = \varphi(t_0)$ получаем $y = \varphi(t)$. Запишем для нее m-стадийный НМРК:

$$\begin{aligned} g_j &= y_n + \sum_{l=1}^{m} a_{j,l}\{\lambda[(g_l - \varphi(t_n + c_l h)] + \varphi'(t_n + c_l h)\}, \\ y_{n+1} &= y_n + \sum_{j=1}^{m} b_j\{\lambda[(g_j - \varphi(t_n + c_j h)] + \varphi'(t_n + c_j h)\}. \end{aligned} \tag{3.55}$$

Локальные погрешности этой схемы получаются, если в данных формулах величины g_j, y_n и y_{n+1} заменить на соответствующие точные значения $\varphi(t_n + c_l h), \varphi(t_n)$ и $\varphi(t_{n+1})$:

$$\begin{aligned} \psi_j^h(t_n) &= \varphi(t_n + c_j h) - \varphi(t_n) - h\sum_{l=1}^{m} a_{j,l}\varphi'(t_n + c_l h), \\ \psi^h(t_n) &= \varphi(t_n + h) - \varphi(t_n) - h\sum_{j=1}^{m} b_{j,l}\varphi'(t_n + c_j h). \end{aligned}$$

Разложение функций в правых частях этих равенств в окрестности точки t_n дает следующие оценки:

$$\psi^h(t_n) = O(h^{p+1}),\ \psi_j^h(t_n) = O(h^{q+1}),$$

где p есть порядок используемой квадратурной формулы с весами b_j и узлами c_j, а величина q, называемая *стадийным порядком*, — это наибольшее число, для которого при всех j выполнено коллокационное условие $C(q)$ в теореме 3.3 Батчера, т. е.

$$\sum_{l=1}^{m} a_{j,l} c_j^{k-1} = c_j^k/k,\ k = 1, ..., q.$$

Наименьшее из чисел p и q, очевидно, определяет порядок НМРК. Исключая из вышеприведенных соотношений внутренние стадии, получим рекуррентные выражения

$$\varphi(t_{n+1}) - y_{n+1} = R(z)(\varphi(t_n) - y_n) + \delta_n^h, \tag{3.56}$$

где
$$\delta_n^h = z\, b^T (I - z\, A)^{-1} \bar{\psi}^h(t_n) + \psi^h(t_n),\ z = \lambda\, h,$$

$$R(z) = 1 + z\, b^T (I - z\, A)^{-1} e,\ \ \bar{\psi}^h = (\psi_1^h, ..., \psi_m^h)^T.$$

Отсюда получаем выражение для глобальной погрешности

$$z_{n+1}^h = \varphi(t_{n+1}) - y_{n+1} = (R(z))^{n+1}(\varphi(t_0) - y_0) + \sum_{j=0}^{n} (R(z))^{n-j} \delta_j^h. \tag{3.57}$$

Для классических (нежестких) ОДУ рассматривается $z = O(h)$, и тогда (при $p = q$) $z_{n+1}^h = O(h^p)$. Для жестких ОДУ необходимо считать одновременно $h \to 0$ и $\lambda h \to -\infty$. Тогда порядок z^h значительно снижается. Например, для методов Гаусса имеем $q = m$, $|R(z)| \leqslant 1$ и $z^h = O(h^m)$, в то время как их локальная ошибка для линейных ОДУ есть $O(h^{2m})$.

Наиболее наглядно понижение порядка (даже локального) для жестких ОДУ можно продемонстрировать для простейших неявных схем Эйлера и средней точки:

$$y_{n+1} = y_n + h\, f(t_{n+1}, y_{n+1}) \quad y_{n+1} = y_n + h\, f\Big(t_n + \frac{h}{2}, \frac{y_n + y_{n+1}}{2}\Big).$$

В предположении достаточной гладкости решения с помощью разложений в ряд Тейлора показывается, что их локальные погрешности суть $O(h^2)$ и $O(h^3)$ соответственно. Однако "жесткий" анализ дает другие результаты. Для схемы Эйлера имеем

$$\psi^h(t_n) = \frac{-1}{2(1-h\lambda)}[h^2\varphi''(t_n) + O(h^3)],$$

т. е. действительно второй порядок. Однако для метода средней точки

$$\psi^h(t_n) = -\frac{h^2}{4}\frac{h\lambda}{1-h\lambda}\varphi''(t_n) + O(h^3), \tag{3.58}$$

откуда при $h\lambda \to -\infty$ (именно в этом заключается "жесткость") будет $\psi^h(t_n) = O(h^2)$.

Теорема 3.16 (о B-согласованности НМРК). *Пусть правая часть ОДУ удовлетворяет одностороннему условию Липшица* (3.42), *матрица A обратима и $\alpha_0(A^{-1}) > 0$, а стадийный порядок НМРК равен q. Тогда для всех n*

$$||\delta_n^h|| \leqslant C\, h^{q+1} M_{q+1} \ \textit{при}\ hl \leqslant \alpha < \alpha_0(A^{-1}).$$

Если $\alpha_D(A^{-1}) = 0$ для некоторой положительно определенной матрицы D и $l < 0$, то

$$||\delta_n^h|| \leqslant C|l|^{-1} h^q M_{q+1},$$

где C зависит только от коэффициентов МРК и $M_{q+1} = \max\limits_{\xi\in[t_n, t_n+h]} ||y^{(q+1)}(\xi)||$.

Замечание 3.1.МРК, удовлетворяющий условиям теоремы, называется *B-согласованным.*

Принципиальным здесь является тот факт, что оценка локальной погрешности зависит только от производных решения, но не от производных функции f (этим свойством обладают отнюдь не все МРК).

Теорема 3.17 (об оценке устойчивости). *В условиях предыдущей теоремы для алгебраически устойчивого МРК при любом $\alpha \in (0, \alpha_0(A^{-1}))$ существует постоянная C такая, что*

$$||\hat{y}_{n+1} - y_{n+1}|| \leqslant (1 + C\,h\nu)||\hat{y}_n - y_n|| \qquad 0 \leqslant h\nu \leqslant \alpha.$$

Определение 3.5.*МРК называется B-сходящимся с порядком r для ОДУ, удовлетворяющих одностороннему условию Липшица, если глобальная погрешность при $h\,l \leqslant \alpha$ допускает оценку*

$$||y_n - y(t_n)|| \leqslant h^r \gamma(t_n - t_0, l) \max_{j=1,\ldots,l} M_j, \tag{3.59}$$

где постоянная α и функция γ зависят только от коэффициентов метода.

Теорема 3.18 (о B-сходимости). *Для глобальной погрешности алгебраически устойчивого МРК с обратимой матрицей A и стадийным порядком $q \leqslant p$, при $h\nu \leqslant \alpha < \alpha_0(A^{-1})$ и выполнении одностороннего условия Липшица, справедлива оценка*

$$||y_n - y(t_n)|| \leqslant h^q \frac{exp(C_1\nu(t_n - t_0)) - 1}{C_1\nu} C_2 M_{q+1},$$

где константы C_1, C_2 *зависят только от коэффициентов МРК.*

Замечание 3.2. Метод трапеций

$$y_{n+1} = y_n + h_n[f(t_n, y_n) + f(t_{n+1}, y_{n+1})]/2$$

не является B-устойчивым, однако для последовательности шагов с условием

$$\sum_{k=0}^{n-1}\{\prod_{j=k+1}^{n-1} \max\{1, h_j/h_{j-1}\}h_k^3\} \leqslant h^2$$

справедлива оценка $||y_n - y(t_n)|| \leqslant C\, M_3 h^2$.

Замечание 3.3. Методы Розенброка не являются B-устойчивыми, так что в общем случае при очень больших λ их порядок равен всего лишь 2. Однако специальные подходы позволяют строить для них "B-сходящиеся" схемы высших порядков.

§ 3.10. Контроль точности и устойчивости одношаговых методов

Практической целью исследования устойчивости и погрешности численных методов является разработка алгоритма для решения конкретного класса задач с требуемой точностью, при минимальных по объему затрачиваемых вычислительных ресурсах. В применении к системам ОДУ будем

считать, что экономичность метода характеризуется количеством вычислений вектора правой части N_f, необходимых для обеспечения заданного уровня ошибки численного решения ε при интегрировании уравнений на интервале длины T. Если предположить для простоты, что вычислительная схема является однородной, то $N_f = n_f \cdot N_e$, где N_e — количество расчетных шагов, а n_f — число вычислений функции $f(t, y)$ на каждом шаге.

Два основных инструмента для минимизации N_f — это выбор порядка численного интегрирования, для одношаговых методов связанного с количеством используемых стадий, а также определение текущего временного шага, верхняя допустимая граница которого может определяться или погрешностью аппроксимации ОДУ, или условиями устойчивости численного решения к привносимым расчетным возмущениям. Наиболее актуальными эти вопросы являются для жестких задач, в которых на переходных участках с большими градиентами решений наиболее критичными являются аппроксимационные аспекты, а на длительных периодах установления – ограничительные условия устойчивости. Успешное использование априорной и/или апостериорной информации для автоматизации подбора счетных параметров во многом обуславливает сравнительные достоинства метода в реальной ситуации.

3.10.1. Характеризация локальной и глобальной ошибок одношаговых схем. Интересующие нас подходы будут рассматриваться для автономной задачи Коши

$$y' = f(y), \;\; y(t_0) = y_0, \;\; t_0 \leqslant t \leqslant t_e, \tag{3.60}$$

численное решение которой ищется с помощью линейных одношаговых методов вида

$$y_{n+1} = y_n + h\varphi_f(t_n, y_n, h), \;\; n = 0, 1, ..., N. \tag{3.61}$$

Линейность понимается в том смысле, что φ_f представляет собой линейную форму от значений неизвестного вектора y, если само уравнение (3.60) является линейным. Шаг сетки $h = t_{n+1} - t_n$ будем считать для простоты постоянным, так что $N_e = (t_e - t_0)/h$, хотя введение переменных h_n принципиальной роли в одношаговых методах не играет, если сетка является квазирегулярной, т. е. отношение максимального шага к минимальному остается ограниченным при $N_e \to \infty$.

Главной целью анализа качества алгоритма является оценка глобальной, или полной, ошибки численного решения ОДУ для аргумента $(t = t_n)$, т. е. величины

$$\varepsilon_n = y(t_n) - y_n, \;\; n = 0, 1, ...,$$

где $y(t_n)$ есть значение точного решения в указанной точке. Основным инструментом при этом является определение погрешности аппроксимации, являющейся фактически функцией невязки расчетной формулы (3.61), после подстановки в нее величин $y(t_n)$ вместо y_n:

$$\delta_{n+1} = y(t_{n+1}) - y(t_n) - h\varphi_f(t_n, y(t_n), h). \tag{3.62}$$

Отметим, что если здесь подставить y_n вместо $y(t_n)$, то получим

$$\delta_{n+1} = y(t_{n+1}) - y_{n+1},$$

вследствие чего δ_{n+1} называется также локальной ошибкой численного решения, т. е. ошибкой, получаемой за один шаг интегрирования.

Глобальная ошибка может быть связана с локальной путем почленного сложения уравнений (3.62) и (3.61):

$$y(t_{n+1}) - y_{n+1} = y(t_n) - y_n + h[\varphi_f(t_n, h) - \varphi_f(t_n, y_n, h))] + \delta_{n+1}.$$

Применяя к выражению в квадратных скобках теорему Лагранжа о среднем

$$\varphi_f(t_n, y(t_n), h) - \varphi_f(t_n, y_n, h) = \frac{\partial \varphi_f(t_n, \xi_n, h)}{\partial y}[y(t_n) - y_n],$$

$$\xi_n \in [y(t_n), y_n],$$

получаем следующее соотношение:

$$\varepsilon_{n+1} = Q_n \varepsilon_n + \delta_{n+1},\ Q_n = I + hJ_n,\ J_n = \frac{\partial \varphi_f(t_n, \xi_n, h)}{\partial y} = \varphi_f(t_n, \xi_n). \tag{3.63}$$

Участвующие здесь матрица перехода Q_n и матрица Якоби J_n определяют устойчивость (или неустойчивость) по норме вычислительной схемы (3.61), вследствие получаемого из (3.63) неравенства

$$||\varepsilon_{n+1}|| \leqslant ||Q_n|| \cdot ||\varepsilon_n|| + ||\delta_{n+1}||,\ \ n = 0, 1, ... \tag{3.64}$$

Если мы здесь положим для всех n

$$||\delta_{n+1}|| \leqslant \delta, \;\; ||Q_n|| \leqslant q, \tag{3.65}$$

то из цепочки рекуррентных неравенств приходим к следующему:

$$||\varepsilon_{n+1}|| \leqslant q^n||\varepsilon_0|| + (1 + q + ... + q^n)\delta.$$

Отсюда при $q = 1$ получаем оценку

$$||\varepsilon_{n+1}|| \leqslant ||\varepsilon_0|| + (n + 1)\delta, \tag{3.66}$$

а при $q \neq 1$ имеем неравенство

$$||\varepsilon_{n+1}|| \leqslant q^n||\varepsilon_0|| + \Big|\frac{q^{n+1} - 1}{q - 1}\Big|\delta. \tag{3.67}$$

В последней формуле выделим специальный случай

$$q = 1 + b/n, \;\; b > 0, \tag{3.68}$$

для которого вследствие справедливости соотношения

$$q^n = ((1 + \frac{b}{n})^{n/b})^b \leqslant e^b$$

приходим к неравенству

$$||\varepsilon_{n+1}|| \leqslant e^b||\varepsilon_0|| + \frac{n}{b}e^b\delta. \tag{3.69}$$

Рассмотренные в (3.65)—(3.69) различные ситуации дают четыре основные характеризации матрицы перехода:

$$||Q_n|| \geqslant c_1 > 1, \;\; ||Q_n|| \leqslant 1 + \frac{b}{n}, \;\; ||Q_n|| \leqslant 1, \; ||Q_n|| \leqslant c_2 < 1,$$

где постоянные b, c_1 и c_2 не зависят от n. Для этих случаев будем называть метод (3.62) неустойчивым, слабо устойчивым, устойчивым и сильно устойчивым по норме соответственно. Как легко заметить, данные ситуации определяют с увеличением n или неограниченный рост ошибки, или ее ограниченное возрастание, или даже уменьшение погрешности.

В общем случае нелинейной системы ОДУ якобиан J_n представляет собой достаточно сложный объект для исследования. Первый шаг к его упрощению — это линеаризация задачи Коши.

Пусть на интервале $[t_a, t_e], \; t_0 \leqslant t_a \leqslant t_e$, вместо ОДУ (3.60) мы имеем возмущенную задачу Коши

$$\tilde{y}' = f(\tilde{y}), \; \tilde{y}(t_a) = \tilde{y}_a. \tag{3.70}$$

Тогда возмущение исходного решения $z(t) = y(t) - \tilde{y}(t)$ удовлетворяет линейному уравнению

$$z' = f(y) - f(\tilde{y}) = A(t)z, \tag{3.71}$$

где матрица $A(t)$ с переменными элементами выражается с помощью интегральной формы теоремы о среднем:

$$A(t) = \int\limits_0^1 f'(t, \theta y(t) + (1-\theta)\tilde{y}(t)) d\, t, \; \theta = \frac{t - t_0}{t_e - t_0}. \tag{3.72}$$

Дальнейшее возможное упрощение задачи — приближенная замена матрицы $A(t)$ в (3.72) постоянной матрицей $\tilde{A} =$

$A(t_b), t_b \in [t_0, t_e]$ с независящими от t (замороженными) элементами, в результате чего для новой неизвестной $\tilde{z}(t)$ получаем уравнение

$$\tilde{z}' = \tilde{A}\tilde{z}. \tag{3.73}$$

Понятно, что приемы линеаризации уравнений и "замораживание" матрицы не являются универсальным средством, особенно для жестких задач, однако изучение ОДУ вида (3.70) или (3.71) является не просто вынужденной мерой, но и практически полезным инструментом для многих актуальных проблем.

3.10.2. Контроль точности численного решения. Классическим способом оценки локальной ошибки одношаговых методов является использование экстраполяции Ричардсона, что более справедливо назвать правилом Рунге. Суть его заключается в двукратном вычислении приближенного решения — с шагом h и $h/2$, в результате чего для метода с порядком p при асимптотически малых h имеем приближенные равенства, см. выше п. 2.1.4:

$$\delta^h_{n,p} = \frac{y_n^{0.5\,h}}{-} y_n^h 1 - 2^{-p},\ \delta^{0.5h}_{n,p} = \frac{y_n^{0.5\,h} - y_n^h}{2^p - 1} = 2^{-p}\delta^h_{n,p}. \tag{3.74}$$

Подчеркнем, что здесь y_n^h и $y_n^{0.5h}$ представляют собой два приближенных решения для одного и того же значения аргумента t_n, вычисленные на вложенных сетках. Достаточная близость этих двух значений свидетельствует о выполнении "асимптотики по h" и о возможности использования выражений (3.74)

для оценки локальной погрешности δ из (3.65). В этом случае с помощью выражений вида (3.66) — (3.69) может быть оценена норма глобальной ошибки и по условию

$$||\varepsilon_{n+1}|| \leqslant \varepsilon$$

выбрана максимально допустимая величина шага сетки h (ε здесь означает заданную гарантированную погрешность численного решения, которую назовем толерантностью).

Однако в описанном подходе используется слишком ограничительное предположение о независимости оценки q в (3.65) от самого решения ОДУ. Получаемый при этом линейный рост $||\varepsilon_{n+1}||$ с увеличением n может реально не выполняться в "плохих" нелинейных задачах, в том числе для жестких систем.

С целью получения "робастного" алгоритма и сокращения вычислительных затрат при автоматизированном выборе шага h рассмотрим подход, предложенный Е. А.Новиковым [48]. Главная идея в данном случае заключается в построении конструктивной взаимосвязи между локальной и глобальной погрешностями численного решения.

Для этого перепишем уравнение (3.63) в следующей форме:

$$\frac{\varepsilon_{n+1} - \varepsilon_n}{n} - \varphi'_n \varepsilon_n = \delta_{n+1}/h. \tag{3.75}$$

Предположим теперь, что метод (3.61) имеет порядок аппроксимации γ, т. е. его локальная ошибка имеет представление

$$\delta_n = h^{\gamma+1}\psi(t_n) + O(h^{\gamma+2}),$$

где функция $\psi(t)$ не зависит от шага интегрирования h.

Тогда при условии устойчивости (хотя бы слабой) алгоритма (3.61) глобальная ошибка также может быть записана в аддитивной форме

$$\varepsilon_n = h^{\gamma} z_n + O(h^{\gamma+1}), \tag{3.76}$$

где величины z_n удовлетворяют следующим из (3.75) уравнениям

$$\frac{z_{n+1} - z_n}{h} - \varphi'_n z_n = \psi(t_n) = \psi_n,\ z_0 = 0. \tag{3.77}$$

Нетрудно увидеть, что алгебраическая система (3.77) аппроксимирует с погрешностью 1-го порядка дифференциальное уравнение

$$z'(t) - f'(t, y(t))z(t) = \psi(t), \tag{3.78}$$

и при $z(0) = 0$ имеет место сходимость по норме

$$||\varepsilon_n|| \equiv ||z(t_n) - z_n|| = O(h),\ n = 1, 2, ... \tag{3.79}$$

Таким образом, если уравнение (3.78) достаточно простое и допускает хотя бы приближенное аналитическое решение (или даже только его оценку), то проблема характеризации функции $\varepsilon_n = \{\varepsilon_i^{(n)}, i = 1, ..., N\}$ в (3.79) решена, и шаг сетки h по известной величине толерантности ε может быть выбран с помощью формулы

$$|\varepsilon_n|_r \equiv \max_i \frac{|\varepsilon_i^{(n)}|}{|y_i^{(n)}| + r} \leqslant \varepsilon, \tag{3.80}$$

где i — номер компоненты векторов y_n и ε_n, а r — некоторый регулируемый положительный параметр. Если для всех i выполняется неравенство $|\varepsilon_i^{(n)}| \ll r$, то фактически контролируется отмасштабированная абсолютная ошибка ($r||\varepsilon_n|| \leqslant \varepsilon$), а в случае $|\varepsilon_i^{(n)}| \gg r$ обеспечивается относительная погрешность ε.

В более реальных ситуациях определение z_n в (3.76) требует решения системы разностных уравнений (3.77), что связано с дополнительной вычислительной работой.

Для повышения экономичности алгоритма выбора шага успешно применяются “вложенные по порядку” численные схемы. Пусть для решения ОДУ имеются два метода $(p-1)$-го и p-го порядков точности, которые реализуют численные решения $y_{n,p-1}$ и $y_{n,p}$ соотвественно. Определим связанную с ними локальную погрешность с помощью формулы

$$\delta_{n,p-1} = y_{n,p} - y_{n,p-1}. \tag{3.81}$$

Предположим, что коэффициенты рассматриваемых вычислительных схем выбраны таким образом, что локальные ошибки согласованы в следующем смысле:

$$\begin{aligned} \delta_{n,p-1} &= c_{p-1} h^p \psi_p(t_n) + O(h^{p+1}), \\ \delta_{n,p} &= c_p h^{p+1} f'(t_n, y(t_n)) \psi_p(t_n) + O(h^{p+2}), \end{aligned} \tag{3.82}$$

где c_{p-1} и c_p — некоторые числовые коэффициенты, а $\psi_p(t_n)$ — не зависящая от h функция, которая в силу соотношения (3.81) может быть вычислена как

$$\psi_p(t_n) = c_{p-1}^{-1} h^{-p}(y_{n,p} - y_{n,p-1}) + O(h). \tag{3.83}$$

Аналогично предыдущему можно показать, что глобальная ошибка

$$\varepsilon_{n,p} = y(t_n) - y_{n,p} \tag{3.84}$$

представима в виде

$$\varepsilon_{n,p} = h^p z(t_n) + O(h^{p+1}), \tag{3.85}$$

где функция $z(t_n)$ есть решение задачи Коши

$$z'(t) - f'(t, y(t))z(t) = -c_p f'(t, y(t))\psi_p(t),\ z(t_0) = 0. \tag{3.86}$$

Важным практическим моментом в данном случае является тот факт, что при условиях (3.82) величина $\psi_{p,n} = c_p\psi_p(t_n)$ может быть определена “бесплатно”, поскольку из (3.83) и (3.85) получаем

$$\varepsilon_{n,p} \approx c_p c_{p-1}^{-1}(y_{n,p} - y_{n,p-1}). \tag{3.87}$$

Таким образом, в описываемом подходе вычисления ведутся по “основной” схеме p-го порядка, а метод $(p-1)$-го порядка используется для контроля точности.

В качестве иллюстрации рассмотрим решение задачи Коши для системы ОДУ с помощью двухстадийного метода Рунге—Кутты

$$y_{n+1} = y_n + p_1 k_1 + p_2 k_2,$$
$$k_1 = h\, f(t_n, y_n),\ k_2 = h\, f(t_n + \beta\, h, y_n + \beta\, k_1), \tag{3.88}$$

где β — неопределенный пока числовой параметр.

Условия второго порядка точности схемы (3.88) имеют вид

$$p_1 + p_2 = 1,\ \beta p_2 = 0.5, \tag{3.89}$$

а локальную ошибку $\delta_{n,2}$ при этом можно записать в форме

$$\delta_{n,2} = \frac{2 - 3\beta}{12} h^3 f'' f + \frac{1}{6} h^3 (f')^2 f + O(h^4). \tag{3.90}$$

Выбирая параметр $\beta = 2/3$, из (3.89) получим коэффициенты схемы (3.88) равными $p_1 = 1/4, p_2 = 3/4$, в результате чего локальная погрешность принимает вид

$$\delta_{n,2} = \frac{1}{3} h\, f' \psi + O(h^4),\ \psi = \frac{1}{2} h^2 f' f. \tag{3.91}$$

Для контроля точности вычислений рассмотренной схемы (3.88) используем вспомогательный метод первого порядка

$$y_{n+1,1} = y_n + k_1, \tag{3.92}$$

локальная ошибка которого имеет вид

$$\delta_{n,1} = \frac{1}{2} \psi + O(h^2). \tag{3.93}$$

С помощью идеи вложенных по порядку методов ошибку $\varepsilon_{n,2}$ алгоритма второго порядка вычисляем по формуле

$$\varepsilon_{n,2} = y_{n+1} - y_{n+1,1} = (p_1 - 1)k_1 + p_2 k_2 = p_2(k_2 - k_1). \tag{3.94}$$

Отметим, что поскольку приращение k_1 зависит от размера шага линейно, повторное вычисление решения в случае нарушения условия (3.80) для требуемой точности будет приводить только к одному дополнительному расчету правой части дифференциальной задачи.

3.10.3. Контроль устойчивости явных вычислительных схем. При численном решении жестких задач Коши для систем обыкновенных дифференциальных уравнений применяются, как правило, A-устойчивые или L-устойчивые (т. е. обладающие неограниченной областью устойчивости) неявные и полунеявные численные методы вида (3.61). Однако при реализации этих алгоритмов в случае большой размерности вычисление и обращение матриц Якоби требует значительных компьютерных ресурсов.

Данного недостатка лишены явные формулы, но они в традиционном применении для решения жестких задач в большинстве своем непригодны. Дело в том, что на участках установления с малыми значениями производных шаг численного интегрирования ОДУ с точки зрения обеспечения точности аппроксимации может быть огромным. Однако при этом будут нарушаться условия устойчивости явных схем, что приведет к катастрофическому накоплению численных погрешностей.

Этого явления можно избежать, если наряду с контролем точности при выборе шага учитывать и условия устойчиво-

сти схемы. Забегая вперед, скажем, что речь не идет о замене неявных A-устойчивых или L-устойчивых методов явными. Реально вопрос ставится о существовании класса умеренно жестких задач, для которых явные схемы с ограниченной областью устойчивости могут быть экономичнее неявных абсолютно устойчивых алгоритмов.

Построение неравенств для контроля устойчивости естественно начать для модельной задачи Коши

$$y' = \lambda\, y,\ \ y(0) = y_0,\ \ t \geqslant 0, \tag{3.95}$$

численное решение которой на равномерной сетке с шагом h представимо в виде

$$y_{n+1} = Q(z) y_n,\ \ z = \lambda\, h. \tag{3.96}$$

Условие устойчивости схемы (3.95) записывается в форме $|Q(z)| \leqslant 1$, а кривая $|Q(z)| = 1$ описывает границу области устойчивости. Условие устойчивости удобно также представить в форме неравенства

$$|\lambda h| \leqslant d, \tag{3.97}$$

где величина d характеризует размер интервала устойчивости. Для линейной задачи $y' = f(t, y)$ с постоянной матрицей Якоби $J = f' \in \mathbb{R}^{m,m}$ условие (3.97) заменяется на

$$|\lambda_i h| \leqslant d,\ \ i = 1, ..., m, \tag{3.98}$$

где λ_i суть собственные числа матрицы Якоби.

В более общих случаях контроль устойчивости можно делать с помощью какой-либо нормы матрицы Якоби

$$h||f'(t, y(t))|| \leqslant d \tag{3.99}$$

или путем оценки ее максимального собственного числа λ_{max}:

$$h|\lambda_{max}| \leqslant d, \tag{3.100}$$

которое можно определить, например, с помощью степенного метода.

Применение неравенств (3.99)—(3.100), как правило, не приводит к значительному увеличению вычислительных затрат.

При сравнении свойств устойчивости различных численных схем удобно пользоваться отношением длины интервала устойчивости к количеству вычислений функции f на каждом шаге интегрирования. Например, в явном методе Эйлера

$$y_{n+1} = y_n + hf(y_n),$$

граница области устойчивости которого есть кривая $|z - 1| = 1$, на одно вычисление правой части f приходится две единицы длины интервала устойчивости.

Для явного двухстадийного метода Рунге—Кутты второго порядка (ЯМРК-2) с функцией устойчивости

$$Q(z) = 1 + z + 0.5z^2$$

интервал устойчивости представляет собой отрезок $[-2, 0]$. Таким образом, в данном алгоритме на одно вычисление функции f (всего их два на каждом шаге) приходится единица длины интервала устойчивости. Отсюда следует, что явный метод Эйлера может быть эффективнее ЯМРК-2 второго порядка, если шаг сетки ограничен только по условию устойчивости.

Построим еще одну схему второго порядка, в которой на одно вычисление правой части ОДУ приходится не менее двух единиц длины интервала устойчивости. А именно, рассмотрим формулу ЯМРК с тремя вычислениями функции f на каждом шаге интегрирования:

$$\begin{aligned} &y_{n+1} = y_n + p_1 k_1 + p_2 k_2 + p_3 k_3, \\ &k_1 = h\, f(y_n),\ k_2 = h\, f(y_n + \beta_{2,1} k_1), \\ &k_3 = h\, f(y_n + \beta_{3,1} k_1 + \beta_{3,2} k_2). \end{aligned} \tag{3.101}$$

Условия второго порядка точности этого алгоритма имеют вид

$$p_1 + p_2 + p_3 = 1,\ \ \beta_{2,1} p_2 + (\beta_{3,1} + \beta_{3,2}) p_3 = 1/2. \tag{3.102}$$

Если дополнительно потребовать выполнение соотношения

$$3\beta_{2,1}^2 p_2 + 3(\beta_{3,1} + \beta_{3,2})^2 p_3 = 1, \tag{3.103}$$

то локальная ошибка метода (3.102) записывается в форме

$$\delta = \frac{1-6\,g}{6} h^3 (f')^2 f + O(h^4), \;\; g = \beta_{2,1}\beta_{2,3} p_3. \tag{3.104}$$

Отметим, что при выполнении условия

$$g = [1 + O(h)]/6 \tag{3.105}$$

трехстадийная схема (3.101) имеет третий порядок, но в общем случае равенства (3.103), (3.104) обеспечивают только 2-й порядок точности.

На примере модельной задачи (3.95) мы покажем, что путем выбора параметра g можно, за счет снижения порядка точности, значительно увеличить длину интервала устойчивости ЯМРК, что в итоге приводит к неожиданному на первый взгляд конечному росту эффективности алгоритма.

Функция устойчивости для рассматриваемой параметризованной схемы описывается формулой

$$Q(z) = 1 + z + 0.5\,z^2 + g\,z^3, \;\; z = \lambda\, h. \tag{3.106}$$

Соответствующая максимальная длина интервала устойчивости рассматриваемой схемы равна $d = 6.26$ и достигается при $g = 1/16$. Это означает, что на одно вычисление функции f приходится более двух единиц длины интервала устойчивости. Однако следует заметить, что для $z = -4$ имеем $Q(-4) = 1$,а это соответствует только слабой устойчивости с возможным наполнением ошибок при наличии возмущений. Во избежание данного эффекта можно положить значение параметра несколько меньшим, например, $g = 1/15$. В этом

случае величина d остается немного больше 6, и полученный метод второго порядка точности будет существенно экономичней явного метода Эйлера.

Чтобы изменение параметра g не приводило к заметному уменьшению длины интервала устойчивости, достаточно, достаточным условием является монотонность $Q(z)$, т. е. ее производная

$$Q'(z) = 3\,g\,z^2 + z + 1$$

не должна менять знак на интервале устойчивости. Учитывая, что корни квадратного уравнения $3\,g\,z^2 + z + 1 = 0$ равны

$$z_{1,2} = -\frac{1}{6g} \pm \sqrt{\frac{1-12g}{36g^2}},$$

имеем, что $Q(z)$ будет монотонной при $g \geqslant 1/12$. В заключение приведем значения коэффициентов (3.102) для трех характерных значений параметра g:

$$g = \frac{1}{12}\colon\ \beta_{2,1} = \beta_{3,1} = \beta_{3,2} = \frac{1}{3},\ p_1 = \frac{1}{4},\ p_2 = 0,\ p_3 = \frac{3}{4};$$

$$g = \frac{1}{15}\colon\ \beta_{2,1} = \frac{1}{3},\ \beta_{3,1} = \beta_{3,2} = \frac{3}{8},\ p_1 = \frac{1}{6},\ p_2 = \frac{3}{10},\ p_3 = \frac{8}{15};$$

$$g = \frac{1}{16}\colon\ \beta_{2,1} = \frac{1}{3},\ \beta_{3,1} = \beta_{3,2} = \frac{7}{18},\ p_1 = \frac{1}{7},\ p_2 = \frac{3}{8},\ p_3 = \frac{27}{56}.$$

§ 3.11. Численное интегрирование гамильтоновых систем

Цель данного параграфа — дать лишь общие представление о симплектических методах решения гамильтоновых систем ОДУ, по которым имеется огромное количество специ-

альной литературы, см., например, монографии [4], [64], [66] и цитируемые там работы.

Рассматриваемые системы имеют четное количество уравнения и записываются в виде

$$\dot{q}_k = \frac{\partial H(p,q)}{\partial p_k}, \quad \dot{p}_k = -\frac{\partial H(p,q)}{\partial q_k}, \ k = 1, ..., l, \tag{3.107}$$

и решаются при начальных данных

$$p_k(t_0) = p_k^0, \ q_k(t_0) = q_k^0,$$

где $q = \{q_k\}$ и $p = \{p_k\}$ суть l-мерные векторные канонические переменные, называемые обобщенными координатами и обобщенными импульсами соответственно. Функционал $H(p,q)$, называемый гамильтонианом, является инвариантом системы (3.107), т. е. на ее решении он не меняется со временем. Более того, дифференциальная гамильтонова система при опредeнных естественных условиях имеет бесчисленное количество инвариантов.

Изучение гамильтоновых систем и методов их решения в значительной степени основывается на аппарате внешней алгебры с дифференциальными и дискретными формами, который представляет относительно новый раздел дифференциальной геометрии, за последние десятилетия активно внедряемый в вычислительную математику. Огромный интерес к данному направлению вызван в значительной степени исследованиями динамических систем, теории хаоса и катастроф, начала которых заложены в трудах А. Пуанкаре.

С точки зрения приближенных алгоритмов трудности решения характерных в данной области задач заключаются в

необходимости численного интегрирования дифференциальных уравнений на длительных временных интервалах, с большим числом периодов (сотни и тысячи) моделируемых колебательных процессов. Традиционные вычислительные подходы здесь встречаются с большими трудностями, и проблема состоит в конструировании методов, которые на дискретном уровне максимально сохраняли свойства решений исходных дифференциальных задач.

3.11.1. Симплектические свойства гамильтоновых систем. Одной из главных особенностей гамильтоновых систем является симплектичность их потоков. Сам термин "симплектичность" был введен Г. Вейлем в 1939 г., а заключается он в сохранении объемов при соответствующих преобразованиях.

Определение 3.6. *Поток гамильтоновой системы — это отображение* $\varphi_t : U \to \mathbb{R}^{2l}$*, которое "продвигает" решение по времени* t*, т. е.*

$$\varphi_t(p_0, q_0) = (p(t, p_0, q_0),\ q(t, p_0, q_0)), \tag{3.108}$$

где $U \subset \mathbb{R}^{2l}$ *есть открытое множество, а* $p(t, p_0, q_0)$ *и* $q(t, p_0, q_0)$ *— решение системы* (3.107)*, соответствующее начальным данным* $p(0) = p_0, q(0) = q_0$*.*

Понятие симплектичности проще всего ввести на геометрических объектах — двумерных параллелограммах, определенных в пространстве $\mathbb{R}^{2l}$. Параллелограмм P можно представить образованным оболочкой двух векторов

$$\xi = \begin{pmatrix} \xi^p \\ \\ \xi^q \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta^p \\ \\ \eta^q \end{pmatrix}$$

в пространстве (p, q), причем $\xi^p, \xi^q, \eta^p, \eta^q \in \mathbb{R}^l$:

$$P = \{t\xi + s\eta \mid 0 \leqslant t \leqslant 1, 0 \leqslant s \leqslant 1\}.$$

В одномерном случае $m = 1$ определим *ориентированную площадь*

$$\omega(p) = \det \begin{pmatrix} \xi^p & \eta^p \\ \\ \xi^q & \eta^q \end{pmatrix} = \xi^p\eta^q - \xi^q\eta^p. \tag{3.109}$$

Для больших размерностей мы обобщаем это понятие путем суммирования ориентированных площадей проекций параллелограммов P на координатные плоскости (p_i, q_i):

$$\omega(\xi, \eta) = \sum_{i=1}^{l} \begin{vmatrix} \xi_i^p & \eta_i^p \\ \\ \xi_i^q & \eta_i^q \end{vmatrix} = \sum_{i=1}^{l} (\xi_i^p\eta_i^q - \xi_i^q\eta_i^p). \tag{3.110}$$

Данное выражение определяет билинейное отображение, действующее на векторы из $\mathbb{R}^{2l}$, которое в матричном обозначении имеет следующую форму:

$$\omega(\xi, \eta) = \xi^t J\eta, \quad J = \begin{bmatrix} 0 & I \\ \\ -I & 0 \end{bmatrix}, \tag{3.111}$$

где I есть единичная матрица порядка l.

Определение 3.7. *Линейное отображение* $A : \mathcal{R}^{2l} \to \mathcal{R}^{2l}$ *называется симплектическим, если*

$$A^T J\, A = J, \tag{3.112}$$

или, что эквивалентно, $\omega(A\xi, A\eta) = \omega(\xi, \eta)$ *для всех* $(\xi, \eta) \in \mathbb{R}^{2l}$.

Геометрическая иллюстрация симплектичности как свойства сохранения площади линейного отображения приводится на рис. 3.1.

Рис. 3.1. Симплектичность (сохранение площади) линейного отображения

В общем случае ($l > 1$) симплектичность означает, что сумма ориентированных площадей проекций параллелограмма P на координатные плоскости (p_i, q_i) та же самая, что и для преобразованных параллелограммов $A(P)$.

Перейдем теперь к нелинейным отображениям. Дифференцируемые функции могут быть аппроксимированы линейными отображениями, что позволяет сформулировать следующее понятие.

Определение 3.8. *Дифференцируемое отображение* $g\colon U \to \mathbb{R}^{2l}$, *где* $U \subset \mathbb{R}^{2l}$ *есть открытое множество, называется симплектическим, если якобиан* $g'(p, q)$ *является всюду симплектическим, т. е.*

$$g'(p,q)^T J\, g'(p,q) = J, \quad \textit{или} \quad \omega(g'(p,q)\xi,\ g'(p,q)\eta) = \omega(\xi, \eta).$$

Геометрическая интерпретация симплектичности для нелинейных отображений может быть представлена следующим образом. Рассмотрим двумерное подмногообразие M $2l$-мерного множества U и предположим, что дан образ $M = \psi(K)$ компактного множества $K \subset \mathbb{R}^2$, где $\psi(s,t)$ есть непрерывно дифференцируемая функция. Тогда многообразие M может рассматриваться как предел объединения небольших параллелограммов, порождаемых векторами

$$\frac{\partial \psi}{\partial s}(s,t)d\,s \quad \text{и} \quad \frac{\partial \psi}{\partial t}(s,t)d\,t.$$

Для каждого такого параллелограмма мы рассматриваем (как и выше) сумму ориентированных площадей его проекций на плоскости (p_i, q_i). Суммируя их затем по всем параллелограммам многообразия, мы в пределе получаем выражение

$$\Omega(M) = \iint\limits_K \omega\Big(\frac{\partial \psi}{\partial s}(s,t),\ \frac{\partial \psi}{\partial t}(s,t)\Big) d\,s\,d\,t. \tag{3.113}$$

Формулы преобразования двойных интегралов демонстрируют, что $\Omega(M)$ не зависит от параметризации ψ на M.

Теорема 3.19. *Если отображение $g\colon U \to \mathcal{R}^{2m}$ симплектчно на U, то оно сохраняет функционал (объем), т. е. соотношение*

$$\Omega(g(M)) = \Omega(M)$$

выполняется для всех двумерных многообразий M, которые могут быть представлены как образы непрерывно дифференцируемых функций ψ.

Если $m = 1$, то M является подмножеством $\mathbb{R}^2$, и мы можем выбрать $K = M$ с тождественным отображением Ψ. В этом случае $\Omega(M) = \iint\limits_M d\,s\,d\,t$ представляет площадь M. Следовательно, теорема 3.19 устанавливает, что все симплектические отображения, в том числе нелинейные, сохраняют ориентированные площади.

Мы теперь установим главный результат данного пункта. Отметим сначала, что, вводя обозначения $y = (p, q)$, мы можем записать гамильтонову систему (3.107) в форме

$$\dot{y} = J^{-1} \nabla H(y), \tag{3.114}$$

где J есть матрица из (3.112) и $\nabla H(y) = H'(y)^T = \left(\frac{\partial H}{\partial p_1}, ..., \frac{\partial H}{\partial p_l}, \frac{\partial H}{\partial q_1}, ..., \frac{\partial H}{\partial q_l}\right)^T$.

Теорема 3.20 (Пуанкаре, 1899). *Пусть $H(p, q)$ есть дважды непрерывно дифференцируемая функция на $U \subset \mathbb{R}^{2l}$. Тогда для каждого фиксированного t поток гамильтоновой системы φ_t есть симплектическое преобразование во всей области его определения.*

Важным свойством симплектических преобразований (установленным еще в знаменитой “Теореме X” Якоби, 1836 г.) является сохранение гамильтоновой структуры дифференциального уравнения. Такие преобразования с XIX в. получили название канонических. Следующая теорема показывает, что канонические и симплектические преобразования — одно и то же.

Теорема 3.21. *Пусть $\psi\colon U \to V$ есть изменение координат такое, что ψ и ψ^{-1} суть непрерывно дифференцируемые*

функции. Если ψ есть симплектическое преобразование, то гамильтонова система $\dot{y} = J^{-1}\nabla H(y)$ в новых переменных $z = \psi(y)$ принимает вид

$$\dot{z} = J^{-1}\nabla K(z), \quad K(z) = H(y). \tag{3.115}$$

И наоборот, если ψ преобразует любую гамильтонову систему в другую гамильтонову систему вида (3.115)*, то ψ является симплектическим отображением.*

3.11.2. Простейшие симплектические интеграторы. Так как симплектичность есть важное свойство гамильтоновых систем, естественно искать такие численные методы их интегрирования (отсюда появилось слово “интеграторы”), которые сохраняют это свойство. Пионерские работы по симплектическим интеграторам были проведены Вогеларе (1956 г.), Руттом (1983 г.) и Ф. Кангом (1985 г.).

Определение 3.9. *Численный метод называется симплектическим, если одношаговое отображение (численный поток)*

$$y_{h+1} = \Phi_h(y_n)$$

является симплектическим, когда оно применяется к гладкой гамильтоновой системе (3.115)*, т. е. якобиан $\Phi_h'(y)$ для всех шагов h удовлетворяет матричному равенству*

$$\Phi_h'(y)^T J \Phi_h'(y) = J.$$

Мы покажем симплектичность нескольких численных методов при использовании их к решению гамильтоновых систем в переменных $y = (p, q)$:

$$\dot{p} = -H_q(p,q), \;\; \dot{q} = H_p(p,q),$$

или

$$\dot{y} = J^{-1}\nabla H(y),$$

где H_p и H_q означают векторы-столбцы частных производных гамильтониона $H(p,q)$ по переменным $p = \{p_k\}, q = \{q_k\}$ соответственно.

Комбинирование явным и неявным методами Эйлера позволяет сконструировать два "сопряженных" симплектических алгоритма, носящих то же самое наименование.

Теорема 3.22 (Вогеларе, 1956). *Так называемые симплектические алгоритмы Эйлера*

$$p_{n+1} = p_n - h\, H_q(p_{n+1}, q_n), \qquad p_{n+1} = p_n - h\, H_q(p_n, q_{n+1}),$$

или

$$q_{n+1} = q_n + hH_p(p_{n+1}, q_n), \qquad q_{n+1} = q_n + hH_p(p_n, q_{n+1}), \tag{3.116}$$

являются симплектическими методами первого порядка.

Рассмотрим для примера только левый алгоритм (3.117). Дифференцирование по переменным p_n, q_n приводит к соотношению

$$\begin{bmatrix} I + h\,H_{q,p}^T & 0 \\ -h\,H_{p,p} & I \end{bmatrix} \begin{bmatrix} \dfrac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \end{bmatrix} = \begin{bmatrix} I & -h\,H_{q,q} \\ 0 & I + h\,H_{q,p} \end{bmatrix}, \tag{3.117}$$

где все матрицы частных производных $H_{q,p}, H_{p,p}, \ldots$ вычисляются в точке (p_{n+1}, q_n). Равенство (3.117) позволяет определить матрицу $A_n = \left[\frac{\partial(p_{n+1},q_{n+1})}{\partial(p_n,q_n)}\right]$ и непосредственно проверить условие симплектичности

$$A_n^T J\, A_n = J.$$

Заметим, что оба классических метода Эйлера, и явный, и неявный, не являются симплектическими.

Методы (3.116) для общего вида гамильтоновых систем являются неявными. Однако если гамильтониан допускает разделение на кинетическую и потенциальную энергию в виде

$$H(p, q) = T(p) + U(q), \tag{3.118}$$

где кинетическая энергия зависит только от обобщенного импульса, а потенциальная — только от координаты, то оба алгоритма являются явными. Интересно отметить, что имеются и более общие ситуации, когда симплектические методы Эйлера оказываются явными. Если

$$\frac{\partial\, H(p, q)}{\partial\, q_i}$$

не зависят от p_j для $j \le i$, то левый метод (3.117) при соответствующем упорядочении неизвестных будет явным и компо-

ненты p_{n+1} могут вычисляться последовательно друг за другом. Если же

$$\frac{\partial H(p,q)}{\partial p_i}$$

не зависят от q_i при $j \leqslant i$, то аналогично правый метод (3.117) становится явным.

Теорема 3.23 (Штермера—Верле). *Вычислительная схема*

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2} H_q(p_{n+1/2}, q_n), \\ q_{n+1} &= q_n + \frac{h}{2}(H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})), \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2} H_q(p_{n+1/2}, q_{n+1}), \end{aligned} \tag{3.119}$$

а также алгоритм

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2} H_q(p_n, q_{n+1/2}), \\ p_{n+1} &= p_n - \frac{h}{2}(H_q(p_n, q_{n+1/2}) + H_q(p_{n+1}, q_{n+1/2})), \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2} H_p(p_{n+1}, q_{n+1/2}) \end{aligned} \tag{3.120}$$

являются симплектическими методами второго порядка.

Доказательство симплектичности непосредственно следует из того, что обе схемы Штермера—Верле являются композициями двух симплектических методов Эйлера. А второй порядок алгоритмов (3.119), (3.120) показывается просто с помощью тейлоровских разложений.

Отметим, что методы Штермера—Верле являются неявными в общем случае.

Теорема 3.24. *Неявный метод средней точки*

$$y_{n+1} = y_n + hJ^{-1}H((y_{n+1} + y_n)/2) \tag{3.121}$$

является симплектическим алгоритмом второго порядка.

Если равенство (3.121) продифференцировать и определить матрицу $A_n = \left(\frac{\partial y_{n+1}}{\partial y_n}\right)$, то мы имеем аналогичное (3.117) соотношение

$$(I - \frac{h}{2}J^{-1}\nabla^2 H)A_n = (I + \frac{h}{2}J^{-1}\nabla^2 H).$$

Отсюда непосредственной подстановкой получаем свойство симплектичности для отображения A_n:

$$A_n^T J A_n = J. \tag{3.122}$$

Метод средней точки (3.121), как и методы Штермера—Верле, являются в общем случае неявными. Однако данные алгоритмы оказываются явными, если гамильтониан системы является сепарабельным, т. е. полная энергия представляется в виде суммы кинетической и потенциальной энергий

$$H(p, q) = E(p) + V(q),$$

где первое слагаемое зависит только от обобщенного импульса, а второе — только от обобщенной координаты.

Определение 3.10. *Численный метод, или интегратор, называется симметричным, если выполняется равенство*

$$\Phi_n(y) = \Phi_{-h}^{-1}(y).$$

Данное свойство означает, что при одновременной замене

$$y_n \longleftrightarrow y_{n+1}, \;\; h \longleftrightarrow -h$$

метод не изменяется. Другими словами, симметричный интегратор позволяет решать ОДУ как для прямого течения времени, так и для обратного.

Нетрудно проверить, что для всех симметричных методов погрешности численных решений имеют четный порядок. В частности, методы Штермера—Верле и средней точки являются симметричными интеграторами второго порядка.

3.11.3. Симплектические методы Рунге—Кутты. Рассмотренные выше симплектические методы Эйлера (3.116) и Штермера—Верле (3.119), (3.120), строго говоря, не относятся к классу схем Рунге—Кутты. Здесь нам потребуется сделать формальное обобщение данного семейства алгоритмов.

Определение 3.11. *Разделяющаяся форма системы ОДУ — это представление дифференциальных уравнений в виде*

$$\dot{y} = f(y, z), \;\; \dot{z} = g(y, z), \tag{3.123}$$

где y и z могут быть векторами различной размерности.

Отметим, что гамильтонова система (3.107) относится к классу разделяющихся ОДУ, в которой каждый из векторов p, q имеет размерность l.

При использовании методов Рунге—Кутты к решению ОДУ не существует принципиальных препятствий к тому,

чтобы в процессе вычислений новых приближений для векторов y и z применять различные значения коэффициентов, которые будем обозначать через $\hat{a}_{i,j}, \hat{b}_i$ и $\check{a}_{i,j}, \check{b}_i$ соответственно.

Определение 3.12. *Пусть $\hat{a}_{i,j}, \hat{b}_i$ и $\check{a}_{i,j}, \check{b}_i$ суть коэффициенты двух МРК. Разделенным, или разделяющимся (partitioned), методом Рунге—Кутты для решения системы ОДУ* (3.123) *будем называть m-стадийный алгоритм следующего вида:*

$$\begin{aligned}
\hat{k}_i &= f(y_0 + h\sum_{j=1}^{m} \hat{a}_{i,j}\hat{k}_j,\ z_0 + h\sum_{j=1}^{m} \check{a}_{i,j}\check{k}_j),\\
\check{k}_i &= g(y_0 + h\sum_{j=1}^{m} \hat{a}_{i,j}\hat{k}_j,\ z_0 + h\sum_{j=1}^{m} \check{a}_{i,j}\check{k}_j),\\
y_1 &= y_0 + h\sum_{i=1}^{m} \hat{b}_i\hat{k}_i,\ \ z_1 = z_0 + h\sum_{i=1}^{m} \check{b}_i\check{k}_i.
\end{aligned} \tag{3.124}$$

Вначале разделенные МРК исследовались для отдельных вычислений жестких и нежестких компонент решений, см. подробнее [64], однако позднее была обнаружена их актуальность при численном интегрировании гамильтоновых систем.

Приведенный ранее симплектический метод Эйлера (3.116) построен на комбинировании неявного и явного методов Эйлера, соответствующим формулам (3.124) с $\hat{b}_1 = 1, \hat{a}_{1,1} = 1$ и $\check{b}_1 = 1, \check{a}_{1,1} = 0$. Метод же Штермера—Верле (3.119) представим с помощью следующих двух МРК-схем Батчера

Табл. 3.10. Схема Батчера для разделенного представления

метода Штермера—Верле

0	0	0	1/2	1/2	0
1	1/2	1/2	1/2	1/2	0
	1/2	1/2		1/2	1/2

Теория методов Рунге—Кутты непосредственным образом переносится и на разделенные методы. В частности, из тейлоровских разложений следует, что если каждый из применяемых МРК имеет порядок $\hat{\gamma}$ или $\check{\gamma}$, то порядок итогового разделенного алгоритма не может превышать $\gamma = \min(\hat{\gamma}, \check{\gamma})$.

Алгебраические уравнения для коэффициентов МРК, обеспечивающие необходимые порядки аппроксимации, имеют обычный вид. В частности, условия обеспечения 1-го порядка записываются в форме

$$\sum_{i=1}^{m} \hat{b}_i = \sum_{i=1}^{m} \check{b}_i = 1. \tag{3.125}$$

Для разделенных алгоритмов 2-го порядка дополнительно должны выполняться равенства

$$\begin{aligned} &\sum_{i=1}^{m} \hat{b}_i \hat{c}_i = \sum_{i=1}^{m} \check{b}_i \check{c}_i = 1/2, \;\; \hat{c}_i = \check{c}_i, \\ &\sum_{i=j=1}^{m} \hat{b}_i \check{a}_{i,j} = \sum_{i=j=1}^{m} \check{b}_i \hat{a}_{i,j} = 1/2, \;\; \hat{c}_i = \sum_{i=1}^{m} a_{i,j}. \end{aligned} \tag{3.126}$$

Если же от метода (3.124) потребовать 3-й порядок точности, то в дополнение к условиям (3.125), (3.126) необходимо выполнение соотношений

$$\sum_{i=1}^{m}\hat{b}_i\hat{c}_i^2=\sum_{i=1}^{m}\check{b}_i\check{c}_i^2=1/3,\ \sum_{i,j=1}^{m}\hat{b}_i\hat{a}_{i,j}\hat{c}_j=\sum_{i,j=1}^{m}\check{b}_i\check{a}_{i,j}\check{c}_j=1/6,$$
$$\sum_{i,j=1}^{m}\hat{b}_i\check{a}_{i,j}\hat{c}_j=\sum_{i,j=1}^{m}\check{b}_i\hat{a}_{i,j}\check{c}_j=1/6.$$

Уравнения для вывода коэффициентов разделенных МРК более высоких порядков, получаемые из тейлоровских разложений с помощью специальных рядов, имеют достаточно сложный вид. Значительно упрощаются технические проблемы при конструировании коллокационных методов Рунге—Кутты, описанных выше в п. 3.3.1.

Идея колокационных методов — достаточно общая и универсальная при решении дифференциальных или интегральных уравнений. Заключается она в построении аппроксимирующего искомое решение *коллокационного многочлена*, коэффициенты которого ищутся по условию удовлетворения исходного уравнения на некотором наборе *коллокационных точек*. В применении к системам ОДУ эта методология рассматривалась еще в 1955 г. П. Хаммером и Дж. Холлингсвортом, а получила серьезное развитие в работах Гуиллоу, Саула (1969 г.) и Райта (1970 г.). Позже коллокационные алгоритмы Рунге—Кутты претерпели значительное обобщение и глубокий алегбраический анализ в исследованиях Ю. И.Кузнецова, итоги которых систематизированы в книге [40].

Определение 3.13. *Коллокационный многочлен метода*

Рунге–Кутты для решения задачи Коши

$$\dot{y}(t) = f(t, y),\ y(0) = y_0 \tag{3.127}$$

есть полином $u(t)$ *степени* m*, удовлетворяющий условиям*

$$\begin{aligned} &u(t_0) = y_0,\ \dot{u}(t_0 + c_i h) = f(t_0 + c_i h, u(t_0 + c_i h)), \\ &i = 1, ..., m,\ \ 0 \leqslant c_i \leqslant 1, \end{aligned} \tag{3.128}$$

и определяющий численное решение $y_1 = u(t_0 + h)$.

Отметим, что данное определение характеризует приближенное решение не как дискретный набор точек, а как непрерывную функцию $u(t)$ на интервале $[t_0, t_1]$ (аналогичным образом — и на остальных интервалах $[t_n, t_{n+1}]$). При $\bar{m} = 1$ коллокационные решения имеют вид

$$u(t) = y_0 + (t - t_0)k,\ \ k = f(t_0 + c_1 h, y_0 + h\, c_1 k)$$

и для значений $c_1 = 0, c_1 = 1, c_1 = 1/2$ представляют собой явный и неявный методы Эйлера, а также метод средней точки. Для коллокационных МРК приниципальное значение имеет следующее утверждение.

Теорема 3.25 (об эквивалентности коллокационных МРК). *Коллокационный метод Рунге—Кутты, определяемый соотношениями* (3.128)*, эквивалентен "классическому"* m*-стадийному МРК* (3.1) *с коэффициентами*

$$a_{i,j} = \int_0^{c_i} L_j(\tau) d\tau,\ b_i = \int_0^1 L_i(\tau) d\tau, \tag{3.129}$$

где $L_i(\tau)$ *есть многочлен Лагранжа*

$$L_i(\tau) = \prod_{l \neq i} (\tau - c_l)/(c_i - c_l).$$

Доказательство следует из того, что если $u(t)$ есть коллокационный многочлен, то из интерполяционной формулы Лагранжа (см. [12], [30]) следует равенство

$$\dot{u}(t_0 + \tau h) = \sum_{j=1}^{m} k_j L_j(\tau), \;\; k_j = \dot{u}(t_0 + c_j h).$$

Отсюда после интегрирования получаем соотношение

$$u(t_0 + c_i h) = y_0 + h \sum_{j=1}^{m} k_j \int_0^{c_i} L_j(\tau) d\tau,$$

с помощью которого и выходят формулы (3.1). □

Из приведенных рассуждений фактически следует и обратное утверждение: m-стадийный метод Рунге—Кутты с определяемыми в соответствии с (3.129) коэффициентами может быть интерпретирован как коллокационный алгоритм. Поскольку

$$\tau^{k-1} = \sum_{j=1}^{m} c_j^{k-1} L_j(\tau), \;\; k = 1, ..., m,$$

то соотношения (3.129) эквивалентны линейным системам

$$C(q) = \sum_{j=1}^{m} a_{i,j} c_j^{k-1} = c_i^k / k, \;\; k = 1, ..., q, \tag{3.130}$$

при всех i и $q = m$, а также

$$B(p) = \sum_{i=1}^{m} b_i c_i^{k-1} = k^{-1}, \;\; k = 1, ..., p, \tag{3.131}$$

при $p = m$. Нетрудно обнаружить, что соотношения (3.130), (3.131) совпадают с алгебраическими условиями сходимости МРК из 2-й теоремы Батчера (теорема 3.3 в п. 3.3.1). Последняя фактически может быть переформулирована следующим образом.

Теорема 3.26 (о сходимости коллокационных МРК). *Если условия* (3.131) *выполняются для некоторого* $p = \gamma \leqslant m$, *то коллокационный МРК имеет порядок* γ, *равный порядку используемой квадратурной формулы при вычислении интеграла в представлении решения*

$$y(t_1) = y_0 + \int_{t_0}^{t_1} f(t, y(t)) d\, t. \tag{3.132}$$

Кроме того, могут быть установлены оценки погрешности приближений не только для самого решения, но и для его производных.

Теорема 3.27 (о погрешности производных в коллокационных МРК). *Коллокационный полином* $u(t)$ *аппроксимирует с порядком* γ *точное решение задачи Коши* (3.127) *на всем сеточном интервале, т. е.*

$$||u(t) - y(t)|| \leqslant C_0 h^{\gamma+1}, \;\; t \in [t_0, t_0 + h],$$

при достаточно малом h. *Более того, для производных* $u^{(k)}(t)$ *при* $t \in [t_0, t_0 + h]$ *справедливы неравенства*

$$||u^{(k)}(t) - y^{(k)}(t)|| \leqslant C_k h^{\gamma+1-k}, \;\; k = 0, ..., m,$$

где постоянные $C_0, ..., C_m$ *не зависят от* h.

Коллокационные методы допускают одно естественное обобщение, освобождающее от обязательного условия непрерывности аппроксимирующего многочлена $u(t)$ в узлах сетки $t_1, t_2...$

Определение 3.14. *Пусть* $0 \leqslant c_2, ..., c_{m-1} \leqslant 1$, *а также* b_1 *и* b_m *— некоторые различные вещественные числа. Назовем разрывным коллокационным МРК алгоритм,который определяется с помощью многочлена* m*-й степени, удовлетворяющего следующим условиям:*

$$\begin{gathered} u(t_0) = y_0 - hb_1(\dot{u}(t_0) - f(t_0, u(t_0)), \\ \dot{u}(t_0 + c_i h) = f(t_0 + c_i h, u(t_0 + c_i h)), \;\; i = 2, ..., m-1, \\ y_1 = u(t_1) - hb_m(\dot{u}(t_1) - f(t_1, u(t_1)). \end{gathered} \tag{3.133}$$

Здесь величины в скобках в первом и третьем соотношениях — это невязки, или дефекты, исходных ОДУ, вычисленных на коллокационном многочлене $u(t)$.

Следующее утверждение устанавливает, что разрывные коллокационные МРК эквивалентны "классическим" неявным алгоритмам Рунге—Кутты.

Теорема 3.28 (об эквивалентности разрывных коллокационных МРК). *Разрывные коллокационные методы Рунге–Кутты* (3.133) *эквивалентны* m*-стадийным МРК*

(3.1), *имеющим коэффициенты* $c_1 = 0, c_m = 1$ *и удовлетворяющим соотношениям*

$$a_{i,1} = b_1,\ a_{i,m} = 0 \text{ для } i = 1, ..., m,$$

а также условиям (3.130), (3.131) *при* $q = m-2$ *и* $p = m-2$.

Наконец, относительно сходимости и оценок погрешности производных приближенных решений в разрывных коллокационных МРК можно сформулировать результат, аналогичный сформулированному в теоремах 3.26 и 3.27 для непрерывных коллокационных алгоритмов.

Теорема 3.29 (о порядке сходимости и оценках погрешностей производных для разрывных коллокационных МРК). *Разрывный коллокационный метод, соответствующий определению* 3.14, *имеет тот же порядок сходимости, что и квадратурные формулы, применяемые для вычисления интеграла* (3.132) *и основанные на аппроксимации решения* $y(t)$ *разрывным коллокационным многочленом* $u(t)$. *Кроме того, производные данного полинома удовлетворяют при достаточно малых* h *следующим оценкам для* $t \in [t_0, t_0 + h]$:

$$||u^{(k)}(t) - y^{(k)}(t)|| \leqslant C_k h^{m-1-k},\ k = 0, 1, ..., m-2,$$

где постоянные C_k *не зависят от* h.

Поскольку в теории численного интегрирования функций хорошо известны квадратурные формулы максимально высокого алгебраического порядка, т .е. точные на многочленах возможно большой степени, то рассмотренные теоремы дают непосредственный способ выбора коллокационных точек,

т. е. оптимальные значения коэффициентов $c_1, ..., c_m$, — это квадратурные узлы формулы Гаусса,а именно корни смещенного многочлена Лежандра степени m

$$L_m(t) = \frac{d^m}{d\,t^m}[t^m(t-1)^m],$$

которые расположены на отрезке $(0, 1)$ симметрично относительно его середины. Соответствующий m-стадийный алгоритм порядка $2m$ будем называть коллокационным методом Рунге—Кутты—Гаусса.

Как известно, существуют "квазиоптимальные" квадратурные формулы А. А.Маркова, в которых один или несколько узлов априори задаются из каких-то "высших" соображений, а остальные – оптимизируются по условиям максимальной алгебраической точности. Порядок таких формул оказывается равным $2m - l$, где l есть количество "замороженных" узлов.

Среди квадратур марковского типа наибольшую известность имеют формулы Радо (левая и правая), в которых фиксированным является один узел, совпадающий с левым или правым концом интервала интегрирования соответственно (в коллокационных МРК это означает $c_1 = 0$ или $c_m = 1$), а также формула Лобатто, использующая в числе своих квадратурных узлов и левый, и правый концы интервала (при этом полагается $c_1 = 0$ и $c_m = 1$).

Получаемые коллокационные алгоритмы будем называть МРК — Радо (I или II, т. е. левый или правый), а также МРК — Лобатто. Они имеют порядки точности $2m - 1$ и $2\,m - 2$, а оптимизируемые значения c_i являются корнями следующих

многочленов Лежандра:

$$\frac{d^{m-1}}{d\,t^{m-1}}[t^m(t-1)^{m-1}],\ \ \frac{d^{m-1}}{d\,t^{m-1}}[t^{m-1}(t-1)^m],$$
$$L_{m-2}(t) = \frac{d^{m-2}}{d\,t^{m-2}}[t^{m-1}(t-1)^{m-1}].$$

§ 3.12. Методы Нюстрема для решения ОДУ 2-го порядка

Задачи Коши для дифференциальных уравнений второго порядка

$$\ddot{y} = g(t, y, \dot{y}),\ \ y(t_0) = y_0,\ \dot{y}(t_0) = \dot{y}_0 \tag{3.134}$$

образуют важный класс проблем в силу того, что согласно второму закону Ньютона силы пропорциональны ускорению, т. е. вторым производным от перемещений. К задачам такого типа сводятся и некоторые гамильтоновы системы. Например, для гармонического осциллятора с сепарабельной функцией Гамильтона

$$H(p, q) = E(p) + V(q) = (p^2 + k^2q^2)/2 \tag{3.135}$$

движение описывается уравнениями

$$\dot{p} = -H_q = -k^2q,\ \ \dot{q} = H_p = p. \tag{3.136}$$

Вводя новую переменную $z = \dot{y}$ для первой производной, или скорости, задачу (3.134) сводим к эквивалентной распадающейся системе первого порядка, см. (3.123) и определение 3.11:

$$\dot y = z,\ \ \dot z = g(t,y,z). \tag{3.137}$$

При использовании m-стадийного разделяющегося метода Рунге—Кутты к решению уравнений (3.137) получаем формулы, см. (3.124) и определение 3.12:

$$\begin{aligned}
&\hat k_i = z_0 + h\sum_{j=1}^{m}\hat a_{i,j}\check k_j,\ \ i=1,...,m,\\
&\check k_i = g(t_0 + c_i h, y_0 + h\sum_{j=1}^{m}\hat a_{i,j}\hat k_j, z_0 + h\sum_{j=1}^{m}\hat a_{i,j}\check k_j),\\
&y_1 = y_0 + h\sum_{i=1}^{m}\hat b_i\hat k_i,\ \ z_1 = z_0 + h\sum_{i=1}^{m}\check b_i\check k_i.
\end{aligned} \tag{3.138}$$

Если в этих соотношениях выражения для $\hat k_i$ из первого равенства подставить в остальные и ввести обозначения

$$\bar a_{i,j} = \sum_{k=1}^{m}\hat a_{i,k}\check a_{k,j},\ \ \bar b_i = \sum_{k=1}^{m}\hat b_k\check a_{k,i},$$

то получим следующие *методы Нюстрема* (1925):

$$\begin{aligned}
&\bar k_i = g(t_0 + c_i h, y_0 + c_i h\dot y_0 + h^2\sum_{j=1}^{m}\bar a_{i,j}\bar k_j, \dot y_0 + h\sum_{j=1}^{m}\check a_{i,j}\bar k_j),\\
&y_1 = y_0 + h\dot y_0 + h^2\sum_{i=1}^{m}\bar b_i\bar k_i,\ \ \dot y_1 = \dot y_0 + h\sum_{i=1}^{m}\check b_i\bar k_i.
\end{aligned} \tag{3.139}$$

Многообразие коэффициентов $c_i, \bar b_i, \check b_i, \bar a_{i,j}$ и $\check a_{i,j}$ дает широкие возможности для построения соответствующих явных или неявных схем по условиям согласования или устойчивости.

Алгоритм Нюстрема называется методом порядка γ, если

$$y(t_0+h)-y_1=O(h^{\gamma+1}),\ \ \dot{y}(t_0+h)-\dot{y}_1=O(h^{\gamma+1}), \tag{3.140}$$

т. е. локальная погрешность имеет порядок $\gamma+1$.

Для актуального во многих приложениях специального случая

$$\ddot{y}=g(t,y), \tag{3.141}$$

где векторное поле сил не зависит от скорости $\dot{y}$, коэффициенты $\check{a}_{i,j}$ определять не требуется.

Примеры явных схем 4-го и 5-го порядков для данного случая приведены в таблице 3.11 (нижние треугольные матрицы здесь содержат значения $\bar{a}_{i,j}$ из 3.139).

Табл. 3.11. Схемы Батчера для явных одношаговых методов Нюстрема 4-го и 5-го порядков

Как отсюда видно, схемы Нюстрема в данном случае имеют существенное преимущество перед методами Рунге—Кутты: для $\gamma=4$ на каждом шаге требуется три вычисления правой части вместо четырех и для $\gamma=5$ нужно четыре вычисления функции g вместо шести в традиционном МРК.

Одношаговым методам Нюстрема, в силу их практической значимости, посвящено большое количество литературы, см. [64] и цитируемые там работы. Однако с теоретической точки зрения они особенного интереса не представляют, поскольку вопросы устойчивости и оценки погрешности численных решений в данном случае — это применение общей методологии

МРК к частному виду задачи Коши.

§ 3.13. МРК для сингулярно-возмущенных и дифференциально-алгебраических уравнений (ДАУ)

Актуальность и имеющиеся особенности данного класса уравнений уже отмечались в п. 1.1.2. Рассмотрим сингулярно-возмущенную задачу (СВЗ) индекса 1

$$y' = f(y,z),\ \ \varepsilon z' = g(y,z),\ \ \varepsilon \ll 1, \tag{3.142}$$

характеризуемую тем свойством, что матрица Якоби $g_z(y,z)$ является обратимой. Соответствующее приведенное ДАУ, т. е. при $\varepsilon = 0$, имеет вид

$$y' = f(y,z),\ \ 0 = g(y,z), \tag{3.143}$$

причем последнее равенство характеризует многообразие M, в котором находится искомое решение. Кроме того, по теореме о неявной функции оно определяет локально единственное решение $z = G(y)$, которое после подстановки в 1-е уравнение СВЗ дает так называемое *уравнение в пространстве состояний*

$$y' = f(y, G(y)). \tag{3.144}$$

Начальные значения считаем согласованными, т. е. $g(y_0, z_0) = 0$.

Рассмотрим сначала для решения приведенного ДАУ так называемый *метод ε-вложения*, заключающийся в том, что предварительно к решению СВЗ (3.142) применяется обычным образом неявный m-стадийный метод Рунге—Кутты:

$$\begin{aligned} y_{n,i} &= y_n + h\sum_{j=1}^{m} a_{i,j} f(y_{n,j}, z_{n,j}), \\ \varepsilon\, z_{n,i} &= \varepsilon\, z_n + h\sum_{j=1}^{m} a_{i,j} g(y_{n,j}, z_{n,j}), \\ y_{n+1} &= y_n + h\sum_{i=1}^{m} b_i f(y_{n,i}, z_{n,i}), \\ \varepsilon\, z_{n+1} &= \varepsilon\, z_n + h\sum_{i=1}^{m} b_i g(y_{n,i}, z_{n,i}), \end{aligned} \tag{3.145}$$

а затем в полученных формулах полагается $\varepsilon = 0$. Предполагая, что матрица $A = \{a_{i,j}\}$ имеет обратную $A^{-1} = \{\bar{a}_{i,j}\}$, из второго уравнения (3.145) имеем $hg(y_{n,i}, z_{n,i}) = \varepsilon\sum_{j=1}^{m} \bar{a}_{i,j}(z_{n,j} - z_n)$. После подстановки этого выражения в четвертое уравнение оказывается, что определение z_{n+1} не зависит от ε. Поэтому без дальнейших затруднений можно положить $\varepsilon = 0$ и получить вместо второго и четвертого уравнений (3.145) следующие:

$$0 = g(y_{n,i}, z_{n,i}), \quad z_{n+1} = \Big(1 - \sum_{i,j=1}^{m} b_i \bar{a}_{i,j}\Big) z_n + \sum_{i,j=1}^{m} b_i \bar{a}_{i,j} z_{n,j}.$$

Отметим, что скобка в последнем равенстве представляет собой величину

$$R(\infty) = 1 - \sum_{i,j=1}^{m} b_i \bar{a}_{i,j} = 1 - b^T A^{-1} e,$$

где $R(z)$ — функция устойчивости НМРК, см. выше определение 2.14 и п. 3.3.2.

Недостатком такого подхода является то, что обычно численное решение (y_{n+1}, z_{n+1}) не принадлежит многообразию $M\colon g(y,z) = 0$. Это, однако, легко поправить, заменив 4-е равенство (3.145) условием

$$g(y_{n+1}, z_{n+1}) = 0. \tag{3.146}$$

Тогда мы будем иметь не только $z_{n,j} = G(y_{n,j})$, но и $z_{n+1} = G(y_{n+1})$. Получаемый прием называется *методом пространства состояний*, поскольку он оказывается идентичен решению уравнения в пространстве состояний (3.144) тем же самым методом Рунге—Кутты.

Для рассматриваемых задач особую роль играют так называемые *жестко точные НМРК*, удовлетворяющие условиям

$$a_{m,i} = b_i, \;\; i = 1, ..., m. \tag{3.147}$$

В этом случае мы имеем $y_{n+1} = y_{n,m}, \;\; z_{n+1} = z_{n,m}$, т. е. полученное численное решение принадлежит многообразию M.

Таким образом, для жестко точных НМРК алгоритмы ε-вложения и пространства состояний эквивалентны. Можно показать, что A-устойчивые методы являются также L-устойчивыми, если они являются жестко устойчивыми.

Возникающие на практике дифференциально-алгебраические задачи зачастую представляются не "полуявной" системой (3.143), а в неявной форме с помощью уравнения

$$B\,u' = \varphi(u), \quad u = (y,z)^T, \quad \varphi = (f,g)^T, \quad y,z \in \mathbb{R}^N, \tag{3.148}$$

где матрица $B \in \mathbb{R}^{2N,2N}$ может быть сингулярной.

Если бы B была постоянной и невырожденной матрицей, то мы могли бы применить какой-либо из m-стадийных методов Рунге—Кутты к решению уравнения, полученного после умножения (3.148) на B^{-1}. Далее, умножая полученные формулы на B, мы приходим к соотношениям

$$\begin{aligned} B(u_{n,i} - u_n) &= h\sum_{j=1}^{s} a_{i,j}\varphi(u_{n,j}), \\ u_{n+1} &= (1 - \sum_{i,j=1}^{s} b_i\bar{a}_{i,j})u_n + \sum_{i,j=1}^{s} b_i\bar{a}_{i,j}u_{n,j}, \end{aligned} \tag{3.149}$$

где величины $\bar{a}_{i,j}$ обозначают, как и выше, элементы матрицы A^{-1}, обратной к $A = \{a_{i,j}\}$. Реализация полученного алгоритма требует на каждом шаге m-кратного решения СЛАУ с матрицей B.

Оказывается, что описанная процедура формально имеет смысл и в том случае, когда B есть сингулярная матрица. Чтобы в этом убедиться, проведем разложение B (например, методом исключения Гаусса с выбором ведущего элемента) и представим ее в виде

$$B = S \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} T,$$

где S и T — обратимые матрицы, а I есть единичная матрица, порядок которой равен рангу B. Подставив это разложение в (3.148), после умножения на S^{-1} и перехода к новым переменным

$$Tu = (\bar{y}, \bar{z})^T$$

получаем уравнение

$$\begin{bmatrix} \bar{y}' \\ 0 \end{bmatrix} = S^{-1}\varphi\Big(T^{-1} \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix}\Big) = \begin{pmatrix} \bar{f} \\ \bar{g} \end{pmatrix}.$$

Таким образом, уравнение (3.148) мы свели к приведенной системе вида (3.144). Начальные данные $u_0 = T^{-1}(\bar{y}_0, \bar{z}_0)$ для нее будут согласованы, если $\varphi(u_0)$ принадлежит образу линейного оператора, представленного матрицей B. А другими словами, векторы $\bar{y}_0, \bar{z}_0$ должны принадлежать многообразию, определяемому соотношением $\bar{g}(\bar{y}_0, \bar{z}_0) = 0$.

Локальные погрешности НМРК (3.145) получаются “почти” стандартным образом после подстановки в формулы алгоритма точных решений (при этом предварительно во втором и четвертом уравнениях проводится подстановка $g = \varepsilon\, z'$, с использованием (3.142), и выполняется сокращение на ε) и дальнейшего их разложения в ряд Тейлора:

$$
\begin{aligned}
y(t_n + c_i h) &= y(t_n) + h\sum_{j=1}^{m} a_{i,j} f(y(t_n + c_j h),\ \ z(t_n + c_j h)) + \psi_{n,j}^{(y)},\\
z(t_n + c_i h) &= z(t_n) + h\sum_{j=1}^{m} a_{i,j} z'(y(t_n + c_j h), z(t_n + c_j h)) + \psi_{n,j}^{(z)},\\
y(t_n + h) &= y(t_n) + h\sum_{i=1}^{m} b_i f(y(t_n + c_i h), z(t_n + c_i h)) + \psi_{n}^{(y)},\\
z(t_n + h) &= z(t_n + h\sum_{i=1}^{m} b_i z'(y(t_n + c_i h), z(t_n + c_i h)) + \psi_{n}^{(z)},
\end{aligned}
\tag{3.150}
$$

где b_i и c_i — веса и узлы квадратурной формулы, используемой при построении НМРК. Здесь мы делаем предположение, что

$$\psi_{n,j}^{(y)}, \psi_{n,j}^{(z)} = O(h^{q+1}),\ \ \psi_{n}^{(y)}, \psi_{n}^{(z)} = O(h^{p+1}),$$

где p есть порядок (b_i, c_i) — квадратурной формулы, а q — стадийный порядок НМРК (при этом предполагается, что выполнено алгебраической условие $C(q)$ 2-й теоремы Батчера, см. выше п. 3.3.1).

Если в рассматриваемой задаче Коши (3.142) $\varepsilon = 0$ и начальные данные согласованы, т. е. $g(y_0, z_0) = 0$, то при корректном сведении к уравнению в пространстве состояний (3.144) мы получаем классическое ОДУ и вопросы сходимости численного решения НМРК решаются стандартным образом, см. выше § 3.6. Аналогичная ситуация имеет место и при использовании жестко точного НМРК, см. выше алгоритм пространства состояний, что дает оценку глобальной ошибки

$$y(t_n) - y_n = O(h^p),\ z(t_n) - z_n = O(h^p),$$

где последнее соотношение следует из равенства $z = G(y)$.

Мы не будем останавливаться на методах Рунге—Кутты для решения СВЗ и ДАУ высших индексов, отошлем заинтересованного читателя к специальной монографии [64].

§ 3.14. Задачи к главе 3

3.14.1. Показать, что у двухстадийного МтР второго порядка

$$y_{n+1} = y_n + (3\,k_1 + k_2)/4, \;\; f'_n = \partial\, f(y_n)/\partial\, y,$$

$$k_1 = h(I - h\, f'_n/4)^{-1} f(y_n), \;\; k_2 = h(I - h\, f'_n/4)^{-1} f(y_n + k_1)$$

функция устойчивости имеет вид $R(z) = (1 + z/4)^2/(1 - z/4)^2$ и что метод является A-устойчивым, но не L-устойчивым.

3.14.2. Показать, что МтР первого порядка точности

$$D_n(y_{n+1} - y_n) = h\, f(t_n, y_n), \;\; D_n = I - a\, h\, f'(t_n, y_n)$$

является A-устойчивым при $a \geqslant 0.5$ и L-устойчивым при $a = 1$.

3.14.3. Показать, что МтР 1-го порядка точности

$$y_{n+1} = y_n + ak_1 + (1 - a)k_2,$$

$$D_n k_1 = h\, f(t_n, y_n), \;\; D_2 k_2 = k_1, \;\; D_n = I - a\, h\, f'(t_n, y_n)$$

является L-устойчивым при $0.5 - 0.25\sqrt{2} < a < 0.5 + 0.25\sqrt{2}$.

3.14.4. Показать, что явный двухстадийный МРК 2-го порядка

$$y_{n+1} = y_n + p_1 k_1 + p_2 k_2,$$

$$k_1 = h\, f(t_n, y_n),\ k_2 = h\, f(t_n + \beta h, y_n + \beta k_1)$$

имеет локальную ошибку

$$\delta_n = \frac{2 - 3\beta}{12} h^3 f'' f + \frac{h^3}{6} (f')^2 + O(h^4).$$

Подсказка: использовать условия второго порядка точности

$$p_1 + p_2 = 1,\ \beta p_2 = 0.5.$$

3.14.5. Доказать, что методы Штермера—Верле (3.120), (3.121) являются симплектическими алгоритмами второго порядка.

3.14.6. Вывести формулы коллокационного МРК—Радо I второго порядка и доказать, что этот метод является симплектическим.

3.14.7. Вывести формулы коллокационного МРК—Лобатто и показать, что данный метод — симплектический.

3.14.8. Показать, что методы Штермера—Верле являются симметричными интеграторами.

3.14.9. Доказать, что для жестко точных m-стадийных НМРК выполняется равенство $y_{n+1} = y_{n,m}$.

3.14.10. Показать, что условие $B(p)$ во 2-й теореме Батчера

$$\sum_{j=1}^{m} b_j c_j^{q-1} = q^{-1},\ q = 1, ..., p,$$

эквивалентно $(p-1)$-й алгебраической степени точности вычисления интеграла $\int\limits_{t_n}^{t_{n+1}} f(t)dt$.

3.14.11. Доказать, что из условий $C(m)$ и $B(m+\xi)$ 2-й теоремы Батчера следует условие $D(\xi)$.

Подсказка: величины $d_j^{(q)} = \sum\limits_{i=1}^{m} b_i c_i^{q-1} a_{i,j} - b_j(1-c_j^q)/q$ являются решениями СЛАУ

$$\sum_{j=1}^{m} d_j^{(q)} c_j^{k-1} = 0 \quad \text{для} \quad k = 1, ..., m \quad \text{и} \quad q = 1, ..., \xi..$$

3.14.12. Показать, что из условий $D(m)$ и $B(m+\eta)$ 2-й теоремы Батчера следует условие $C(\eta)$.

Подсказка: величины $e_i^{(q)} = \sum\limits_{j=1}^{m} a_{i,j} c_j^{q-1} - c_i^{(q)}/q$ является решениями СЛАУ

$$\sum_{i=1}^{m} b_i c_i^{k-1} e_i^{(q)} = 0 \quad \text{для} \quad k = 1, ..., m;\ q = 1, ..., \eta$$

.

3.14.13. Выписать и исследовать условия согласования для двухстадийного ЯМРК.

3.14.14. Показать, что никакой ЯМРК не может быть A-устойчив.

3.14.15. Как с использованием двух НМРК — методов средней точки и трапеций — повысить точность результата и получить гарантированную оценку ошибки?

3.14.16. Определить интервал устойчивости для ЯМРК второго порядка.

3.14.17. Показать, что методы типа Розенброка не являются B-устойчивыми.

3.14.18. Доказать, что метод трапеций не является B-устойчивым.

3.14.19. Показать, что симплектические методы Эйлера оказываются явными в случае разделяющегося гамильтониана.

3.14.20. Доказать второй порядок методов Штермера—Верле.

3.14.21. Доказать, что симплектический метод средней точки является явным в случае сепарабельности гамильтониана.

3.14.22. Построить двухстадийный метод Нюстрема для решения ОДУ 2-го порядка.

3.14.23. Вывести формулу локальной ошибки для неявного метода средней точки.

3.14.24. Вывести формулу локальной ошибки для неявного метода трапеций.

3.14.25. Построить функцию устойчивости для неявного метода средней точки и определить его интервал устойчивости.

3.14.26. Построить функцию устойчивости для неявного метода трапеций и определить его интервал устойчивости.

3.14.27. Построить функцию устойчивости для одностадийного МтР второго порядка и определить его интервал устойчивости.

0	$\gamma = 4$		
1/2	1/8		
1	0	1/2	
$\bar{b}_i$	1/6	1/3	0
$\check{b}_i$	1/6	4/6	1/6

0	$\gamma = 5$			
1/5	1/50			
2/3	-1/27	7 7/27		
1	3/10	-2/35	3/35	
$\bar{b}_i$	14/336	100/336	54/336	0
$\check{b}_i$	14/336	125/336	162/336	35/336

Глава 4

МНОГОШАГОВЫЕ МЕТОДЫ (ММ)

Если одношаговые методы строятся практически одинаково для постоянного и переменного шагов h_n, то далее в этой главе нам будет гораздо проще рассматривать равномерную сетку. Существуют варианты многошаговых методов и с переменными шагами, но они менее популярны, в силу своей повышенной технической сложности, и мы на них останавливаться не будем. В данной главе рассматриваются только так называемые *линейные многошаговые методы* (ЛММ), т. е. такие, в которых для случая систем линейных ОДУ получаемая система уравнений для коэффициентов схемы также является линейной. В силу этого слово "линейные" мы будем опускать и употреблять соответственно укороченное обозначение ММ.

Напомним, что построение ЛММ повышенной точности основывается на использовании нескольких значений решения с предыдущих, уже рассчитанных, шагов алгоритма. В силу этого такие методы принципиально являются неоднородными, так как реализация первых нескольких шагов должна осуществляться по какой-либо другой вычислительной схеме. Очевидная альтернатива здесь — это или применение одношаговых МРК, или многошаговых алгоритмов низкого поряд-

ка (возможно, с использованием более мелкой сетки). Однако этим вопросам мы в данной главе уделяем относительно меньшее внимание, акцентируя внимание на изучение главным образом классических вопросов аппроксимации, устойчивости и сходимости (линейной и нелинейной) именно в условиях многошаговости алгоритмов.

В этой же главе мы рассмотрим и так называемые общие линейные методы, объединяющие идеи многостадийных и многошаговых алгоритмов, среди которых наибольшую известность имеют многошаговые методы Рунге—Кутты. Будет уделено также внимание и дифференциально-алгебраическим уравнениям, или сингулярно-возмущенным задачам, имеющим особую актуальность в проблемах теории и практики управления сложными системами.

§ 4.1. Явные методы Адамса

Как и МРК, многошаговые методы выводятся из интегрального представления решения ОДУ. Однако если ранее из соотношения

$$y(t_{n+1}) = y(t_n) + \int\limits_{t_n}^{t_{n+1}} f(t, y(t))dt \tag{4.1}$$

численные схемы конструировались путем аппроксимации интеграла с помощью квадратурных формул, использующих вспомогательные узлы только из интервала $[t_n, t_{n+1}]$, то теперь будем строить ММ на основе интерполяционного приближения подынтегральной функции по ее значениям в k точках

$f_j = f(t_j, y_j),\ j = n-k+1, ..., n$, которые предполагаются уже известными. Эта идея очень естественна — использовать хотя бы частично имеющуюся информацию.

Если применить интерполяционный многочлен Ньютона для равномерной сетки [30], выражаемый с помощью конечных разностей:

$$L_{k-1}(t) = L_{k-1}(t_n + s\,h) = \sum_{j=0}^{k-1}(-1)^j \binom{-s}{j} \nabla^j f_n,$$

$$\nabla^0 f_n = f_n,\ \nabla^{j+1} f_n = \nabla^j f_n - \nabla^j f_{n-1},$$

$$\binom{-s}{j} = \frac{(-1)^j}{j!}(s-2)(s-1)...(s+j-3),$$

а потом точно вычислить интеграл, то вместо точного равенства (4.1) получаем для разных k приближенные формулы вида

$$y_{n+1} = y_n + h\sum_{j=0}^{k-1}\gamma_j \nabla^j f_n,\ \gamma_j = (-1)^j \int_0^1 \binom{-s}{j} d\,s, \qquad (4.2)$$

которые составляют семейство *явных методов Адамса* (появившихся задолго до МРК, еще в 1855 г., и иногда называемых *методами Адамса—Башфорта*). Коэффициенты явных методов Адамса удовлетворяют рекуррентному соотношению (Хенричи, 1962 г.)

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \frac{1}{3}\gamma_{m-2} + ... + \frac{1}{m+1}\gamma_0 = 1. \qquad (4.3)$$

Первые значения последовательности γ_j, например, равны $1,\ 1/2,\ 5/12,\ 3/8,\ 251/720, ...$

Формула (4.3) получается с помощью эйлерового метода производящих функций. Обозначим через $G(t)$ ряд

$$G(t) = \sum_{j=0}^{\infty} \gamma_j t^j. \tag{4.4}$$

Из определения γ_j в (4.2) и применения квадратурных формул имеем

$$G(t) = \sum_{j=0}^{\infty} \binom{-s}{j} d\, s = \int_0^1 (1-t)^{-s} d\, s = -\frac{t}{(1-t)\log(1-t)}.$$

Непосредственно отсюда следует равенство

$$\Big(1 + \frac{1}{2}t + \frac{1}{3}t^3 + ...\Big)(\gamma_0 + \gamma_1 t + ...) = 1 + t + t^2 + ...$$

Сравнивая в его левой и правой частях коэффициенты при t^m, приходим к (4.3).

Из формулы (4.2) при $k = 1$ мы имеем явную схему Эйлера, а для $k = 2, 3, 4$ явные методы Адамса записываются в следующем виде:

$$k = 2\colon\ y_{n+1} = y_n + \frac{h}{2}(3\, f_n - f_{n-1}),$$

$$k = 3\colon\ y_{n+1} = y_n + \frac{h}{12}(23\, f_n - 16\, f_{n-1} + 5\, f_{n-1}),$$

$$k = 4\colon\ y_{n+1} = y_n + \frac{h}{24}(55\, f_n - 59\, f_{n-1} + 37\, f_{n-1} - 9\, f_{n-2}). \tag{4.5}$$

Естественно, данные формулы неприменимы к вычислению стартовых значений y_1, y_2 и т. д. (в зависимости от числа

используемых шагов k), и для их нахождения надо применять одношаговые методы.

Очевидно, что для всех этих алгоритмов сумма знакопеременных множителей в скобках равна единице, а сумма их модулей — значительно больше.

Интуитивно ясно, что последний факт приведет к неустойчивости алгоритмов, поскольку при наличии погрешностей, или возмущений, в значениях $f_n, f_{n-1}, ...$ они будут накапливаться при вычислении y_{n+1}. В подтверждение можно привести следующие простые рассуждения. Тогда k-шаговый явный метод Адамса реализуется по формуле

$$y_{n+1} = y_n + h(\beta_0 f_n + ... + \beta_k f_{n-k}).$$

Рассмотрим теперь реализацию данного алгоритма с возмущениями, обусловленными, например, неточностью выполнения арифметических операций при наличии машинных округлений. Пусть вместо величин $y_n, f_n, ..., f_{n-k}$ нам известны их возмущенные значения $\tilde{y}_n, \tilde{f}_n, ..., \tilde{f}_{n-k}$. Тогда новое возмущенное решение $\tilde{y}_{n+1}$ определяется как

$$\tilde{y}_{n+1} = \tilde{y}_n + h(\beta_0 \tilde{f}_n + ... + \beta_k \tilde{f}_{n-k}) + \delta_n,$$

где δ_n — погрешность машинной реализации данной формулы. Отсюда для ошибок $\Delta_n = y_n - \tilde{y}_n$ получаем соотношения

$$|\Delta_{n+1}| \leqslant |\Delta_n| + h\beta|\delta f_n| + |\delta_n|, \tag{4.6}$$

$$\delta f_n = \max_k\{|f_{n-k} - \tilde{f}_{n-k}|\}, \;\; \beta = \textstyle\sum_{i=0}^{k} |\beta_k|.$$

Таким образом, при прочих равных условиях накопление ошибок определяется суммой модулей коэффициентов $\beta = |\beta_0| + ... + |\beta_k|$, которую можно назвать коэффициентом усиления ошибки. Более тонкий анализ устойчивости требует анализа корней характеристических многочленов, на чем мы остановимся позже.

§ 4.2. Неявные методы Адамса

В явных методах Адамса фактически используется экстраполяция подынтегральной функции на интервале $[t_n, t_{n+1}]$. Если для повышения точности (и устойчивости) к ранее рассчитанным точкам добавить новую точку — t_{n+1}, то после интегрирования нового интерполяционного многочлена

$$L_k(t) = L_k(t_n + s\,h) = \sum_{j=0}^{k} (-1)^j \binom{-s+1}{j} \nabla^j f_{n+1}$$

получим семейство *неявных методов Адамса* (иногда называемых *методами Адамса–Мултона*)

$$y_{n+1} = y_n + h \sum_{j=0}^{k} \hat{\gamma}_j \nabla^j f_{n+1},\ \hat{\gamma}_j = (-1)^j \int_0^1 \binom{-s+1}{j} d\,s,$$

где новые коэффициенты $\hat{\gamma}_j$ при $m \geqslant 1$ удовлетворяют рекуррентному соотношению

$$\hat{\gamma}_m + \frac{1}{2}\hat{\gamma}_{m-1} + \frac{1}{3}\hat{\gamma}_{m-2} + ... + \frac{1}{m+1}\hat{\gamma}_0 = 0,\ \hat{\gamma}_0 = 1. \qquad (4.7)$$

Данный результат устанавливается по аналогии с (4.3) с помощью производящей функции, представляемой рядом, который похож на (4.4):

$$\hat{G}(t) = \sum_{j=0}^{\infty} \hat{\gamma}_j t^j. \qquad (4.8)$$

Первые значения этой последовательности равны: $1,\ -1/2,\ -1/12,\ -1/24, ...$ Для $k = 0, 1$ при этом получаем

неявный метод Эйлера и *метод трапеций*, а для $k = 2, 3$ — алгоритмы

$$y_{n+1} = y_n + h(5\, f_{n+1} + 8\, f_n - f_{n-1})/12,$$
$$y_{n+1} = y_n + h(9\, f_{n+1} + 19\, f_n - 5\, f_{n-1} + f_{n-2})/24. \tag{4.9}$$

Данные соотношения представляют собой нелинейные уравнения, которые можно решать, например, с помощью простой итерации

$$y^m_{n+1} = y_n + h(\beta_k f(t_{n+1}, y^{m-1}_{n+1}) + \beta_{k-1} f_n + ... + \beta_0 f_{n-k+1}). \tag{4.10}$$

Естественно, что при достаточно малых h этот итерационный процесс будет сходиться. В качестве начального значения y^0_{n+1} можно взять какую-то близкую величину, например, y_n. Если же за начальное приближение принять "прогностический" результат $\hat{y}_{n+1}$, получаемый с помощью соответствующего явного метода Адамса (4.2) (возможно, и более низшего порядка), а затем его уточнить (скорректировать) с помощью одной итерации предыдущей формулы, то получим алгоритм, называемый *методом прогноза и коррекции*, или методом *предиктор-корректор*. Такие алгоритмы иногда называются методами Адамса—Башфорта—Мултона.

Понятно, что с увеличением числа шагов k неустойчивость, т. е. величина β в (4.4), методов Адамса будет усугубляться как для неявных, так и явных вариантов.

Для гарантии точности, конечно, целесообразно сделать несколько итераций (с контролем сходимости), причем в це-

лях их ускорения можно применять метод Ньютона. В качестве примера запишем метод предиктор-корректор

$$\begin{aligned} \hat{y}_{n+1} &= y_n + h\, f_n, \\ y_{n+1} &= y_n + h[f_n + f(t_{n+1}, \hat{y}_{n+1})]/2, \end{aligned} \tag{4.11}$$

полностью совпадающий с явным двухстадийным МРК второго порядка.

За счет уменьшения экстраполяционного вклада при вычислении интеграла в (4.1) неявные методы Адамса являются более устойчивыми, чем явные. Конкретно это отражается в уменьшении величины β — суммы модулей коэффициентов. Например, в неявных формулах (4.5) она равна $7/6$ и $17/12$, а в явных формулах (4.3) — 2 и $21/6$ для $k = 2, 3$ соответственно.

§ 4.3. Многошаговые методы Нюстрема и Милна

Отправляясь от другого вида интегрального представления

$$y(t_{n+1}) = y(t_{n-1}) + \int\limits_{t_{n-1}}^{t_{n+1}} f(t, y(t))dt, \tag{4.12}$$

после интерполяции подынтегральной функции по уже посчитанным значениям $f_n, f_{n-1}, ...$, можно получить новые формулы ММ. При этом явные k-шаговые алгоритмы записываются в форме

$$y_{n+1} = y_{n-1} + h\sum_{j=1}^{k-1} \varkappa_j \nabla^j f_n$$

и называются *методами Нюстрема*, коэффициенты которых определяются следующим образом:

$$\varkappa_j = (-1)^j \int_{-1}^{1} \binom{-s}{j} d\, s. \tag{4.13}$$

Первые члены этой последовательности такие: $2,\ 0,\ \dfrac{1}{3},\ \dfrac{1}{3}, \ldots$ При $k = 1$ получаем так называемое *правило средней точки*:

$$y_{n+1} = y_{n-1} + 2\, h\, f_n, \tag{4.14}$$

заключающееся в аппроксимации интеграла формулой центрального прямоугольника. Случай $k = 2$ нового ничего не дает, поскольку $\varkappa_1 = 0$, а при $k = 3$ имеем схему

$$y_{n+1} = y_{n-1} + h\Big(\frac{7}{3} f_n - \frac{2}{3} f_{n-1} + \frac{1}{3} f_{n-2}\Big). \tag{4.15}$$

Если же для интерполяции подынтегральной функции в (4.8) дополнительно использовать значение f_{n+1}, то приходим к семейству неявных *методов Милна*:

$$y_{n+1} = y_{n-1} + h\sum_{j=0}^{k} \hat{\varkappa}_j \nabla^j f_{n+1}, \tag{4.16}$$

коэффициенты которых определяются как

$$\hat{\varkappa}_j = (-1)^j \int_{-1}^{1} \binom{-s+1}{j} d\,s, \tag{4.17}$$

а их первые значения равны $2,\ -2,\ \frac{1}{3},\ 0,\ -\frac{1}{90}, \ldots$ В частных случаях мы имеем: $k = 0$ — неявный метод Эйлера с шагом $2h, k = 1$ — совпадение с соответствующим явным методом Нюстрема (правило средней точки), $k = 2$ — популярный метод, основанный на квадратурной формуле Симпсона:

$$y_{n+1} = y_{n-1} + \frac{h}{3}(f_{n+1} + 4\,f_n + f_{n-1}). \tag{4.18}$$

Случай $k = 3$ опять приводит к схеме (4.18), поскольку $\hat{\varkappa}_2 = 0$, а при $k = 4$ получаем схему

$$y_{n+1} = y_{n-1} + h(29\,f_{n+1} + 144\,f_n + 4\,f_{n-1} + 4\,f_{n-2} - f_{n-3})/90.$$

Поскольку использование дополнительных значений $f_{n-2}, f_{n-3}, \ldots$ для аппроксимации интеграла в (4.12) формально можно рассматривать как обобщение правила Симпсона, то получаемые при этом схемы называют методами Милна—Симпсона.

Несложно показать, что определяемые в (4.13) и (4.17) коэффициенты $\varkappa_j$ и $\hat{\varkappa}_j$ связаны с величинами γ_j и $\hat{\gamma}_j$ из (4.2) и (4.7) следующими простыми соотношениями при $j \geqslant 0$:

$$\varkappa_j = 2\gamma_j - \gamma_{j-1}, \quad \hat{\varkappa}_j = 2\hat{\gamma}_j - \hat{\gamma}_{j-1},$$

где полагается $\gamma_{-1} = \hat{\gamma}_{-1} = 0$.

На основе комбинирования явных и неявных аппроксимаций интегрального уравнения (4.12) можно также строить

различные схемы типа прогноза и коррекции с одной, двумя или большим числом итераций на каждом шаге.

§ 4.4. Формулы дифференцирования назад (ФДН)

Все рассмотренные выше схемы строились на основе применения квадратурных формул для приближенного вычисления интеграла от функции f в (4.1) или (4.8). Альтернативный подход заключается в использовании численного дифференцирования, основанном на интерполировании искомого решения $y(t)$.

Пусть нам известно $k+1$ значение (приближенное) решения ОДУ: $y_n, y_{n-1}, ..., y_{n-k+1}$. Тогда по точкам (t_j, y_j) строим полином Ньютона с использованием разностей назад:

$$P_k(t) = P_k(t_n + s\,h) = \sum_{j=0}^{k} (-1)^j \binom{-s+1}{j} \nabla^j y_{n+1}.$$

Определим теперь неизвестное значение y_{n+1} так, чтобы многочлен $P_k(t)$ удовлетворял исходному дифференциальному уравнению по крайней мере в одном узле сетки, т. е. для некоторого r

$$P_k'(t_{n+1-r}) = f(t_{n+1-r}, y_{n+1-r}). \tag{4.19}$$

В этом случае для $r = 1$ получим *явные ФДН*. При $k = 1$ и $k = 2$ они эквивалентны явному методу Эйлера и правилу средней точки соответственно, а при $k = 3$ имеем

$$\frac{1}{3} y_{n+1} + \frac{1}{2} y_n - y_{n-1} + \frac{1}{6} y_{n-2} = h\, f_n. \tag{4.20}$$

Однако эта формула неустойчива, как и все остальные явные алгоритмы ФДН для $k > 3$. Действительно, формулу (4.20) можно переписать в виде

$$y_{n+1} = 3\,y_{n-1} - \frac{3}{2}y_n - \frac{1}{2}y_{n-2} + 3\,h\,f_n,$$

где сумма модулей коэффициентов $\beta = \sum |\beta_i|$ (коэффициент усиления ошибки) в правой части равна 5.

Если же в "коллокационном" уравнении (4.19) положить $r = 0$, то получим *неявные формулы дифференцирования назад*

$$\sum_{j=0}^{k} \check{\varkappa}_j \nabla^j y_{n+1} = h\,f_{n+1} \tag{4.21}$$

с множителями $\check{\varkappa}_j = (-1)^j \dfrac{d}{d\,s}\binom{-s+1}{j}\Big|_{s=1}$, которые после дифференцирования биномиального коэффициента приводятся к виду $\check{\varkappa}_0 = 0$ и $\check{\varkappa}_j = j^{-1}$ при $j \geqslant 1$. Отсюда получаем популярное семейство *неявных ФДН*, введенных в работе Кертиса и Хиршфельдера (1952 г.), широко применяемых для решения жестких ОДУ:

$$\sum_{j=1}^{k} \frac{1}{j} \nabla^j y_{n+1} = h\,f_{n+1}. \tag{4.22}$$

Для $k = 1$ отсюда следует неявный метод Эйлера, а для $k = 2$ имеем

$$\frac{3}{2}y_{n+1} - 2\,y_n + \frac{1}{2}y_{n-1} = h\,f_{n+1}. \tag{4.23}$$

Переписывая эту формулу в виде

$$y_{n+1} = \frac{4}{3}y_n - \frac{1}{3}y_{n-1} + \frac{2h}{3}f_{n+1},$$

получаем значение коэффициента подавления (а точнее — усиления) ошибки $\beta = \frac{5}{3}$, т. е. значительно меньше, чем в (4.20). Для полноты изложения приведем еще неявные формулы, получаемые из (4.22) при $k = 3$ и $k = 4$:

$$11\,y_{n+1} - 18\,y_n + 9\,y_{n-1} - 2\,y_{n-2} = 6\,h\,f_{n+1},$$

$$25\,y_{n+1} - 48\,y_n + 36\,y_{n-1} - 16\,y_{n-2} + 3\,y_{n-3} = 12\,h\,f_{n+1}.$$

Отметим, что в отличие от рассмотренных выше в §§ 4.1 — 4.3 многошаговых методов, формулы дифференцирования назад используют на каждом шаге вычисление правой части ОДУ только в одной точке. Такие многошаговые методы называются *одноопорными*, и им посвящен далее § 4.10.

§ 4.5. Локальные и глобальные погрешности ММ

Рассмотрим сеточное многоточечное уравнение общего вида

$$\alpha_k y_{n+k} + ... + \alpha_0 y_n = h(\beta_k f_{n+k} + ... + \beta_0 f_n), \tag{4.24}$$

в котором предполагается $\alpha_k \neq 0,\ |\alpha_0| + |\beta_0| > 0$. Перепишем его в операторной форме

$$L^h(y^h, f^h, h) \equiv \sum_{j=0}^{k} (\alpha_j y_{n+j} - h\beta_j f_{n+j}) = 0, \tag{4.25}$$

где $y^h = \{y_{n+j}\}$ и $f^h = \{f(t_{n+j}, y_{n+j})\}$ — сеточные векторы размерности $k+1$, составленные из значений численного решения и определяемых через них функций f при $j = 0, ..., k$.

Пусть теперь векторы $(y)^h = \{y(t_{n+j})\}$ и $(f)^h = \{f(t_{n+j}, y(t_{n+j}))\}$ имеют компонентами точные значения решения $y(t)$ и правой части $f(t)$ системы ОДУ для аргументов $t_{n+j}, j = 0, 1, ..., k$. Тогда величина

$$\psi^h_{n+k} = L^h_n(y^h, f^h, h) = \sum_{j=0}^{k} [\alpha_j y(t_{n+j}) - h\beta_j f(t_{n+j}, y(t_{n+j}))] \tag{4.26}$$

представляет собой локальную погрешность k-шагового метода в точке $t = t_n$. Обозначим далее через $\bar{y}_{n+k}$ и $\bar{f}_{n+k}$ величины, определяемые из аналога уравнения (4.24) при точных значениях остальных членов: $y(t_n), ..., y(t_{n+k-1})$, $f(t_n, y(t_n)), ..., f(t_{n+k-1}, y(t_{n+k-1}))$. Вычитая получаемое при этом уравнение из уравнения (4.26) и применяя теорему о среднем

$$f(t_{n+j}, y(t_{n+k})) - f(t_{n+k}, \bar{y}_{n+k}) = \frac{\partial f}{\partial y}(t_{n+k}, \eta)[y(t_{n+k}) - \bar{y}_{n+j}],$$

$\eta \in [y(t_{n+k}), \bar{y}_{n+k}]$,

мы приходим к равенству

$$y(t_{n+k}) - \bar{y}_{n+k} = \left[\alpha_k I - h\beta_k \frac{\partial f}{\partial y}(t_{n+k}, \eta)\right]^{-1} \psi^h_{n+k}, \tag{4.27}$$

которое иногда принимается за другое определение локальной ошибки. Очевидно, что такие два понятия эквивалентны

по порядку, если величина квадратной скобки в (4.27) ограничена снизу не зависящей от h константой.

Необходимо отметить, что в определении локальной погрешности $\psi^h = \{\psi^h_{n+k}\}$ из (4.26) имеется произвол, поскольку в соотношениях (4.24), (4.25) коэффициенты α_j и β_j определены с точностью до множителя. Естественно выбрать такую их нормировку, чтобы левая и правая части равенства (4.24), после их деления на h, аппроксимировали величины $\dot{y}(t)$ и $f(t)$ соответственно. Для этого достаточно положить $\beta_k + ... + \beta_0 = 1$, хотя формально это не обязательно. Далее будем считать, что коэффициент $\alpha_k = 1$ или $\alpha_k = O(1)$. В этом случае будем говорить, что порядок ММ равен γ, если $||\psi^h|| = 0(h^{\gamma+1})$. Заметим, что левая часть (4.27) от нормировки α_j, β_j не зависит. Если y, f суть векторные функции, то $\dfrac{\partial f}{\partial y}(t_n, \eta)$ — матрица Якоби, строки которой вычислены, возможно, при разных значениях, принадлежащих $[y(t_n), y_n]$.

Теорема 4.1 (об условиях порядка ММ). *Многошаговый метод имеет порядок γ тогда и только тогда, когда выполняется одно из следующих эквивалентных условий*:

а) $\sum\limits_{j=1}^{k}\alpha_j = 0$ и $\sum\limits_{j=0}^{k}\alpha_j j^q = \sum\limits_{j=0}^{k}\beta_j j^{q-1}$ при $q = 1, ..., \gamma$,

б) $\rho(e^h) - h\sigma(e^h) = O(h^{\gamma+1})$ при $h \to 0$,

в) $\dfrac{\rho(\theta)}{\log\theta} - \sigma(\theta) = O((\theta-1)^\gamma)$ при $\theta \to 1$,

где $\rho(\theta) = \alpha_k\theta^k + ... + \alpha_0$, $\sigma(\theta) = \beta_k\theta^k + ... + \beta_0$ – соответственно *первый и второй характеристические, или производящие,*

многочлены.

Доказательство п.(а) теоремы следует из применения разложения в ряд Тейлора членов $y(t+jh)$ и $y'(t+jh)$ в выражении для локальной погрешности (4.26), которую можно записать в следующем виде:

$$L(y,t,h)=\sum_{j=0}^{k}[\alpha_j y(t+jh)-h\beta_j y'(t+jh)]=\sum_{j=0}^{k}\Big[\alpha_j\sum_{q=0}\frac{j^q}{q!}h^q y^{(q)}(t)-h\beta_j\sum_{r}$$
$$=y(t)\sum_{j=0}^{k}\alpha_j+\sum_{q\geqslant 0}\tfrac{h^q}{q!}y^{(q)}(t)\Big[\sum_{j=0}^{k}\alpha_j j^q-q\sum_{j=0}^{k}\beta_j j^{q-1}\Big]. \tag{4.28}$$

Отсюда видно, что при первом условии теоремы 4.1 для всех достаточно гладких функций $y(t)$ будет

$$L(y,t,h)=C_{\gamma+1}h^{\gamma+1}y^{(\gamma+1)}(t)+O(h^{\gamma+2}), \tag{4.29}$$

где постоянная $C_{\gamma+1}$ вычисляется по формуле

$$C_{\gamma+1}=\frac{1}{(\gamma+1)!}\Big[\sum_{j=0}^{k}\alpha_j j^{\gamma+1}-(\gamma+1)\sum_{j=0}^{k}\beta_j j^{\gamma}\Big]. \tag{4.30}$$

Другими словами, данные соотношения представляют собой рассмотренные в замечании 2.2 условия коллокаций. Эквивалентность условий (а) и (б) следует из тождества

$$L^h(e^h,he^h,h)=\rho(e^h)-h\sigma(e^h),$$

а эквивалентность 3-го условия проистекает из преобразования $\theta=e^h$, или $h=\log\theta$. □

Замечание 4.1. Для многошаговых методов условия первого порядка ($\gamma=1$), обычно называемые условиями *согласованности*, можно также записать в виде

$$\rho(1) = 0, \;\; \rho'(1) = \sigma(1). \tag{4.31}$$

Данные равенства получаются при $x \to 0$ из соотношения

$$\rho(x) - \log x \sigma(x) = O((\log x)^{\gamma+1}),$$

которое следует из преобразования $x = e^h$ в условии (б) теоремы 4.1.

Рассмотренные явные и неявные ММ будут точны, когда используемые интерполяционные многочлены имеют нулевую погрешность. Отсюда, например, следует, что явные k-шаговые методы Адамса обеспечивают (при точной арифметике) нулевую ошибку, если функция $f(t, y(t))$ является многочленом от t степени не выше $k-1$, а само решение удовлетворяет ОДУ

$$y' = q\, t^{q-1} \qquad q = 0, ..., k.$$

Таким образом, их порядок не меньше k, поскольку для таких многочленов погрешность аппроксимации равна нулю вследствие соотношения (4.26) и равенства

$$L_n^h(t^q, qt^{q-1}, h) = h^q \sum_{i=0}^{k} (\alpha_i i^q - q\beta_i i^{q-1}) = 0.$$

А то, что он не больше k, легко обнаружить, если вычислить локальную погрешность для полиномиального решения $y = t^{k+1}$:

$$y(t_{n+k}) - \bar{y}_{n+k} = h\gamma_k \nabla^k f(t_{n+k-1}, y(t_{n+k-1})) = h^{k+1}\gamma_k y^{(k+1)}(t_{n+k}).$$

Аналогично легко заметить, что k-шаговые неявные методы Адамса имеют порядок $k+1$, поскольку они используют на единицу большие порядки многочленов Ньютона и числа точек. Таким же образом получаем, что k-шаговые явные методы Нюстрема и неявные методы Милна имеют порядки k и $k+1$ соответственно, а неявные алгоритмы ФДН — k-го порядка.

Рассмотрим теперь глобальную погрешность, масштабированную на h^γ:

$$\bar{z}_n = (y(t_n) - y_n)/h^\gamma.$$

Вычитая (4.24) из (4.26) и учитывая (4.28), получим

$$\sum_{j=0}^{k}\alpha_j\bar{z}_{n+j} = h^{1-\gamma}\sum_{j=0}^{k}\beta_j\Big[f(t_{n+j}y(t_{n+j})) - f(t_{n+j},y_{n+j})\Big]+$$

$$+C_\gamma h y^{(\gamma+1)}(t_n) + O(h^2). \tag{4.32}$$

Используя далее равенство

$$y^{(\gamma+1)}(t_n) = \frac{1}{\sigma(1)}\sum_{j=0}^{k}\beta_j y^{(\gamma+1)}(t_{n+1}) + O(h)$$

и линеаризацию выражения

$$f(t_{n+j},y(t_{n+j}))-f(t_{n+j},y_{n+j}) = \frac{\partial f}{\partial y}(t_{n+j},y(t_{n+j}))h^\gamma\bar{z}_{n+j}+O(h^{2\gamma}),$$

после их подстановки в (4.32) мы получаем формулу, которую после отбрасывания членов $O(h^2)$ и $O(h^{2\gamma})$ можно рассматривать как многошаговый метод, примененный к решению задачи Коши

$$\bar{z}'(t) = \frac{\partial f}{\partial y}(t, y(t))\bar{z}(t) + Cy^{(\gamma+1)}(t), \;\; \bar{z}(t_0) = 0, \tag{4.33}$$

где

$$C = C_{\gamma+1}/\sigma(1) \tag{4.34}$$

называется *константой погрешности ММ*, поскольку решение ОДУ (4.33), очевидно, пропорционально C.

Константы погрешностей вида (4.34) позволяют сравнивать глобальные погрешности различных многошаговых методов не только по порядку, но и количественно. В таблице 4.1 мы приводим для рассмотренных выше в данной главе алгоритмов значения порядков и констант погрешностей.

Табл. 4.1. Порядки и константы погрешностей k-шаговых методов

Метод	порядок	константа C
явный Адамса	k	γ_k
неявный Адамса	$k+1$	$\hat{\gamma}_{k+1}$
Нюстрема, $k > 2$	k	$\varkappa_k/2$
Милна, $k > 3$	$k+1$	$\hat{\varkappa}_{k+1}/2$
ФДН	k	$-1/(k+1)$

Необходимо подчеркнуть, что рассмотренные категории порядка и константы погрешности имеют асимптотический

характер, т. е. адекватно определяют точность численного решения только при достаточно малых шагах сетки. Для конечных значениях h, однако, отбрасываемые нами члены высших порядков могут играть основную роль. Для более аккуратного анализа локальной ошибки ее можно представить в интегральной форме.

Теорема 4.2 (об интегральном представлении локальной погрешности ММ). *Если k-шаговый метод (4.24) имеет порядок γ, а q есть некоторое целое число $(1 \leqslant q \leqslant \gamma)$, то для любой функции $y(t)$, непрерывно дифференцируемой $q+1$ раз, выполняется равенство*

$$L(y,t,h) = h^{q+1} \int_0^k K_q(s) y^{(q+1)}(t+sh) d\, s, \tag{4.35}$$

где выражение

$$K_q(s) = \frac{1}{q!} \sum_{j=0}^{k} \alpha_j (j-s)_+^q - \frac{1}{(q-1)!} \sum_{j=0}^{k} \beta_j (j-s)_+^{q-1},$$

$$(j-s)_+^r = \begin{cases} (j-s)^r & \text{при } j-s>0, \\ 0 & \text{при } j-s \leqslant 0 \end{cases}$$

называется ядром Пеано данного многошагового метода.

Доказательство данного утверждения следует из ряда Тейлора с интегральным представлением остаточного члена:

$$y(t+jh) = \sum_{r=0}^{q} \frac{ir}{r!} h^r y^{(r)}(t) + h^{q+1} \int_0^j \frac{(j-s)^q}{q!} y^{(q+1)}(t+sh) d\, s,$$

$$hy'(t+jh) = \sum_{r=0}^{q} \frac{i^{r-1}}{(r-1)!} h^r y^{(r)}(t) + h^{q+1} \int_0^j \frac{(j-s)^{q-1}}{(q-1)!} y^{(q+1)}(t+sh)\,ds.$$

Подставляя эти выражения в представление $L(y,t,h)$ из (4.28), мы получим, что при $q \leqslant \gamma$ стоящие перед интегралом многочлены сокращаются, в силу чего с помощью равенства

$$\int_0^j \frac{(j-s)^q}{q!} y^{(q+1)}(t+s\,h)\,ds = \int_0^k \frac{(j-s)_+^q}{q!} y^{(q+1)}(t+s\,h)\,ds$$

выводится формула (4.35). □

Ядра Пеано обладают рядом полезных свойств, позволяющим их использовать для более детального исследования ошибки.

а. $K_q(s) = 0$ при $s \in (-\infty, 0) \cup [k, \infty)$ и $q = 1, ..., \gamma$. При $s \geqslant k$ и $q \leqslant \gamma$ данное утверждение следует непосредственно из определения ядра Пеано. А для $s < 0$ равенство вытекает из следующего равенства, справедливого для $y(t) = (t-s)^q$ при $q \leqslant \gamma$:

$$L(y, 0, h) = \sum_{j=0}^{k} \alpha_j (j-s)^q - q \sum_{j=0}^{k} \beta_j (j-s)^{q-1} = 0.$$

б. $K_q(s)$ — непрерывно дифференцируемая $q-2$ раза функция и $K_q'(s) = -K_{q-1}(s)$ при $q = 2, ..., \gamma$ (этот факт устанавливается интегрированием по частям в (4.35)); $K_1(s)$ есть кусочно-линейная функция, имеющая в точке j скачок β_j; на интервале $(j-1, j)$ ее угловой коэффициент равен $-(\alpha_j + ... + \alpha_k)$.

в. Константа погрешности (4.30) имеет простое выражение через ядро Пеано:

$$C_{\gamma+1} = \int_0^k K_\gamma(s)d\,s.$$

§ 4.6. Устойчивость многошаговых методов

Некоторые общие понятия устойчивости уже рассматривались нами в § 2.2. Основная задача данного параграфа — рассмотреть их конкретизацию в применении к описанным в § § 4.1 — 4.4 алгоритмам.

Общая форма линейного ММ на равномерной сетке имеет вид

$$y_n = S_n(h),\ n = 0, ..., k-1,$$

$$\alpha_k y_{n+k} + ... + \alpha_0 y_n = h(\beta_k f_{n+k} + ... + \beta_0 f_n),\ n = 0, ..., N-k, \tag{4.36}$$

где $S_n(h)$ означает некоторую стартовую процедуру для определения первых k значений численного решения. Если отвлечься от вычисления начальных значений $y_1, ..., y_{k-1}$, то в применении к автономному уравнению Далквиста $y' = \lambda\, y$ из (4.36) получаем соотношения

$$\alpha_k y_{n+1} + ... + \alpha_0 y_n = \lambda\, h(\beta_k y_{n+k} + ... + \beta_0 y_n).$$

Многочлены k-го порядка, составленные из коэффициентов данных соотношений: $\rho(\theta) = \sum\limits_{j=0}^{k}\alpha_j\theta^j$, $\sigma(\theta) = \sum\limits_{j=0}^{k}\beta_j\theta^j$ и

$\pi_z(\theta) = \rho(\theta) - z\sigma(\theta)$ называются первым производящим, вторым производящим и характеристическим полиномами ММ соответственно, причем многочлены π_z содержат параметр $z = \lambda h$.

Если производящие многочлены $\rho(\theta)$ и $\sigma(\theta)$ имеют общий делитель $\varphi(\theta)$, то можно определить полиномы более низкого порядка

$$\bar\rho(\theta) = \rho(\theta)/\varphi(\theta), \;\; \bar\sigma(\theta) = \sigma(\theta)/\varphi(\theta), \tag{4.37}$$

которые являются производящими многочленами для некоторого нового, более простого многошагового метода.

С помощью оператора сдвига

$$Ey_n = y_{n+1}, \quad \text{или} \quad Ey(t) = y(t+h),$$

полученный многошаговый алгоритм можно записать в компактной форме

$$\bar\rho(E)y_n = h\bar\sigma(E)f_n. \tag{4.38}$$

Очевидно, что после умножения обеих частей этого равенства на оператор $\varphi(E)$ мы придем с учетом (4.37) к исходному уравнению (4.36). Таким образом, любое решение уравнения (4.38) является также решением (4.36), т. е. обе вычислительные схемы в сущности совпадают.

Отсюда следует, что при рассмотрении многошаговых методов можно ограничиться только теми, у которых производящие многочлены $\rho(\theta)$ и $\sigma(\theta)$ не имеют общих множителей, и такие ММ называются *неприводимыми*.

При $h \to 0$ уравнение (4.36) сводится к равенству

$$\alpha_k y_{n+k} + ... + \alpha_0 y_n = 0, \tag{4.39}$$

его можно рассматривать как аппроксимацию ОДУ $y' = 0$, решением которого является константа.

Подставив в (4.39) многочленное представление решения $y_{n+k} = \theta^{n+k}$ и поделив на θ^n, получим характеристическое уравнение

$$\rho(\theta) = \alpha_k \theta^k + \alpha_k \theta^{k-1} + ... + \alpha_0 = 0. \tag{4.40}$$

Если $\rho(\theta)$ имеет корни $\theta_1, ..., \theta_l$ кратностей $m_1, ..., m_l$ соответственно ($l \leqslant k$, $m_1 + ... + m_l = k$), то общее решение уравнения (4.39) задается формулой

$$y_n = P_1(n)\theta_1^n + ... + P_l(n)\theta_l^n, \tag{4.41}$$

где $P_j(n)$ — многочлены степеней $m_j - 1$. Отсюда видно, что для ограниченности y_n при $n \to \infty$ необходимо, чтобы корни уравнения (4.40) лежали в единичном круге, а корни, принадлежащие единичной окружности, были простыми. Данные требования задают корневые условия ММ, устанавливающие устойчивость метода (или нуль-устойчивость, или D-устойчивость — свойства, определенные ранее в главе 2).

Напомним еще, что в § 2.2 мы рассматривали и понятие сильной устойчивости, связанное с дополнительным условием, что на единичной окружности находится только один корень (равный единице).

Рассмотрим несколько примеров. Для явного и неявного методов Адамса имеем $\rho(\theta) = \theta^k - \theta^{k-1}$. Этот многочлен имеет

нулевой корень кратности $k-1$ и простой корень, равный 1, т. е. методы Адамса являются сильно устойчивыми. Для методов Нюстрема и Милна многочлен $\rho(\theta)=\theta^k-\theta^{k-2}$ имеет нулевой корень кратности $k-2$ и простые корни, равные 1 и -1. Таким образом, данные методы D-устойчивы для всех k, но не являются сильно устойчивыми.

Теорема 4.3 (об устойчивости ФДН-методов). *Явные и неявные k-шаговые методы ФДН устойчивы при $k \leqslant 6$ и неустойчивы при $k \geqslant 7$.*

Доказательство следует из того, что характеристическое уравнение равенства $\nabla^j y_{n+k}=0$ есть $\theta^{k-j}(\theta-1)^j=0$, а первый производящий многочлен для ФДН равен

$$\rho(\theta)=\sum_{j=1}^{k}\frac{1}{j}(\theta-j)^j\theta^{k-j}, \tag{4.42}$$

непростое исследование корней которого дает требуемое утверждение. □

§ 4.7. Наивысший достижимый порядок устойчивых многошаговых методов

Из условий аппроксимации следует, что для обеспечения порядка γ коэффициенты ММ должны удовлетворять $\gamma+1$ линейному алгебраическому уравнению, которые получаются из (4.24) после подстановки

$$y_{n+k}=t_{n+k}^q,\ \ f_{n+k}=qt_{n+k}^{q-1},\ \ q=0,1,...,\gamma.$$

Поскольку k-шаговая схема содержит $2k+1$ свободный параметр (без ограничений общности можно считать $\alpha_k=1$), ее

наивысший достижимый порядок равен $2k$. Однако высокий порядок не имеет значения, если метод неустойчив.

Теорема 4.4 (первый барьер Далквиста). *Порядок γ устойчивого k-шагового метода подчиняется следующим ограничениям:*

а) $\gamma \leqslant k+2$ *для четных* k;$\gamma \leqslant k+1$ *для нечетных* k;

б) $\gamma \leqslant k$ *при* $\beta_k/\alpha_k \leqslant 0$ *(в частности, для явных схем).*

Доказательство основывается на громоздкой технике теории функции комплексного переменного. Побочный результат этого анализа заключается в том, что многошаговые методы максимально возможного порядка $k+2$ имеют специальную структуру.

Теорема 4.5 (о симметричности ММ). *Устойчивые k-шаговые методы порядка $k+2$ являются симметричными, т. е.*

$$\alpha_j = -\alpha_{k-j},\ \beta_j = \beta_{k-j} \text{ при всех } j.$$

Замечание 4.2. Естественность, или закономерность, теоремы 4.5 можно объяснить с помощью так называемого "греко-римского" преобразования комплексной переменной

$$\xi = \frac{z+1}{z-1} \quad \text{или} \quad z = \frac{\xi+1}{\xi-1},$$

которое отображает круг $|\xi| < 1$ в полуплоскость $\operatorname{Re} z < 0$, верхнюю полуплоскость $\operatorname{Im} z > 0$ — в нижнюю, окружность $|\xi| = 1$ — в мнимую ось, точку $\xi = 1$ – в $z = \infty$, а точку $\xi = -1$

— в точку $z = 0$. Из определения симметричности ММ следует равенство $\rho(\xi) = -\xi^k \rho(1/\xi)$, а отсюда получаем, что если ξ_i есть простой корень многочлена $\rho(\xi)$, то ξ_i^{-1} также является корнем этого многочлена, т. е. все такие корни простые и лежат на единичной окружности $|\xi| = 1$.

Напомним (см. § 2.2), что в силу теоремы эквивалентности Лакса устойчивый ММ с локальной погрешностью порядка γ является сходящимся с тем же порядком, т. е. его глобальная ошибка есть $O(h^\gamma)$.

Теорема 4.6 (о необходимых условиях сходимости ММ). *Если многошаговый метод (4.24) является сходящимся, то он обязательно является устойчивым и согласованным, т. е. выполняется условие первого порядка (4.31).*

Доказательство этой теоремы проводится с помощью трех простых примеров. Применение ММ (4.24) к задаче Коши $y' = 0,\ y(0) = 0$ дает разностное уравнение (4.40). Предположим противное, т. е. что $\rho(\theta)$ имеет корень θ_1 с модулем $|\theta_1| > 1$ или кратный корень на единичной окружности. Тогда θ_1^n и $n\theta_2^n$ будут расходящимися решениями (4.40), а функции $y_n(t) = \sqrt{h}\theta_1^{t/h}$ и $y_h(t) = (t/\sqrt{h})\theta_2^{t/h}$ при $h \to 0$ сходятся к $y_0(0) = 0$, но останутся расходящимися при любом фиксированном $t > 0$, что подтверждает необходимость устойчивости.

Теперь рассмотрим задачу $y' = 0$ при $t > 0$ и $y'(0) = 0$ с точным решением $y(t) = 1$. Ей также соответствует уравнение (4.40), которое можно записать в виде

$$\alpha_k y_h(t + kh) + \ldots + \alpha_0 y_h(t) = 0.$$

Отсюда при подстановке $h \to 0$ из сходимости следует первое

условие согласованности $\rho(1) = 0$. И наконец, применим метод (4.24) к задаче $y' = 1, y(0) = 0$ с точным решением $y(t) = t$. Поскольку мы уже установили равенство $\rho(1) = 0$, легко проверить, что частное решение уравнения (4.36) имеет вид $y_n = n\,h\,b$, или $y_h(t) = t\,b$, где $b = \sigma(1)/\rho'(1)$. Отсюда видно, что для сходимости необходимо $b = 1$, т. е. $\sigma(1) = \rho'(1)$. □

§ 4.8. Области абсолютной устойчивости ММ

Применение общей схемы ММ (4.22) к модельному уравнению Далквиста $y' = \lambda y$ дает однородное уравнение

$$(\alpha_k - z\beta_k)y_{n+k} + ... + (\alpha_0 - z\beta_0)y_n = 0,\ z = \lambda\,h, \tag{4.43}$$

решение которого, как и ранее, ищем с помощью подстановки $y_{n+k} = \zeta^{n+k}$ и последующего сокращения на ζ^n. В результате получаем характеристическое уравнение k-го порядка

$$\pi_z(\zeta) = \rho(\zeta) - z\sigma(\zeta) = (\alpha_k - z\beta_k)\zeta^k + ... + (\alpha_0 - z\beta_0) = 0, \tag{4.44}$$

зависящее от комплексного параметра $z = \lambda\,h$.

Его решение устойчиво тогда и только тогда, когда все корни по модулю не превосходят единицы, а кратные — строго меньше единицы.

Рассмотренная в предыдущих двух параграфах устойчивость по Далквисту фактически предполагала асимптотику не только в смысле $h \to 0$, но и $z = \lambda h \to 0$, поскольку величина λ считалась конечной и не зависящей от h. Однако в

жестких задачах с очень большими значениями λ такое допущение является неприемлемым, так как реальными становятся ситуации, когда величинами $O(h)$ можно пренебрегать, а члены порядка $O(\lambda h)$ необходимо учитывать в анализе. Поэтому изучаемые ниже понятия, хотя формально и имеют общий характер, практически составляют специальный раздел теории устойчивости жестких ОДУ.

Определение 4.1. *Областью абсолютной устойчивости многошагового метода* (4.41) *(в отличие от нуль— или D— устойчивости) называется множество $S = \{z \in \mathbb{C}\}$, где $|\zeta_j(z)| \leqslant 1$ для всех корней уравнения* (4.44) *и $|\zeta_j(z)| < 1$ для кратных корней.*

Замечание 4.3. Если $S \supset \mathbb{C}_-$, то ММ является A-устойчивым, см. § 2.2.

Замечание 4.4. При $z \to \infty$ корни (4.44) стремятся к корням уравнения $\sigma(\zeta) = 0$.

Замечание 4.5. При $z = 0$ уравнение (4.44) переходит в $\rho(\zeta) = 0$.

Таким образом, нуль-устойчивость метода эквивалентна требованию $0 \in S$.

Теорема 4.6 (об абсолютной устойчивости). *ММ является абсолютно устойчивым (все численные решения ограничены) тогда и только тогда, когда $z = h\lambda \in S$ для всех значений λ модельного уравнения $y' = \lambda y$.*

Замечание 4.6. В определении 4.1 допускается наличие нескольких простых корней на единичной окружности, т. е. удовлетворяющих условию $|\zeta_j(z)| = 1$. Если же наложить дополнительное требование, что такой корень только один

(равный единице), то соответствующий ММ можно назвать сильно абсолютно устойчивым в S.

В силу линейности характеристического многочлена (4.44) относительно z обратное отображение $\zeta \to z$ легко вычисляется и является однозначным:

$$z = \rho(\zeta)/\sigma(\zeta). \tag{4.45}$$

Внешность единичного круга в ζ–плоскости, будучи отображенной по этой формуле обратно в z–плоскость, дает запрещенные значения z, для которых хотя бы один из корней $\zeta_j(z)$ порождает неустойчивость. Образ границы единичного круга $\zeta = e^{i\theta},\ 0 \leqslant \theta \leqslant 2\pi$, называется *кривой локуса корней* (геометрическое место точек — граничных значений корней). Ее следует рассматривать как ориентированную кривую, причем область абсолютной устойчивости S (если она не пуста) должна лежать слева от нее.

На основе формулы (4.45) могут быть рассмотрены области устойчивости различных многошаговых методов.

Явные методы Адамса для модельного уравнения Далквиста дают соотношение (см. § 4.1)

$$y_{n+1} = y_n + z\sum_{j=0}^{k-1}\gamma_j\nabla^j y_n,\ \ \gamma_0 = 1,\ \ \gamma_1 = \frac{1}{2}, ..., \tag{4.46}$$

откуда следует характеристическое уравнение

$$\zeta - 1 = z(\gamma_0 + \gamma_1(1-\zeta^{-1}) + \gamma_2(1-\zeta^{-1})^2 + ...)$$

и формула кривой локуса

$$z = (\zeta - 1)/\sum_{j=0}^{k-1} \gamma_j (1 - \zeta^{-1})^j,\ \zeta = e^{i\theta},\ \ 0 \leqslant \theta \leqslant 2\pi. \qquad (4.47)$$

При $k = 1$ мы отсюда получаем кривую локуса корней для явного метода Эйлера, представляющую собой единичную окружность с центром в точке -1. Таким образом, в соответствии с определением 2.11, интервалом абсолютной устойчивости (пересечение области устойчивости S с вещественной осью) данного метода является отрезок $[-1, 0]$.

На основе анализа выражения (4.47) не очень просто, но можно установить, что области абсолютной устойчивости явных методов Адамса с ростом k достаточно быстро уменьшаются. В частности, левые границы интервалов абсолютной устойчивости $[-a_k, 0]$ для некоторых значений k приведены в таблице 4.2.

Табл. 4.2. Левые границы интервалов абсолютной устойчивости явных методов Адамса

k	1	2	3	4	5	6
$-a_k$	-2	-1	-0.545	-0.3	-0.163	-0.088

Неявные методы Адамса приводят к соотношению (см. § 4.2)

$$y_{n+1} = y_n + z\sum_{j=0}^{k} \hat{\gamma}_j \nabla^j y_n,\ \ \hat{\gamma}_0 = 1,\ \ \hat{\gamma}_1 = -\frac{1}{2}, ..., \qquad (4.48)$$

после подстановки в которое $y_n = \zeta^n$ и деления на ζ^{n+1} вместо (4.47) получаем формулу

$$z = (1-\zeta)/\sum_{j=0}^{k} \hat{\gamma}_j (1-\zeta^{-1})^j, \;\; \zeta = e^{i\theta}, \;\; 0 \leqslant \theta \leqslant 2\pi. \qquad (4.49)$$

Отсюда при $k = 1$ имеем A-устойчивый неявный метод трапеций. С увеличением количества шагов k области устойчивости, хотя и остаются больше, чем у соответствующих явных алгоритмов, но также быстро уменьшаются и не покрывают левую полуплоскость $\mathbb{C}_-$. Это означает, что такие методы не являются A-устойчивыми. Левые границы интервалов $[-a_k, 0]$ абсолютной устойчивости неявных методов Адамса приведены в таблице 4.3.

Табл. 4.3. Левые границы интервалов абсолютной устойчивости для неявных методов Адамса

k	1	2	3	4	5	6
$-a_k$	$-\infty$	-6	-3	-1.84	-1.18	-0.769

Анализ таблиц 4.2 и 4.3 позволяет утверждать, что как явные, так и неявные методы Адамса повышенных порядков точности практически неприемлемы для решения жестких ОДУ.

Явный метод Нюстрема при $k = 1, 2$ дает правило средней точки

$$y_{n+1} = y_{n-1} + 2\, z\, y_n,$$

которому соответствует формула кривой локуса корней

$$z = (e^{i\theta} - e^{-i\theta})/2 = i \sin\theta. \tag{4.50}$$

Отсюда область устойчивости S — это отрезок мнимой оси между точками $+i$ и $-i$, а интервал абсолютной устойчивости — точка начала координат.

Неявный метод Милна при $k = 2$ и $k = 3$ (см. § 4.3) имеет вид

$$y_{n+1} = y_{n-1} + z(y_{n+1} + 4y_n + y_{n-1})/3,$$

что определяет кривую локуса

$$z = \frac{3(e^{i\theta} - e^{-i\theta})}{e^{i\theta} + 4 + e^{-i\theta}} = 3\,i\frac{\sin\theta}{\cos\theta + 2}. \tag{4.51}$$

Здесь область устойчивости — отрезок мнимой оси $[-i\sqrt{3}, i\sqrt{3}]$, т. е. только немного больше, чем у метода Нюстрема. Для методов Нюстрема и Милна более высоких порядков области устойчивости сужаются до одной точки — начала координат.

Таким образом, оба семейства алгоритмов, основанных на аппроксимации интегрального уравнения (4.12), не являются A-устойчивыми и неприемлемы к численному решению жестких задач.

В неявных алгоритмах ФДН формулы кривых локуса, как следует из (4.22), имеют вид

$$z = \sum_{j=1}^{k} (1 - \zeta^{-1})^j / j = \sum_{j=1}^{k} (1 - e^{-i\theta})^j / j. \tag{4.52}$$

При $k = 1$ мы имеем неявный метод Эйлера с областью устойчивости

$S = \{z\colon |z-1| \leqslant 1\}$. Для $k=2$ величина $\mathrm{Re}(z) = \frac{3}{2} - 2\cos\theta + \frac{1}{2}\cos 2\theta$ на кривой локуса корней неотрицательна при любых значениях θ. Следовательно, данный метод второго порядка является A-устойчивым. Таким же свойством обладают формулы дифференцирования назад и при $k=3$. Однако с дальнейшим ростом k методы ФДН все больше и больше теряют устойчивость, а при $k \geqslant 7$ они неустойчивы всюду, даже в начале координат.

Из рассмотренных в § 2.1 методов предиктор-корректор после подстановки формулы предиктора

$$\hat{y}_{n+1} = y_n + z(\bar{\gamma}_0 y_n + \bar{\gamma}_1(y_n - y_{n-1}) + \bar{\gamma}_2(y_n - 2y_{n-1} + y_{n-2}) + \ldots)$$

в формулу корректора общего вида

$$y_{n+1} = y_n + z(\gamma_0 \hat{y}_{n+1} + \gamma_1(\hat{y}_{n+1} - y_n) + \gamma_2(\hat{y}_{n+1} - 2y_n + y_{n-1}) + \ldots),$$

где выражения для коэффициентов $\bar{\gamma}_0, \bar{\gamma}_1, \ldots$ $\check{\gamma}_0, \check{\gamma}_1, \ldots$ определяются видами конкретных применяемых многошаговых схем, при выводе областей устойчивости мы получаем достаточно сложное соотношение, которое после использования представления $y_n = \zeta^n$ и последующего деления на ζ^n принимает вид квадратного уравнения относительно z:

$$A\,z^2 + B\,z + C = 0. \tag{4.53}$$

Здесь коэффициенты A, B, C имеют громоздкие выражения, на которых мы не останавливаемся. В данном случае важным новым моментом является то, что при каждом

значении $\zeta = e^{i\theta}$ уравнение (4.53) имеет два корня, что порождает две кривые локуса корней, которые и определяют область абсолютной устойчивости. Мы остановились здесь на одном из многих возможных вариантов (типа PEC — один предиктор + один корректор), однако такой подход дает принципиальную возможность исследования других схем.

§ 4.9. Второй барьер Далквиста. $A(\alpha)$-устойчивость и жесткая устойчивость

При рассмотрении областей устойчивости многошаговых методов в предыдущих параграфах мы не встретили ни одной -устойчивой схемы, которая обладала бы порядком погрешности выше $\gamma \geqslant 3$. И тот факт, что это не случайно, устанавливается следующим знаменитым утверждением.

Теорема 4.6 (второй барьер Далквиста). *Любой A-устойчивый многошаговый метод должен иметь порядок не более $\gamma \leqslant 2$.*

Доказательство теоремы требует погружения в теорию функций комплексного переменного, на чем мы останавливаться не будем. Отметим только, что попутно здесь устанавливается два интересных факта. Во-первых, постоянная погрешности для метода второго порядка должна удовлетворять неравенству $C \leqslant -1/12$. Во-вторых, существует только один такой метод, для которого $C = -1/12$ — это метод трапеций, оказывающийся, таким образом, оптимальным из всех A-устойчивых многошаговых алгоритмов.

Чтобы сгладить данный пессимистический вывод, вводят-

ся в рассмотрение "почти" A-устойчивые многошаговые методы. Основанием к этому служит то, что многие важные классы практических задач не требуют устойчивости во всей левой полуплоскости $\mathbb{C}_-$.

Определение 4.2 (Видлунд, 1967). *Сходящийся ММ называется $A(\alpha)$-устойчивым при $0 < \alpha < \dfrac{\pi}{2}$, если*

$$S \supset S_\alpha = \{z : \ |\arg(-z)| < \alpha, \ z \neq 0\}.$$

Геометрический смысл области устойчивости такого метода очевиден: это симметричный относительно вещественной оси угол в $\mathbb{C}_-$ с вершиной в начале координат. В частности, многошаговый метод называется $A(0)$-устойчивым, если он $A(\alpha)$-устойчив при сколь угодно малом $\alpha > 0$. Очевидно, что интервал устойчивости $A(\alpha)$-устойчивого метода не зависит от значения α и представляет собой вещественную полуось $[-\infty, 0]$.

Определение 4.3 (Гир, 1971). *Многошаговый метод называется жестко устойчивым, если выполняется условие*

$$S\{z; Re\, z < -d\}$$

при некотором $d > 0$.

Интервал устойчивости жестко устойчивого метода — это отрезок $[-\infty, -d]$. Различными авторами вводились и другие виды устойчивости, но они получили меньшее распространение.

Для иллюстрации свойств $A(\alpha)$-устойчивости и жесткой

устойчивости в таблице 4.4 приводятся значения α и d из определений 4.2 и 4.3 для различных k-шаговых методов ФДН.

Табл. 4.4. Значения α и d для методов ФДН

k	1	2	3	4	5	6
α	90°	90°	86.03°	73.35°	51.84°	17.84°
d	0	0	0.083	0.667	2.327	6.075

Как и следовало ожидать, с ростом k области абсолютной устойчивости сужаются, однако еще остается возможность применимости этих алгоритмов для определенного класса задач.

Но все же существует и альтернативный результат, позволяющий оптимистически смотреть на методы повышенных порядков.

Теорема. 4.7. *Для любых* $0 < \alpha < \dfrac{\pi}{2}$ *и целых* k *существует* $A(\alpha)$*-устойчивый* k*-шаговый метод порядка* $\gamma = k$.

Имеется достаточно большое количество публикаций, в которых конструируются и изучаются алгоритмы сверхвысоких порядков (до десятого и более). Построение формул для их коэффициентов и для областей устойчивости в таких случаях становится уже слишком громоздким и осуществляется с помощью имеющихся программ автоматизации аналитических выкладок.

В определенном смысле такие исследования имеют спортивный характер, однако они могут сыграть и практическую роль в отыскании оптимального по трудозатратам алгоритма

для решения конкретного класса задач с требуемой точностью.

§ 4.10. Одноопорные многошаговые методы и G-устойчивость

В предыдущих параграфах данной главы мы рассматривали устойчивость многошаговых методов только для линейных ОДУ. В 3 главе устойчивость одношаговых алгоритмов для нелинейных уравнений

$$y' = f(t, y) \tag{4.54}$$

исследовалась для функций f, удовлетворяющих одностороннему условию Липшица

$$((f(t, y) - f(t, z)),\ (y - z)) \leqslant l||y - z||^2 \tag{4.55}$$

или, в комплекснозначном случае,

$$\mathrm{Re}((f(t, y) - f(t, z)),\ (y - z)) \leqslant l||y - z||^2, \tag{4.56}$$

где константа Липшица $l(t)$ — некоторое вещественное число, определенное для каждого значения параметра t.

Напомним, что свойство (4.55), (4.56) при $l(t) \leqslant 0$ обеспечивает контрактивность решений нелинейных ОДУ вида (4.54), т. е. разность двух произвольных решений с ростом t не увеличивается, см. (1.64).

В данном параграфе мы исследуем устойчивость многошаговых методов для нелинейного уравнения (4.54) при условиях (4.55), (4.56), но только для специального класса так называемых одноопорных методов. Во-первых, эти методы имеют самостоятельный интерес в силу своей экономичной реализации, требующей на каждом шаге вычисления только одного значения правой части. А во-вторых, их изучение проводится значительно проще.

4.10.1. Определение одноопорного многошагового метода (ОММ). Рассмотрим неявный многошаговый метод

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j}), \tag{4.57}$$

у которого все коэффициенты положительны, $\alpha_k \neq 0$ и $\beta_k \neq 0$, а производящие полиномы

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \ \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j$$

не имеют общих делителей, причем $\sigma(1) = 1$.

Определение 4.4. *Многошаговый метод с одним вычислением правой части*

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\, f\Big(\sum_{j=0}^{k} \beta_j t_{n+j}, \sum_{j=0}^{k} \beta_j y_{n+j}\Big) \tag{4.58}$$

называется одноопорным методом, соответствующим ММРК (4.57).

Замечание 4.6. Для линейного автономного ОДУ $y' = A\,y$ методы (4.57) и (4.58) совпадают, но в общем случае их

решения отличаются, хотя и связаны между собой определенным образом.

Пример 4.1. Правило трапеций является двухопорным (двухшаговым) методом

$$y_{n+1} - y_n = h[f(t_n, y_n) + f(t_{n+1}, y_{n+1})]/2, \tag{4.59}$$

а соответствующий ему одноопорный метод определяется правилом средней точки

$$y_{n+1} - y_n = h\, f\Big(\frac{t_n + t_{n+1}}{2}, \frac{y_n + y_{n+1}}{2}\Big). \tag{4.60}$$

Непосредственной подстановкой можно проверить, что если совокупность $\{\hat{y}_n\}$ есть решение (4.60), то пары значений

$$\hat{y}_n = 0.5(y_n + y_{n+1}),\ \hat{t}_n = 0.5(t_n + t_{n+1})$$

удовлетворяют соотношениям (4.59).

Рассмотренная в приведенном примере для $k = 2$ эквивалентность решений ООММ и "классического" многошагового метода обобщается и на случай произвольного k.

Теорема 4.8 (об эквивалентности ООММ и ММ, Далквист). *Пусть* $\{y_n\}$ *и* $t_n = t_0 + n\,h$ *удовлетворяют системе уравнений* (4.58) *для одноопорного метода. Тогда величины*

$$\hat{y}_n = \sum_{j=0}^{k} \beta_j y_{n+j},\ \ \hat{t}_n = \sum_{j=0}^{k} \beta_j t_{n+j} \tag{4.61}$$

удовлетворяют уравнениям (4.57) для многошагового метода.

Доказательство данного утверждения несложно осуществляется следующим образом: умножим уравнение (4.58) на β_i, заменим n на $n+i$ и просуммируем полученное равенство от $i=0$ до $i=k$. Тогда с заменой порядка суммирования и y_{n+j}, t_{n+j} на $\hat{y}_{n+j}, \hat{t}_{n+j}$) получаем (4.57).

В силу эквивалентности решений для обычных многошаговых схем и одноопорных методов анализ локальных погрешностей аппроксимаций для ООММ специального рассмотрения не требует, поскольку может проводиться по общей методологии ММ.

Замечание 4.7. Вообще говоря, верно и обратное утверждение: если y_n есть решение уравнения ММ (4.57), то ему соответствуют некоторые величины $\hat{y}_n, \hat{t}_n$, удовлетворяющие одноопорному уравнению (4.58), но нам данный факт не потребуется.

4.10.2. Существование и единственность численного решения. Многошаговый метод (4.57) можно записать в виде

$$y - \eta = h\, f(t,y)\beta_k/\alpha_k, \tag{4.62}$$

где $t = t_{n+k}, y = y_{n+k}$, а η — вектор, составленный из известных, т. е. уже посчитанных и не зависящих от y величин. Одноопорную формулу (4.58) с помощью преобразования $y = \beta_k y_{n+k} + ... + \beta_0 y_n$ можно также привести к виду (4.57), так что решения уравнений (4.57) и (4.58) однозначно взаимосвязаны, и последующие утверждения справедливы для обоих соответствующих методов. При этом величину $\tilde{h} = h\beta_k/\alpha_k$ можно рассматривать как новый шаг сетки,

и тогда (4.62) формально представляет собой неявный метод Эйлера, к которому можно применить теоремы для неявных МРК. В частности, легко определяется условие, достаточное для существования и единственности решения данного нелинейного уравнения.

Теорема 4.9 (Далквист). *Пусть f — непрерывно дифференцируемая функция, удовлетворяющая одностороннему условию Липшица с константой l. Если $hl \leqslant \alpha_k/\beta_k$, то нелинейное уравнение* (4.62) *имеет единственное решение.*

Можно также получить оценку возмущенного решения.

Теорема 4.10. *Пусть y есть решение уравнения* (4.62), *а $\tilde{y}$ — решение уравнения*

$$\tilde{y} - \eta = h\, f(t, \tilde{y})\beta_k/\alpha_k + \delta,$$

где δ — некоторое возмущение правой части. Тогда в предположениях теоремы 4.9 *справедливо неравенство*

$$||y - \tilde{y}|| \leqslant ||\delta||/[1 - hl(\beta_k/\alpha_k)].$$

4.10.3. G-устойчивость. Будем рассматривать далее многошаговый метод как отображение $\mathbb{R}^{Nk} \to \mathbb{R}^{Nk}$, введем обозначение для “супервектора”

$$Y_n = (y_{n+k-1}, ..., y_n)^T$$

и определим в $\mathbb{R}^{Nk}$ нормы (напомним, что N есть порядок системы ОДУ)

$$||Y_n||_G^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} g_{i,j}(y_{n+i-1}, y_{n+j-1}),$$

где $(,)$ есть скалярное произведение в $\mathbb{R}^N$, а $g_{i,j}$ — элементы некоторой вещественной симметричной положительно определенной (с.п.о.) матрицы $G \in \mathbb{R}^{k,k}$.

Определение 4.5. *Одноопорный метод* (4.58) *называется G-устойчивым, если существует вещественная с.п.о. матрица G такая, что для различных решений $\{y_n\}$ и $\{\hat{y}_n\}$ данной схемы выполняется неравенство*

$$||Y_{n+1} - \hat{Y}_{n+1}||_G \leqslant ||Y_n - \hat{Y}_n||_G \tag{4.63}$$

для любой величины шага h и для всех ОДУ, удовлетворяющих одностороннему условию Липшица с константой $l \leqslant 0$.

Можно отметить, что свойство G-устойчивости метод наследует в определенном смысле контрактивность решений ОДУ (см.(1.64)), т. е. разность двух численных решений не увеличивается с ростом n. Очевидно, что из G-устойчивости следует A-устойчивость, поскольку ОДУ $y' = \lambda y$ при $\mathrm{Re}\lambda \leqslant 0$ удовлетворяет одностороннему условию Липшица с $l = 0$. Справедлив и более сильный результат, устанавливающий, что G-устойчивость является не только достаточным, но необходимым условием свойства A-устойчивости.

Теорема 4.11 (об эквивалентности A- и G-устойчивости). *Многошаговый метод с производящими многочленами ρ и σ, не имеющими общих делителей, явля-*

ется A-устойчивым тогда и только тогда, когда соответствующий одноопорный метод G-устойчив.

Поскольку выше уже указывалось о возможности сопоставления численных решений, полученных с помощью одноопорных алгоритмов и общего вида ММ, то теорема 4.11 может применяться для исследований различных многошаговых методов, в частности, неявных методов Адамса, Милна и ФДН.

Пример 4.2. Рассмотрим двухшаговый ФДН-алгоритм

$$1.5\, y_{n+2} - 2\, y_{n+1} + 0.5\, y_n = h\, f(t_{n+2}, y_{n+2}). \tag{4.64}$$

Пусть $\Delta y_n = y_n - \hat{y}_n$ есть разность невозмущенного и возмущенного численных решений, полученных по формуле (4.64). Применяя к ней одностороннее условие Липшица при $l = 0$, получаем неравенство

$$\mathrm{Re}((f(t_{n+2}, y_{n+2}) - f(t_{n+2}, \hat{y}_{n+2})),\ (y_{n+2} - \hat{y}_{n+2})) \leqslant 0,$$

которое в рассматриваемом конкретном случае принимает вид

$$F = \mathrm{Re}((1.5\Delta y_{n+2} - 2\Delta y_{n+1} + 0.5\Delta y_n),\ \Delta y_{n+2}) \leqslant 0. \tag{4.65}$$

Если ввести обозначение $\Delta Y_n = (\Delta y_{n+1}, \Delta y_n)^T$, то функционал F из (4.65) можно привести к форме

$$F = ||\Delta Y_{n+1}||_G^2 - ||\Delta Y_n||_G^2 + ||0.5\Delta y_{n+2} + \Delta y_{n+1} + 0.5\Delta y_n||^2,$$

где матрица G второго порядка определяется как

$$G = \frac{1}{4}\begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix} \tag{4.66}$$

и является симметричной положительно определенной. Таким образом, в силу (4.65) неравенство (4.63) выполняется и метод (4.64) является G-устойчивым с матрицей (4.66) и, следовательно, он обладает также A-устойчивостью.

Поскольку выполнение одностороннего условия Липшица позволяет, в соответствии с теоремой 4.10, оценивать возмущение численного решения многошаговой схемы, то свойство G-устойчивости одноопорного метода позволяет оценивать глобальную ошибку соответствующего A-устойчивого ММ.

Рассмотрим линейный многошаговый метод

$$\sum_{i=0}^{k} \alpha_i \hat{y}_{n+i} = h \sum_{i=0}^{k} \beta_i f(\hat{t}_{n+i}, \hat{y}_{n+i}). \tag{4.67}$$

В соответствии с теоремой 4.8, величины $\hat{t}_n, \hat{y}_n$ с помощью формул (4.61) выражаются через решения соответствующей одноопорной схемы

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = hf\Big(\sum_{i=0}^{k} \beta_i t_{n+i}, \sum_{i=0}^{k} \beta_i y_{n+i}\Big),$$

в которой стартовые значения $y_0, y_1, ..., y_{2k-1}$ можно предварительно определить из решения СЛАУ порядка $2\,k$

$$\sum_{i=0}^{k} \beta_i y_{j+1} = \hat{y}_j,\ \sum_{i=0}^{k} \alpha_i y_{j+i} = hf(\hat{t}_j, \hat{y}_j),\ j = 0, ..., k-1.$$

Теорема 4.12. *Если линейный многошаговый метод* (4.67) *является A-устойчивым и имеет локальный порядок погрешности γ, а функция f удовлетворяет одностороннему условию Липшица* (4.55) *или* (4.56), *то существует постоянная $C_0 > 0$ такая, что при $hl \leqslant C_0$ выполняются неравенства*

$$||\hat{y}_n - y(\hat{t}_n)|| \leqslant C\Big(\max_{0\leqslant j,k} ||\hat{y}_j - y(\hat{t}_j)|| + h \max_{0\leqslant j<k} ||f(\hat{t}_j, \hat{y}_j) - y'(\hat{t}_j)||\Big) + Mh^{\gamma},$$

где константа C зависит от метода и (при $l \geqslant 0$) от длины $t_n - t_0$ интервала интегрирования, а постоянная M зависит, кроме того, от нормы γ-й и $(\gamma+1)$-й производных точного решения.

Отметим, что в приведенной оценке первые два слагаемых правой части определяют погрешности стартового участка численного интегрирования (который может реализовываться каким-либо из подходящих алгоритмов), а последнее слагаемое отвечает непосредственно за ошибку метода (4.67).

Удобно рассмотреть свойства одноопорных методов более подробно на использованном уже в предыдущей главе простом примере линейной скалярной задачи Протеро—Робинсона:

$$y' = \lambda y + g(t), \;\; y(t_0) = y_0. \tag{4.68}$$

Применяя к задаче (4.68) классический многошаговый метод с первым и вторым характеристическими многочленами $\rho(\theta)$ и $\sigma(\theta)$, который для краткости будем называть (ρ, σ)-метод, получим формулу

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h\lambda \sum_{i=0}^{k} \beta_i y_{n+i} + h \sum_{i=0}^{k} \beta_i g(t_{n+i}). \tag{4.69}$$

Нетрудно заметить, что глобальная погрешность

$$z_n = y(t_n) - y_n$$

удовлетворяет разностному уравнению

$$\sum_{i=0}^{k} (\alpha_i - h\lambda\beta_i) z_{n+i} = \psi(t_n), \tag{4.70}$$

где локальная ошибка определяется соотношением

$$\psi(t) = \sum_{i=0}^{k} \alpha_i y(t + ih) - h \sum_{i=0}^{k} y'(t + ih). \tag{4.71}$$

Отметим, что правая часть (4.71) не зависит от параметра жесткости λ и равна $O(h^{\gamma+1})$, если классический порядок метода есть γ.

Если мы применим алгоритм в его одноопорном варианте, то получим формулу

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h\lambda \sum_{i=0}^{k} \beta_i y_{n+i} + h\, g(t_n + \beta h), \tag{4.72}$$

где $\beta = \sum\limits_{i=0}^{k} \beta_i i$ и $\sum\limits_{i=0}^{k} \beta_i = 1$. В этом случае глобальную погрешность удобно определить по — другому:

$$z_n^* = y(t_n + \beta h) - \sum_{i=0}^{k} \beta_i y_{n+i}. \tag{4.73}$$

Заменяя в (4.70) n на $n + j$, после умножения на β_j и суммирования получаем уравнение для глобальной погрешности

$$\sum_{i=0}^{k}(\alpha_i - h\lambda\beta_i)z^*_{n+i}) = \psi(t_n + \beta h), \tag{4.74}$$

очень похожее на (4.69).

Для оценки ошибки удобно, как мы делали уже раньше, представить рассматриваемый алгоритм в виде одношагового метода, введя вектор

$$Z_n = (z^*_{n+k-1}, ..., z^*_{n+1}, z^*_n)^T \in \mathbb{R}^k,$$

матрицу перехода

$$B = \begin{bmatrix} b_{k-1}(\mu) & \cdots & b_1(\mu) & b_0(\mu) \\ 1 & \cdots & 0 & 0 \\ & \ddots & \vdots & \vdots \\ 0 & & 1 & 0 \end{bmatrix}, \quad b_j(\mu) = -\frac{\alpha_j - \mu\beta_j}{\alpha_k - \mu\beta_k},$$

и вектор правой части

$$\Psi_n = \Big(\frac{\psi(t_n)}{(\alpha_k - \mu\beta_k)}\Big),\ 0, ..., 0)^T, \quad \mu = h\lambda.$$

В результате соотношение (4.73) принимает вид

$$Z_{n+1} = B(\mu)Z_n + \Psi_n, \tag{4.75}$$

откуда получаем итоговое выражение для глобальной ошибки

$$Z_{n+1} = B^{n+1}(\mu)Z_0 + \sum_{j=0}^{n} B^{n-j}(\mu)\Psi_j, \qquad (4.76)$$

где первое слагаемое правой части характеризует фактически вклад стартовой процедуры в итоговую ошибку.

Легко видеть, что для получения оценки необходимо, чтобы степени матрицы B были равномерно ограничены по параметру $\mu = h\lambda$ (напомним, что величины $||\psi_j|| = O(h^{\gamma+1})$). Данный факт устанавливается с помощью двух следующих утверждений.

Теорема 4.13 (Крайсс, 1962). *Для квадратной матрицы B "степенное условие"*

$$||B^n|| \leqslant C_1, \;\; n = 0, 1, 2, ... \qquad (4.77)$$

эквивалентно "резольвентному условию"

$$||(B - zI)^{-1}|| \leqslant \frac{C_2}{|z| - 1} \;\; при \;\; |z| > 1, \qquad (4.78)$$

где C_1 и C_2 — некоторые постоянные.

Теорема 4.14 (Като, 1960). *Для евклидовой нормы невырожденной матрицы*
$B \in \mathbb{R}^{k,k}$ справедливо неравенство

$$||B^{-1}|| \leqslant ||B^{k-1}||/(detB).$$

Из последней теоремы следует неравенство

$$||(B(\mu) - zI)^{-1}|| \leqslant \frac{(||B(\mu)|| + |z|)^{k-1}}{|\det(B(\mu) - zI)|},$$

причем справедливо равенство

$$\det(B(\mu) - zI) = \prod_{j=1}^{k}(z - \theta_j(\mu)),$$

где $\theta_j(\mu)$ — собственные значения матрицы $B(\mu)$, т. е. корни характеристического уравнения k-го порядка

$$\sum_{i=0}^{k}(\alpha_i - \mu\beta_i)\theta^i = 0.$$

Теперь предположим, что параметр $\mu = h\lambda$ принадлежит S — области устойчивости (ρ, σ)-метода, — т. е. все значения $\theta_j(\mu)$ принадлежат замкнутому единичному кругу, причем те из них, которые лежат на единичной окружности, отделены друг от друга. Тогда для любого фиксированного $\mu \in S$ можно показать выполнимость неравенства (4.78) и, следовательно, (4.77), что обеспечивает ограниченность степеней $B(\mu)$ в (4.75) и сходимость метода.

Отметим еще одно достаточно естественное и распространенное понятие, которое позволяет переносить результаты о сходимости на нелинейные системы ОДУ: (ρ, σ)-метод называется *A-контрактивным* в норме $||\cdot||_G$, если при $\mathrm{Re}\mu \leqslant 0$ выполняется неравенство $||B(\mu)||_G \leqslant 1$, где $B(\mu)$ — матрица перехода в (4.75).

§ 4.11. Общие линейные методы

До сих пор мы рассматривали или многостадийные одношаговые методы, или многошаговые, но одностадийные. Естественно также рассмотреть их обобщение, т. е. гибридные ме-

тоды, использующие для построения численного решения и несколько стадий, и несколько предыдущих шагов.

В частности, для решения нелинейных ОДУ с правыми частями, удовлетворяющими одностороннему условию Липшица, в предыдущей главе вводилось понятие -устойчивости методов Рунге—Кутты, а в данной главе — G-устойчивость одноопорных многошаговых схем. Представляет значительный интерес вопрос, удается ли сохранить положительные качества как многостадийных, так и многошаговых методов при объединении используемых в них разных подходов.

Предложенное за примечательно короткое время (1964 — 1966 гг.) и независимо друг от друга Грэггом, Штеттером, Батчером и Гиром новое семейство алгоритмов, называемое *общими линейными методами*, можно записать в следующем виде:

$$y_i^{(n+1)} = \sum_{j=1}^{k} a_{i,j} y_j^{(n)} + h \sum_{j=1}^{s} b_{i,j} f(t_n + c_j h, v_j^{(n)}), \;\; i = 1, ..., k, \tag{4.79}$$

$$v_i^{(n)} = \sum_{j=1}^{k} \tilde{a}_{i,j} y_j^n + h \sum_{j=1}^{s} \tilde{b}_{i,j} f(t_n + c_j h, v_j^n), \;\; i = 1, ..., s. \tag{4.80}$$

Если ввести матрицы $A = \{a_{i,j}\} \in \mathbb{R}^{k,k}$ и $B = \{b_{i,j}\} \in \mathbb{R}^{k,s}$ и "супервекторы" $y^{(n)} = \{y_1^{(n)}, ..., y_k^{(n)}\}$, $f^{(n)} = \{f_1^{(n)} = f(t_n + c_1 h, v_1^{(n)}), ..., f_s^{(n)} = f(t_n + c_s h, v_s^{(n)})\}$ (напомним, что каждый из векторов $y_j^{(n)}, v_i^{(n)}$ имеет размерность N, равную порядку системы ОДУ), то соотношение (4.79) можно переписать в форме

$$y^{(n+1)} = A\, y^{(n)} + h\, B\, f^{(n)}, \tag{4.81}$$

которая внешне напоминает представление неявного линейного многошагового метода. Однако при вычислении каждого из k шагов необходимо еще реализовывать s стадий с определением величин $v_1^{(n)}, ..., v_s^{(n)}$ из решения уравнений (4.80), которые в общем случае являются нелинейными и требуют проведения итераций.

Алгоритмы виды (4.79), (4.80) содержат большое число параметров $a_{i,j}, b_{i,j}, \tilde{a}_{i,j}, \tilde{b}_{i,j}$ и c_j, наличие которых внушает надежду на возможность, за счет их соответствующего выбора, достичь и высоких порядков точности, и устойчивости.

Обнадеживающим обстоятельством здесь явился результат Батчера, построившего в 1965 г. двухшаговый двухстадийный алгоритм 5-го порядка:

$$\hat{y}_{n+1/2} = y_{n-1} + \frac{h}{8}(9\, f_n + 3\, f_{n-1}),$$

$$\hat{y}_{n+1} = \frac{1}{5}(28\, y_n - 23\, y_{n-1}) + \frac{h}{6}(32\, f_{n_1/2} - 60\, f_n - 26\, f_{n-1}),$$

$$y_{n+1} = \frac{1}{31}(32\, y_n - y_{n_1}) + \frac{h}{93}(64\, f_{n+1/2} + 12\, f_n - f_{n-1}).$$

К этому классу алгоритмов относятся и кратко рассмотренные ранее схемы “предиктор-корректор”. А одним из частных случаев общих линейных методов являются основанные на некотором упрощении формул (4.79) *многошаговые методы* Рунге—Кутты (ММРК):

$$\begin{aligned} y_{n+1} &= \sum_{j=1}^{k} a_j y_{n+j-1} + h \sum_{j=1}^{s} b_j f(t_n + c_j h, v_j), \\ v_j &= \sum_{l=1}^{k} a_{j,l} y_{n+l-1} + h \sum_{l=1}^{k} b_{j,l} f(t_n + c_l h, v_l). \end{aligned} \tag{4.82}$$

В данном случае также можно записать представление (4.81), если определить вектор $y^{(n)} = (y_n, y_{n-1}, ..., y_{n-k+1})^T$ и матрицы

$$A = \begin{bmatrix} a_1 & \dots & a_k \\ 1 & \dots & 0 \\ \ddots & & \vdots \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 & \dots & b_s \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

Было даже высказано предположение, что для произвольных значений k и s существуют нуль-устойчивые ($A(0)$-устойчивые) многошаговые методы Рунге—Кутты, но это оказалось не так. Более того, справедливо следующее утверждение.

Теорема 4.15. *Наивысший порядок A-устойчивого k-шагового s-стадийного ММРК равен $2\,k$.*

Тем не менее, очевидно, что наличие свободных параметров в ММРК, как и в общих линейных методах, дает возможности повышения как областей устойчивости, так и порядков точности для $A(\alpha)$-устойчивых и жестко устойчивых алгоритмов.

§ 4.12. Многошаговые методы решения ОДУ 2-го порядка

С. Штермер в своих воспоминаниях пишет, что в 1904 г. он разработал достаточно простой метод для вычисления траекторий заряженных частиц в магнитном поле, которые описываются задачей Коши для системы ОДУ второго порядка

$$y'' = f(t, y, y'), \;\; y|_{t=0} = y_0, \;\; y'|_{t=0} = y'_0. \tag{4.83}$$

В предыдущей главе мы уже видели, что применение специальных одношаговых алгоритмов с учетом особенностей дифференциального уравнения оказывается более эффективным, чем использование классических методов Рунге—Кутты.

В данном параграфе мы рассмотрим для решения задачи (4.83) возможную специфику построения многошаговых методов, которые после переписывания исходного уравнения в форме системы уравнений 1-го порядка приводятся к виду

$$\begin{aligned} &\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i y'_{n+i}, \\ &\sum_{i=0}^{k} \alpha_i y'_{n+i} = h \sum_{i=0}^{k} \beta_i f(t_{n+i}, y_{n+i}, y'_{n+i}). \end{aligned} \tag{4.84}$$

Здесь можно выделить такой характерный для уравнений движения случай, когда действующие силы не зависят от скорости, т. е. правая часть ОДУ не зависит от производной y':

$$y'' = f(t, y). \tag{4.85}$$

В таких ситуациях естественно обращаться к методам, не содержащим первой производной. Исключая из (4.84) величины y'_n, мы получаем формулу

$$\sum_{i=0}^{2k} \hat{\alpha}_i y_{n+i} = h^2 \sum_{i=0}^{2\,k} \hat{\beta}_i f(t_{n+i}, y_{n+i}), \tag{4.86}$$

где новые коэффициенты $\hat{\alpha}_i, \hat{\beta}_i$ определяются через "старые" значения α_i, β_i соотношениями

$$\sum_{i=0}^{2k} \hat{\alpha}_i \xi^i = \Big(\sum_{i=0}^{k} \alpha_i \xi^i\Big)^2, \quad \sum_{i=0}^{2k} \hat{\beta}_i \xi^i = \Big(\sum_{i=0}^{k} \beta_i \xi^i\Big)^2. \tag{4.87}$$

Вообще говоря, коэффициенты многошаговой схемы (4.86) можно искать из условий устойчивости и согласования с исходным уравнением (4.85) без удовлетворения равенствам (4.87).

Наиболее высокие порядки при этом получаются, естественно, на равномерной сетке, в силу легко проверяемого с помощью тейлоровских разложений следующего соотношения:

$$\begin{aligned} h^2 \Delta y(t_n) &= y(t_n + h) - 2\, y(t_n) + y(t_n + h) = \\ &= h^2 y''(t_n) + \tfrac{h^4}{12} y^{(4)}(t_n) + \tfrac{h^6}{360} y^{(6)}(t_n) + \ldots \end{aligned} \tag{4.88}$$

Отсюда следует семейство трехточечных уравнений

$$y_{n+1} - 2y_n + y_{n-1} = h^2\Big(f_n + \frac{1}{12}\Delta f_{n-1} - \frac{1}{240}\Delta^2 f_{n-2} + ...\Big), \tag{4.89}$$

в которых после отбрасывания тех или других малых членов в правой части получаем какой-то из вариантов схемы Штермера. В частности, если в скобке в (4.89) ограничиться первыми двумя членами, то придем к популярной аппроксимации четвертого порядка, носящей имя своего исследователя Б. Нумерова (1924 г.).

Построение схем различных порядков можно делать на основе выражения, получаемого после двукратного интегрирования уравнения (4.85):

$$y(t+h) - 2\,y(t) + y(t-h) = h^2 \int_0^1 (1-s)[f(t+sh, y(t+sh)) + f(t-sh, y($$

(4.90)

Если теперь в правой части (4.90) функцию f заменим интерполяционным многочленом Ньютона (построенным по точкам $f_n,\ f_{n-1}, ...$)

$$P(t) = P(t_n + s\,h) = \sum_{j=0}^{k-1} (-1)^j \begin{pmatrix} -s \\ \\ j \end{pmatrix} \nabla^j f_n,$$

где ∇^j — определенные в § 4.1 левые конечные разности, то аналогично тому, как это делалось при выводе формул Адамса (4.2), мы получим следующее семейство *явных k-шаговых методов Штермера*:

$$y_{n+1} - 2\,y_n + y_{n-1} = h^2 \sum_{j=0}^{k-1} \sigma_j \nabla^j f_n,$$
$$\sigma_j = (-1)^j \int_0^1 (1-s) \Big[\begin{pmatrix} -s \\ \\ j \end{pmatrix} \Big] ds. \tag{4.91}$$

Числовые значения коэффициентов σ_j для различных j приводятся в таблице 4.5.

Табл. 4.5. Коэффициенты явных методов Штермера

j	0	1	2	3	4	5	6
σ_j	1	0	1/12	1/12	19/240	3/40	863/12096

Если в (4.91) конечные разности выразить через значения правых частей, то для частных случаев $k = 2, 3, 4$ получаем следующие формулы:

$$y_{n+1} - 2\,y_n + y_{n-1} = h^2 g_n^{(k)},$$
$$g_n^{(2)} = f_n,$$
$$g_n^{(3)} = (13\,f_n - 2\,f_{n-1} + f_{n-2})/12, \tag{4.92}$$
$$g_n^{(4)} = (14\,f_n - 5\,f_{n-1} + 4\,f_{n-2} - f_{n-3})/12.$$

С целью повышения порядка аппроксимации и устойчивости от явных схем (4.91), (4.92) можно уйти, если к интерполяционным узлам при построении многочлена Ньютона добавить точку (t_{n+1}, f_{n+1}), в результате чего получаем семейство *неявных методов Штермера*

$$
\begin{aligned}
& y_{n+1} - 2\,y_n + y_{n+1} = h^2 \sum_{j=0}^{k} \hat{\sigma}_j \nabla^j f_{n+1}, \\
& \hat{\sigma}_j = (-1)^j \int_0^1 (1-s) \Big[\begin{pmatrix} -s+1 \\ \\ j \end{pmatrix} + \begin{pmatrix} s+1 \\ \\ j \end{pmatrix} \Big] ds.
\end{aligned}
\tag{4.93}
$$

Значения коэффициентов $\hat{\sigma}_j$ для $j \leqslant 6$ приведены в таблице 4.6.

Табл. 4.6. Коэффициенты неявных методов Штермера

j	0	1	2	3	4	5	6
$\hat{\sigma}_j$	1	-1	1/12	0	$-1/240$	$-1/240$	$-221/60480$

Из соотношений (4.93), в частности при $k = 2$ (а также при $k = 3$), получается упомянутый выше метод Нумерова

$$y_{n+1} - 2\,y_n + y_{n-1} = h^2(f_{n+1} + 10\,f_n + f_{n-1})/12. \qquad (4.94)$$

Используя идеи Нюстрема и Милна из § 4.3 для ОДУ первого порядка, можно вывести и другие алгоритмы. При формальных заменах в формуле (4.90) h на $2h$, $2s$ на s и t на $t-h$ мы приходим к интегральному соотношению

$$y(t+h) - 2\,y(t-h) + y(t-3\,h) =$$

$$h^2 \int_0^2 (2-s)[f(t+(s-1)h),\ \ y(t+(s-1)h)) + \qquad (4.95)$$

$$+ f(t+(s+1)h),\ \ y(t+(s+1)h))]\,d\,s.$$

Если здесь подынтегральную функцию аппроксимировать интерполяционным многочленом Ньютона (с использованием или без использования точки (t_{n+1}, f_{n+1})), то получим соответственно новые семейства неявных или явных методов.

Рассмотрим теперь другой возможный подход к построению метода решения уравнения (4.85). Будем искать вычислительную неявную схему следующего вида:

$$y_{n+1} = \alpha_0 y_n + \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \\ + h^2(\beta_{-1} y''_{n+1} + \beta_0 y''_n + \beta_1 y''_{n-1} + \beta''_2 y_{n-2}) + \psi_h, \tag{4.96}$$

где коэффициент α_2 используем в качестве свободного параметра, а остальные шесть множителей $\alpha_0, \alpha_1, \beta_{-1}, ..., \beta_2$ находим по принципу коллокаций, т. е. чтобы уравнение (4.96) удовлетворялось точно (с $\psi_h = 0$) для одночленов $1, t, ..., t^5$. Саму же величину ψ_h будем называть главным членом локальной ошибки и определяем как результат подстановки в (4.96) функции $y = t^6$. В этом случае мы получаем соотношения

$$\alpha_0 = 2 + \alpha_2, \;\; \alpha_1 = -(1 + 2\alpha_2), \;\; \beta_{-1} = 1/12,$$

$$\beta_0 = (10 - \alpha_2)/12, \;\; \beta_1 = (1 - 10\alpha_2)/12, \;\; \beta_2 = -a_2/12,$$

$$\psi_h = (-3 + 3\alpha_2) h^6 y^{(6)}/720. \tag{4.97}$$

При этом корнями характеристического многочлена

$$\rho(\theta) = \theta^3 - \alpha_0 \theta^2 - \alpha_1 \theta - \alpha_2 = 0$$

оказываются значения $\theta_1 = \theta_2 = 1$ и $\theta_3 = \alpha_2$, откуда следует, что общее решение уравнения (4.96) при $\alpha_2 \neq 1$ имеет вид

$$y_n = C_1 + C_2 n + C_3 (\alpha_2)^n, \tag{4.98}$$

где C_1, C_2, C_3 — некоторые постоянные.

Однако в случае трехкратного корня $\theta_1 = \theta_2 = \theta_3 = 1$ вместо (4.98) имеем

$$y_n = C_1 + C_2 n + C_3 n^2, \tag{4.99}$$

что свидетельствует о наличии сильной неустойчивости. Таким образом, значение $\alpha_2 = 1$, минимизирующее локальную ошибку ($\psi_h = 0$ в (4.97)) , оказывается неприемлемым. Наиболее привлекательным выглядит вариант с $\alpha_2 = \beta_2 = 0$, что приводит к неявной схеме Нумерова (4.94), в которой, как следует из (4.97), главный член погрешности равен $\psi_h = -\frac{h^6 y^{(6)}}{240}$.

При итерационной реализации неявного метода Нумерова необходимо выбрать какое-то начальное приближение для члена f_{n+1}, для чего рекомендуется использовать

$$\hat{f}_{n+1} = f(t_{n+1}, \hat{g}(t_{n+1})),$$

где величина $\hat{y}_{n+1}$ определяется из явной формулы прогноза с тем же порядком:

$$\hat{y}_{n+1} = 2\,y_{n-1} - y_{n-3} + \frac{4\,h^2}{3}(f_n + f_{n-1} + f_{n-2}) + \hat{\psi}_h, \;\; \hat{\psi}_h = \frac{16}{240} h^6 y^{(6)}. \tag{4.100}$$

Заметим, что здесь последний член $\hat{\psi}_h$ имеет локальную ошибку другого знака. Таким образом, если формулу Нумерова считать корректором, то она в совокупности с предиктором (4.100) образует двустороннее приближение, при условии, когда производная $y^{(6)}$ не меняет знак на интервале $[t_{n-3}, t_{n+1}]$. В частности, отсюда следует апостериорная оценка ошибки

$$\max\{|y(t_{n+1}) - y_{n+1}|, |y(t_{n+1}) - \hat{y}_{n+1}|\} \leqslant \frac{17}{240} h^6 |y^{(6)}|.$$

§ 4.13. Явные схемы с переменным шагом для решения жестких задач

В данном параграфе мы изложим оригинальные идеи В. И.Лебедева по устойчивому решению жестких задач с помощью простейших явных схем, экономичность которых обеспечивается специальным выбором переменных шагов сетки, см. статью В. И.Лебедева "Явные разностные схемы для решения жестких задач с комплексным или разделимым спектром", ЖВММФ, т. 40, № 12, 2000, 1801—1812, а также цитируемые там авторские работы.

Рассмотрение ради простоты линейную задачу Коши

$$y' = -A\,y, \;\; y|_{t=0} = y_0, \;\; 0 \leqslant t \leqslant t_e < \infty, \tag{4.101}$$

с постоянной матрицей A, собственные числа которой лежат в круге K_r минимального радиуса r в комплексной плоскости:

$$K_r = \{z : |z - r| \leqslant r\}.$$

Предполагается, что решение задачи (4.101) существует и оно достаточно гладкое. Пусть также для заданного $\varepsilon > 0$ (точности локальной аппроксимации) известно $h > 0$ такое, что при $h_n \leqslant h$

$$y'(t_n) = \frac{y(t_n + h_n) - y(t_n)}{h_n} + \delta_n, \ \ |\delta_n| \leqslant \varepsilon \tag{4.102}$$

для $t_0 \leqslant t_n \leqslant T$. Задачу Коши будем называть жесткой при выполнении условия

$$t_e r \gg 1. \tag{4.103}$$

Пусть $A \in -R^{N,N}$ — нормальная матрица с собственными числами λ_i и базисом из собственных векторов φ_i. Раскладывая вектор начальных данных по базису:

$$y_0 = \sum_{i=1}^{N} a_i \varphi_i,$$

решение исходной задачи мы можем представить в форме

$$y(t) = \exp(-A\,t) y_0 = \sum_{i=1}^{N} \exp(-\lambda_i t) a_i \varphi_i.$$

Для решения задачи (4.101) мы проводим явные методы вида

$$y_{n+1} = (I - h_{n+1} A) y_n. \tag{4.104}$$

Если рассмотреть частный случай (4.104) — явный метод Эйлера с постоянным шагом $h_{n+1} = h$

$$y_{n+1} = y_n - h\,A\,y_n, \tag{4.105}$$

то естественным требованием для обеспечения устойчивости является условие на спектральный радиус матрицы перехода $B = I - h\,A$:

$$\rho(B) \leqslant \max\{|1 - 2\,h\,r|, 1\} \leqslant 1. \tag{4.106}$$

Для его выполнения достаточно неравенства

$$h \leqslant \mathrm{cour} = r^{-1}, \tag{4.107}$$

где величина *cour* называется *числом Куранта*. Отсюда следует, что для устойчивого использования метода (4.105) на интервале интегрирования $t = [0, T]$ необходимо число шагов

$$N_e = t_e/h \geqslant t_e r \gg 1. \tag{4.108}$$

В. И.Лебедевым был сформулирован и решен следующий вопрос: существуют ли явные устойчивые алгоритмы с переменными шагами h_{n+1} вида (4.104), интегрирующие жесткую задачу Коши с заданной точностью при затрате существенно меньшего количества шагов? Более того, им было построено семейство методов, оптимизированных под различные типы спектров оператора A и реализованных в составе авторской программы DUMKA, успешное применение которой продемонстрировано во многих публикациях.

Анализ данной проблемы базируется на полиномиальном рассмотрении численного решения

$$y_{N_e} = P_{N_e}(A) y_0, \tag{4.109}$$

где $P_{N_e}(A)$ есть многочлен порядка N_e

$$P_{N_e}(z) = \prod_{k=1}^{N_e}(1 - h_k z) = 1 - t_{N_e} z + \sum_{i,k} h_i h_k z^2 + ..., \quad (4.110)$$

удовлетворяющий очевидным условиям

$$P_{N_e}(0) = 1, \;\; -P'_{N_e}(0) = t_{N_e} = \sum_{k=1}^{N_e} h_k, \quad (4.111)$$

причем корни этого многочлена — это обратные значения h_k^{-1}.

Сравнивая разложение экспоненты

$$\exp(-Tz) = 1 - Tz + (Tz)^2/2 + ...$$

с представлением (4.110), мы можем записать

$$|\exp(-t_{N_e} z) - P_{N_e}(z)| = O((t_{N_e} z)^2),$$

т. е. аппроксимация оператора $\exp(-At)$ многочленом $P_{N_e}(A)$ выполняется.

Более конкретно по заданной локальной точности ε условие аппроксимации, в соответствии с (4.102) при проведении N_e шагов мы будем определять как

$$\max_{1 \leqslant k \leqslant N_e} \{h_k\} = h, \;\; h \leqslant t_{N_e} \leqslant b\, h, \quad (4.112)$$

где $b > 1$ есть независящая от N_e величина. Сделав N_e шагов, мы сможем переопределить значения h, N_e и продолжить процесс численного интегрирования.

Условие спектральной устойчивости искомого алгоритма и, следовательно, многочлена $P_{N_e}(z)$ имеет вид

$$\rho(P_{N_e}(A)) = \sup_{z\in R} |P_{N_e}(z)| \leqslant q_{N_e} \leqslant 1, \tag{4.113}$$

где R — некоторое множество из K_r, содержащее спектр матрицы, т. е. $\mathrm{Sp}(A)$.

Таким образом, для наибольшего продвижения по времени за N_e шагов надо решить следующую *обобщенную задачу Маркова*: среди всех многочленов порядка N_e вида (4.110), удовлетворяющих условию аппроксимации (4.112) и условию устойчивости (4.113), найти такой, который обладает максимальной по модулю производной в нуле, т. е.

$$t_{N_e} = \sup_{P_{N_e}}(-P'_{N_e}(0)), \;\; P_{N_e}(z) = \mathrm{argmax}(\sup_{P_{N_e}}(-P'_{N_e}(0)). \tag{4.114}$$

Данную проблему мы не будем исследовать в общем виде, а ограничимся простым случаем, когда матрица A имеет положительные собственные числа

$$0 < m \leqslant \lambda(A) \leqslant M < \infty. \tag{4.115}$$

Рассмотрим многочлен порядка N_e

$$Q_{N_e}(z) = T_{N_e}(x)/T_{N_e}(\gamma), \tag{4.116}$$

где $T_{N_e}(x)$ — *полином Чебышева 1-го рода*,

$$\begin{gathered} x = (M + m - 2\,z)/(M - m), \\ \gamma = (M + m)/(M - m) > 1. \end{gathered} \tag{4.117}$$

Как известно, многочлен (4.115) наименее уклоняется от нуля на интервале $[m, M]$ среди всех многочленов, нормированных условием $Q_{N_e}(0) = 1$. Кроме того, он обладает следующими свойствами:

$$\max_{z\in[m,M]}\{|Q_{N_e}(z)|\} = \eta = T_{N_e}^{-1}(\gamma) < 1,$$

$$Q_{N_e}(m) = \eta, \;\; \eta \leqslant |Q_{N_e}(z)| < 1 \;\; \text{при} \;\; z \in (0, m),$$

а его корни выражаются формулой

$$z_k = \Big[M + m - (M - m)\cos\frac{(2k-1)\pi}{2N_e}\Big]/2, \;\; k = 1, ..., N_e. \tag{4.118}$$

На основе свойств чебышевских многочленов можно показать (технические детали мы опускаем), что

$$-Q'_{N_e}(0) = N_e\sqrt{(1-\eta^2)/m\,M},$$

откуда получается следующий результат:

$$\max_{\eta\in(0,1)} |Q'_{N_e}(0)| = 2\,N_e^2/M.$$

Таким образом, на основе соотношений (4.110) получаем, что при использовании в методе (4.104) шагов сетки $h_k = z_k^{-1}$, где z_k — определяемые из (4.118) корни чебышевского многочлена (4.116), (4.117), за N_e шагов мы можем провести устойчивое численное интегрирование задачи (4.101) с матрицей (4.115) до момента времени

$$t_e = 2\,N_e^2/M.$$

При этом, если использовать явный метод Эйлера (4.105) с постоянным шагом h, интервал устойчивого интегрирования будет в N_e раз меньше, т. е.

$$t_e \leqslant 2\,N_e/M,$$

поскольку число Куранта в данном случае равно

$$\mathrm{cour} = 2/M.$$

Отметим, что в работах В. И.Лебедева исследованы и реализованы в программе DUMKA устойчивые явные алгоритмы и для более сложных распределений собственных чисел оператора A.

§ 4.14. Многошаговые методы на неравномерной сетке

В отличие от рассмотренных в предыдущем параграфе явных схем В. И.Лебедева первого порядка с переменными шагами сетки далее мы остановимся на более общих многошаговых методах. Конкретнее, будут исследованы достаточно кратко алгоритмы Адамса и неявные формулы дифференцирования назад, включая вопросы их аппроксимации, устойчивости и сходимости по норме.

Совершенно понятно, что с алгоритмической точки зрения построение многошаговых методов для решения задач Коши в случае сеток с переменными шагами не представляет ничего принципиально нового. Для аппроксимации подынтегральной функции требуется только применить соответствующие интерполяционные формулуы Ньютона с использованием разделенных разностей [12, 30].

4.14.1. Методы Адамса с переменным шагом. Построим с помощью интерполяционного многочлена Ньютона

$(k-1)$-го порядка по точкам $t_{n-k+1}, ..., t_n$:

$$L_n(t) = \sum_{i=0}^{k-1} \prod_{j=0}^{i-1} (t - t_{n-i}) f(t_n, ..., t_{n-j}), \tag{4.119}$$

где величины $f(t_n, ..., t_{n-j})$ являются разделенными разностями и выражаются с помощью рекуррентных формул

$$f(t_n) = f_n, \quad f(t_n, ..., t_{n-j}) = \frac{f(t_n, ..., t_{n-j+1}) - f(t_{n-1}, ..., t_{n-j})}{t_n - t_{n-j}}. \tag{4.120}$$

Для практических расчетов формулу (4.119) удобно переписать в виде

$$\begin{aligned} L_n(t) &= \sum_{i=0}^{k-1} \frac{t - t_{n-j}}{t_{n+1} - t_{n-i}} P_i(n), \\ P_i(n) &= \prod_{j=0}^{i-1} (t_{n+1} - t_{n-j}) f(t_n, ..., t_{n-j}). \end{aligned} \tag{4.121}$$

Подставляя (4.121) в приближенное интегральное представление решения

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} L_n(t) d\, t, \tag{4.122}$$

получаем обобщение явного метода Адамса на случай переменного шага:

$$y_{n+1} = y_n + h_n \sum_{i=0}^{k-1} g_i(n) P_i(n), \tag{4.123}$$

где числовые множители имеют вид

$$g_i(n) = \frac{1}{h_n} \int\limits_{t_n}^{t_{n+1}} \prod_{j=0}^{i-1} \frac{t - t_{n-j}}{t_{n+1} - t_{n-j}} d\,t. \tag{4.124}$$

Отметим, что для постоянного шага формулы (4.121), (4.124) переходят в соотношения из § 4.1:

$$g_i(n) = \gamma_i, \quad P_i(n) = \nabla^i f_n. \tag{4.125}$$

Подобным образом можно вывести и неявные методы Адамса для неравномерной сетки. Обозначим через $\check{L}_n(t)$ интерполяционный многочлен Ньютона степени k, проходящий через точки (t_j, f_j), $j = n-k+1, ..., n, n+1$:

$$\check{L}_n(t) = L_n(t) + \prod_{i=0}^{k-1} (t - t_{n-i}) f(t_{n+1}, t_n, ..., t_{n-k+1}),$$

где $L_n(t)$ — определяемый из (4.121) многочлен $(k-1)$-го порядка, а величина f_{n+1} является неизвестной.

Тогда численное решение, определяемое формулой

$$y_{n+1} = y_n + \int\limits_{t_n}^{t_{n+1}} \check{L}_n(t) d\,t,$$

можно представить в виде

$$y_{n+1} = \check{y}_{n+1} + h_n g_k(n) P_k(n+1). \tag{4.126}$$

В данном случае через $\check{y}_{n+1}$ переобозначен результат, получаемый из формулы (4.123) для явного метода Адамса, а также определена величина

$$P_n(n+1) = \sum_{i=0}^{k-1}(t_{n+1} - t_{n-i}) f(t_{n+1}, t_n, ..., t_{n-k+1}).$$

Построение конкретных формул неявных методов Адамса мы оставляем для упражнений, см. ниже п.4.15.29.

Как нетрудно увидеть, реализация методов Адамса с переменными шагами гораздо сложнее, чем с постоянными. Для более экономичного вычисления значений $g_k(n)$ и $P_k(n)$ для них можно вывести удобные рекуррентные соотношения, но мы на этих деталях не останавливаемся, см. подробнее [64].

4.14.2. Формулы дифференцирования назад с переменным шагом. Рассмотренные в § 4.4 ФДН — методы можно также обобщить на случай неравномерной сетки. Обозначим через $\check{L}_n(t)$ интерполяционный полином Ньютона степени k, проходящий через точки $(t_i, y_i), i = n+1, n, ..., n-k+1$:

$$\check{L}_n(t) = \sum_{j=0}^{k}\prod_{i=0}^{j-1}(t - t_{n+1-i}) y(t_{n+1}, t_n, ..., t_{n-j+1}). \qquad (4.127)$$

Удовлетворяя коллокационному условию

$$\check{L}'_n(t_{n+1}) = f(t_{n+1}, y_{n+1}),$$

получаем неявный метод ФДН с переменным шагом

$$\sum_{j=1}^{k} h_n \prod_{i=1}^{j-1}(t_{n+1} - t_{n+1-i}) y(t_{n+1}, ..., t_{n+1-j}) = h_n f(t_{n+1}, y_{n+1}). \qquad (4.128)$$

Вычисление коэффициентов при его реализации гораздо проще, чем в методах Адамса. В частности, двухшаговая ФДН имеет вид

$$(1+2\omega_n)y_{n+1}-(1+\omega_n)^2y_n+\omega_n^2y_{n-1}=h_n(1+\omega_n)f_{n+1}, \tag{4.129}$$

где используется обозначение $\omega_n = h_n/h_{n-1}$.

4.14.3. Погрешность аппроксимации многошаговых методов на неравномерной сетке. Рассмотрим формулу для многошагового алгоритма общего вида:

$$y_{n+k}+\sum_{j=0}^{k-1}\alpha_{j,n}y_{n+j}=h_{n+k-1}\sum_{j=0}^{k}\beta_{j,n}f_{n+j}, \tag{4.130}$$

где коэффициенты $\alpha_{j,n}$ и $\beta_{j,n}$ зависят от отношений шагов $\omega_i = h_i/h_{i-1},\ i = n+1, ..., n+k-1$. Отметим, что запись (4.130) отличается от (4.129), а также от (4.24) из § 4.5 формально тем, что коэффициент при y_{n+k} отнормирован на единицу.

В соответствии с принципом коллокаций введем следующее понятие.

Определение 4.6. *Метод* (4.130) *называется согласованным с порядком q, если для всех многочленов $P_q(t)$, степень которых не превосходит q, во всех узлах сетки t_j выполняется равенство*

$$P_q(t_{n+k})+\sum_{j=0}^{k-1}\alpha_{j,n}P_q(t_{n+j})=h_{n+k-1}\sum_{j=0}^{k}\beta_{j,n}P_q'(t_{n+j}.). \tag{4.131}$$

Согласно построению рассмотренных в п. 4.14.1 явного и неявного методов Адамса, а также алгоритмов ФДН из п. 4.14.2 для них порядки согласования равны $k, k+1$ и k соответственно.

Напомним, что в § 4.5 условие согласованности первого порядка ($\gamma = 1$) называется просто условием согласованности, которое имеет вид (4.31).

В соответствии с определением локальной ошибки из § 2.1, или погрешности аппроксимации, для уравнения вида (4.130) она имеет вид

$$\psi^h_{n+k} = y(t_{n+k}) + \sum_{j=0}^{k-1} \alpha_{j,n} y(t_{n+j}) - h_{n+k-1} \sum_{j=0}^{k} \beta_{j,n} f(t_{n+j}), \tag{4.132}$$

где величины $y(t_{n+j})$ суть значения точного решения исходного ОДУ в узлах сетки. Легко увидеть, что если искомое решение является достаточно гладким, а коэффициенты $\alpha_{j,n}$ и $\beta_{j,n}$ ограничены для всех j, n, то согласованный с порядком γ метод имеет локальную погрешность того же порядка, т. е. при $h \to 0$

$$\psi^h_{n+1} = O(h^{\gamma+1}). \tag{4.133}$$

Более строгий результат в отношении погрешности аппроксимации заключается в получении для нее оценки в какой-либо векторной норме:

$$||\psi^h_{n+1}|| \le C_\gamma h^{\gamma+1}, \tag{4.134}$$

где постоянная C_γ не зависит от шага сетки и определяется, как правило, гладкостью искомого решения.

При теоретическом исследовании вычислительных схем на последовательности сгущающихся дискретизаций мы будем предполагать, что сетки являются регулярными, т. е. при

$h \to 0$ отношение максимального шага к минимальному является ограниченным:

$$h/h_{min} \le \omega < \infty. \tag{4.135}$$

Напомним, что в случае неравномерной сетки h обозначает ее максимальный шаг.

Нетрудно показать, что условие регулярности сетки (4.135) является достаточным для ограниченности коэффициентов согласованной вычислительной схемы.

4.14.4. Устойчивость и сходимость по норме многошаговых методов с переменным шагом. В данном пункте мы рассмотрим сначала упрощенную постановку на примере скалярного однородного уравнения $y' = 0$, когда формула k-шагового метода имеет вид

$$y_{n+k} + \sum_{j=0}^{k-1} \alpha_{j,n} y_{n+j} = 0. \tag{4.136}$$

Вводя вектор k-го порядка

$$Y_n = (y_{n+k-1}, ..., y_n),$$

запишем (4.136) формально в виде одношаговой схемы

$$Y_{n+1} = B_n Y_n, \tag{4.137}$$

где матрицы перехода B_n зависят от отношений шагов $\omega_n, ..., \omega_{n+k-1}$ и имеют вид

$$B_n = \begin{bmatrix} -\alpha_{k-1,n} & \cdots & -\alpha_{1,n} & -\alpha_{0,n} \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{bmatrix}. \tag{4.138}$$

Очевидно, что решение уравнений (4.136), (4.137) будет ограниченным в какой-то из векторных норм, если для соответствующих подчиненных матричных норм при всех n и $j \geq 0$ выполняются неравенства

$$||B_{n+j}B_{n+j-1}\cdots B_{n+1}B_n|| \leq c < \infty. \tag{4.139}$$

Как мы увидим ниже, данное условие можно взять за определение устойчивости по норме вместо использованного в § 2.2 условия (2.24). В частности, если для всех n выполняется неравенство

$$||B_n|| \leq 1 + h\,b, \;\; 0 < b < \infty, \tag{4.140}$$

то вследствие соотношения $(1 + h\,b)^{n+1} \leq \exp\{b(t_e - t_0)\}$ в (4.139) можно положить

$$c = \exp\{b(t_e - t_0)\}.$$

Таким образом, условие (4.140) является достаточным для выполнения (4.139). В то же время легко показать, что если

соотношения (4.139) выполняются для всех n и j, а начальный вектор однородного уравнения (4.137) удовлетворяет неравенству

$$||Y_0|| \leqslant \delta < \infty,$$

то для векторов Y_{n+1} при всех n имеем

$$||Y_{n+1}|| = ||B_n \cdots B_0 Y_0|| \leqslant c\delta < \infty. \tag{4.141}$$

Данный результат и означает устойчивость схемы (4.136) в том смысле, что определяемое ею численное решение исходного однородного ОДУ при $n \to \infty$ остается ограниченным (хотя и может расти).

Для рассмотрения сходимости вычислительной схемы перейдем к задаче Коши для системы ОДУ N-го порядка

$$\dot{y} = f(t, y(t)), \;\; y(t_0) = y_0, \;\; t \in [t_0, t_e], \;\; y(t), \; f(t) \in \mathbb{R}^N. \tag{4.142}$$

Формулу многошагового метода будем использовать в виде (4.130), т. е. с нормировке коэффициента при y_{n+k} на единицу. Подставляя в это уравнение вместо численного решения значения точного решения исходной задачи Коши в соответствующие моменты времени, получаем вместо (4.130) соотношение

$$y(t_{n+k}) + \sum_{j=0}^{k-1} \alpha_{j,n} y(t_{n+j}) = h_{n+k-1} \sum_{j=0}^{k} \beta_{j,n} f(t_{n+j}, y(t_{n+j})) + \psi_{n+k}, \tag{4.143}$$

где ψ_{n+k} есть вектор погрешности аппроксимации.

Вычитая теперь почленно уравнения (4.143) и (4.130), для векторов ошибок численного решения (см. (2.15) в определении 2.3)

$$z_{n+j} = y(t_{n+j}) - y_{n+j}, \;\; j = 0, ..., k,$$

получаем равенства

$$z_{n+k} + \sum_{j=0}^{k-1} \alpha_{j,n} z_{n+j} = h_{n+k-1} \sum_{j=0}^{k} \beta_{j,n} g_{n+j} + \psi_{n+k}. \qquad (4.144)$$

Здесь введено обозначение $g_{n+j} = f(t_{n+j}) - f_{n+j}$, где, в свою очередь, необходимо отметить следующие различия двух последних членов:

$$f(t_{n+j}) = f(t_{n+j}, y(t_{n+j})), \;\; f_{n+j} = f(t_{n+j}, y_{n+j}).$$

При ограниченности первых производных функций $f(t, y^1, ..., y^N) = \{f^i(t, y)\}$, $y = \{y^i\}$, $i = 1, ..., N$, по y^i для компонент вектор–функций $g_{n+j} = \{g^i_{n+j}\}$ справедливо соотношение

$$g^i_{n+j} = f^i(t_{n+j}, y^1(t_{n+j}), ...y^N(t_{n+j}) - f^i(t_{n+j}, y^1_{n+j}, ..., y^N_{n+j}) =$$

$$= \frac{\partial f^i}{\partial y^1}(y^1(t_{n+j}) - y^1_{n+j}) + ... + \frac{\partial f^i}{\partial y^N}(y^N(t_{n+j}) - y^N_{n+j}), \qquad (4.145)$$

где производные $\dfrac{\partial f^i}{\partial y^j}$ берутся в подходящих промежуточных точках $(t_{n+j}, y^*_{n+j}), y^*_{n+j} \in [y(t_{n+j}), y_{n+j}]$. Применяя далее неравенство Коши—Буняковского, из (4.145) получаем выражение

$$|g^i_{n+j}| \leqslant \Big(\sum_{i=1}^{N} q^2_{i,j}\Big)^{1/2} \Big(\sum_{i=1}^{N} (z^i_{n+j})^2\Big)^{1/2}, \tag{4.146}$$

в котором используются обозначения

$$q_{i,j} = \partial f^i/\partial y^j, \;\; z^i_{n+j} = y^i(t_{n+j}) - y^i_{n+j}.$$

Теперь из скалярных неравенств (4.146) для евклидовых норм векторных функций мы можем выписать соотношение

$$\begin{gathered} ||f(t_{n+j}, y(t_{n+j})) - f(t_{n+j}, y_{n+j})||_2 \leqslant L||z_{n+j}||_2, \\ L = \Big(\sum_{i=1}^{N} q^2_{i,k}\Big)^{1/2}, \end{gathered} \tag{4.147}$$

представляющее собой обобщение, или следствие, введенного в (1.27) условия Липшица с постоянной L, которое можно также принять в конечномерном случае за эквивалентное определение условия Липшица. С другой стороны, вектор-функцию g_{n+j} можно представить с помощью матрицы Якоби:

$$g_{n+j} = J_{n+j} z_{n+j}, \;\; J_{n+j} = \Big\{\Big(\frac{\partial f^i}{\partial y^k}\Big)_{n+j}\Big\} \in \mathbb{R}^{N,N}. \tag{4.148}$$

А поскольку из равенства (4.148) для любой из векторных норм и соответствующей подчиненной матричной нормы следует неравенство

$$||g_{n+j}|| \leqslant ||J_{n+j}|| \cdot ||z_{n+j}||, \tag{4.149}$$

то отсюда можно заключить, что в качестве постоянной Липшица L в (4.147) можно брать норму якобиана J_{n+j}.

Таким образом, из (4.144), (4.147) и (4.149) можно установить справедливость неравенства

$$||z_{n+k}|| \leqslant ||B_n|| \sum_{j=0}^{k-1} ||z_{n+j}|| + h_{n+k-1} L\, c_n \sum_{j=0}^{k} ||z_{n+j}|| + ||\psi_{n+k}||, \tag{4.150}$$

где норма определенной в (4.138) матрицы и константы c_n определяется как

$$||B_n|| = \sum_{j=0}^{k-1} |\alpha_{j,n}|, \;\; c_n = \sum_{j=0}^{k} |\beta_{j,n}|.$$

Для упрощения дальнейшего анализа сходимости рассматриваемых алгоритмов мы перепишем их, как это уже делали раньше, в форме одношаговых методов. Вводя "супервектор"

$$Y_n = (y_{n+k-1}, ..., y_{n+1}, y_n)^T \in \mathbb{R}^{kN},$$

перепишем схему (4.130) в виде

$$Y_{n+1} = (B_n \bigotimes I) Y_n + h_{n+k-1} \bar{F}_n(t_n, Y_n), \tag{4.151}$$

где матрица B_n k-го порядка определена в (4.138), $I \in \mathbb{R}^{N,N}$ есть единичная матрица, а значок $\bigotimes$ означает кронекерово матричное произведение, которое для матрицы $B = \{b_{i,j}\} \in \mathbb{R}^{k,k}$ определяет "большую" блочную матрицу порядка kN:

$$\bar{B} = B \bigotimes I = \{\bar{B}_{i,j} = b_{i,j} I \in \mathbb{R}^{N,N}\} \in \mathbb{R}^{kN,kN}.$$

В последнем члене из (4.151) вектор $\bar{F}_n \in \mathbb{R}^{kN}$ определяется как

$$\bar{F}_n = \bar{C}_n \cdot F_n(t_n, Y_n) + (e_1 \bigotimes I)\beta_{k,n} f_{n+k},$$

$$\bar{C}_n = C_n \bigotimes I \in \mathbb{R}^{kN,kN}, \;\; F_n = \{f_{n+j} : i = 1, ..., N; j = 0, ..., k-1\} \in \mathbb{R}^{kN},$$

$$C_n = \begin{bmatrix} \beta_{k-1,n} & \cdots & \beta_{1,n} & \beta_{0,n} \\ 1 & 0 & \cdots & 0 \\ & \ddots & & \\ 0 & & 1 & 0 \end{bmatrix}, \;\; e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^k. \tag{4.152}$$

Подставляя теперь в (4.151) вместо векторов Y_{n+1}, Y_n соответствующие векторы $Y(t_{n+1}), Y(t_n)$ с компонентами $y(t_{n+j})$ вместо численных решений y_{n+j}, получим следующий векторный аналог уравнения (4.143):

$$Y(t_{n+1}) = \bar{B}_n Y(t_n) + h_{n+k-1}\bar{F}_n(t_n, Y(t_n)) + \Psi_n, \tag{4.153}$$

$$\bar{B}_n = B_n \bigotimes I, \;\; \Psi_n = (\psi_{n+k-1}, ..., \psi_{n+1}, \psi_n)^T.$$

Вычитая теперь почленно уравнения (4.153) и (4.152), с учетом используемых в (4.152) обозначений, для “больших” векторов ошибки

$$Z_{n+1} = Y(t_{n+1}) - Y_{n+1}$$

получаем двучленные соотношения

$$Z_{n+1} = \bar{B}_n Z_n + h_{n+k-1}(\bar{C}_n \bar{J}_n Z_n + \beta_{k,n} \bar{J}_{n,1} Z_{n+1}) + \Psi_n, \tag{4.154}$$

в которых введены квадратные матрицы-якобианы порядка kN

$$\bar{J}_n = J_n \bigotimes I,$$

а также используется обозначение $\bar{J}_{n,1}$ для блочной квадратной матрицы порядка kN, у которой отличен от нуля только левый верхний блок порядка N, равный J_n.

Уравнение (4.154) можно переписать в виде

$$(\bar{I} - h_{n+k-1}\beta_{k,n}\bar{J}_{n,1})Z_{n+1} = (\bar{B}_n + h_{n+k-1}\bar{C}_n\bar{J}_n)Z_n + \Psi_n,$$

где $\bar{I}$ есть единичная матрица порядка kN.

Последнее уравнение можно привести к "канонической" форме

$$\begin{aligned} Z_{n+1} &= \tilde{B}_n Z_n + \tilde{\Psi}_n, \\ \tilde{B}_n &= (\bar{I} - h_{n+k-1}\beta_{k,n}\bar{J}_{n,1})^{-1}(\bar{B}_n + h_{n+k-1}\bar{C}_n\bar{J}_n), \\ \tilde{\Psi}_n &= (I - h_{n+k-1}\beta_{k,n}\bar{J}_{n,1})^{-1}\Psi_n. \end{aligned} \tag{4.155}$$

Отсюда вследствие легко получаемых рекуррентных соотношений для норм

$$\begin{aligned} &||Z_{n+1}|| \leqslant \rho^{n+1}||Z_0|| + (1 + \rho + ... + \rho^n)||\Psi||, \\ &\rho = \max_k\{||\tilde{B}_k||\}, \;\; ||\Psi|| = \max_k\{||\tilde{\Psi}_k||\}, \end{aligned} \tag{4.156}$$

получаем, по аналогии с теоремой 2.1, при условии $\rho \leqslant 1 + b\,h$ $||\Psi|| = O(h^{\gamma+1})$ сходимость по норме порядка γ.

Таким образом, формула (4.156) позволяет оценить вклад в глобальную ошибку Z_{n+1} как погрешности начальных данных Z_0, так и локальных погрешностей $\tilde{\Psi}_k$, если только определить соответствующие нормы операторов $\tilde{B}_k$.

§ 4.15. Решение ОДУ с помощью разрывных методов Галеркина

Разрывные методы Галеркина (РМГ, или DG, или DGM — от Discontinuous Galerkin Methods) впервые предложены Б. Ридом и Т. Хиллом в 1973 г. для решения уравнений в частных производных, описывающих перенос нейтронов. После некоторого перерыва они стали достаточно широко использоваться при решении эллиптических краевых задач. И только позже они стали применяться к решению задач Коши для ОДУ. Мы не ставим себе целью дать развернутый обзор или систематизацию алгоритмов данного класса, пока находящегося еще в стадии становления, а только представим иллюстрацию свежих идей этого безусловно перспективного направления.

4.15.1. Предварительные сведения о РМГ. В этом пункте мы приведем краткое описание идеи данных алгоритмов на простом примере, следуя работе Б. Кокбурна 2003 года (B. Cockburn, Discontinuous Galerkin Methods. Z. Angew. Math. Mech., 83, № 11, 2003, 731—754).

Рассмотрим для простоты линейную скалярную задачу

$$\frac{d\,y}{d\,t} = b(t)y(t),\ \ t \in (0, t_e),\ \ y(0) = y_0, \tag{4.157}$$

где $b(t)$ есть достаточно гладкая функция. В соответствии с методом Галеркина рассмотрим *слабую*, или *вариационную*, постановку задачи (4.157). Для этого умножим обе части данного уравнения на некоторую достаточно гладкую (пробную) функцию $v(t)$ и проинтегрируем результат по t. После применения интегрирования по частям получаем

$$-\int\limits_0^{t_e} y(t)\frac{d\,v(t)}{d\,t}d\,t + y\,v|_0^{t_e} = \int\limits_0^{t_e} b(t)y(t)v(t)d\,t \tag{4.158}$$

для любой функции $v(t)$ из некоторого класса $\mathcal{P}^{(k)}$. В этой вариационной (от английского vary — изменяться) формулировке за $\mathcal{P}^{(k)}$ можно взять пространство полиномов степени не выше k.

Задачу (4.158) дискретизируем следующим образом. Расчетную область $[0, t_e]$ разобьем на N_e интервалов $I_n = (t_n, t_{n+1}),\ n = 0, ..., N_e - 1$, т. е. построим неравномерную сетку с шагами $h_n = t_{n+1} - t_n,\ t_0 = 0,\ t_{N_e} = t_e$. Выпишем систему N_e уравнений, каждое из которой отличается от (4.158) интервалом интегрирования — I_n вместо $[0, t_e]$:

$$-\int\limits_{I_n} y_h(t)\frac{d\,v(t)}{d\,t}d\,t + \hat{y}_h v|_{t_n}^{t_{n+1}} = \int\limits_{t_n}^{t_{n+1}} b(t)y(t)v(t)d\,t,\ \ \forall\, v(t) \in \mathcal{P}^{(k_n)}, \tag{4.159}$$

где $y_h(t)$ на каждом из интервалов I_n есть "свой" многочлен из $\mathcal{P}^{(k_n)}$, т. е. представляет собой разрывное кусочно-полиномиальное приближение искомой функции $y(t)$, а $\hat{y}_h$

есть след $y_h(t)$ в точках сетки t_n. Этот след, вообще говоря, можно определить различным образом, и один из естественных — это следующий:

$$\hat{y}_h(t_n) = \begin{cases} y_0, & t_n = 0, \\ y_h(t_n^-) \equiv \lim\limits_{\varepsilon \to 0} y_h(t_n - \varepsilon), & t_n > 0. \end{cases} \tag{4.160}$$

Соотношения (4.159), (4.160), с точностью до порядка выбираемого полинома и выбора квадратурной формулы в правой части (4.159), определяют численный метод решения задачи Коши (4.157). Возможны и другие определения следа $\hat{y}_h(t)$, от которых будут зависеть свойства согласованности, устойчивости и точности вычислительной схемы.

Отметим, что приведенная выше дискретная слабая постановка является согласованной в том смысле, что при замене в (4.159) величин $y_h(t)$ и $\hat{y}_h$ на соответствующие точные значения, т. е. на решение исходной задачи (4.157), рассматриваемые вариационные уравнения также будут выполняться.

Важным свойством данных методов является локальная консервативность, или балансность. А именно, если в соотношениях (4.159) положить $v(t) \equiv 1$, то мы получаем

$$\hat{y}_h|_{t_n}^{t_{n+1}} = \int\limits_{y_n}^{t_{n+1}} f(t) d\, t. \tag{4.161}$$

Если же первый член в уравнении (4.159) проинтегрировать по частям, то приходим к следующему равенству:

$$\int\limits_{I_n} \frac{dy_h(t)}{d\,t} v(t) d\,t + (\hat{y}_h - y_h) v \Big|_{t_n}^{t_{n+1}} = \int\limits_{I_n} f(t) y(t) v(t) d\,t,$$

которое можно переписать в виде

$$\int\limits_{I_n} r(t) v(t) d\,t = (y_h - \hat{y}_h) v \Big|_{t_n}^{t_{n+1}},$$

где $r(t) = \dot{y}_h - b\,y_h$ обозначает невязку исходного уравнения. В случае $v(t) = 1$ отсюда получаем соотношение

$$\int\limits_{I_n} r(t) d\,t = [y_h]_n. \tag{4.162}$$

Здесь введено обозначение для скачка функции

$$[y_h]_n = y_h^-(t_{n+1}) - y_h^+(t_n), \;\; y_h^\pm(t) = y(t^\pm) = \lim_{\varepsilon \to 0} y_h(t \pm \varepsilon). \tag{4.163}$$

Таким образом, скачок приближенного решения на сеточном интервале I_n равен соответствующему интегралу от невязки.

Рассмотрим теперь вопросы, касающиеся устойчивости полученной вычислительной схемы. При умножении исходного ОДУ (4.157) на $y(t)$ и интегрирования по расчетной области приходим к равенству

$$\frac{1}{2}\Big[y^2(t_e) - y^2(t_0)\Big] = \int\limits_0^{t_e} b\,(t) y^2(t) d\,t.$$

Аналогичный результат можно получить и для схемы Галеркина. Для этого мы полагаем в (4.159) $v = y_h$, интегрируем получаемое соотношение по частям и суммируем по n:

$$\sum_{n=0}^{N-1}\Big(-\frac{1}{2}y_h^2+\hat{y}_h y_h\Big)\Big|_{t_n}^{t_{n+1}}=\frac{1}{2}y_h^2(t_e^-)+T_h(t_e)-\frac{1}{2}y_0^2=$$
$$=\int_0^{t_e}b(t)y_h^2(t)\,d\,t, \tag{4.164}$$

где

$$T_h(t_e)=-\frac{1}{2}y_h^2(t_e^-)+\sum_{n=0}^{N-1}\Big(-\frac{1}{2}y_h^2+\hat{y}_h y_h\Big)\Big|_{t_n}^{t_{n+1}}+\frac{1}{2}y_0^2.$$

Отметим, что если бы величина $T_h(t_e)$ была неотрицательная, то равенство (4.164) могло бы рассматриваться как свойство устойчивости данной схемы Галеркина. Другими словами, ее устойчивость гарантируется, если определение следа $\hat{y}_h$ обеспечивает неравенство $T_h(t_e)\geqslant 0$. Например, при $b=0$ в данном случае из (4.164) следует невозрастание неотрицательной величины $y_h^2(t_e^-)$.

Полагая $y_h(t)=y_0$ при $t<0$ и используя обозначения (4.163), а также

$$\{y_h\}=(y_h^++y_h^-)/2,\ \ [y_h]=y_h^--y_h^+,$$

для $T_h(t_e)$ получаем следующее представление:

$$T_n(t_e) = -\frac{1}{2}y_h^2(t_e^-) + \big(-\frac{1}{2}y_h^2(t_e^-) + \hat{y}_h(t_e)y_h(t_e^-)\big) +$$

$$+\sum_{n=1}^{N-1}\big(-\frac{1}{2}[y_h^2] + \hat{y}_h[y_h]\big)(t^n) - \big(-\frac{1}{2}y_h^2(0^+) + \hat{y}_h(0)y_h(0^+)\big) + \frac{1}{2}y_0^2 =$$

$$= \big(\hat{y}_h(t_e) - y_h(t_e^-)\big)y_h(t_e^-) + \sum_{n=1}^{N-1}\big((\hat{y}_h - \{y_h\})\big)[y_h](t_n) -$$

$$-(\hat{y}_h(0) - y_h(0^+) + \frac{1}{2}[y_h^2](0).$$

При выводе этого выражения использовано соотношение

$$[y_h^2] = 2\{y_h\}[y_h].$$

Далее легко проверить, что если мы положим

$$\hat{y}_h(t_n) = \begin{cases} y_0, & \text{при} \quad t_n = 0, \\ (\{y_h\} + C_n[y_h])(t_n), & \text{при} \quad t_h \in (0, t_e), \\ y(t_e^-), & \text{при} \quad t_n = t_e, \end{cases}$$

где $C_n \geqslant 0$ суть некоторые постоянные, то при $C_0 = 1/2$ будем иметь

$$T_h(t_e) = \sum_{n=0}^{N-1} C_n[y_h]^2(t_n) \geqslant 0,$$

что мы и хотели получить для обеспечения устойчивости.

Отметим, что выбор $C_n \equiv 1/2$ соответствует определению следа в формуле (4.160). Важно также, что разрывная схема Галеркина при условии $C_n \geqslant 0$ не только устойчива, но и

согласована, поскольку в данном случае условие $\hat{y} = y$ удовлетворяется.

Можно показать (доказательство, ввиду его громоздкости, опускаем), что если $C_n = \dfrac{1}{2}$ для всех n, то порядок погрешности РМГ в точках t_n будет $2k + 1$, а при $C_n = 0$ ошибка равна $O(h^{2k+2})$. Однако в последнем случае, как и при других значениях $C_n \neq 1/2$, алгоритм приводит к необходимости вычисления y_h одновременно (неявно) для всей области $[0, t_e]$. И только при $C_n \equiv 1/2$ нахождение решения может реализовываться последовательно шаг за шагом.

4.15.2. Вопросы A-устойчивости РМГ. В данном пункте мы рассмотрим модельную задачу Далквиста

$$y(t) = \lambda\, y, \quad y(0) = 1, \tag{4.165}$$

где λ есть комплексная постоянная. Напомним, что областью абсолютной устойчивости алгоритма с постоянным шагом h есть множество значений величины $z = \lambda\, h$ в комплексной плоскости, для которых численное решение задачи Коши (4.165) остается ограниченным. Если область абсолютной устойчивости включает всю левую полуплоскость $\mathrm{Re}(\lambda\, h) < 0$ (отметим, что при $\lambda < 0$ как решение исходной дифференциальной задачи, так и численное решение, стремятся к нулю при $t \to \infty$), то метод называется A-устойчивым.

При решении жестких систем ОДУ, у которых различные компоненты вектора $y(t) = \{y_i(t)\}$ имеют сильно отличающиеся временные масштабы изменений, для соответствующих алгоритмов важное значение имеет свойство жесткой

A-устойчивости, характеризуемое поведением численных решений при $\mathrm{Re}(\lambda h) \to -\infty$. А именно, если при $\mathrm{Re}(z) \to -\infty$ $y_i^{n+1}/y_i^n \to 0$ для всех i, то для быстро убывающих компонент исходного решения ОДУ будут быстро стремиться к нулю и соответствующие компоненты численного решения, и такой метод называется жестко A-устойчивым.

Отметим, что для скалярной задачи (4.165) точное решение имеет вид $y(t) = e^{\lambda t}$, а численное решение для линейного метода можно записать в виде

$$
\begin{aligned}
&y^n = R_{p,q}(\lambda h) y^{n-1}, \quad n = 1, 2, ..., \\
&R_{p,q}(z) = \frac{P_p(z)}{Q_q(z)} = \frac{a_0 + a_1 z + ... + a_p z^p}{b_0 + b_1 z + ... + b_q z^q},
\end{aligned}
\tag{4.166}
$$

где $R_{p,q}(z)$ есть аппроксимация Паде экспоненты e^z. При $p = q$ аппроксимация называется диагональной, а при $p = q - 1$ — поддиагональной (предполагается, что $a_p a_q \neq 0$). Приведем некоторые свойства аппроксимаций Паде:

$|R_{p,q}(z)| < 1$ при $\mathrm{Re}(z) < 0$ для $p = q,\ q-1$ или $q-2$,

$R_{p,p}(z) \to (-1)^p$ при $\mathrm{Re}(z) \to -\infty$,

$R_{p,q}(z) \to 0$ при $\mathrm{Re}(z) \to -\infty$ для $p = q,\ q-1$ или $q-2$,

$\min\limits_{a_i, b_j} |R_{p,q}(z) - e^{\lambda z}| = O(z^{p+q+1})$ при $|z| \to 0$.

Последнее равенство означает, что экспонента аппроксимируется рациональной функцией с порядком $p+q+1$.

Разрывный метод Галеркина для модельного уравнения (4.165) можно сформулировать следующим образом: найти функцию $y_h \in \mathcal{P}^{(k)}(I_n)$ для $n = 1, 2, ..., N$ такую, что при любом $v_h \in \mathcal{P}^{(k)}(I_n)$ выполняются равенства

$$B_n(v_h, y_h) \equiv \int\limits_{I_n} (\dot{y}_h v_h - \lambda\, y_h v_h) d\,t + y_{h+}^{n-1} v_{h+}^{n-1} = \tag{4.167}$$

$$= y_{h-}^{n-1} v_{h+}^{n-1} \equiv L_n(v_h), \quad y_{h-}^0 = 1.$$

Полагая в (4.167) $v_h = y_h$ и используя формулы интегрирования по частям

$$\int\limits_{I_n} \dot{y}_h y_h d\,t = y_h^2\Big|_{t_+^{n-1}}^{t_-^n} - \int\limits_{I_n} y_h \cdot \dot{y}_h d\,t =$$

$$= \frac{1}{2}((y_{h-}^n)^2 - (y_{h+}^{n-1})),$$

в итоге получаем соотношения

$$-\mathrm{Re}(\lambda) \int\limits_{I_n} y_h^2 d\,t + \frac{1}{2}(y_{h-}^n)^2 + \frac{1}{2}(y_{h+}^{n-1})^2 =$$

$$= y_{h-}^{n-1} y_{h+}^{n-1} \leqslant \frac{1}{2}((y_{h-}^{n-1})^2 + (y_{h+}^{n-1})^2).$$

Отсюда следуют неравенства

$$-\mathrm{Re}(\lambda) \int\limits_{I_n} y_h^2 d\,t + \frac{1}{2}(y_{h-}^n)^2 \leqslant \frac{1}{2}(y_{h-}^{n-1})^2,$$

которые означают, что для $\mathrm{Re}(\lambda) < 0$ имеем $(y^n_{h-})^2 \leqslant (y^{n-1}_{h-1})^2$, т. е.

$$|y^n_{h-}| \leqslant |y^{n-1}_{h-}|.$$

Таким образом, для любых порядков РМГ(к) обеспечивается A-устойчивость.

Отметим еще, что для вещественных отрицательных λ решение модельной задачи Далквиста является монотонным, т. е.

$$y(t+\delta) < y(t) \text{ при } \delta > 0.$$

Естественно, желательно, чтобы свойство монотонности наследовалось и для численного метода. Рассмотрим теперь простейшие примеры разрывных методов Галеркина для уравнения Далквиста (4.165).

Вычисление линейной и билинейной форм на отрезке $I_{n-1} = (t_{n-1}, t_n)$ для различных порядков аппроксимаций удобно проводить при отображении его на "стандартный" интервал $I^\xi_n = (-1, 1)$ с помощью линейного преобразования

$$\xi(t) = c_1 + c_2 t,$$

где постоянные c_1, c_2 определяются из условий

$$\xi(t_{n-1}) = -1, \quad \xi(t_n) = 1,$$

откуда следуют простые формулы

$$\xi(t) = \frac{2t - t_{n-1} - t_n}{h}, \;\; t(\xi) = \frac{h\xi + t_{n-1} + t_n}{2}.$$

При этом вариационные уравнения (4.167) принимают вид

$$\int_{-1}^{1} y'(\xi)v(\xi)d\xi - \frac{\lambda h}{2}\int_{-1}^{1} y(\xi)v(\xi)d\xi + y_+^{n-1}v_+^{n-1} = y_-^{n-1}v_+^{n-1}, \tag{4.168}$$

а сами величины $y(\xi), v(\xi)$ выражаются через полиномиальные базисные функции $\varphi_i(\xi) \in \mathcal{P}^{(k)}(-1,1), i = 0, ..., k$. Разрывные методы Галеркина порядка ,как k, будем в дальнейшем обозначать через РМГ(k) или DGM(k).

Для метода нулевого порядка имеем $k = 0$, $\varphi_0(\xi) = 1$, $\varphi_0'(\xi) = 0$ и $y_-^n = y_+^n = y^n$, $v_-^n = v_+^{n-1} = v^n$ $y(\xi) = y^n$, $y'(\xi) = 0$, $v(\xi) = v^n$.

Таким образом, уравнение (4.168) принимает вид

$$(y^n - \lambda\, h\, y^n)v^n = y^{n-1}v^n,$$

откуда, в силу произвольности v^n, получаем

$$y^n = R_{0,1}y^{n-1} = (1 - \lambda\, h)^{-1}y^{n-1}. \tag{4.169}$$

Последнее означает, что РМГ(0) представляет собой неявный метод Эйлера со свойствами

$$0 < R_{0,1}(\lambda\, h) \leqslant 1, \text{ при } \mathrm{Re}(\lambda\, h) \leqslant 0,$$

$$|R_{0,1}(\lambda\, h)| \to 0 \text{ при } \mathrm{Re}(\lambda\, h) \to -\infty,$$

$$|R_{0,1}(\lambda\, h) - e^{\lambda\, h}| = O(h^2).$$

Кроме того, поскольку для вещественных отрицательных λ значения $R_{0,1}(\lambda)$ положительны, данная схема обладает

свойством монотонности при любых шагах сетки h, т.е. является абсолютно монотонной и $y^n < y^{n-1}$.

Рассмотрим теперь схему первого порядка РМГ(1), в которой используются следующие базисные функции и их производные:

$$\begin{aligned} &\varphi_1(\xi) = (1-\xi)/2, \quad \varphi_1'(\xi) = -1/2, \\ &\varphi_2(\xi) = (1+\xi)/2, \quad \varphi_2'(\xi) = 1/2, \end{aligned} \tag{4.170}$$

обладающие свойствами

$$\varphi_1(-1) = \varphi_2(1) = 1, \quad \varphi_1(1) = \varphi_2(-1) = 0,$$

и для которых выполняются интегральные соотношения

$$\int_{-1}^{1} \varphi_1\varphi_1'\,d\xi = \int_{-1}^{1} \varphi_1'\,d\xi = -1/2, \quad \int_{-1}^{1} \varphi_2\varphi_2'\,d\xi = \int_{-1}^{1} \varphi_1\varphi_2'\,d\xi = 1/2,$$
$$\int_{-1}^{1} \varphi_1^2\,d\xi = \int_{-1}^{1} \varphi_1\varphi_2\,d\xi = 1/3, \quad \int_{-1}^{1} \varphi_2^2\,d\xi = 2/3.$$

С помощью этих функций выписываются разложения

$$y(\xi) = \varphi_1(\xi)y_+^{n-1} + \varphi_2(\xi)y_-^n, \quad y'(\xi) = \varphi_1'(\xi)y_+^{n-1} + \varphi_2'(\xi)y_-^n,$$

$$v(\xi) = \varphi_1(\xi)v_+^{n-1} + \varphi_2(\xi)v_-^n,$$

а также выражаются интегралы, входящие в уравнение (4.167):

$$\int_{-1}^{1} y'(\xi)v(\xi)\,d\xi = \Big(y_+^{n-1}\int_{-1}^{1}\varphi_1\varphi_1'\,d\xi + y_-^{n}\int_{-1}^{1}\varphi_1\varphi_2'\,d\xi\Big)v_+^{n-1}+$$

$$+\Big(y_+^{n-1}\int_{-1}^{1}\varphi_2\varphi_1'\,d\xi + y_-^{n}\int_{-1}^{1}\varphi_2\varphi_2'\,d\xi\Big)v_-^{n} =$$

$$= \Big((-y_+^{n-1}+y_-^{n})v_+^{n-1} + (-y_+^{n-1}+y_-^{n})v_-^{n}\Big)/2,$$

$$\frac{\lambda h}{2}\int_{-1}^{1} y(\xi)v(\xi)\,d\xi = \frac{\lambda h}{2}\Big(\big(y_+^{n-1}\int_{-1}^{1}\varphi_1^2\,d\xi + y_-^{n}\int_{-1}^{1}\varphi_1\varphi_2\,d\xi\big)v_+^{n-1}+$$

$$+\big(y_+^{n-1}\int_{-1}^{1}\varphi_1\varphi_2\,d\xi + y_-^{n}\int_{-1}^{1}\varphi_2^2\,d\xi\big)v_-^{n}\Big) =$$

$$= \frac{\lambda h}{6}\big[(2y_+^{n-1}+y_-^{n})v_+^{n-1} + (y_+^{n-1}+2y_-^{n})v_-^{n}\big].$$

Отсюда вариационные соотношения (4.167) принимают вид

$$(a_{1,1}y_+^{n-1} + a_{1,2}y_-^{n})v_+^{n-1}+$$

$$+(a_{2,1}y_+^{n-1} + a_{2,2}y_-^{n})v_-^{n} = y_-^{n-1}v_+^{n-1},$$

$$a_{1,1} = a_{2,2} = 1/2 - \lambda h/3, \quad a_{1,2} = 1/2 - \lambda h/6, \quad a_{2,1} = -1/2 - \lambda h/6.$$

В силу произвольности величин v_+^{n-1}, v_-^{n} далее для каждого интервала I_n получаем два уравнения:

$$a_{1,1}y_+^{n-1} + a_{1,2}y_-^{n} = y_-^{n-1}, \quad a_{2,1}y_+^{n-1} + a_{2,2}y_-^{n} = 0,$$

из которых после исключения y_+^{n-1} имеем

$$y_-^n = R_{1,2}(\lambda h) y_-^{n-1}, \quad R_{1,2}(z) = \frac{2(3+z)}{z^2 - 4z + 6}. \tag{4.171}$$

Вследствие очевидных свойств функции устойчивости $R_{1,2}(z)$ (см. определение 2.13)

$$R_{1,2}(z) \to 0 \ \text{ при } \ \mathrm{Re}(z) \to -\infty,$$

$$|R_{1,2}(z) - e^z| = O(h^4),$$

метод РМГ(1) является жестко A-устойчивым и имеет третий порядок.

Числитель $R_{1,2}(z)$ имеет нуль при $z = -3$ и является отрицательным при $z < -3$, а знаменатель положителен для всех вещественных $z < 0$. Таким образом, $R_{1,2}(z) < 0$ для вещественных отрицательных λ, если шаг сетки $h > -3/\lambda$, вследствие чего последовательные численные решения ОДУ будут осциллировать между положительными и отрицательными значениями. Таким образом, схема РМГ(1) является условно монотонной, а точнее, монотонной только при $h < -3/\lambda$.

В заключение данного пункта кратко представим РМГ(2). Лагранжевы базисные функции второго порядка в локальных координатах и их производные при этом имеют вид

$$\varphi_1(\xi) = (\xi^2 - \xi)/2, \quad \varphi_2(\xi) = 1 - \xi^2, \quad \varphi_3(\xi) = (\xi + \xi^2)/2,$$

$$\varphi_1'(\xi) = \xi - 1/2, \quad \varphi_2'(\xi) = -2\xi, \quad \varphi_3'(\xi) = \xi + 1/2, \tag{4.172}$$

который определяется по условиям

$$\varphi_1(-1) = \varphi_2(0) = \varphi_3(1) = 1,$$

$$\varphi_1(0) = \varphi_1(1) = \varphi_2(-1) = \varphi_2(1) = \varphi_3(-1) = \varphi_3(0) = 0.$$

На интервале I_n^ξ искомое приближенное решение $y(\xi)$ и пробная функция $v(\xi)$ выражаются через пробные функции следующим образом:

$$y(\xi) = \varphi_1 y_+^{n-1} + \varphi_2 y^{n-1/2} + \varphi_3 y_-^n,$$

$$v(\xi) = \varphi_1 v_+^{n-1} + \varphi_2 v^{n-1/2} + \varphi_3 v_-^n.$$

Опуская промежуточные выкладки, приведем получаемые при этом выражения для интегральных членов в вариационных равенствах (4.168):

$$\int_{-1}^{1} y'(\xi)v(\xi) = \Big(-\frac{1}{2}y_+^{n-1} + \frac{2}{3}y^{n-1/2} - \frac{1}{6}y^n\Big)v_+^{n-1} +$$

$$+\Big(-\frac{2}{3}y_+^{n-1} + \frac{2}{3}y_-^n\Big)v^{n-1/2} + \Big(\frac{1}{6}y_+^{n-1} - \frac{2}{3}y^{n-1/2} + \frac{1}{2}y_-^n\Big)v_-^n,$$

$$\int_{-1}^{1} y(\xi)v(\xi)\,d\xi = \Big(\frac{4}{15}y_+^{n-1} + \frac{2}{15}y^{n-1/2} - \frac{1}{15}y_-^n\Big)v_+^{n-1} +$$

$$+\Big(\frac{2}{15}y_+^{n-1} + \frac{16}{15}y^{n-1/2} + \frac{2}{15}y_-^n\Big)v^{n-1/2} + \Big(-\frac{1}{15}y_+^{n-1} + \frac{2}{15}y^{n-1/2} + \frac{4}{15}y_-^n\Big)v$$

Подставляя далее эти представления в (4.168) и используя произвольность значений $v_+^{n-1}, v^{n-1/2}, v_-^n$ относительно неиз-

вестных значений $y_+^{n-1}, y^{n-1/2}, y_-^n$, для каждого интервала получаем СЛАУ третьего порядка, которую можно записать в следующей векторно-матричной форме ($z = \lambda\, h$):

$$\begin{bmatrix} 15-4z & 20-2z & -5+z \\ -10-z & -8z & 10-z \\ 5+z & -20-2z & 15-4z \end{bmatrix} \begin{bmatrix} y_+^{n-1} \\ y^{n-1/2} \\ y_-^n \end{bmatrix} = \begin{bmatrix} 30y_+^{n-1} \\ 0 \\ 0 \end{bmatrix}. \tag{4.173}$$

Из решения этой системы получаем рекуррентные соотношения

$$y_-^n = R_{2,3}(z) y_-^{n-1},$$

где $R_{2,3}$ есть поддиагональная аппроксимация экспоненты

$$R_{2,3}(z) = \frac{3z^2 + 8z + 20}{60 - 36z + 9z^2 - z^3},$$

которая обладает следующими легко проверяемыми свойствами:

$$|R_{2,3}(z)| \to 0 \ \text{ при } \ \mathrm{Re}(z) \to -\infty, \quad |R_{2,3}(z) - e^z| = O(h^5).$$

Таким образом, схема РМГ(2) имеет пятый порядок точности и является жестко A-устойчивой. Кроме того, $R_{2,3}(0) = 1$ и $R_{2,3}(z) > 0$ для всех вещественных отрицательных z, что означает абсолютную монотонность данного алгоритма ($y_-^n < y_-^{n-1}$).

Аналогично можно показать, что разрывный метод Галеркина третьего порядка (РГМ(3)) является жестко A-

устойчивым, имеет глобальную погрешность $O(h^7)$, но не обладает свойством абсолютной монотонности.

4.15.3. Мульти-адаптивные методы Галеркина. В данном пункте мы будем следовать оригинальной работе А. Логга (A. Logg. Multi–adaptive Galerkin methods for ODEs, SIAM J. Sci. Comput., vol. 24, № 6, 2003, 1879–2002), где предложены непрерывные и разрывные методы Галеркина для решения систем ОДУ, в которых для различных искомых функций $y_i(t)$ при численном интегрировании адаптивным образом выбираются разные шаги сетки $h_{n_i} = t_{n_i} - t_{n_i-1},\ n_i = 1, 2, ..., N_i^e$. Данная возможность имеет очень большое значение во многих актуальных приложениях, когда различные компоненты решений имеют сильно отличающиеся временные масштабы динамических процессов. В частности, сюда относятся типичные задачи химической кинетики.

Рассмотрим задачу Коши достаточно общего вида

$$\dot{y}(t) = f(t, y(t)), \quad t \in (0, t_e], \quad y(0) = y^0 \in \mathbb{R}^N, \tag{4.174}$$

где $f\colon (0, t_e] \times \mathbb{R}^N$ есть заданная вектор-функция размерности N, удовлетворяющая условию Липшица. При численном интегрировании задачи (4.174) каждую из компонент искомого векторного решения $y = \{y_i(t)\} \in \mathbb{R}^N$ будем определять в различных точках t_{n_i}, которые, однако, совпадают в некоторых точках синхронизации $t^{(l)},\ l = 0, 1, ..., M$:

$$t^{(0)} = t_0 < t^{(1)} < ... < t^{(M)} = t_e.$$

Совокупность временных точек $t^{(l)}$ и расстояния между ними

$$H_l = t^{(l)} - t^{(l-1)}, \quad l = 1, ..., M,$$

можно назвать макроузлами и макрошагами макросетки Ω_H, которая является общей для дискретизации всех искомых функций $y_i(t)$. В свою очередь, если мы обозначим через $n_{i,l}$ номер узла i-й сетки, совпадающий с l-м макроузлом, т. е. $t_{n_i,l} = t^{(l)}$, то каждый макрошаг содержит

$$N_{i,l} = n_{i,l} - n_{i,l-1}, \quad i = 1, ..., N; \quad j = 1, ..., M,$$

"обычных" шагов i-й сетки, причем $N_{i,M} = N_i^l$.

Каждую из функций $y_i(t)$ аппроксимируем на "своих" интервалах $I_{n_i} = (t_{n_i-1}, t_{n_i})$, причем методы приближений, вообще говоря, для различных i и n_i могут отличаться.

Непрерывный метод Галеркина k-го порядка НМГ(k) для решения задачи (4.174) формируется следующим образом: найти функцию y^h из некоторого пространства W (которое будем называть пространством тестовых функций) с условием $y^h(0) = y^0$ такую, что для любой пробной функции $v \in V$ выполняется равенство

$$\int_0^{t_e} (\dot{y}^h, v)\, d\, t = \int_0^{t_e} (f(t, y^h(t))), v)\, d\, t, \tag{4.175}$$

где пространства V и W суть

$$V = \{v \in [C([0, t_e])]^N \colon v_i\big|_{I_{n_i}} \in \mathcal{P}^{(k_{n_i}-1)}(I_{n_i}),\ \ i = 1, ..., N\},$$

$$W = \{w \colon w_i\big|_{I_{n_i}} \in \mathcal{P}^{(k_{n_i})}(I_{n_i}),\ \ i = 1, ..., N;\ n = 1, ...\}.$$

Здесь $\mathcal{P}^{(k_{n_i})}(I_{n_i})$ означает линейное пространство многочленов степени не выше k_{n_i}, определенных на соответствующем интервале I_{n_i}.

Таким образом, пробные функции из V представляют собой кусочно-непрерывные многочлены, имеющие локальные порядки $k_{n_i} - 1$, а тестовые функции из W имеют непрерывные кусочно-полиномиальные представления с локальными порядками k_{n_i}.

Поскольку пробные функции v являются разрывными и разными на I_{n_i}, глобальная вариационная задача (4.175) может быть преобразована в совокупность последовательных локальных задач для соответствующих компонент: для $i = 1, ..., N$, $n_i = 1, ..., N_i^e$, найти функции $y_i^h\big|_{I_{n_i}} \in \mathcal{P}^{(k_{n_i})}(I_{n_i})$ с заданными значениями $y_i^h(t_{n_i-1})$ такие, что для каждого $v \in \mathcal{P}^{(k_{n_i}-1)}(I_{n_i})$ выполняются вариационные соотношения

$$\int\limits_{I_{n_i}} \dot{y}_i^h v \, d\, t = \int\limits_{I_{n_i}} f_i(t, y^h(t)) v \, d\, t. \tag{4.176}$$

Если мы определим невязку приближенного решения

$$r(t, y^h(t)) = f(t, y^h(t)) - \dot{y}_i^h, \tag{4.177}$$

то вариационная постановка (4.176) переформулируется как свойство галеркинской ортогональности НМГ(k):

$$\int\limits_{I_{n_i}} r_i(t, y^h(t)) v \, d\, t = 0, \quad \forall v \in \mathcal{P}^{(k_{n_i})}(I_{n_i}). \tag{4.178}$$

Отметим, что в правой части (4.176) функция f_i зависит, в принципе, от всех компонент $y_i^h(t)$ искомого вектора решения. Таким образом, локальная задача для интервала I_{n_i}, вообще говоря, неявно определяет связанные между собой компоненты, для которых требуется решить систему нелинейных уравнений.

Для реализации алгоритма каждую компоненту искомого приближенного решения $y_i^h(t)$ на локальном интервале I_{n_i} раскладываем по базисным функциям пространства $\mathcal{P}^{(k_{n_i})}(I_{n_i})$, а затем для вычисления интегралов в (4.176) или (4.178) применяем какие-либо квадратурные формулы. При этом в качестве базисных функций удобно брать фундаментальные интерполяционные многочлены Лагранжа [30], построенные по квадратурным узлам соответствующей формулы интегрирования.

В случае непрерывного метода Галеркина естественно использовать методы численного интегрирования максимальной алгебраической точности, при закреплении двух квадратурных узлов в концевых точках интервала интегрирования I_{n_i}. Такой алгоритм носит имя Лобатто и является частным случаем метода Маркова [30]. На стандартном интервале интегрирования $x \in [0, 1]$ получаемые квадратурные узлы (в количестве $k + 1$) являются корнями многочленов (также называемых иногда именем Лобатто), определяемых рекуррентными соотношениями

$$P_{k+1}(x) = xL_k(x) - L_{k-1}(x),$$

где $L_k(x)$ суть ортогональные на данном отрезке многочле-

ны Лежандра. Напомним, что погрешность соответствующей квадратурной формулы есть $O(h^{2k})$.

Решение рассматриваемой задачи можно представить следующим образом. Пусть $\{s_j; j = 0, 1, ..., k\}$ есть множество точек на интервале $[0, 1]$, $s_0 = 0$, $s_k = 1$, $s_2, ...$, представляющих собой узлы Лобатто. Обозначим через $\tau_{n_i}(t)$ линейное отображение интервала $I_{n_i} = [t_{n_i-1}, t_{n_i}]$ на отрезке $(0, 1]$, определяемое как

$$\tau_{n_i}(t) = \frac{t - t_{n_i-1}}{t_{n_i - t_{n_i-1}}}.$$

Лагранжевые базисные функции пространства многочленов $\mathcal{P}^{(k)}([0, 1])$, построенные по точкам s_j, записываются в виде

$$\lambda_j^{(k)}(s) = \frac{(s - s_0) \cdots (s - s_{j-1})(s - s_{j+1}) \cdots (s - s_k)}{(s_j - s_0) \cdots (s_j - s_{j-1})(s_j - s_{j+1}) \cdots (s_j - s_k)}.$$

Функцию $y_i^h(t)$ на интервале I_{n_i} можно представить в форме полинома k_{n_i}-го порядка

$$y_i^h(t) = \sum_{j=0}^{k_{n_i}} \beta_{n_i,j} \lambda_j^{(k_{n_i})}(\tau_{n_i}(t)). \tag{4.179}$$

Выбирая теперь в качестве пробных функций $v(t) \in V$ базисные многочлены $\lambda_j^{(k_{n_i}-1)}$, локальную вариационную задачу (4.176) можно представить таким образом: найти значения $\beta_{n_i,j}, j = 0, 1, ..., k_{n_i}$, такие, что для всех $m = 0, 1, ..., k_{n_i} - 1$ выполняются равенства

$$\int\limits_{I_{n_i}} \sum_{j=0}^{k_{n_i}} \beta_{n_i,j} \frac{d}{d\,t} [\lambda_j^{(k_{n_i})}(\tau_{n_i}(t))] \lambda_m^{(k_{n_i}-1)}(\tau_{n_i}(t)) d\,t = \\ = \int\limits_{I_{n_i}} f_i \lambda_m^{(k_{n_i}-1)}(\tau_{n_i}(t)) d\,t. \tag{4.180}$$

Поскольку в НМГ(k) искомые решения $y_i^h(t)$ являются непрерывными, то в разложениях (4.179) надо положить

$$\beta_{n_i,k_{n_i}} = \beta_{n_i+1,0}, \quad n_i = 1, 2, ..., N_i^l - 1. \tag{4.181}$$

Таким образом, соотношения (4.180) представляют собой систему нелинейных уравнений (СНАУ) относительно "галеркинских" коэффициентов $\beta_{i,j}$, поскольку функция f зависит в общем случае от всех компонент искомого решения $y_i(t)$.

В силу условий непрерывности (4.181) СНАУ (4.180) может последовательно решаться по макрошагам, причем на каждом l-м макрошаге порядок такой системы, т. е. количество уравнений и неизвестных равен

$$N_l^H = \sum_{i=1}^{N} \sum_{n_i=n_i,l-1}^{n_i=n_{i,l}} k_{n_i}.$$

В некотором смысле данный алгоритм можно считать обобщением неявного метода Рунге–Кутты с использованием квадратурных узлов Лобатто.

Разрывный метод Галеркина в локальной форме определяется следующим образом: для $i = 1, ..., N,\ n_i = 1, ..., N_i^e$ найти функцию $y_i^h\big|_{I_{n_i}} \in \mathcal{P}^{(k_{n_i})}(I_{n_i})$ такую, что при всех $v \in \mathcal{P}^{(k_{n_i})}(I_{n_i})$ выполняются вариационные равенства

$$[y_i^h]_{n_i-1} v(t^+_{n_i-1}) + \int\limits_{I_{n_i}} \dot{y}_i^h v \, d\, t = \int\limits_{I_{n_i}} v \, f_i(t, y_i^h) d\, t, \qquad (4.182)$$

где $[\cdot]$ означает скачок, т. е. $[v]_{n_i} = v(t^+_{n_i}) - v(t^-_{n_i})$, а начальные условия для всех i определяются как $y_i^h(0^-) = y_i(0)$. На глобальном уровне пространства пробных и тестовых функций в этом случае совпадают и определяются как

$$V = W = \{v : v_i\big|_{I_{n_i}} \in \mathcal{P}^{(k_{n_i})}(I_{n_i}), \;\; i = 1, ..., N; \; n_i = 1, ..., N_i^e\}, \qquad (4.183)$$

где функции v на границах интервалов I_{n_i}, вообще говоря, терпят разрывы.

Если мы теперь определим невязку приближенного решения в соответствии с (4.177), как и для непрерывного метода Галеркина, то равенство (4.182) можно записать в виде

$$\int\limits_{I_{n_i}} r_i(t, y^h(t)) v \, d\, t - [y_i^h]_{n_i-1} v(t^+_{n_i-1}) = 0, \;\; \forall v \in \mathcal{P}^{(k_{n_i})}(I_{n_i}).$$

Формально данное соотношение можно также считать условием галеркинской ортогональности, если мы обобщим понятие интеграла, добавив к интервалу I_{n_i} его левую концевую точку (напомним, что производная разрывной функции является дельта-функцией Дирака).

Для приближенного вычисления интегралов в уравнениях (4.182), содержащих разрывные функции, применяются квадратурные формулы максимальной алгебраической точности,

в которых зафиксирован только один квадратурный узел, совпадающий с правым (для определенности) концом интервала I_{n_i}. Эти формулы носят имя Радо [30], а их узлы являются корнями многочленов (k_{n_i+1})-го порядка

$$R_{k_{n_i}+1}(t) = L_{k_{n_i}}(t) + L_{k_{n_i}+1}(t).$$

Порядок погрешности этой формулы есть $O(h^{2k_{n_i}+1})$, т. е. на единицу выше, чем в формуле Лобатто.

В данном случае мы для реализации алгоритма также применяем разложение (4.179) искомого приближенного решения по фундаментальным интерполяционным многочленам Лагранжа $\lambda_j^{k_{n_i}}$. Однако для РМГ(k) на каждом интервале I_{n_i} неизвестных коэффициентов $\beta_{n_{i,j}}$ теперь на единицу больше и равно $k_{n_i}+1$, поскольку условие непрерывно (4.181) не выполняются.

После подстановки (4.179) в соотношение (4.182) получаем аналогичную (4.180) СНАУ:

$$\begin{aligned}
&\Big(\sum_{j=0}^{k_{n_i}} \beta_{n_i,j}\lambda_j^{(k_{n_i})} - \beta^-_{n_i,0}\Big)\lambda_m^{(k_{n_i})}(0) + \\
&+ \int\limits_{I_{n_i}} \sum_{j=0}^{k_{n_i}} \Big[\beta_{n_i,j}\frac{d}{d\,t}\lambda_j^{(k_{n_i})}(\tau_{n_i}(t))\Big]\lambda_m^{(k_{n_i})}(\tau_{n_i}(t))d\,t = \qquad (4.184)\\
&= \int\limits_{I_{n_i}} f_i(t, y^h(t))\lambda_m^{(k_{n_i})}(\tau_{n_i}(t))d\,t, \;\; m = 0, 1, ..., k_{n_i}.
\end{aligned}$$

Здесь $\beta^-_{n_i,0} = \beta_{n_i-1,k_{n_i-1}}$ обозначает значение последнего коэффициента в правой точке предыдущего интервала. Решение нелинейной системы (4.184), как и в случае НМГ, может

последовательно реализовываться от макрослоя к макрослою. Число неизвестных в каждой l-й системе равно

$$N_l^H = \sum_{i=1}^{N} \sum_{n_i=n_{i,l-1}}^{n_i=n_i,l} (k_{n_i+1}).$$

Мы ограничимся изложением только принципов методов Галеркина для решения ОДУ, отсылая читателя для их более глубокого исследования к цитируемым выше работам, а также к многочисленным статьям, имеющимся по данной теме в Интернете.

Следует сказать, что в рассматриваемом научном направлении многие вопросы находятся пока на стадии динамического развития и еще ждут своих обобщений и систематизаций. Сюда относятся, в частности, условия разрешимости нелинейных систем для нахождения коэффициентов $\beta_{n_i,j}$, устойчивости и сходимости алгоритмов, а также подходы к автоматизации выбора адаптивных шагов сетки и порядков аппроксимации.

§ 4.16. Задачи к главе 4

4.16.1. Вывести явную схему Адамса 3-го порядка на равномерной сетке.

4.16.2. Вывести неявную схему Адамса 3-го порядка на равномерной сетке.

4.16.3. Построить многошаговую схему Нюстрема 3-го порядка.

4.16.4. Построить трехшаговую схему Милна.

4.16.5. Вывести явную трехшаговую формулу дифференцирования назад (ФДН).

4.16.6. Вывести неявную трехшаговую формулу дифференцирования назад.

4.16.7. Вывести выражение локальной погрешности для k-шагового неявного метода Адамса.

4.16.8. Построить оценку локальной погрешности k-шагового метода Нюстрема.

4.16.9. Вывести представление локальной ошибки k-шагового метода Милна.

4.16.10. Вывести выражение для локальной ошибки k-шагового метода ФДН.

4.16.11. Показать справедливость следующих соотношений для коэффициентов схем Адамса и Нюстрема:

$$\varkappa_j = \gamma_j + \hat{\gamma}_j, \;\; \hat{\gamma}_j = \gamma_j - \gamma_{j-1}.$$

4.16.12. Показать, что не существует k-шаговых методов порядка $2\,k + 1$.

4.16.13. Доказать, что все симметричные ММ имеют четный порядок (подсказка — использовать замечание 4.1).

4.16.14. Показать, что не существует $A(0)$-устойчивых явных методов.

4.16.15. Доказать, что условие $\beta_k/\alpha_k > 0$ является необходимым для $A(\alpha)$-устойчивости многошагового метода.

4.16.16. Показать, что область устойчивости k-шаговых неявных методов Адамса имеет конечный размер при любом $k \geqslant 2$.

Указание: покажите, что $(-1)^k\sigma(-1) < 0$ и, таким обра-

зом, σ имеет действительный отрицательный корень, меньший чем -1.

4.16.17. Показать, что все двухшаговые методы 2-го порядка задаются многочленами

$$\rho(\xi) = (\xi - 1)(\alpha\xi + 1 - \alpha),$$

$$\sigma(\xi) = (\xi - 1)^2\beta + (\xi - 1)\alpha + (\xi + 1)/2,$$

непрерывными при $\alpha \neq 2\beta$ и являющимися A-устойчивыми тогда и только тогда, когда выполнены условия

$$\alpha \geqslant 1/2, \;\; \beta > \alpha/2.$$

Указание: $\frac{\sigma(\xi)}{\rho(\xi)} = \frac{1}{2}\,\frac{\xi+1}{\xi-1} + (\beta - \frac{\alpha}{2})\frac{\xi-1}{\alpha\xi+1-\alpha}$.

4.16.18. Показать, что для k-шаговых явных методов Адамса граница интервала устойчивости $[-a_k, 0]$ определяется равенством

$$a = 2/b_k, \;\; b_k = \sum_{j=0}^{k-1} 2^j\gamma_j,$$

причем $b_1 = 1, b_2 = 2, b_3 = 11/3, b_4 = 20/3, ...$

Указание: подставить $\theta = \pi$ в формулу кривой локуса корней (4.47).

4.16.19. Вычислить значения a границы интервала устойчивости $[-a_k, 0]$ для k-шаговых неявных методов Адамса.

4.16.20. Показать, что одноопорный многошаговый метод можно записать в форме общего линейного метода.

4.16.21. Доказать, что многошаговый (ρ, σ)-метод A-контрактивен при некоторой с.п.о. матрице G в том и только в том случае, если он A-устойчив.

4.16.22. Построить точную трехточечную схему для задачи Коши

$$y'' = t^2, \;\; y|_{t=0} = 0, \;\; y'|_{t=0} = 1.$$

4.16.23. Вывести трехточечную явную схему Штермера для ОДУ 2-го порядка.

4.16.24. Построить и проанализировать схему предиктор-корректор Адамса—Башфорта—Мултона, совпадающую с явным двухстадийным МРК второго порядка.

4.16.25. Построить и проанализировать явную трехточечную ФДН.

4.16.26. Вывести выражение для константы погрешности (4.34) неявного метода ФДН.

4.16.27. Вывести формулу взаимосвязи численных решений метода трапеций (4.59) и метода средней точки (4.60).

4.16.28. Вывести формулу Штермера—Нумерова 4-го порядка точности для решения ОДУ второго порядка.

4.16.29. Показать, что двухшаговый неявный метод Адамса

$$y_{n+1} = y_n + \frac{h_n}{6(1+\omega_n)}(3+2\omega_n)f_{n+1} +$$

$$(3+\omega_n)(1+\omega_n)f_n - \omega_n^2 f_{n-1}, \;\; \omega_n = h_n/h_{n-1},$$

имеет локальную погрешность второго порядка.

4.16.30. Показать неустойчивость численного решения для следующей явной 2-шаговой схемы третьего порядка (пример Далквиста):

$$y_{n+2} + 4y_{n+1} - 5y_n = h(4f_{n+1} + 2f_n).$$

Указание: рассмотреть задачу Коши

$$y' = y, \;\; y(0) = 1$$

с точным решением $y(t) = \exp(t)$ и взять точные стартовые значения $y_0 = 1, \;\; y_1 = \exp(h)$.

Глава 5

МЕТОДЫ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ОДУ

Содержание данной главы практически не зависит от предыдущих разделов и представляет собой ,по сути , введение в такой значительный раздел вычислительной математики, как методы решения дифференциальных уравнений в частных производных. Мы остановимся на всех основных моментах данного направления и изложим в достаточно компактной форме обширный спектр алгоритмических проблем: методы конечных разностей, конечных объемов и конечных элементов, разрывные методы Галеркина для построения приближений как классических, так и обобщенных решений, алгебраические свойства получаемых сеточных уравнений и оценки их погрешности в различных нормах, а также алгоритмы решения соответствующих ленточных систем линейных алгебраических уравнений.

§ 5.1. Классификация и свойства одномерных краевых задач

Краевые задачи для ОДУ описываются одномерными дифференциальными уравнениями (или системами уравнений), решения которых зависят от одного аргумента (который мы будем обозначать через x) и подчиняются заданным граничным (краевым) условиям в нескольких точках. Формально такую краевую задачу можно представить в виде

$$L\,u(x) = f(x), \quad x \in \Omega = (a, b), \tag{5.1}$$

$$l_a u = g_a|_{x=a}, \quad l_b u = g_b|_{x=b}. \tag{5.2}$$

Здесь L, l_a и l_b — дифференциальные операторы (в общем случае нелинейные), определяющие исходное уравнение в расчетной области Ω и краевые условия. Их порядки и конкретные выражения должны быть согласованы по условиям существования и единственности искомого решения $u(x)$ в соответствующих функциональных пространствах, см. [15], [54], [55].

Естественно рассматривать метод вычисления решения только тогда, когда оно существует. Однако сам факт существования решения должен определяться не абстрактно, а в каком-то функциональном пространстве. Например, если решение поставленной задачи имеет разрывные первые производные, то для его отыскания бессмысленно строить алгоритм повышенной точности, предполагающий ограниченность старших производных.

Мы предполагаем, как правило, что в (5.1), (5.2) $u(x)$ есть скалярная функция, хотя во многих практических областях требуется решать краевые задачи для систем ОДУ, когда ре-

шение есть вектор-функция $u = (u_1(x), ..., u_m(x))$, а L, l_a и l_b — некоторые матричные операторы.

Отметим еще, что в рассматриваемых граничных условиях (5.2), вообще говоря, значения искомых функций и их производных в точке $x = a$ могут быть связаны с соответствующими значениями в точке $x = b$. Такие краевые условия называются *нераспадающимися*, а в противном случае — *распадающимися*. Мы в данной главе будем ограничиваться только последним простейшим случаем.

Краевая задача может иметь не одно, а несколько решений. Более того, их количество может зависеть от значений некоторых параметров исходной постановки. В данной главе такие случаи не исследуются, т. е. мы строим методы приближенного построения только единственных решений, а для более общих случаев, зачастую очень актуальных, рекомендуем читателю специальную литературу, см. [13], [24] и цитируемые там работы.

Вообще говоря, формулы (5.1), (5.2) характеризуют только частный случай одномерных краевых задач, а именно *двухточечных*. Для некоторых приложений требуется решать *многоточечные задачи*, в которых граничные условия задаются в большем числе точек. Кроме того, рассматриваемая область Ω может представлять не один отрезок (a, b), а быть многосвязной. Однако на таких вопросах мы также не останавливаемся.

Совокупность граничных точек (в нашем случае $x = a$ и $x = b$) составляют границу Γ области Ω, а ее замыкание $\bar{\Omega} = \Omega \cup \Gamma$ будем называть *расчетной областью*, в которой требуется найти приближенное решение исходной задачи.

Рассмотрим сначала скалярное линейное дифференциальное уравнение второго порядка

$$L\,u \equiv -(p(x)u')' + q(x)u' + r(x)u = f(x), \quad x \in (a,b), \tag{5.3}$$

для которого на концах отрезка (a,b) ставятся граничные условия вида

$$\begin{gathered} l_a u \equiv \alpha_a u + \beta_a u' = g_a|_{x=a}, \;\; l_b u \equiv \alpha_b u + \beta_b u'|_{x=b} = g_b, \\ |\alpha_a| + |\beta_b| \neq 0, \;\; \alpha_a \cdot \beta_a \geqslant 0, \;\; \alpha_b \cdot \beta_b \leqslant 0. \end{gathered} \tag{5.4}$$

Первый член левой части уравнения (5.3) имеет так называемую дивергентную форму $(-\mathrm{div}(p\mathrm{grad}u))$ и в физическом смысле ответственен за диффузионный перенос субстанции, следующий член — за конвективный перенос, а последний играет роль источника (так же, как и правая часть уравнения $f(x)$). Соответственно, множитель $p(x)$ называется коэффициентом диффузии (или теплопроводности, если u — это температура, а само уравнение описывает явление теплопереноса) и величина $q(x)$ играет роль скорости конвективного переноса.

В отдельных случаях уравнения вида (5.3) имеют распространенные специальные названия:

- $q(x) = r(x) = 0$ — *диффузионное уравнение*,
- $q(x) = r(x) = 0$, $p(x) = 1$ — *уравнение Пуассона*,
- $q(x) = r(x) = f(x) = 0$, $p(x) = 1$ — *уравнение Лапласа*,

- $q(x) = 0$, $r(x) = \text{const}, p(x) = 1$ — *уравнение Гельмгольца.*

Если в граничных соотношениях (5.4) $\beta_a = 0$ или $\beta_b = 0$, то соответствующее краевое условие называется *условием* 1*-го рода, или Дирихле*. И наоборот, при $\alpha_a = 0$ или $\alpha_b = 0$ имеем *условия* 2*-го рода, или Неймана.* Если же оба коэффициента ненулевые ($\alpha_a \cdot \beta_a > 0$ и/или $\alpha_b \cdot \beta_b < 0$), то соответствующие граничные условия называются *условиями* 3*-го рода, или Робена* (иногда также — условиями Ньютона). Когда на всей границе (в данном случае в точках $x = a$ и $x = b$) заданы граничные условия одного типа, то говорят соответственно о краевой задаче 1-го, 2-го или 3-го рода (или о *задаче Дирихле*, *Неймана* и *Робена* для соответствующего уравнения).

Уравнение (5.3) предполагает, хотя это и не оговаривалось выше, использование декартовой системы координат. Значительный практический интерес представляют также краевые задачи в цилиндрической и сферической системах координат. Мы ограничимся иллюстрацией уравнения Пуассона

$$\Delta_r^{(\alpha)} u \equiv -\frac{1}{r^\alpha}\frac{\partial}{\partial r}\left(r^\alpha \frac{\partial u}{\partial r}\right) = f(r),\ 0 \leqslant R_0 \leqslant r \leqslant R_1 < \infty,$$

где $\alpha = 1, 2$ для цилиндрической и сферической систем координат соответственно. Если здесь $R_0 > 0$, т. е. фактически это уравнение описывает осесимметричное или центрально-симметричное решение в бесконечном цилиндрическом или сферическом слое, то граничные условия могут использоваться для любого из рассмотренных выше типов. А при $R_0 = 0$ (расчетная область включает ось или центр симметрии) кор-

ректная постановка всегда предполагает задание условия симметрии $u_r'(0) = 0$, обеспечивающего ограниченность решения.

Если порядки дифференциальных уравнений повышаются, то для однозначности решения требуется повышать количество граничных условий. Например, для уравнения 4-го порядка (такие уравнения в многомерных случаях при $f(x) = 0$ называются бигармоническими)

$$\frac{d^4u}{d\,x^4} = f(x), \quad x \in (a, b) \tag{5.5}$$

необходимо поставить четыре краевых условия. В частности, задача с условиями

$$u|_a = u_a, \quad u'|_a = u_a', \quad u|_b = u_b, \quad u'|_b = u_b', \tag{5.6}$$

при заданных величинах u_a, u_a', u_b, u_b' называется *задачей Дирихле для уравнения* (5.5).

Кроме категорий существования и единственности, важной характеристикой краевой задачи является ее корректность.

Определение 5.1. *Краевая задача называется корректной, если ее решение существует, единственно и непрерывным образом зависит от исходных данных.*

В рассматриваемых случаях — это коэффициенты уравнения (5.1) и граничных условий (5.2) (или (5.3), (5.4)). Альтернативой корректным (т. е. “хорошим”) задачам являются так называемые или *некорректные*, или *условно корректные*, или *плохо обусловленные задачи*. Не останавливаясь на детальных свойствах таких “плохих” (но зачастую очень важных)

задач, чему посвящаются отдельные монографии и учебники (см.[33], [41], [59]), мы только отметим тот принципиальный факт, что понятие обусловленности задачи является не абсолютным, а относительным. Например, изменение в уравнении (5.3) диффузионного коэффициента $p(x)$ в 2 раза может привести к изменению решения $u(x)$ и в 10, и в 100, и в 10^6 раз, в зависимости от других исходных данных. И здесь трудно определить границу, при которой задача из хорошо обусловленной переходит в плохо обусловленную. Строго говоря, можно только сравнивать различные краевые задачи по качествам обусловленности, т. е. установить, что одна задача хуже или лучше обусловлена, чем другая.

С рассмотренными свойствами тесно связано и такое понятие, как устойчивость краевой задачи, а точнее говоря — *устойчивость решения краевой задачи*. Качество устойчивости определяется оценкой возмущения решения, вызванного возмущением исходных данных задачи. Сильно неустойчивая задача — это плохо обусловленная задача, или другими словами, имеющая очень большое число обусловленности. Последнее как раз определяет коэффициент, во сколько раз норма возмущения задачи может превосходить норму соответствующего возмущения исходных данных.

Для построения приближенного решения краевой задачи требуются предположения о достаточной гладкости ее искомого решения. Такой гипотезой может быть принадлежность $u(x)$ банаховому пространству функций $C^p(\Omega)$, непрерывных в Ω со своими производными до p-го порядка включительно, для которых отрезок ряда Тейлора

$$u(x+h) = u(x)+h\,u'(x)+...+\frac{1}{p!}h^p u^{(p)}(x)+\frac{h^{p+1}}{(p+1)!}u^{(p+1)}(\xi), \quad \xi \in [x, x+h],$$

при любом $x \in \Omega$ и достаточно малом h имеет погрешность $\psi(x)$ с порядком $0(h^{p+1})$:

$$|\psi(x)| \leqslant \frac{h^{p+1}}{(p+1)!}M_{p+1}, \quad M_{p+1} = \max_{x\in\Omega}|u^{(p+1)}(x)| \equiv ||u^{(p+1)}||_\infty.$$

Очевидно, что для обеспечения определенной гладкости $u(x)$ необходима некоторая соответствующая гладкость коэффициентов $p(x), q(x), r(x)$ и правой части $f(x)$ уравнения (5.3). Конкретные требования и свойства здесь — это прерогатива теории дифференциальных уравнений, см., например, [35], [50]. Во многих приложениях коэффициенты являются кусочно-гладкими, что порождает аналогичные качества искомых решений. Типичный пример — это разрывы коэффициентов диффузии $p(x)$ во внутренних точках Γ_k расчетной области, приводящие к следующим условиям сопряжения для решения диффузионно-конвективного уравнения:

$$p^-\frac{\partial u}{\partial x}|_{\Gamma_k^-} = p^+\frac{\partial u}{\partial x}|_{\Gamma_k^+}, \quad u|_{\Gamma_k^-} = u|_{\Gamma_k^+}, \tag{5.7}$$

характеризующим непрерывность решения и потока субстанции $\Phi = -p\dfrac{\partial u}{\partial x}$. Значки "$-$," здесь означают односторонние значения функций и их производных, а само соотношение (5.7), вообще говоря, является следствием исходного уравнения (5.3).

Особый актуальный случай представляют *сингулярно-возмущенные задачи*, для которых в окрестности определенных точек производные решения стремятся к бесконечности.

К такому классу относятся уравнения с малыми коэффициентами при старших производных, например,

$$-\varepsilon u'' + q\,u' + r\,u = f(x), \quad \varepsilon \ll 1, \quad x \in (a,b). \tag{5.8}$$

Это уравнение при $\varepsilon \to 0$ имеет предельный первый порядок, и для однозначного определения его решения достаточно одного граничного условия, определяемого значениями q, r, f. Однако при конечном ε решение $u(x)$ определено при любых двух краевых условиях (например, Дирихле), которые могут быть "несогласованы" с остальными данными задачи. В силу этого в окрестности одной из граничных точек (какой именно — зависит от знака q, определяющего направление скорости) возникает *пограничный слой* с резким изменением решения и его производных. Если же конвективный коэффициент $q(x)$ меняет знак в некоторых точках $x_k \in \Omega$, то в их окрестностях появляются аналогичные сингулярности с *внутренними слоями*. Построение приближенных решений в таких случаях требует знания асимптотического поведения искомого решения в окрестности сингулярных точек.

Если какие-то из коэффициентов в (5.3), (5.4) зависят от искомого решения, то мы приходим к нелинейным краевым задачам, свойства которых гораздо богаче, а построение и обоснование численных решений — значительно труднее, чем в линейных случаях. Фактически результаты здесь являются "штучными", т. е. достаточно полные исследования удается проводить только для конкретных частных ситуаций. В качестве иллюстрации приведем нелинейное граничное условие радиационного излучения для уравнения теплопроводности,

когда $u(x)$ в (5.3) есть температура:

$$u'|_{x=a} = \alpha(u_a^4 - u_0^4). \tag{5.9}$$

Здесь u_0 — заданная температура внешней среды, а α — радиационный коэффициент. В уравнении теплопроводности при высоких температурах коэффициент теплопроводности $p(x)$ является функцией температуры, как правило, определяемой эмпирически.

Из рассмотренных выше дифференциальных постановок краевых задач могут следовать различные интегральные соотношения, имеющие физический смысл законов сохранения. Так, интегрируя уравнение диффузии ($q = r = 0$ в (5.3)) по некоторому внутреннему интервалу $(x', x'') \in \Omega$, получаем равенство

$$-p(x)u'(x)|_{x'}^{x''} = \int_{x'}^{x''} f(x)d\,x, \tag{5.10}$$

выражающее закон сохранения массы на отрезке: разность потоков на границе равна интегралу от функции источника. Такого типа соотношения являются самостоятельной основой для построения приближенных решений краевых задач *методами конечных объемов* (МКО), название которых связано с необходимостью аппроксимации появляющихся интегралов по достаточно малым, но конечным интервалам.

В отличие от МКО при непосредственной аппроксимации производных в исходном дифференциальном уравнении мы приходим к *методам конечных разностей* (МКР), которые порождают систему дискретных алгебраических уравнений,

имеющие как общие качества, так и существенные различия с уравнениями МКО. Последние, вообще говоря, на заре своего развития назывались *балансными*, или *консервативными, конечно-разностными схемами*, но позже сформировались в самостоятельное направление. Наибольшие различия в методологиях МКО и МКР наблюдаются в многомерных задачах для уравнений в частных производных, не являющихся предметом наших рассмотрений в данной книге.

МКО и МКР относятся к классу сеточных методов, основанных на дискретизации расчетной области Ω.

Построение сетки в одномерном случае не представляет большой проблемы и заключается в определении множества узлов

$$\begin{aligned} &x_{i+1} = x_i + h_i, \quad i = 0, 1, ..., N, \\ &x_0 \leqslant a, \quad x_{N+1} \geqslant b \end{aligned} \tag{5.11}$$

таким образом, чтобы сеточная расчетная область $[x_0, x_{N+1}]$ содержала в себе $\bar{\Omega}$.

Альтернативой, в некотором смысле, МКО и МКР являются *методы конечных элементов* (МКЭ), заключающиеся в построении приближений не для рассмотренных классических, а для обобщенных, или вариационных, или слабых, постановок краевых задач.

Рассмотрим для простоты однородную смешанную краевую задачу для уравнения Пуассона:

$$\begin{aligned} -u'' &= f(x), \quad x \in (a,b), \\ u(a) &= 0, \quad u'(b) = 0. \end{aligned} \tag{5.12}$$

Если $u(x)$ есть решение (5.12), а $v(x)$ — достаточно гладкая функция, удовлетворяющая условию $v(0) = 0$, то мы можем определить *скалярное произведение* функции $(,)$ и *билинейную форму* $a(,)$ с помощью интегрирования по частям следующим образом:

$$\begin{aligned} f(v) = (f,v) \equiv \int\limits_a^b f(x)v(x)d\,x = -\int\limits_a^b u''(x)v(x)d\,x = \\ = \int\limits_a^b u'(x)v'(x)d\,x \equiv a(u,v). \end{aligned} \tag{5.13}$$

Если мы определим формально пространство функций

$$V = \{v \in L^2(a,b)\colon a(v,v) < \infty, \quad v(a) = 0\}, \tag{5.14}$$

то постановку задачи (5.12) можно привести к следующему виду:

$$u \in V \quad , \quad a(u,v) = f(v), \ \ \forall v \in V, \tag{5.15}$$

который называется *вариационной*, или *слабой, постановкой*. Термин “вариационная” происходит от того, что соотношение (5.14) выполняется для любой (vary по-английски) функции $v(x)$.

Отметим, что при определении V обозначение $L^2(a,b)$ есть гильбертово пространство функций, интегрируемых с квадратом на отрезке (a,b), т. е. имеющих ограниченную норму $||v||_2 \equiv (v,v)^{1/2}$.

Укажем также, что определенное в (5.14) пространство V, которое называется *пространством пробных функций*, является также гильбертовым пространством, а роль скалярного произведения в нем играет билинейная форма $a(,)$.

Методически важно, что мы в (5.13) автоматически подразумевали классическое определение производной, т. е.

$$u'(x) = \lim_{h \to 0} \frac{u(x+h) - u(x)}{h}.$$

На самом деле слабые постановки допускают определение обобщенных производных, на которых мы пока не останавливаемся. Другой актуальный вопрос: являются ли классическая и вариационная постановка эквивалентными? Следующее утверждение дает здесь положительный ответ, но при дополнительных предположениях.

Теорема 5.1. *Пусть в* (5.15) *выполняются условия* $f \in C^0[a,b]$, *а* $u \in C^2[a,b]$. *Тогда* $u(x)$ *является решением задачи* (5.12).

Замечание 5.1. Следует отметить также существенное отличие при использовании условий Дирихле и Неймана в вариационной постановке. А именно, условие $u(a) = 0$ явно указывалось при определении пространства V, а участвующее в классической постановке (5.12) условие $u'(b) = 0$ в формулировке (5.15) присутствует только неявно. В силу этого условие Дирихле называется *существенным*, *или главным*, краевым

условием, а условие Неймана — *естественным*.

Принципиальное отличие классической и слабой постановок — существенное снижение требований к гладкости функций в последних. В частности, можно показать, что теорема 5.1 остается справедливой и при менее ограничительных предположениях относительно функции $f(x)$.

Отметим, что зачастую краевая задача может быть факторизована и сведена к решению последовательности задач Коши. Например, краевая задача для линейного уравнения второго порядка

$$u'' + q(t)u' + r(t)u = f(t), \;\; t_0 \leqslant t \leqslant t_e,$$

$$u'(t_0) - \alpha_0 u(t_0) = \alpha_1, \;\; u'(t_e) - \beta_0 u(t_e) = \beta_2$$

формально записывается в факторизованном виде

$$\Big[\frac{d}{d\,t} + \mu(t)\Big]\Big[\frac{d}{d\,t} + \nu(t)\Big]u = f(t),$$

который естественно сводится к системе двух ОДУ первого порядка

$$\Big[\frac{d}{d\,t} + \mu(t)\Big]v = f(t), \;\; \Big[\frac{d}{d\,t} + \nu(t)\Big]u = v(t).$$

Чтобы факторизованная и исходная постановка были эквивалентны, необходимо выполнение условия

$$u'' + (\mu + \nu)u' + (\nu + \mu\nu)u = u'' + q\,u' + r\,u,$$

т. е. функции $\mu(t)$ и $\nu(t)$ должны удовлетворять уравнениям

$$\mu(t) + \nu(t) = q(t), \;\; \nu'(t) + \mu(t)\nu(t) = r(t).$$

Отсюда для нахождения $\nu(t)$ получаем задачу Коши

$$\nu'(t) + q(t)\nu(t) - \nu^2(t) = r(t), \;\; \nu(t_0) = -\alpha_0.$$

Далее функции $v(t)$ и $u(t)$ находятся из следующих задач Коши:

$$v' + [q(t) - \nu(t)]v = f(t), \;\; v(t_0) = \alpha_1,$$

$$u' + \nu(t)u = v(t), \;\; u(t_1) = u_1,$$

причем значение u_1 определяется из системы уравнений

$$u'(t_1) - \beta_0 u(t_1) = \beta_1, \;\; u'(t_1) + \nu(t_1)u(t_1) = v(t_1).$$

Такой аналитический подход называется методом *дифференциальной прогонки*, и мы на нем останавливаться не будем.

Отдельный параграф мы посвящаем краткому изложению относительно нового аппроксимационного подхода, появившегося около 40 лет назад и активно развиваемого в последние десятилетия, — семейству разрывных методов Галеркина (РМГ, или DGM — от Discontinuous Galerkin Methods). Данные алгоритмы пока еще не нашли своего систематизированного описания в монографической литературе, а в методологическом плане они прокладывают мостик между МКО и МКЭ.

С одной стороны, они являются обобщением методов конечных объемов, которые выводятся из законов сохранения,

формально получаемых из интегрирования исходных дифференциальных уравнений, предварительно умноженных на финитные базисные функции нулевого порядка. В РМГ же эта идея развивается путем использования пробных функций различных порядков.

С другой стороны, разрывные схемы Галеркина,по существу, являются смешанными схемами конечных элементов, основанными на сведении дифференциальных уравнений второго или более высоких порядков к системе уравнений первого порядка. Кроме того, в отличие от рассмотренной в (5.15) вариационной постановки, решение ищется не в пространстве пробных функций V, а в некотором пространстве тестовых функций W, отличающемся от V.

§ 5.2. Конечно-разностные методы аппроксимации краевых задач

В данном параграфе мы последовательно рассмотрим методы конечных разностей для аппроксимации производных, дифференциальных уравнений и краевых задач с граничными условиями различных типов.

5.2.1. Конечно-разностные аппроксимации производных. Обозначим через Ω_h введенную в (5.11) сетку, которая характеризуется своими узлами x_i (пронумерованными в порядке возрастания координат) и шагами $h_i > 0$. В дальнейшем через h обозначаем или шаг равномерной сетки ($h = h_i = (x_{N+1} - x_0)/(N+1)$), или максимальный шаг неравномерной сетки ($h = h_{max} = \max\limits_i h_i$). Сетки предполагаем

регулярными или *квазиравномерными* в том смысле, при котором при рассмотрении последовательностей сгущающихся сеток ($h \to 0$) всегда считается, что отношение максимального шага к минимальному остается конечным ($h_{\max}/h_{\min} < \infty$).

Будем обозначать далее через $u_i^{(p)} = u^{(p)}(x_i), p = 0, 1, ...,$ значения функции $u(x)$ и ее производных p -го порядка в узлах сетки, а требуемая гладкость будет указываться в конкретных случаях.

Рассмотрим сначала конечно-разностные аппроксимации производной первого порядка. Введем определения *конечных разностей первого порядка*

$$\begin{aligned} \nabla_h^+ u_i &\equiv \frac{u(x_{i+1}) - u(x_i)}{h_i} = u'(x_i) + \psi_i^+, \\ \nabla_h^- u_i &\equiv \frac{u(x_i) - u(x_{i-1})}{h_{i-1}} = u'(x_i) + \psi_i^-. \end{aligned} \tag{5.16}$$

Первая из них называется *правой разностью*, а вторая — левой. Употребляются еще наименования *разность вперед* и *разность назад*, а также общее для них название — *односторонние разности*. Погрешности аппроксимации этих разностей легко выводятся из разложения величин $u(x_{i\pm1})$ в ряд Тейлора относительно точки x_i:

$$\begin{aligned} \psi_i^+ &= \frac{1}{2} h_i u_i''(\xi_i^+), \;\; \xi_i^+ \in [x_i, x_{i+1}], \\ \psi_i^- &= \frac{-1}{2} h_{i-1} u_i''(\xi_i^-), \;\; \xi^- \in [x_{i-1}, x_i]. \end{aligned} \tag{5.17}$$

Отсюда видно, что односторонние разности первого порядка (слово “конечные” здесь и далее для краткости опускаем)

аппроксимируют первую производную на функциях из класса $C^2[a,b]$ с погрешностью первого порядка: $|\psi_i^+| = O(h),\ |\psi_i^-| = O(h)$.

Нетрудно увидеть, что для аппроксимации $u'(x_i)$ можно применять и так называемую *двустороннюю разность первого порядка*

$$\bar{\nabla}_h u_i = \frac{u_{i+1} - u_{i-1}}{h_i + h_{i-1}} = u_i' + \bar{\psi}_i, \tag{5.18}$$

погрешность которой имеет вид

$$\bar{\psi}_i = \frac{1}{2}(h_i - h_{i-1})u_i'' + \frac{1}{6}(h_i^2 - h_i + h_{i-1}^2)u'''(\bar{\xi}_i), \quad \bar{\xi}_i \in [x_{i-1}, x_{i+1}].$$

Отсюда следует, что двусторонняя разность аппроксимирует первую производную на неравномерной сетке также с погрешностью $O(h)$. Однако в частном случае $h_i = h_{i-1} = h$ имеем $\bar{\psi}_i = \dfrac{h^2}{6}u'''(\xi_i)$, а получаемое при этом из (5.18) выражение

$$\bar{\nabla}_h u_i \equiv \frac{u_{i+1} - u_{i-1}}{2\,h} \tag{5.19}$$

называется *центральной разностью* и на функциях из $C^3[a,b]$ обладает погрешностью $O(h^2)$.

Поскольку для погрешностей $\psi_i^{\pm}$ из (5.16) при использовании трех членов в отрезке ряда Тейлора можно получить вместо (5.17) новые соотношения

$$\psi_i^+ = \frac{1}{2}h_i u_i'' + \frac{1}{6}h_i^2 u'''(\xi_i^+),\ \ \xi_i^+ \in [x_i, x_{i+1}],$$

$$\psi_i^- = \frac{-1}{2}h_{i-1} u_i'' + \frac{1}{6}h_{i-1}^2 u'''(\xi_i^-),\ \ \xi^- \in [x_{i-1}, x_i]$$

(здесь величины $\xi_i^{\pm}$ отличаются, строго говоря, от предыдущих, участвующих в (5.17)), то понятно, что линейная комбинация односторонних разностей дает новую — *трехточечную* в общем случае — *аппроксимацию первой производной*:

$$\nabla_h u_i = \frac{1}{h_i + h_{i-1}}(h_{i-1}\nabla_h^+ u_i + h_i \nabla_h^- u_i) =$$

$$= \frac{1}{h_i + h_{i-1}}\Big[\frac{h_{i-1}}{h_i}u_{i+1} - \Big(\frac{h_{i-1}}{h_i} - \frac{h_i}{h_{i-1}}\Big)u_i - \frac{h_i}{h_{i-1}}u_{i-1}\Big] =$$

$$= u_i' + \frac{h_i h_{i-1}}{6}u'''(\xi_i), \;\; \xi_i \in [x_{i-1}, x_{i+1}], \tag{5.20}$$

которая имеет уже погрешность второго порядка, как только функция $u(x)$ обладает ограниченной третьей производной. Если сетка равномерная, то выражение $\nabla_h u_i$ становится двухточечным:

$$\nabla_h u_i \equiv \frac{u_{i+1} - u_{i-1}}{2h} = u_i' + \psi_i, \;\; |\psi_i| \leqslant \frac{h^2}{6}M_3,$$

и совпадает с центральной разностью (5.19) относительно точки x_i (величина M_3 — это максимум модуля третьей производной на интервале $[x_{i-1}, x_{i+1}]$).

Трехточечная аппроксимация первой производной (5.20) может быть выведена также с помощью применения интерполяции и центральных разностей. Для этого нужно воспользоваться выражением для линейной интерполяции

$$u(x) = \frac{(x-a)u(b) + (b-x)u(a)}{b-a} + \psi^1, \;\; \psi^1 = \frac{(x-a)(x-b)}{2}u''(\xi), \;\; \xi \in [a,$$

а также аппроксимациями первых производных:

$$\bar{\Delta}_h u_{i+1/2} \equiv \frac{u_{i+1}-u_i}{h_i} = u'_{i+1/2} + \bar{\psi}_{i+1/2},$$
$$\bar{\psi}_{i+1/2} = \frac{h_i^2}{24} u'''(\xi_{i+1/2}), \;\; \xi_{i+1/2} \in [x_i, x_{i+1}].$$

Интерполируя теперь значение u'_i по величинам $u'_{i\pm1/2}$, т. е. полагая $a = x_{i-1/2}$ и $b = x_{i+1/2}$, с помощью этих формул получаем соотношение

$$\bar{\nabla}_h u_i = \frac{1}{x_{i+1/2} - x_{i-1/2}} \Big[(x_i - x_{i-1/2})(\bar{\Delta}_h u - \bar{\psi})_{i-1/2} +$$
$$+ (x_i - x_{i+1/2})(\bar{\Delta}_h u - \bar{\psi})_{i+1/2} \Big] + \psi_i^1,$$

которое после проведения простых преобразований приводит к (5.20).

Перейдем теперь к конечно-разностным выражениям второго порядка. Через конечные разности первого порядка рекуррентно могут быть определены разности второго и более высокого порядков, аппроксимирующие различные производные. Аналогично предыдущему на неравномерной сетке *разность второго порядка* записывается в виде

$$\begin{aligned} \Delta_h u_i &\equiv \bar{\nabla}_h \underset{i+1/2}{u} - \bar{\nabla}_h u_{i-1/2}(h_i + h_{i-1})/2 = \\ &= \frac{2\,u_{i-1}}{h_{i-1}(h_i + h_{i-1})} - \frac{2\,u_i}{h_i h_{i-1}} + \frac{2\,u_{i+1}}{h_i(h_i + h_{i-1})} = \\ &= u''_i + \frac{h_i - h_{i-1}}{3} u'''_i + \frac{h_i^2 - h_i h_{i-1} + h_{i-1}^2}{12} u^{(4)}(\xi_i), \end{aligned} \tag{5.21}$$

$\xi_i \in [x_{i-1}, x_{i+1}]$.

Здесь $\Delta_h u_{i\pm1/2}$ суть центральные разности относительно срединных точек сетки $x_{i\pm1/2} = \frac{1}{2}(x_i + x_{i\pm1})$. Как видно, вторая производная аппроксимируется соответствующей разностью на функциях из класса $C^4[x_{i-1}, x_{i+1}]$ с погрешностью порядка $O(h)$. Если же сетка равномерная, то

$$\Delta_h u_i = \frac{u_{i-1} - 2\,u_i + u_{i-1}}{h^2} = u_i'' + \frac{h^2}{12} u^{(4)}(\xi_i) \tag{5.22}$$

и погрешность аппроксимации имеет уже второй порядок, но только если функция $u(x)$ обладает ограниченной производной четвертого порядка.

Таким же образом, т. е. последовательной формальной заменой производных разностями, строятся конечно-разностные приближения более сложных дифференциальных выражений. Так, если $p(x)$ — некоторая дифференцируемая функция, то можно вывести аппроксимацию для "диффузионного члена"

$$(p(x)u')_i' = \frac{(pu')_{i+1/2} - (pu')_{i-1/2}}{(h_i + h_{i-1})/2} + \psi_i^{(1)} = \frac{2\,p_{i-1/2}}{h_{i-1}(h_i + h_{i-1})} u_{i-1} -$$
$$- \frac{2}{h_i + h_{i-1}} \Big(\frac{p_{i-1/2}}{h_{i-1}} + \frac{p_{i+1/2}}{h_i} \Big) u_i + \frac{2\,p_{i+1/2}}{h_i(h_i + h_{i-1})} u_{i+1} + \psi_i. \tag{5.23}$$

Здесь $\psi_i^{(1)}$ — погрешность аппроксимации "внешней" производной, равная

$$\psi_i = \frac{h_i - h_{i-1}}{4} (pu')_i'' + 0(h^2),$$

а ψ_i — полная погрешность, в которой дополнительно присутствуют ошибки приближения величин $(pu)'_{i\pm1/2}$ с помощью

центральных разностей:

$$\psi_i = \frac{h_i - h_{i-1}}{12}[pu''' + 3(pu')'']_i + 0(h^2). \tag{5.24}$$

Отсюда, в частности, следует, что на равномерной сетке погрешность аппроксимации в случае переменного коэффициента диффузии $p(x)$ есть величина второго порядка, как и для второй производной. Более того, если сетка "почти равномерная", т. е. $h_i - h_{i-1} = O(h^2)$, то погрешность аппроксимации также имеет второй порядок.

Легко проверить, что для функций $u(x)$, обладающих ограниченными производными до шестого порядка включительно, вторую производную можно аппроксимировать с погрешностью $O(h^4)$ при помощи пятиточечной схемы

$$\Delta_{h,5}u_i \equiv \frac{1}{12\,h^2}\Big[- u_{i-2} - u_{i+2} + 16(u_{i-1} + u_{i+1}) - 30\,u_i\Big] =$$
$$= u_i'' + \frac{h^4}{72}u^{(6)}(\xi_i), \;\; \xi_i \in [x_{i-2}, x_{i+2}]. \tag{5.25}$$

Очевидно, что с помощью аппарата конечных разностей высших порядков и техники интерполяции можно строить различные многоточечные аппроксимации повышенной точности и для диффузионных выражений вида (5.23).

Остановимся далее на аппроксимациях производных высших порядков. Используя рекуррентное определение конечных разностей высших порядков через низшие, можно строить аппроксимации и для производных более высоких порядков. Например, определяя разность третьего порядка в i-й

точке через правые разности от вторых разностей в точках x_{i+1} и x_i неравномерной сетки, мы получим следующее четырехточечное выражение, приближающее третью производную с погрешностью $O(h)$:

$$L_h^{(3)}u_i \equiv \frac{1}{h_i}(\nabla_h u_{i+1} - \nabla_h u_i) =$$
$$= \frac{2}{h_i}\Big\{\frac{u_{i+2}}{h_{i+1}(h_{i+1}+h_{i+2})} - \frac{h_{i-1}+h_i+h_{i+1}}{h_i}\Big[\frac{u_{i+1}}{h_{i+1}(h_{i-1}+h_i)} - \frac{u_i}{h_{i-1}(h_i+h$$
$$-\frac{u_{i-1}}{h_{i-1}(h_i+h_{i-1})}\Big\} = u_i''' + \psi_i^{(3)},\ \ \psi_i^{(3)} = O(h). \tag{5.26}$$

Если же сетка равномерная, то на функциях $u \in C^5[a,b]$ можно аппроксимировать 3-ю производную в i-й точке со вторым порядком погрешности, определяя ее с помощью центральной разности от вторых производных в точках $x_{i\pm1}$:

$$L_n^{(3)}u_i \equiv \frac{1}{2h^3}[u_{i+2} - u_{i-2} - 2(u_{i+1} - u_{i-1})] = u_i^{(3)} + O(h^2).$$

На равномерной сетке также легко аппроксимируется 4-я производная, причем разность четвертого порядка строится как "вторая разность в квадрате":

$$\Delta_h^2 u_i \equiv \Delta_h \frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} = \frac{u_{i-2} - 4\,u_{i-1} + 6\,u_i - 4\,u_{i+1} + u_{i+2}}{h^4} =$$
$$= u_i^{(4)} + \psi_i,\ \ |\psi_i| \leqslant \frac{h^2}{6}M_6,\ \ M_6 = \max_{x\in[x_{i-2},x_{i+2}]}\{|u^{(6)}(x)|\}.$$

Совокупность узлов, используемых при аппроксимации производной в одной точке, называется *сеточным шаблоном*. Формально его можно определить перечислением набо-

ра участвующих индексов и обозначать, например, для сеточных выражений (5.31) или (5.33) как $\omega_i^{(3)} = \{i-1, i, i+1\}$ и $\omega_i^{(5)} = \{i-2, i-1, i, i+1, i+2\}$ соответственно. По количеству участвующих узлов такие шаблоны, как и сами разностные выражения, называются трехточечными или пятиточечными.

5.2.2. Конечно-разностные аппроксимации дифференциальных выражений, уравнений и краевых условий. После рассмотрения принципов конечно-разностных представлений производных нетрудно перейти и к аппроксимации краевых задач в целом. Эта проблема состоит из двух частей: приближение дифференциального уравнения и граничных условий. Обе они сводятся к сеточной аппроксимации дифференциальных выражений, которая очевидным образом применяется к каждому из слагаемых, с последующим использованием принципа аддитивности. Теперь мы можем сформулировать достаточно общее понятие.

Определение 5.2. *Пусть $Lu(x)$ — некоторое дифференциальное выражение, определенное в области изменения аргумента Ω, а $L_h u_i$ — определенное в точке x_i из Ω_h конечно-разностное выражение. Если при достаточно малых h имеют место соотношения*

$$(L_h u)_i = L\,u(x_i) + \psi_i^u, \;\; |\psi_i^u| \leqslant C_i h^\gamma, \;\; \gamma > 0, \tag{5.27}$$

где C_i — не зависящая от h постоянная, то ψ_i^u называется погрешностью аппроксимации разностного выражения в точке x_i на функции $u(x)$, а γ — порядком погрешности.

Поскольку выражение $(L_h u)_i$ на множестве сеточных

функций $u_h = \{u_i\}$ определяет *разностный или конечно-разностный оператор* L_h, то предыдущему определению эквивалентно следующее: при выполнении условий (5.27) *разностный или конечно-разностный оператор* L_h аппроксимирует дифференциальный оператор L на функции $u(x)$ с погрешностью аппроксимации $\psi_h = \{\psi_i\}$ порядка γ. Соответственно можно записать

$$L_h(u)_h = (L\,u)_h + \psi_h^u, \tag{5.28}$$

где $(u)_h = \{u(x_i)\}$ означает сеточную функцию или вектор с элементами, равными значениям точного решения в узлах сетки.

Заметим, что здесь при обозначении погрешности мы использовали верхний индекс "u" для акцентирования того, что аппроксимация рассматривается не на абстрактном классе функций, а на конкретной функции $u(x)$ (и выражается именно через ее производные).

Введем теперь оператор аппроксимации правой части дифференциального уравнения (5.1):

$$(P_h f)_i = f_i + \psi_i^f, \tag{5.29}$$

где ψ_i^f есть погрешность P_h на функции $f(x)$ в i-м узле.

Теперь с использованием (5.27) и (5.29) мы можем записать конечно-разностное уравнение

$$(L_h u)_i = (P_h f)_i + \psi_i, \;\; \psi_i = \psi_i^u - \psi_i^f, \tag{5.30}$$

аппроксимирующее исходное дифференциальное уравнение на функциях u, f в точке x_i.

Рассмотрим в качестве примера аппроксимации уравнения Пуассона из (5.12). Если на неравномерной сетке оператор Лапласа (в нашем случае — вторую производную) дискретизировать с помощью трехточечного представления (5.21), а оператор P_h выбрать единичным, т. е. $(P_h f)_i = f_i$ и $\psi_i^f = 0$, то получаем разностное уравнение

$$(\Delta_h u)_i \equiv \frac{2}{h_i + h_{i-1}} \Big[\frac{u_{i-1}}{h_{i-1}} - \Big(\frac{1}{h_{i-1}} + \frac{1}{h_i} \Big) u_i + \frac{u_{i+1}}{h_i} \Big] =$$

$$= f_i + \psi_i, \;\; \psi_i = O(h), \tag{5.31}$$

имеющее погрешность первого порядка на функциях $u \in C^3[a, b]$.

Очевидно, что на равномерной сетке уравнение (5.31) принимает вид

$$(\Delta_h u)_i \equiv \frac{1}{h_2}(u_{i-1} - 2\, u_i + u_{i+1}) = f_i + \psi_i, \;\; |\psi_i| \leqslant \frac{h^2 M_4}{12}, \tag{5.32}$$

погрешность которого есть $O(h^2)$ при $u \in C^4[a, b]$. Если же использовать пятиточечную аппроксимацию, то получаем разностное уравнение Пуассона

$$\frac{1}{12\, h^2}(-u_{i-2} + 16\, u_{i-1} - 30\, u_i + 16\, u_{i+1} - u_{i+2}) = f_i + \psi_i \tag{5.33}$$

с погрешностью $O(h^4)$ на функциях $u \in C^6[a, b]$.

До сих пор мы рассматривали аппроксимации дифференциальных выражений и операторов на функциях $u(x)$, от которых требовались только свойства гладкости. Перейдем теперь к аппроксимации дифференциальных уравнений

$$L\,u = f, \tag{5.34}$$

с учетом того, что $u(x)$ есть решение именно этого уравнения. Как мы сейчас увидим, использование данного факта позволяет существенно улучшить оценку погрешности аппроксимации.

Пусть $u_h = \{u_i\}$ есть решение уравнения (5.34), "оно"аппроксимируется системой сеточных уравнений

$$L_h u_i = f_i + \psi_i,\ L_h u_h = f_h + \psi_h,\ \psi_h = \{\psi_i\}, \tag{5.35}$$

в которых оператор P_h из (5.29) для простоты берется единичным. Предположим теперь, что погрешность аппроксимации ψ_h, представляющая собой сеточные значения какой-то функции непрерывного аргумента, имеет следующий специальный вид:

$$\psi_h = h^{\gamma}(Q\,L\,u)_h + O(h^{\gamma_1});\ \gamma_1 > \gamma > 0, \tag{5.36}$$

где Q — некоторый дифференциальный оператор, который можно аппроксимировать на функции $f(x) = L\,u(x)$ некоторым сеточным оператором Q_h с погрешностью порядка γ_2. Тогда из (5.36) мы получаем

$$\psi_h = h^\gamma [Q_h f_h + O(h^{\gamma_2})] + O(h^{\gamma_1}),\ \gamma_2 > 0,$$

откуда следует справедливость следующего сеточного уравнения:

$$L_h u_h = f_h + h^\gamma Q_h f_h + \bar{\psi}_h,$$

$$\bar{\psi}_h = O(h^{\gamma+\gamma_2}) + O(h^{\gamma_1}) = O(h^{\bar{\gamma}}),\ \bar{\gamma} = \min(\gamma + \gamma_2, \gamma_1). \tag{5.37}$$

Если теперь здесь отбросить члены малого порядка, то приходим к сеточному уравнению (с некоторым отличным от u_h сеточным решением v_h)

$$L_h v_h = f_h + h^\gamma Q_h f_h, \tag{5.38}$$

с повышенным порядком аппроксимации $\bar{\gamma} > \gamma$.

Перейдем далее к особенностям конечно-разностных аппроксимаций граничных условий краевой задачи. Пусть нам заданы на левой границе условия Дирихле вида $x_0 \leqslant a \leqslant x_1,\ u|_a = g_a$. Рассмотрим последовательно следующие возможные случаи:

- *интерполяция нулевого порядка* с погрешностью $O(h)$ —

$$u_0 = u_a + (x_0 - a)u'(\xi_0) \Rightarrow v_0 = g_a, \tag{5.39}$$

- *линейная интерполяция* (Коллатца)—

$$v_0 = \frac{g_a(x_1 - x_0) - v_1(a - x_0)}{x_1 - a},\ \psi = 0(h^2), \qquad (5.40)$$

- *квадратичная интерполяция* по точкам $x_0 < a < x_1$ с погрешностью ($\psi = 0(h^3)$)—

$$v_0 = g_a + (u_a - u_1)\frac{a - x_0}{a - x_2} + \frac{(a - x_0)^2}{2} \times$$
$$\Big[\frac{g_a}{(x_1 - a)(x_2 - a)} + \frac{v_2}{(x_2 - x_1)(x_2 - a)} - \frac{2\,v_1}{(x_1 - a)(x_2} - x_1)\Big]. \qquad (5.41)$$

  Некоторым минусом данного сеточного уравнения является то, что в нем используется "лишняя" точка x_2, а это несколько усложняет и алгоритм решения, и его теоретическое исследование;

- *аппроксимация Шортли—Уэллера*: данный способ основан на трехточечной аппроксимации (5.21) второй производной по точкам $x_0 = a, x_1 = x_0 + h_0, x_2 = x_1 + h_1$, которая формально имеет только 1-й порядок, т. е. $\psi_1 = O(h)$:

$$v_1/h_0h_1 - v_2/(h_0 + h_1)h_1 = g_0/(h_0 + h_1).$$

  Однако это уравнение обладает свойством строгого диагонального преобладания, что, как мы увидим позже, дает свои большие достоинства.

Рассмотрим теперь условия 2-го и/или 3-го рода: $\alpha u + \beta u' = \gamma$ при $x = a$. Ограничиваясь простейшей аппроксимацией этого уравнения, т. е. заменой u_a с помощью линейной

интерполяции по значениям u_0, u_1 и приближением производной разностью 1-го порядка, мы приходим к уравнению

$$\alpha \frac{v_0(x_1 - a) + v_1(a - x_0)}{x_1 - x_0} - \beta \frac{v_1 - v_0}{x_1 - x_0} = \gamma,$$

которое имеет погрешность в общем случае $\psi = O(h)$ и $\psi = O(h^2)$ при $a = (x_0 + x_1)/2$.

Рассмотренные сеточные уравнения могут быть как линейные, так и нелинейные, в зависимости от свойств исходной дифференциальной краевой задачи. Например, коэффициенты $p_{i\pm1/2}$ в (5.23), а также величины g_a и α, β, γ в приведенных выше соотношениях могут зависеть от самого решения. Формально этот факт никак не влияет на аппроксимационные свойства разностных уравнений.

В заключение данного пункта рассмотрим особенности аппроксимации уравнения Пуассона в цилиндрической системе координат на равномерной сетке. Применение стандартной трехточечной схемы при $r_i > 0$ дает сеточный оператор

$$\begin{aligned}(\Delta_r^h u)_i &\equiv \frac{-1}{h^2 r_i}[r_{i+1/2}(u_i - u_{i+1}) + r_{i-1/2}(u_i - u_{i-1})] = \\ &= \frac{1}{r_i}(u' + ru'' + \frac{h^2}{6}u''' + rh^2 u^{(4)}) + O(h^4),\end{aligned} \tag{5.42}$$

который аппроксимирует "цилиндрический- оператор Лапласа $\frac{1}{r}(ru')' = \frac{1}{r}(u' + ru'')$ с погрешностью второго порядка.

Данный подход неприменим для построения сеточных уравнений на оси симметрии, и при $r_i = 0$ необходимо использовать специальные приемы. Один из возможных подходов заключается в построении сетки таким образом, чтобы

точка симметрии $r = 0$ попадала в середину первого интервала, т. е. $x_0 = -h/2, x_1 = h/2$. В этом случае с учетом условия симметрии $u_0 = u_1$ сеточный оператор (5.42) имеет погрешность аппроксимации $O(h^2)$.

Второй способ использует тот факт, что при $r \to 0$ уравнение Пуассона приобретает форму

$$\Delta^{(1)}_{r,0} u = -2u_{rr} = f(0).$$

Принимая во внимание свойство симметрии решения рассматриваемой задачи, мы определяем на оси сеточный оператор

$$(\Delta^h_{r,0} u)_0 \equiv 4\frac{u_0 - u_1}{h^2} = (\Delta^{(1)}_{r,0} u)_0 + \frac{h^2}{12}(u^{(4)})_0 + O(h^4),$$

который имеет погрешность аппроксимации $O(h^2)$.

Остановимся теперь подробнее на некоторых особенностях конечно-разностных аппроксимаций диффузионно-конвективного уравнения (5.3), предполагая для простоты заданными краевые условия Дирихле.

Если в (5.3) конвективный член аппроксимируется правой разностью, а диффузионный — в соответствии с (5.23), то для i-го узла сетки получаем *трехточечное разностное уравнение* вида

$$-a_i v_{i-1} + b_i v_i - c_i v_{i+1} = f_i,$$

коэффициенты которого описываются формулами

$$a_i = \frac{p_{i-1/2}}{h^2}, \quad c_i = \frac{p_{i+1/2}}{h^2} - \frac{q_i}{h}, \quad b_i = \frac{p_{i-1/2} + p_{i+1/2}}{h^2} - \frac{q_i}{h} + r_i, \tag{5.43}$$

а погрешность аппроксимации имеет порядок $O(h)$. Коэффициент b_i при неизвестной величине будем называть диагональным, а остальные коэффициенты ($-a_i$, $-c_i$ в данном случае) — внедиагональными.

Определение 5.3. *Разностное уравнение называется уравнением положительного типа, если диагональный коэффициент является положительным, внедиагональные — неположительными и имеется диагональное преобладание (в общем случае — модуль диагонального коэффициента уравнения не меньше суммы модулей внедиагональных).*

Свойство положительности типа, как мы увидим, является очень полезным как в теоретическом плане, так и в практическом. Для уравнения (5.43) оно сводится к неравенствам

$$b_i > 0, \qquad a_i, c_i \geqslant 0, \qquad b_i \geqslant a_i + c_i.$$

Если в *неразложимой системе* все уравнения положительного типа, то она называется *системой положительного типа.*.

Неразложимой называется система уравнений, в которой не может быть выделена подсистема меньшего порядка такая, что соответствующие искомые компоненты могут быть найдены вне зависимости от остальных неизвестных.

Система уравнений положительного типа является частным случаем более широкого класса монотонных систем уравнений.

Определение 5.4. *Система линейных алгебраических*

уравнений с матрицей L_h называется монотонной, если из векторного неравенства $L_h v_h \geqslant 0$ следует $v_h \geqslant 0$ (здесь и далее векторные неравенства понимаются покомпонентно, как и равенства, а нуль по контексту понимается как вектор соответствующей размерности с нулевыми компонентами).

Заметим, что определения 5.3 и 5.4 содержат методологическое отличие: понятие положительного типа существует и для отдельного уравнения, а понятие монотонности – только для системы. Оба свойства не зависят от правых частей уравнений и характеризуются только коэффициентами при неизвестных. Доказательство того, что система положительного типа является монотонной, может быть проведено от противного. Если система разностных уравнений положительного типа (монотонная), то определяемый ею разностный оператор также положительного типа (монотонный). Другими словами, в этом случае матрица системы также называется монотонной, но данное понятие мы рассмотрим подробнее позже.

Определение 5.5. *Система разностных уравнений (разностная схема) называется абсолютно монотонной, если она монотонна при любых шагах сетки, и условно монотонной, если она монотонна при выполнении некоторых условий на шаги сетки.*

Отметим, что понятие положительности типа можно ввести не только для уравнения, но и для разностного выражения.

Аналогично интуитивно оправдано и употребление термина *монотонная аппроксимация* (или разностная схема, под

этим фактически будет подразумеваться, что соответствующие разностные уравнения являются уравнениями положительного типа). Более строго категория "монотонность" характеризуется в теории матриц, и ниже мы активно будем ее эксплуатировать при алгебраическом анализе разностных методов.

Из приведенных формул видно, что если $r_i \geqslant 0$ (это мы будем предполагать и в дальнейшем), то уравнение является абсолютно монотонным при $q_i \leqslant 0$, а при положительных q_i монотонность достигается только в случае выполнения условия

$$h \leqslant \frac{p_{i+1/2}}{q_i},$$

которое означает, что при относительно больших конвективных членах шаг сетки должен быть достаточно мал. Важно подчеркнуть, что это требование на "малость" h может быть гораздо жестче, чем это следует из оценки погрешности аппроксимации. А использование чрезмерно большого числа узлов — значительное удорожание вычислительного алгоритма.

Если в дифференциальном уравнении (5.3) $q(x_i) \geqslant 0$, то получить абсолютно монотонное разностное уравнение очень легко: достаточно для аппроксимации u' использовать не правую, а левую разность. В результате этого формулы для разностных коэффициентов заменятся с (5.43) на аналогичные:

$$a_i = \frac{p_{i-1/2}}{h^2} + \frac{q_i}{h}, \quad c_i = \frac{p_{i+1/2}}{h^2}, \quad b_i = \frac{p_{i-1/2} + p_{i+1/2}}{h^2} + \frac{q_i}{h} + r_i.$$

Погрешность аппроксимации здесь будет тоже первого порядка, а условие монотонности при $q_i < 0$ принимает форму

$$h \leqslant \frac{p_{i-1/2}}{|q_i|}.$$

Когда конвективный коэффициент $q(x)$ является знакопеременной функцией, то легко догадаться, как сконструировать и здесь абсолютно монотонную аппроксимацию. Это делается с помощью так называемых *направленных односторонних разностей* — в точках с $q(x_i) \leqslant 0$ используется правая разность, а при $q_i > 0$ — левая. Получаемые разностные уравнения записываются формально следующим единым образом:

$$a_i = \frac{p_{i-1/2}}{h^2} + \frac{|q_i| + q_i}{2\,h}, \qquad c_i = \frac{p_{i+1/2}}{h^2} + \frac{|q_i| - q_i}{2\,h},$$
$$b_i = \frac{p_{i-1/2} + p_{i+1/2}}{h^2} + \frac{|q_i|}{h} + r_i.$$

Так как во всех трех случаях погрешность аппроксимации имеет первый порядок, то совершенно естественно попытаться ее повысить в результате применения не односторонних разностей, а центральной. Тогда уравнение имеет по-прежнему трехточечный вид (5.43), а его коэффициенты -

$$a_i = \frac{p_{i-1/2}}{h^2} + \frac{q_i}{2\,h}, \quad c_i = \frac{p_{i+1/2}}{h^2} - \frac{q_i}{2\,h}, \quad b_i = \frac{p_{i-1/2} + p_{i+1/2}}{h^2} + r_i.$$

Достигаемый здесь эффект — второй порядок погрешности — омрачается тем обстоятельством, что разностное уравнение при любом знаке q_i становится условно монотонным. Правда, условия при этом в два раза “смягчаются”:

$$h \leqslant \frac{2\,p_{i\pm1/2}}{|q_i|}.$$

5.2.3. Погрешность аппроксимации и ошибка сеточной краевой задачи. Запишем дифференциальную краевую задачу (5.1), (5.2) в форме одного операторного уравнения

$$\bar{L}\bar{u}(x) = \bar{f}(x), \;\; x \in \bar{\Omega} = \Omega \cup \Gamma, \tag{5.44}$$

где функция $\bar{f}(x)$ включает в себя и $f(x)$, и g_a, и g_b, а $\bar{L}$ — операторы L, l_a и l_b.

Соответственно, систему сеточных уравнений, получаемую после аппроксимации дифференциальных уравнений и граничных условий, представим в алгебраическом виде

$$\bar{L}_h \bar{v}_h = \bar{P}_h \bar{f}_h; \;\; \bar{v}_h, \bar{f}_h \in \mathbb{R}^{\bar{N}}; \bar{L}_h, \bar{P}_h \in \mathbb{R}^{\bar{N},\bar{N}}. \tag{5.45}$$

Здесь $\bar{v}_h$ и $\bar{f}_h$ — сеточные функции, или векторы, $\bar{L}_h$ и $\bar{P}_h$ — конечно-мерные разностные операторы, или квадратные матрицы. Размерность задачи (5.45) равна $\bar{N} = N + 2$, т. е. общему количеству узлов, в которых определены компоненты векторов $\bar{v}_h, \bar{f}_h$. Мы не будем останавливаться на примерах конкретного представления операторов $\bar{L}_h, \bar{P}_h$ или сеточных функций, но сделаем два следующих замечания.

Замечание 5.2. Строго говоря, если при аппроксимации граничных условий используются узлы сетки, выходящие за пределы расчетной области $\bar{\Omega} = [a, b]$ (например, $x_0 \notin (a, b)$ в (5.40)), то в (5.44) необходимо использовать гладкое продолжение (экстраполяцию) решения за пределы $\bar{\Omega}$ и соответствующее "небольшое" расширение расчетной области.

Замечание 5.3. При аппроксимации краевых условий 1-го рода в (5.45) могут входить тривиальные граничные уравнения вида $v_0 = g_a$ в (5.39). В таком случае из алгебраической системы естественно исключить "лишние" неизвестные переменные, в результате чего ее порядок понизится, и вместо (5.44) мы запишем

$$L_h v_h = P_h f_h; \;\; v_h, f_h \in \mathbb{R}^N, \;\; L_h \in \mathbb{R}^{N,N}. \tag{5.46}$$

Строго говоря, раньше N означало число внутренних узлов сетки, и (5.46) соответствует задаче Дирихле. Однако далее по контексту соотношение (5.46) мы будем иногда употреблять и для смешанной краевой задачи, когда после исключения условий Дирихле $\bar{N} = N + 1$. Отметим также, что зачастую $\bar{P}_h$ есть единичный оператор, и тогда мы его в тексте просто опускаем.

Если в уравнении (5.45) мы вместо компонент вектора $\bar{v}_h = \{v_i\}$ подставим соответствующие сеточные значения точного решения $v(x_i)$ и проведем разложения в ряд Тейлора, то получим равенство

$$\bar{L}_h \bar{u}_h = P_h \bar{f}_h + \psi_h, \;\; \psi_h = \psi_h^u - \psi_h^f = O(h^\gamma). \tag{5.47}$$

Определение 5.6. *Порядок (погрешности) аппроксимации дифференциальной краевой задачи есть γ, если $||\psi_h|| \leqslant Ch^\gamma$, где постоянные $\gamma > 0$ и C не зависят от h.*

Остановимся теперь временно на рассмотрении линейных краевых задач. Вычитая уравнения (5.47) и (5.45) почленно, для *вектора ошибки сеточного решения*

$$z_h = \bar{u}_h - \bar{v}_h = \{u_i - v_i\} \tag{5.48}$$

приходим к уравнению

$$\bar{L}_h z_h = \psi_h. \tag{5.49}$$

Предполагая в (5.49) матрицу $\bar{L}_h$ невырожденной, из очевидных соотношений

$$||z_h|| = ||\bar{L}_h^{-1}\psi_h|| \leqslant ||\bar{L}_h^{-1}|| \cdot ||\psi_h||$$

мы получаем утверждение, фактически соответствующее теореме эквивалентности Лакса.

Теорема 5.2. *Пусть выполняются неравенства* $||\psi_h|| \leqslant C_1 h^\gamma$, $||L_h^{-1}|| \leqslant C_2$.

Тогда справедлива оценка

$$||z_h|| \leqslant C_1 C_2 h^\gamma.$$

Отметим, что константы C_1 и C_2, которые предполагаются не зависящими от h, характеризуют соответственно аппроксимацию и устойчивость разностной задачи.

При исследовании порядка скорости сходимости сеточных решений при $h \to 0$ достаточно показать существование констант C_1 и C_2. Более детальный анализ предполагает получение априорных оценок ошибки z_h в зависимости от исходных данных, и результат в таком случае, как правило, зависит от типа используемых норм, которые в самой теореме 5.2 не конкретизируются.

5.2.4. Компактные разностные схемы повышенной точности. Как уже отмечалось в п. 5.2.2, погрешность аппроксимации дифференциального уравнения можно повысить при использовании того факта, что сама аппроксимация строится не на произвольной функции, а на решении исходного уравнения. На этом принципе конструируются так называемые *компактные разностные схемы* повышенной точности, название которых происходит от того, что улучшение аппроксимации происходит без расширения сеточного шаблона.

В качестве простейшего примера использования этой идеи рассмотрим построение трехточечных схем высокой точности для уравнения Пуассона

$$u'' = f(x), \;\; u \in C^p, \;\; f \in C^{p-2}.$$

На равномерной сетке с шагом h с помощью тейлоровского разложения и использования исходного уравнения имеем следующее соотношение:

$$\frac{1}{2}(u_{i-1} + u_{i+1}) = u_i + \frac{h^2}{2}u_i'' + ... + \frac{h^p}{p!}u^{(p)}(\xi_i),$$

где в правой части участвуют производные только четных порядков.

Отсюда получаем, что погрешность аппроксимации может быть выражена через производные правой части:

$$\begin{aligned} &\frac{u_{i-1} - 2\,u_i + u_{i+1}}{2} = u_i'' + ... + \frac{2\,h^p}{p!}u^{(p)}(\xi_i) = \\ &= f_i + \frac{2h^2}{4!}f_i'' + ... + \frac{2h^p}{p!}f^{(p-2)}(\xi_i), \;\; \xi_i \in [x_{i-1}, x_{i+1}]. \end{aligned} \tag{5.50}$$

Таким образом, если производные функции $f(x)$ могут быть явно вычислены, то уравнение (5.50) позволяет построить аппроксимации любого порядка, который только позволяет гладкость искомого решения $u(x)$. Более того, производные от f могут быть аппроксимированы со вторым или более высоким порядком погрешности, и при этом итоговая ошибка компактной схемы останется того же порядка.

Аналогично можно построить компактную схему для уравнения Пуассона в цилиндрической системе координат на равномерной сетке. Как следует из (5.42), при $r_i > 0$ трехточечный сеточный оператор обладает следующими аппроксимационными свойствами:

$$(\Delta_r^h u)_i \equiv \frac{-1}{r_i}[r_{i+1/2}(u_i - u_{i+1}) + r_{i-1/2}(u_i - u_{i-1})] =$$
$$= (\Delta_r^{(1)} u)_i - \frac{h^2}{12}(\Delta_r^{(1)}\Delta_r^{(1)} u)_i - \frac{h^2}{12 r_i}\Big(\frac{1}{r}u'\Big)'_i.$$

Отсюда с учетом того, что u есть решение исходного уравнения Пуассона, после трехточечной аппроксимации дополнительного дифференциального члена получаем компактную схему

$$-\frac{1}{r_i}[r^{(1)}_{i+1/2}(u_i - u_{i+1}) + r^{(1)}_{i-1/2}(u_i - u_{i-1})] = f_i - \frac{h^2}{12}(\Delta_r^h f)_i,$$
$$r^{(1)}_{i\pm1/2} = r_{i\pm1/2} - \frac{h^2}{12 r_{i\pm1/2}}.$$

Аналогично из (5.43) с учетом соотношения $(u^{(4)})_0 = \frac{3}{8}(\Delta^{(1)}_{r,0}\Delta^{(1)}_{r,0} u)_0$ получаем компактную схему на оси

$$4\frac{u_0 - u_1}{h^2} = f_0 - \frac{1}{4}(f_0 - f_1) + O(h^4).$$

Перейдем теперь к более сложной задаче — построению компактной конечно-разностной схемы повышенного порядка для диффузионно-конвективного уравнения с переменными коэффициентами

$$L\,u \equiv -(p(x)u')' + q(x)u' = f(x) - r(x)u. \tag{5.51}$$

Отправной точкой для нас будет "классическая"трехточечная аппроксимация 2-го порядка

$$\begin{aligned}(L^h u)_i \equiv \frac{1}{h^2}[p_{i+1/2}(u_i - u_{i+1}) + p_{i-1/2}(u_i - u_{i-1})] + \\ + \frac{q_i}{2h}(u_{i+1} - u_{i-1}),\end{aligned} \tag{5.52}$$

важной особенностью которой является следующее специальное представление:

$$L^h u = L\,u - \frac{h^2}{12} R\,u + 0(h^4). \tag{5.53}$$

Здесь Ru есть дифференциальное выражение вида

$$R\,u = p\frac{d^4u}{dx^4} + 2\frac{d\,p\,d^3u}{d\,x\,d\,x^3} + \frac{3}{2}\frac{d^2p}{d\,x^2}\frac{d^2u}{d\,x^2} + \frac{1}{2}\frac{d^3p}{d\,x^3}\frac{d\,u}{d\,x} - 2\,q\frac{d^3u}{d\,x^3}.$$

Отметим, что если пытаться непосредственно аппроксимировать член Ru, содержащий старшие производные u, то получим многоточечное выражение, включающее минимум 5 узлов. Поэтому будем искать разностный трехточечный оператор L_1^h, для которого на достаточно гладких функциях u, p справедливо представление

$$L_1^h u = L\,u - \frac{h^2}{12} L\Big(\frac{1}{p} L\,u\Big) + 0(h^4). \tag{5.54}$$

Для этого сначала выпишем несложное проверяемое соотношение

$$L\Big(\frac{1}{p} L\,u\Big) = \Big(\frac{d}{d\,x} p \frac{d}{d\,x} + q \frac{d}{d\,x}\Big) \frac{1}{p} \Big(\frac{d}{d\,x} p \frac{d\,u}{d\,x} + q \frac{d\,u}{d\,x}\Big) = R\,u + M\,u,$$

где последний дифференциальный член имеет вид

$$\begin{aligned} M\,u = \frac{d}{d\,x}\Big[\Big(\frac{1}{2}\frac{d^2 p}{d\,x^2} - \frac{1}{p}\Big(\frac{d\,p}{d\,x}\Big)^2 - \frac{d\,q}{d\,x}\Big)\frac{d\,u}{d\,x}\Big] + \\ + q \frac{d}{d\,x}\Big(\frac{q}{p}\frac{d\,u}{d\,x}\Big) - \frac{d\,q}{d\,x}\frac{d^2 u}{d\,x^2}. \end{aligned} \tag{5.55}$$

Отсюда с помощью (5.53) и (5.54) получаем, что искомый разностный оператор может быть записан в форме

$$L_1^h u = L^h u - \frac{h^2}{12} M^h u, \tag{5.56}$$

где $M^h u$ есть легко конструируемая трехточечная аппроксимация второго порядка дифференциального выражения $M\,u$ из (5.55).

В итоге искомая компактная схема приобретает внешне простой вид

$$(L_1^h)_i u = \frac{1}{h^2}[\check{p}_{i-1/2}(u_i - u_{i-1}) + \hat{p}_{i+1/2}(u_i - u_{i+1})] = f_i - r_i u_i - \frac{h^2}{12} L^h\Big(\frac{f -$$

(5.57)

Однако участвующие в ней коэффициенты описываются достаточно длинными формулами:

$$\hat{p}_{i+1/2} = p_{i+1/2} - \frac{q_i}{2}h + \frac{1}{12}\Big[d_{i+1/2} + h^2 q_i \Big(\frac{q}{p}\Big)_{i+1/2} - \frac{h}{2}(q_{i+1} - q_{i-1})\Big],$$

$$\check{p}_{i-1/2} = p_{i-1/2} + \frac{q_i}{2}h + \frac{1}{12}\Big[d_{i-1/2} + h^2 q_i \Big(\frac{q}{p}\Big)_{i-1/2} - \frac{h}{2}(q_{i+1} - q_{i-1})\Big]. \tag{5.58}$$

Здесь, в свою очередь, величины $d_{i+1/2}$ выражаются следующими представлениями:

$$d_{i+1/2} = 2(p_i - 2p_{i+1/2} + p_{i+1}) - \frac{(p_{i+1} - p_i)^2}{p_{i+1/2}} - h_x(q_{i+1} - q_i),$$

а $d_{i-1/2}$ выписывается аналогично.

Нетрудно убедиться, что схема (5.57) имеет погрешность аппроксимации $O(h^4)$. Конкретные необходимые для этого свойства функций u, p, q, r, f могут быть выведены из анализа участвующих в приведенных выкладках дифференциальных выражений.

Заметим, что полученная компактная схема 4-го порядка является далеко не единственной. Действительно, если вместо коэффициентов $\hat{p}_{i\pm1/2}$ из (5.57), (5.58) возьмем какую-то из их возможных аппроксимаций

$$\hat{p}^{(k)}_{i\pm1/2} = \hat{p}_{i\pm1/2} + h^2\varphi_k(x_{i\pm1/2}) \tag{5.59}$$

с достаточно гладкой функцией $\varphi_k(x)$, то легко убедиться в справедливости следующих соотношений:

$$(\bar{L}^h_k u) \equiv -\frac{1}{h^2}[\hat{p}^k_{i+1/2}(u_{i+1} - u_i) - \hat{p}^k_{i-1/2}(u_i - u_{i-1})] =$$

$$(\bar{L}^h_1 u_i) - \frac{h^2}{12}(Q_k u)_i, \;\; Q_k u = \frac{d}{d\,x}\Big(\varphi_k \frac{d\,u}{d\,x}\Big).$$

Отсюда несложно конструируется семейство трехточечных операторов L_k^h четвертого порядка

$$L_k^h u \equiv \bar{L}_k^h + \frac{h^2}{12}(Q_k^h - M^h)u = L_1 u - \frac{h^2}{12}(R+M)u + O(h^4).$$

В частности, при отсутствии конвективного члена ($q(x) = 0$) А. А.Самарским предложена схема с коэффициентами

$$\begin{aligned} &\hat{p}^{(2)}_{i\pm1/2} = 6(p_i^{-1} + 4\,p^{-1}_{i\pm1/2} + p^{-1}_{i\pm1}) = \\ &p_{i\pm1/2} + \frac{h^2}{12}\Big[\frac{1}{2}\frac{d^2p}{d\,x^2} + \frac{1}{p}\big(\frac{d\,p}{d\,x}\big)^2\Big]_{i\pm1/2} + O(h^4). \end{aligned} \tag{5.60}$$

Если же использовать значения $p_{i\pm1/2}$ неудобно, то можно вместо (5.60) применять следующее параметризованное представление (ниже $\theta \in [0,1]$):

$$\begin{aligned} &\hat{p}^{(3)}_{i\pm1/2} = 2[\theta p_i^{-1} + 4(1-\theta)(p_i + p_{i\pm1})^{-1} + \theta p^{-1}_{i\pm1}] = \\ &= p_{i+1/2} + \frac{h^2}{4}\Big[\frac{1}{2}\frac{d^2p}{d\,x^2}\Big]_{i\pm1/2} + O(h^4). \end{aligned}$$

§ 5.3. Методы конечных объемов

Методы дискретизации исходных краевых задач, рассматриваемые выше, основаны на непосредственной аппроксимации дифференциальных уравнений конечно-разностными уравнениями. Однако такой подход не является единственным, так как практически для каждой задачи математической физики известны интегральные соотношения, отражающие законы сохранения (баланса) той или иной субстанции

(массы, тепла, энергии и т. д.). Эти интегральные равенства могут быть также аппроксимированы путем замены участвующих в них производных конечными разностями и применения квадратурных формул для приближения интегралов.

Поскольку аппроксимируемые при этом балансные соотношения рассматриваются на элементарных ячейках сетки, для такого подхода сложилось название, получившее широкое распространение в последние годы, — *метод конечных объемов* (МКО)..

Получающиеся сеточные алгебраические уравнения иногда совпадают с "обычными" конечно-разностными, а иногда несколько отличаются. Теоретический анализ для *балансных*, или *консервативных, разностных схем*, как их иногда называют, также почти не отличается, и обычно методическое преимущество здесь заключается в меньших требованиях к гладкости исходного решения, так как в законах сохранения содержатся производные более низких порядков, чем в эквивалентных дифференциальных уравнениях.

При общности методологических принципов балансных методов здесь имеются и существенные различия в тактике реализации, и на них мы остановимся отдельно.

Главным принципиальным качеством методов конечных объемов является возможность применения поэлементной технологии аппроксимации и формирования алгебраической системы, что приближает их в этом плане к методам конечных элементов, которые опираются на понятие локальной матрицы жесткости и рассматриваются в следующем параграфе. Аналогично в МКО могут быть определены *ло-*

кальные матрицы баланса, из которых собирается (ассемблируется) *глобальная матрица* системы уравнений. Независимое друг от друга вычисление локальных матриц баланса не только намного упрощает вычислительную сложность реализации многомерных задач, но и значительно повышает эффективность распараллеливания алгоритмов.

5.3.1. Интегральное тождество Г. И. Марчука и его применение. Изложение данного подхода наиболее наглядно представлено на примере одномерного уравнения диффузии

$$-\frac{\mathrm{d}}{\mathrm{d}x}p(x)\frac{\mathrm{d}u}{\mathrm{d}x}+r(x)u=f(x),\quad p(x)>0,\quad a<x<b. \tag{5.61}$$

Будем считать, что функции p, r и f — кусочно-непрерывные на отрезке $[a,b]$ с возможными точками разрыва первого рода. Решение $u(x)$ при этом в точке ξ разрыва функции p удовлетворяет так называемым условиям сопряжения $u|_{\xi^+}=u|_{\xi^-}$, $p^+u'|_{\xi^+}=p^-u'|_{\xi^-}$, где знаковые индексы относятся к односторонним предельным значениям функций. Для построения аппроксимации возьмем сетку с множеством узлов

$$a=x_0<x_1<\ldots<x_N<x_{N+1}=b, \tag{5.62}$$

включающим все точки разрыва.

Проинтегрируем уравнение (5.61) по x в пределах $(x_{i-1/2},x_{i+1/2})$, $x_{i\pm1/2}=(x_i+x_{i\pm1})/2$, в результате чего получим соотношение баланса

$$J_{i+1/2} - J_{i-1/2} + \int\limits_{x_{i-1/2}}^{x_{i+1/2}} (r\,u - f)\mathrm{d}\,x = 0, \tag{5.63}$$

где введены обозначения $J_{i\pm1/2} = J(x_{i\pm1/2})$ функции потока

$$J(x) = -p(x)\frac{\mathrm{d}\,u}{\mathrm{d}x}. \tag{5.64}$$

Для нахождения $J_{i-1/2}$ проинтегрируем исходное уравнение в пределах $(x_{i-1/2}, x)$, что дает соотношение

$$-p\frac{\mathrm{d}\,u}{\mathrm{d}x} = J_{i-1/2} - \int\limits_{x_{i-1/2}}^{x} (ru - f)\mathrm{d}\,x'. \tag{5.65}$$

Поделив последнее соотношение на p (напомним, что этот коэффициент диффузии предполагается строго положительным) и произведя затем интегрирование на интервале (x_{i-1}, x_i), после несложных преобразований находим выражение для потока

$$J_{i-1/2} = \left[-(u_i - u_{i-1}) + \int\limits_{x_{i-1}}^{x_i} \frac{\mathrm{d}\,x}{p} \int\limits_{x_{i-1/2}}^{x} (r\,u - f)\mathrm{d}\,x'\right] \Bigg/ \int\limits_{x_{i-1}}^{x_i} \frac{\mathrm{d}\,x}{p}. \tag{5.66}$$

Отсюда заменой индекса i на $i+1$ получается формула для $J_{i+1/2}$. Подставляя значения потоков в уравнение (5.63), приходим к *интегральному тождеству Г. И.Марчука*

$$
\begin{aligned}
&-\frac{u_{i+1}-u_i}{\int\limits_{x_i}^{x_{i+1}}\frac{dx}{p}}+\frac{u_i-u_{i-1}}{\int\limits_{x_{i-1}}^{x_i}\frac{dx}{p}}+\int\limits_{x_{i-1/2}}^{x_{i+1/2}}(ru-f)dx\\
&=\left(\int\limits_{x_i}^{x_{i+1}}\frac{dx}{p}\right)^{-1}\int\limits_{x_i}^{x_{i+1}}\frac{dx}{p}\int\limits_{x}^{x_{i+1/2}}(ru-f)dx' \qquad (5.67)\\
&+\left(\int\limits_{x_{i-1}}^{x_i}\frac{dx}{p}\right)^{-1}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{p}\int\limits_{x_{i-1/2}}^{x}(ru-f)dx'.
\end{aligned}
$$

Это точное соотношение может служить основой для вывода разностных уравнений с погрешностями различных порядков. В частности, если $r(x)\equiv 0$, а интегралы от функций f и $1/p$ вычисляются точно, то из (5.67) следуют трехточечные разностные уравнения с нулевой погрешностью.

Правая часть (5.67) является величиной порядка $O(h^3)$, если только все участвующие в ней подынтегральные функции внутри интервалов (x_i, x_{i+1}) гладкие. Это следует просто из того, что она обращается в нуль после применения квадратурной формулы центральных прямоугольников, которая, как хорошо известно, имеет ошибку интегрирования третьего порядка.

Если этой величиной пренебречь, а оставшиеся интегралы вычислить приближенно по той же формуле центральных прямоугольников, то на равномерной сетке получим "классическое" разностное уравнение диффузии (естественно, после перенормировки, сводящейся к делению на h) с ошибкой второго порядка. На неравномерной сетке (при использовании приближения $\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f\,\mathrm{d}\,x \approx f_i(h_i+h_{i-1})/2$) получаем трехто-

чечное уравнение, которое после деления на $\frac{h_i+h_{i-1}}{2}$ также переходит в обычное разностное уравнение, имеющее ошибку $O(h)$.

Естественно, что если функция $1/p$ точно не интегрируется, то приближенное вычисление соответствующих интегралов вносит дополнительную погрешность. Если $p(x)$ задана аналитически, то это, как правило, не привносит существенных трудностей. На практике часто значения функции p известны только в узлах сетки, и тогда требуется дополнительно привлекать интерполяцию (линейную или более высоких порядков в зависимости от требуемой точности).

В принципе , все интегралы в тождестве (5.67) при использовании соответствующих квадратурных формул могут быть определены со сколь угодно малой ошибкой (если $r \neq 0$, то придется искомую функцию интерполировать предварительно по значениям u_i). В итоге это позволяет построить разностные уравнения с произвольно высоким порядком аппроксимации. Если $r(x) \equiv 0$, то все они будут оставаться трехточечными.

Рассмотренный подход применим и для диффузионно-конвективного уравнения вида (5.3), в котором член $q(x)u'$ заменим на $\frac{\mathrm{d}}{\mathrm{d}\,x}(p\frac{\mathrm{d}s}{\mathrm{d}\,x}u)$, что соответствует введению переменной s, удовлетворяющей уравнению $\mathrm{pus}' = q'$ (если p и q — константы, то $s = (l\,n\,u)q/p$). Тогда после обобщения понятия потока (содержащего диффузионную и конвективную составляющие)

$$J(x) = -p\Big(\frac{\mathrm{d}\,u}{\mathrm{d}\,x} - \frac{\mathrm{d}\,s}{\mathrm{d}\,x}u\Big) = -pe^s\frac{\mathrm{d}}{\mathrm{d}\,x}(e^{-s}u) \tag{5.68}$$

уравнение (5.3) сводится к форме

$$\frac{\mathrm{d}J}{\mathrm{d}\,x} + r\,u = f(x). \tag{5.69}$$

Далее выписывается соотношение баланса (5.63) и с небольшими изменениями повторяются выкладки (5.62)—(5.66). Например, вместо (5.67) после отбрасывания малой правой части получается уравнение

$$-\frac{(e^{-s}u)_{i+1} - (e^{-s}u)_i}{\int\limits_{x_i}^{x_{i+1}}} \frac{\mathrm{d}\,x}{p\,e^s} + \frac{(e^{-s}u)_i - (e^{-s}u)_{i-1}}{\int\limits_{x_{i-1}}^{x_i}} \frac{\mathrm{d}\,x}{p\,e^s} + \int\limits_{x_{i-1/2}}^{x_{i+1/2}} (ru - f)\mathrm{d}\,x = 0 \tag{5.70}$$

Как и в предыдущей главе, отметим основные алгебраические свойства получаемых уравнений. Во-первых, если s не зависит от x и $r(x) \equiv 0$, то при любых приближениях оставшихся интегралов уравнения являются симметричными и положительного типа. Если $r \neq 0$, то при использовании простейшей квадратурной формулы прямоугольников эти свойства сохраняются, а при более сложных — теряются. Присутствие конвективного члена в (5.68), т. е. $s(x) \neq \mathrm{const}$, нарушает как свойство положительности типа уравнений (нет диагонального преобладания), так и свойство симметричности. Однако, как легко заметить, оба эти свойства возвращаются, если ввести новую неизвестную переменную $w_i = e^{-s_i}u_i$.

Из соотношения (5.70) следует одна аппроксимация, заслуживающая пристального внимания. В данном случае достаточно остановиться на частном случае $r(x) \equiv 0$, $p(x) \equiv 1$ и равномерной сетке. Проведя линейную интерполяцию для

$s(x)$ и вычисляя затем точно интегралы в знаменателях , получаем разностное уравнение

$$\frac{1}{h^2}\Big[-\frac{s_i-s_{i-1}}{e^{s_i-s_{i-1}}-1}v_{i-1}+\left(\frac{s_i-s_{i-1}}{1-e^{s_{i-1}-s_i}}+\frac{s_{i+1}-s_i}{e^{s_{i+1}-s_i}-1}\right)v_i$$
$$-\frac{s_{i+1}-s_i}{1-e^{s_i-s_{i+1}}}v_{i+1}\Big]=f_i, \tag{5.71}$$

исследование и применение которого связаны с именами А. М.Ильина, Шарфеттера—Гуммеля и многих других. Главная его особенность заключается в том, что оно имеет погрешность $O(h^2)$ и является абсолютно монотонным (более конкретно, уравнение (5.71) — положительного типа при сколь угодно большой величине конвективного члена). Наибольшее значение такого типа разностные схемы имеют при решении задач с так называемыми пограничными, или внутренними, слоями, характеризующимися наличием локальных подобластей с резко меняющимися свойствами решения.

Рассмотренные разностные уравнения являются балансными в том смысле, что сохраняют дискретный аналог свойства баланса субстанции для исходной дифференциальной задачи. Если в (5.69) для простоты взять $r=0$, то для любых x', x'' из интервала $[a,b]$ справедливо соотношение

$$J(x'')-J(x')=\int\limits_{x'}^{x''}f(x)\mathrm{d}\,x. \tag{5.72}$$

Вводя обозначение для сеточного потока

$$J^h_{i+1/2} = -\left[(e^{-s}u)_{i+1} - (e^{-s}u)_i\right]\left(\int\limits_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{p\,e^s}\right)^{-1} \tag{5.73}$$

и суммируя уравнения (5.70) в пределах от некоторого индекса i_1 до i_2, получаем разностный баланс (дискретный эквивалент (5.72))

$$J^h_{i_1+1/2} - J^h_{i_2-1/2} = \sum_{i=i_1}^{i_2} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} f\,\mathrm{d}\,x. \tag{5.74}$$

Важно отметить, что после замены интегралов в (5.74) на их дискретные аналоги получаемое равенство выполняется точно, какие бы квадратурные формулы для вычисления интегралов в (5.73), (5.74) не применялись.

Сделаем еще одно замечание об учете граничных условий при использовании интегрального тождества Г. И. Марчука. Если на концах отрезка $[a, b]$ заданы условия Дирихле, то вопросов нет, так как уравнения $v_0 = u_a$, $v_{N+1} = u_b$ на сетке (5.62) учитывают их точно. Когда же, например, в точке $x = a$ задан поток J вида (5.64) или (5.68) (наиболее естественное в математической физике условие), то целесообразно сетку выбрать так, чтобы середина первого интервала $x_{1/2}$ совпадала с границей. Тогда в балансном соотношении (5.63) при $i = 1$ поток $J_{1/2}$ учитывается точно и его не надо аппроксимировать. То же самое верно и при определении потока на правом конце отрезка.

В качестве эффективного применения интегрального тождества (5.67) мы покажем, как из него с помощью использования соответствующих квадратурных формул построить на

равномерной сетке

$$x_i = a + ih,\ \ i = 0, 1, ..., N+1,\ \ h = \frac{b-a}{N+1},\ \ x_{N+1} = b$$

схему четвертого порядка точности, если все участвующие в исходном уравнении (5.61) функции являются достаточно гладкими. Для аппроксимации последнего интеграла в левой части (5.67) с помощью построения отрезка ряда Тейлора мы можем написать

$$\begin{aligned}\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f dx &= \int\limits_{x_{i-1/2}}^{x_{i+1/2}} \Big(f_i + (x-x_i) f_i' + \frac{(x-x_i)^2}{2} f_i'' + \\ &\frac{(x-x_i)^3}{6} f_i''' \Big) dx + O(h^5) \\ &= \frac{h}{24}(f_{i-1} + 22 f_i + f_{i+1}) + O(h^5).\end{aligned} \tag{5.75}$$

Двойные интегралы в правой части (5.67) можно вычислить с хорошей точностью путем применения квадратурной формулы Симпсона (или парабол, см., например, |30|) к внешнему интегралу:

$$\begin{aligned}&\int\limits_{x_i}^{x_{i+1}} \frac{dx}{p} \int\limits_{x}^{x_{i+1/2}} f dx = \frac{h}{6}\Big(\frac{1}{p_i} \int\limits_{x_i}^{x_{i+1/2}} f dx + \frac{1}{p_{i+1}} \int\limits_{x_{i+1}}^{x_{i+1/2}} f dx \Big) + O(h^5) = \\ &\frac{h^2}{12}\Big[\frac{1}{p_i}(f_i + \frac{f_i + f_{i+1}}{2}) - \frac{1}{p_{i+1}}(f_{i+1} + \frac{f_i + f_{i+1}}{2}) \Big] + O(h^5) = \\ &= \frac{h^2}{24}\Big(\frac{3f_i + f_{i+1}}{p_i} - \frac{3f_{i+1} + f_i}{p_{i+1}} \Big) + O(h^5).\end{aligned} \tag{5.76}$$

Выражения, аналогичные(5.75), (5.76), легко выписываются и для подынтегральной функции ru. Вычисление интегралов в знаменателях левой части (5.67) осуществляем также с помощью формулы Симпсона следующим образом:

$$\Big(\int\limits_{x_i}^{x_{i+1}} \frac{dx}{p}\Big)^{-1} = \bar{p}_{i+1/2}/h + O(h^5), \tag{5.77}$$

где введено обозначение

$$\bar{p}_{i+1/2} = [(p_i^{-1} + 4p_{i+1/2}^{-1} + p_{i+1}^{-1})/6]^{-1}. \tag{5.78}$$

В итоге после подстановки соотношений вида (5.75)—(5.78) в (5.67) и отбрасывания остаточных членов получаем трехточечное уравнение четвертого порядка точности:

$$\begin{aligned}
&\bar{p}_{i-1/2}(u_i - u_{i-1}) + \bar{p}_{i+1/2}(u_i - u_{i+1}) + \\
&+\frac{h^2}{24}(r_{i-1}u_{i-1} + 22r_iu_i + r_{i+1}u_{i+1}]) = \frac{h^2}{24}\Big\{f_{i-1} + 22f_i + f_{i+1} + \\
&+h\Big[\bar{p}_{i+1/2}\big(\frac{3\varphi_i + \varphi_{i+1}}{p_i} - \frac{3\varphi_{i+1} + \varphi_i}{p_{i+1}}\big) + \\
&\bar{p}_{i-1/2}\big(\frac{3\varphi_i + \varphi_{i-1}}{p_i} - \frac{3\varphi_{i-1} + \varphi_i}{p_{i-1}}\big)\Big]\Big\},
\end{aligned} \tag{5.79}$$

где для краткости используется обозначение $\varphi_i = r_iu_i - f_i$.

Рассмотрим теперь, как для этих уравнений аппроксимировать краевые условия 3-го рода

$$
\begin{aligned}
&(-s_a u' + t_a u)|_a = g_a, \;\; s_a \geq 0, \;\; t_a \geq 0, \;\; s_a + t_a > 0, \\
&(s_b u' + t_b u)|_b = g_b, \;\; s_b \geq 0, \;\; t_b \geq 0, \;\; s_b + t_b > 0,
\end{aligned}
\tag{5.80}
$$

из которых формально следуют условия Дирихле и Неймана как частные случаи (при нулевых значениях s_a или s_b и t_a или t_b соответственно). Ниже условия 1-го рода ($s_a = 0$ или $s_b = 0$) мы можем из рассмотрения исключить, так как их учет легко производится без внесения какой-либо дополнительной погрешности (при этом уравнения (5.79) строятся только для внутренних узлов сетки $i = 1, ..., N$, к которым добавляются точные равенства $u_0 = u_a = g_a/t_a$ и/или $u_{N+1} = u_b = g_b/t_b$).

Остановимся подробнее на учете первого из условий(5.80). В силу определения потока (5.64)справедливо равенство

$$
-\int\limits_{x_0}^{x_1} u' d\, x = u_0 - u_1 = \int\limits_{x_0}^{x_1} J \frac{dx}{p},
$$

после применения к которому формулы Симпсона получаем соотношшение

$$
u_0 - u_1 = \frac{h}{6}\Big(\frac{J_0}{p_0} + 4\frac{J_{1/2}}{p_{1/2}} + \frac{J_1}{p_1}\Big) + O(h^5).
$$

Отсюда с помощью простых преобразований имеем

$$u_0 - u_1 - h\bar{p}_{1/2}^{-1} J_0 = [4p_{1/2}^{-1}(J_{1/2} - J_0) + p_1^{-1}(J_1 - J_0)]h/6 + O(h^5) =$$
$$= [4p_{1/2}^{-1} \int_{x_0}^{x_{1/2}} \varphi(x)dx + p_1^{-1} \int_{x_0}^{x_1} \varphi dx]h/6 + O(h^5), \quad \varphi = f - ru,$$

$$\bar{p}_{1/2}^{-1} = (p_0^{-1} + 4p_{1/2}^{-1} + p_1^{-1})/6. \tag{5.81}$$

Для участвующих в данном уравнении интегралов с помощью конечно-разностных аппроксимаций производных

$$\varphi_0' = \frac{4\varphi_1 - 3\varphi_0 - \varphi_2}{2h} + O(h^2), \quad \varphi_0'' = \frac{\varphi_0 - 2\varphi_1 + \varphi_2}{h^2} + O(h)$$

можно построить следующие квадратурные формулы:

$$\int_{x_0}^{x_{1/2}} \varphi \, d\, x = \varphi_0 + \frac{h}{2}\varphi_0' + \frac{h^2}{2}\varphi_0'' + O(h^4) = \frac{h}{24}(8\varphi_0 + 5\varphi_1 - \varphi_2) + O(h^4),$$
$$\int_{x_0}^{x_1} \varphi \, d\, x = \frac{h}{12}(5\varphi_0 + 8\varphi_1 - \varphi_2) + O(h^4).$$

Подставляя данные выражения в (5.81), а также учитывая следующее из (5.80) равенство

$$J_0 = p_0(g_a - t_a u_0)/s_a,$$

приходим к трехточечному уравнению с погрешностью четвертого порядка:

$$u_0 - u_1 - h\bar{p}_{1/2}^{-1} p_0(g_a - t_a u_0)/s_a =$$
$$= \frac{h^2}{36p_{1/2}}(8\varphi_0 - 5\varphi_1 - \varphi_2) + \frac{h^2}{72p_1}(5\varphi_0 + 8\varphi_1 - \varphi_2) + O(h^5).$$

Совершенно аналогичным образом аппроксимация второго из краевого условия (5.80) приводит к уравнению с погрешностью $O(h^5)$ относительно неизвестных u_{N+1}, u_N, u_{N-1}.

И последний комментарий данного пункта: наличие известных особенностей подынтегральных функций в тождестве (5.67) не препятствует, вообще говоря, построению сеточных уравнений сколь угодно высокой точности. Для этого достаточно только применить известные квадратурные формулы для вычисления сингулярных интегралов,см.[30]

5.3.2. Интегро-балансные аппроксимации. Данная группа методов построения балансных (консервативных) разностных уравнений базируется, как и предыдущая, на использовании интегральных законов сохранения субстанции, которые для уравнения диффузии (5.61) записываются в виде (5.63) или подобных ему выражений. Однако дальнейшая идея заключается в конструировании аппроксимации разностей потоков, связанных с решением, вследствие (5.64), интегральным соотношением

$$u(x'') - u(x') = -\int\limits_{x'}^{x''} J(x)\frac{\mathrm{d}\,x}{p(x)}, \qquad x', x'' \in [a, b]. \tag{5.82}$$

Если здесь при $x' = x_i$, $x'' = x_{i+1}$ под интегралом поток проинтерполировать (отсюда и название метода — *интегро-интерполяционный*,, принадлежащее А. А. Самарскому [55]), то величины $J_{i\pm1/2}$ легко выражаются через u_i:

$$J_{i+1/2} = -(u_{i+1} - u_i) \Big/ \int\limits_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{p} - \varphi_{i+1/2}, \tag{5.83}$$

где остаточный член $\varphi_{i\pm1/2}$ записывается в виде

$$\varphi_{i+1/2} = \left[J'_{i+1/2} \int\limits_{x_i}^{x_{i+1}} \frac{(x - x_{i+1/2})\mathrm{d}\,x}{p} + \frac{1}{2} \int\limits_{x_i}^{x_{i+1}} (x - x_{i+1/2})^2 J''(\xi_{i+1/2}) \frac{\mathrm{d}\,x}{p} \right] \Big/ \int\limits_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{p} \tag{5.84}$$

и имеет порядок $O(h^2)$, если функция $p(x)$ достаточно гладкая на интервале $[x_i, x_{i+1}]$, а $\xi_{i+1/2}$ — некоторая точка из этого интервала.

После подстановки (5.76) в балансное соотношение (5.63) получаем равенство

$$-\frac{u_{i+1} - u_i}{\int\limits_{x_i}^{x_{i+1}}} \frac{\mathrm{d}\,x}{p} + \frac{u_i - u_{i-1}}{\int\limits_{x_{i-1}}^{x_i}} \frac{\mathrm{d}\,x}{p} + \int\limits_{x_{i-1/2}}^{x_{i+1/2}} (\mathrm{r}u - f)\mathrm{d}\,x = \varphi_{i+1/2} - \varphi_{i-1/2}, \tag{5.85}$$

левая часть которого та же, что и в интегральном тождестве (5.67). Отсюда следует, что правые части у них тоже должны совпадать, хотя внешне они существенно отличаются: в (5.67) остаточные члены выписываются через искомое решение, а в (5.78) — через производные от потока.

С точки зрения минимизации ошибки $\varphi_{i+1/2}$ применение формул (5.67) и (5.76) не является, вообще говоря, оптимальным. Как известно из теории гауссовских квадратурных формул (см., например, [12], [30]), при заданных пределах интегрирования и числе квадратурных узлов их координаты и коэффициенты однозначно определяются весовой функцией. А

именно, абсциссы формул Гаусса являются корнями многочленов соответствующих порядков, ортогональных на интервале интегрирования с заданной весовой функцией. Поэтому вместо $x_{i+1/2}$ оптимальной является некоторая точка $\xi_{i+1/2}$, зависящая от функции $p(x)$, и вместо (5.76) можно записать

$$u_{i+1} - u_i = -J_{\xi_{i+1/2}} \int\limits_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{p} + \Phi_{i+1/2}, \tag{5.86}$$

где ошибка $\Phi_{i+1/2}$ имеет вид

$$\Phi_{i+1/2} = \frac{J''(\eta_{i+1/2})}{2} \int\limits_{x_i}^{x_{i+1}} \frac{(x - \xi_{i+1/2})^2}{p(x)} \mathrm{d}\,x, \qquad \eta_{i+1/2} \in [x_i, x_{i+1}].$$

Если пользоваться выражением (5.79), то в соотношении баланса в качестве x', x'' надо брать $\xi_{i-1/2}$, $\xi_{i+1/2}$ и последние значения необходимо использовать в (5.78) вместо $x_{i-1/2}$, $x_{i+1/2}$ соответственно. При этом в определенном смысле "наилучшая" для $r = 0$ схема второго порядка на неравномерной сетке имеет вид

$$\begin{gathered} \frac{-p_{i+1/2}(v_{i+1} - v_i)}{h_i} + \frac{p_{i-1/2}(v_i - v_{i-1})}{h_{i-1}} = \int\limits_{\xi_{i-1/2}}^{\xi_{i+1/2}} f(x)\mathrm{d}\,x, \\ \xi_{i+1/2} = \int\limits_{x_i}^{x_{i+1}} x\frac{\mathrm{d}\,x}{p} \Big/ \int\limits_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{p}, \quad p_{i+1/2} = h_i \Big/ \int\limits_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{p(x)}. \end{gathered} \tag{5.87}$$

Эту схему можно конкретизировать после выбора способа вычисления $\xi_{i+1/2}$, а также линейной или квадратичной интерполяции функции $f(x)$ под интегралом (если он не вычисляется точно).

Рассмотренный подход, который будем называть *интегро-балансным* , можно развивать следующим образом. Для аппроксимации интеграла в (5.75) естественно попытаться использовать формулы численного интегрирования более высоких порядков. Обозначая через n_i, $\alpha_{i,k}$ и $J_{i,k}$ число квадратурных узлов, коэффициенты и значения потока в точках $x_{i,k} \in [x_i, x_{i+1}]$, мы можем записать

$$\begin{aligned} &\int\limits_{x_i}^{x_{i+1}} J(x)\frac{\mathrm{d}\,x}{p(x)} = \frac{h_i}{p_{i+1/2}}\sum_{k=1}^{n_i}\alpha_{i,k}J_{i,k} + \varphi_{i+1/2}, \\ &\sum_{k=1}^{n_i}\alpha_{i,k} = 1, \qquad J_{i,k} = J(x_{i,k}), \\ &\varphi_{i+1/2} = C_i h_i^{\beta}\frac{\partial^{\beta} J(\eta_i)}{\partial\, x^{\beta}}, \qquad \eta_i \in [x_i, x_{i+1}]. \end{aligned} \tag{5.88}$$

Здесь не зависящая от h_i константа C_i и показатель степени β определяются конкретной формулой интегрирования. Например, при использовании гауссовых квадратур для $p(x) \equiv 1$ имеем

$$\varphi_{i+1/2} = \frac{h_i (n_i!)^4}{[(2\,n_i)!]^3(2n+1)} J^{(2\,n_i)}(\eta_i).$$

Подчеркнем, что в результате выбора n_i, $\alpha_{i,k}$ и $x_{i,k}$ порядок погрешности β в формулах (5.81) может быть сколь угодно большим, техника приближенного интегрирования позволяет это сделать даже при наличии известной особенности функции потока $J(x)$.

Выписывая аналогичные (5.75) балансные соотношения на интервалах $[x_{i-1}, x_i]$ и $[x_i, x_{i+1}]$, с помощью выражений (5.81) после несложных выкладок получаем

$$-p_{i+1/2}\frac{u_{i+1}-u_i}{h_i}+p_{i-1/2}\frac{u_i-u_{i-1}}{h_{i-1}}$$
$$=\sum_{k=1}^{n_i}\alpha_{i,k}J_{i,k}-\sum_{k=1}^{n_{i-1}}\alpha_{i-1,k}J_{i-1,k}-\frac{p_{i+1/2}}{h_i}\varphi_{i+1/2}+\frac{p_{i-1/2}}{h_{i-1}}\varphi_{i-1/2}. \tag{5.89}$$

На различных сеточных интервалах значения квадратурных коэффициентов, да и числа n_i , могут отличаться. Но в силу их одинаковых нормировок $\sum\limits_{k=1}^{n_i}\alpha_{i,k}=1$ соотношения (5.82) можно переписать в виде

$$-p_{i+1/2}\frac{u_{i+1}-u_i}{h_i}+p_{i-1/2}\frac{u_i-u_{i-1}}{h_{i-1}}=\sum_{l=1}^{m_i}\gamma_{i,l}(J_{i,l}-J_{i-1,l})+\hat{\psi}_i,$$

$$m_i=\sup\{n_i,n_{i-1}\},$$

$$\hat{\psi}_i=\frac{p_{i-1/2}}{h_{i-1}}\varphi_{i-1/2}-\frac{p_{i+1/2}}{h_i}\varphi_{i+1/2},\qquad\sum_{l=1}^{m_i}\gamma_{i,l}=1. \tag{5.90}$$

Здесь значения $J_{i,l}=J(x_{i,l})$ берутся в точках из того же набора квадратурных узлов, что и в формулах (5.82), но конкретные величины $x_{i,l}$ и $x_{i,k}$ могут отличаться даже при совпадении $k=l$ (в частности, в (5.83) одному $J_{i,l}$ могут соответствовать в (5.82) значения потока в разных точках $x_{i,k'}$ и $x_{i,k''}$).

Поскольку из балансного соотношения (5.72) имеем

$$J_{i,l}-J_{i-1,l}=\int\limits_{x_{i-1,l}}^{x_{i,l}}f(x)\mathrm{d}\,x,$$

то выражения (5.83) приводятся к виду

$$-p_{i+1/2}\frac{u_{i+1}-u_i}{h_i}+p_{i-1/2}\frac{u_i-u_{i-1}}{h_{i-1}}=\sum_{l=1}^{m_i}\left(\gamma_{i,l}\int\limits_{x_{i-1,l}}^{x_{i,l}}f(x)\mathrm{d}\,x\right)_i+\hat{\psi}_i. \tag{5.91}$$

Если функция $f(x)$ точно не интегрируется, то приближенное вычисление правой части в (5.84) вносит дополнительную погрешность $\check{\psi}_i$, но она может быть сделана сколь угодно малой (при достаточном числе узловых точек) за счет выбора способа интерполяции и соответствующих квадратурных формул. Предполагая для простоты, что эти формулы строятся для каждого i по одинаковому числу узлов r, приходим в итоге к выражениям

$$-p_{i+1/2}\frac{u_{i+1}-u_i}{h_i}+p_{i-1/2}\frac{u_i-u_{i-1}}{h_{i-1}}=\sum_{l=1}^{m_i}\gamma_{i,l}\sum_{k=1}^{r}q_{i,l,k}f(x_{i,l,k})+\hat{\psi}_i+\check{\psi}_i.$$

Отбрасывая теперь погрешность аппроксимации, получаем сеточное уравнение

$$-p_{i+1/2}\frac{v_{i+1}-v_i}{h_i}+p_{i-1/2}\frac{v_i-v_{i-1}}{h_{i-1}}=\sum_{l=1}^{m_i}\gamma_{i,l}\sum_{k=1}^{r}q_{i,l,k}f_{i,l,k}. \tag{5.92}$$

Рассмотрим теперь, как данный подход можно применить к аппроксимации граничных условий, не привнося никаких дополнительных погрешностей. Запишем краевое условие на левом конце отрезка $x=a$ в формально обобщенном виде

$$\alpha_a u(a)+\beta_a J(a)=g_a, \tag{5.93}$$

где при обращении чисел α_a или β_a в нуль получаем частные случаи условий Неймана или Дирихле.

Сетку выбираем с условием $x_0 = a$ и вместо соотношения (5.82) при $i = 0$ выписываем точное равенство, вытекающее из (5.75), (5.81):

$$-p_{1/2}\frac{u_1 - u_0}{h_0} = \sum_{k=1}^{n_0}\alpha_{0,k}J_{0,k} + \frac{p_{1/2}}{h_0}\psi_{1/2}, \tag{5.94}$$

в котором все обозначения соответствуют (5.81). Умножая равенство (5.87) на β_a и прибавляя к нему почленно (5.86), получаем

$$\alpha_a u_0 - \beta_a p_{1/2}\frac{u_1 - u_0}{h_0} = \beta_a\sum_{k=1}^{n_0}\alpha_{0,k}[J_{0,k} - J(a)] + \beta_a\frac{p_{1/2}}{h_0}\psi_{1/2} + g_a. \tag{5.95}$$

Заменяя далее разности потоков с помощью балансных соотношений вида (5.72), имеем

$$\begin{aligned}\alpha_a u_0 - \beta_a p_{1/2}\frac{u_1 - u_0}{h_0} &= \beta_a\sum_{k=1}^{n_0}\alpha_{0,k}\int_a^{x_{0,k}} f(x)\,\mathrm{d}\,x + \hat{\psi}_0 + g_a \\ &= \beta_a\sum_{l=1}^{m_0}\gamma_{0,l}\sum_{k=1}^{r}q_{0,l,k}f(x_{0,l,k}) + \hat{\psi}_0 + \check{\psi}_0 + g_a,\end{aligned} \tag{5.96}$$

где в данном случае $m_0 = n_0$, $\gamma_{0,l} = \alpha_{0,l}$, $\hat{\psi}_0 = \beta_a\dfrac{p_{1/2}}{h_0}\psi_{1/2}$, а $\check{\psi}_0$, как и в (5.84), есть суммарная погрешность вычисления m_0 интегралов от функции $f(x)$.

В итоге для $i = 0$ получаем аналогичное по обозначениям с (5.85) уравнение

$$\alpha_a v_0 - \beta_a p_{1/2}\frac{v_1 - v_0}{h_0} = \beta_a\sum_{l=1}^{m_0}\gamma_{0,l}\sum_{k=1}^{r}q_{0,l,k}f_{0,l,k} + g_a. \tag{5.97}$$

Очевидно, что при $\beta_a \neq 0$ для сохранения симметричности системы последнее уравнение надо поделить на β_a (если же $\beta_a = 0$, то проведенные после (5.86) выкладки не нужны, и к уравнениям (5.85) достаточно добавить из условия Дирихле тривиальное равенство $v_0 = g_a/\alpha_a$).

Если на правом конце задано также граничное условие $\alpha_b u(b) + \beta_b J(b) = g_b$, то сетка выбирается с правым узлом $x_{N+1} = x_b$ и из аналогичных построений получаем уравнение, замыкающее систему (5.85), (5.90):

$$\alpha_b v_{N+1} + \beta_b p_{N+1/2} \frac{v_{N+1} - v_N}{h_N} = \sum_{l=1}^{m_N} \gamma_{N,l} \sum_{k=1}^{r} q_{N,l,k} f_{N,l,k} + g_b. \tag{5.98}$$

Легко увидеть, что если $\alpha_a\,\beta_a > 0$, $\alpha_b\,\beta_b > 0$, то уравнения (5.90), (5.91) являются абсолютно монотонными, конкретнее — при любых шагах сетки являются уравнениями положительного типа (напомним, что функция $p(x)$ по изначальному предположению строго положительна). Если же одна из пар граничных коэффициентов имеет разные знаки, то для положительности типа соответствующего разностного граничного уравнения требуется выполнение неравенства типа $|\beta_a p_{1/2}| > \alpha_a h_0$.

Построенные сеточные уравнения являются балансными в двояком смысле. Во-первых, они выведены из соотношений "детального" баланса (5.75) (кстати, при наличии граничного условия Неймана, т. е. при $\alpha_a\,\alpha_b = 0$, значения потоков $J_{i,l}$ легко вычисляются рекуррентным образом). Во-вторых, для численных усредненных потоков, определяемых как

$$J_{i+1/2} = \sum_{k=1}^{n_i} \alpha_{i,k} J_{i,k} = \sum_{l=1}^{m_i} \gamma_{i,l} J_{i,l} = -p_{i+1/2} \frac{v_{i+1} - v_i}{h_i},$$

выполняются (точно!) суммарные соотношения баланса

$$J_{i''+1/2} - J_{i'-1/2} = \sum_{i=i'}^{i''} \sum_{l=1}^{m_i} \sum_{k=1}^{r} q_{i,l,k} f_{i,l,k}, \qquad (5.99)$$

$$0 \leqslant i' < i'' \leqslant N.$$

Аналогичные разностные аппроксимации произвольного порядка точности могут быть построены и для диффузионно-конвективного уравнения (5.3). Вводя определение потока вида (5.68), (5.69) и повторяя предыдущие выкладки, мы вместо (5.85) при $r = 0$ получим уравнение

$$-p_{i+1/2} \frac{(e^{-s}v)_{i+1} - (e^{-s}v)_i}{h_i} + p_{i-1/2} \frac{(e^{-s}v)_i - (e^{-s}v)_{i-1}}{h_{i-1}} = \sum_{l=1}^{m_i} \gamma_{i,l} \sum_{k=1}^{r} q_{i,l,k} f(x_{i,l,k}), \qquad (5.100)$$

где значения $x_{i,l,k}$, $\gamma_{i,l}$ и $q_{i,l,k}$ будут, вообще говоря, отличаться от предыдущих, а величины $p_{i+1/2}$ имеют вид

$$p_{i+1/2} = h_i \Big/ \int_{x_i}^{x_{i+1}} \frac{\mathrm{d}\,x}{pe^s}.$$

Важно отметить, что полученная сеточная *аппроксимация экспоненциального типа* является абсолютно монотонной, т. е. при ее записи в форме (5.43) внедиагональные коэффициенты a_i, c_i являются неотрицательными для любых соотношений шагов и скоростей $q(x)$. Кроме того, уравнения (5.93)

при любых знаках $q(x)$ обладают свойством диагонального преобладания, но не по строкам, как обычно, а по столбцам, т. е.

$$b_i \geqslant a_{i+1} + c_{i-1}.$$

5.3.3. Аппроксимации на элементно-ориентированных сетках. В этом пункте будут рассмотрены фактически такие же интегробалансные аппроксимации, как и в предыдущем. Главное отличие носит (в некотором смысле) философско-методологический оттенок: что при построении сетки считать первичным — точки (узлы) или ячейки (конечные объемы, элементы).

В первом случае (*точечно-ориентированные сетки*, рассмотренные выше) конструирование сетки предполагает сначала построение множества узлов, а затем — при использовании балансных аппроксимаций — формирование около каждого из них соответствующей ячейки. Альтернативный вариант (*элементно-ориентированные сетки*) предполагает первоначальное разбиение расчетной области на конечные объемы, после чего внутри каждого из них определяется сеточный узел. Иногда эти различающиеся изначально подходы дают одинаковые итоговые результаты, но могут наблюдаться и существенные отличия.

Сеточные методы на элементно-ориентированных сетках сейчас получают все большее распространение и обнаруживают необычные свойства: при ухудшении аппроксимационных свойств в “классическом” смысле они обеспечивают точность $O(h^2)$ даже на неравномерных сетках. Впервые этот факт доказан для одномерных краевых задач в монографии

В. В.Смелова [58].

В качестве простейшего примера рассмотрим балансную аппроксимацию одномерного уравнения диффузии (5.61) при $r(x) = 0$ на элементно-ориентированной сетке. Пусть h_i (рис. 5.1) означает конечный отрезок $[x_{i-1/2},\ x_{i+1/2}]$, длина которого имеет то же обозначение, а узел x_i располагается в центре этого отрезка.

Рис. 5.1. *Элементно-ориентированная сетка*

Функцию $p(x)$ мы допускаем на интервале [а,в] кусочно-непрерывной, но точки ее разрыва предполагаются расположенными только на границах $x_{i+1/2} = x_i + \dfrac{h_i}{2} = x_{i+1} - \dfrac{h_{i+1}}{2}$, $x_{i-1/2} = x_i - \dfrac{h_i}{2} = x_{i-1} + \dfrac{h_{i-1}}{2}$ отрезков h_i. При этом в таких точках должны выполняться условия сопряжения

$$p^+ \frac{\mathrm{d}\, u^+}{\mathrm{d}\, x}\bigg|_{x_{i\pm1/2}} = p^- \frac{\mathrm{d}\, u^-}{\mathrm{d}\, x}\bigg|_{x_{i\pm1/2}}, \quad u^+_{i\pm1/2} = u^-_{i\pm1/2}, \qquad (5.101)$$

где верхние значки "$\pm$" означают предельные значения функций справа и слева при стремлении к точкам $x_{i\pm1/2}$.

Расстояние между соседними узлами определяется очевидным равенством $x_i - x_{i-1} = \frac{1}{2}(h_i + h_{i-1})$. Точное соотношение баланса на конечном объеме h_i принимает вид

$$-p^- \frac{d\, u^-}{d\, x}\bigg|_{x_i} + \frac{h_i}{2} + p^+ \frac{d\, u^+}{d\, x}\bigg|_{x_i} - \frac{h_i}{2} = \int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x) d\, x. \qquad (5.102)$$

Для получения разностных уравнений в правой части (5.95) проще всего применить квадратурную формулу центральных прямоугольников, которая имеет достаточно высокий — третий — порядок погрешности на функциях $f(x) \in C^2(h_i)$, даже при наличии разрывов в точках $x_{i\pm1/2}$:

$$\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\mathrm{d}\,x = h_i f_i + \psi_i^f, \quad |\psi_i^f| \leqslant \frac{h_i^3}{24} M_2^f, \qquad M_2^f = \max_{x\in h_i}\left\{\left|\frac{\mathrm{d}^2 f}{\mathrm{d}\,x^2}\right|\right\}. \tag{5.103}$$

С аппроксимацией левой части балансного соотношения (5.95) дело обстоит несколько сложнее, и для начала мы выпишем приближения для правой и левой производных в точке $x_{i-1/2}$:

$$\left.\frac{\mathrm{d}\,u^+}{\mathrm{d}\,x}\right|_{i-1/2} = 2\frac{u_i - u_{i-1/2}}{h_i} + \frac{h_i}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi_i'), \;\; \xi_i' \in [x_{i-1/2}, x_i],$$
$$\left.\frac{\mathrm{d}\,u^-}{\mathrm{d}\,x}\right|_{i-1/2} = 2\frac{u_{i-1/2} - u_{i-1}}{h_{i-1}} + \frac{h_{i-1}}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi_{i-1}''), \;\; \xi_{i-1}'' \in [x_{i-1}, x_{i-1/2}]. \tag{5.104}$$

Чтобы построить уравнения относительно значений решения в узлах, из последних равенств необходимо исключить величину $u_{i-1/2}$, что мы и сделаем с помощью условий сопряжения (5.94). После подстановки в них (5.97) и несложных преобразований получаем

$$
\begin{aligned}
&u_{i-1/2} = \frac{1}{\alpha_i + \beta_{i-1}}(\alpha_i u_i + \beta_{i-1} u_{i-1} - \psi_{i-1/2}),\\
&\alpha_i = \frac{2p^+_{i-1/2}}{h_i}, \qquad \beta_{i-1} = \frac{2\,p^-_{i-1/2}}{h_{i-1}},\\
&\psi_{i-1/2} = -p^+_{i-1/2}\frac{h_i}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi'_i) + p^-_{i-1/2}\frac{h_{i-1}}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi''_{i-1}).
\end{aligned}
\tag{5.105}
$$

Отсюда после подстановки $u_{i-1/2}$ в (5.97) следует выражение

$$
\begin{aligned}
p^+\frac{d\,u^+}{d\,x}\bigg|_{i-1/2} &= \alpha_i(u_i - u_{i-1/2}) - p^+_{i-1/2}\frac{h_i}{4}\frac{d^2 u}{d\,x^2}(\xi'_i)\\
&= \frac{\alpha_i\beta_{i-1}}{\alpha_i + \beta_{i-1}}(u_i - u_{i-1}) + \hat{\psi}_i,\\
\hat{\psi}_i &= \frac{\alpha_i}{\alpha_i + \beta_{i-1}}p^-_{i-1/2}\frac{h_{i-1}}{4}\frac{d^2 u}{d\,x^2}(\xi''_{i-1}) - \frac{\beta_{i-1}}{\alpha_i + \beta_{i-1}}p^+_{i-1/2}\frac{h_i}{4}\frac{d^2 u}{d\,x^2}(\xi'_i).
\end{aligned}
\tag{5.106}
$$

Таким же образом с помощью условий сопряжения в точке $x_{i+1/2}$ и аналогичных (5.97) аппроксимаций односторонних производных на стыке отрезков h_i, h_{i+1} получаем

$$
\begin{aligned}
p^-\frac{\mathrm{d}\,u^-}{\mathrm{d}\,x}\bigg|_{i+1/2} &= \beta_i(u_{i+1/2} - u_i) + p^-_{i+1/2}\frac{h_i}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi''_i)\\
&= \frac{\alpha_{i+1}\beta_i}{\alpha_{i+1} + \beta_i}(u_{i+1} - u_i) + \check{\psi}_i,\\
\check{\psi}_i &= \frac{\alpha_{i+1}}{\alpha_{i+1} + \beta_i}p^-_{i+1/2}\frac{h_i}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi''_i) - \frac{\beta_i}{\alpha_{i+1} + \beta_i}p^+_{i+1/2}\frac{h_{i+1}}{4}\frac{\mathrm{d}^2 u}{\mathrm{d}\,x^2}(\xi'_{i+1}).
\end{aligned}
\tag{5.107}
$$

Наконец, используя (5.96), (5.99) и (5.100), формируем дискретное представление балансного равенства (5.95):

$$
\begin{aligned}
&-a_{i+1}(u_{i+1}-u_i)+a_i(u_i-u_{i-1})=h_i f_i+\psi_i,\\
&a_i=\tfrac{\alpha_i\beta_{i-1}}{\alpha_i+\beta_{i-1}},\qquad \psi_i=\hat{\psi}_i-\check{\psi}_{i-1}+\psi_i^f.
\end{aligned}
\tag{5.108}
$$

Отсюда после отбрасывания погрешности аппроксимации получаем итоговое разностное уравнение

$$
-a_{i+1}(v_{i+1}-v_i)+a_i(v_i-v_{i-1})=h_i f_i. \tag{5.109}
$$

Интересно сравнить этот результат с "обычным" разностным уравнением, которое получается из балансного соотношения (5.63) (при $r\equiv 0$) на точечно-ориентированной элементной сетке, если производные в потоках $J_{i\pm1/2}$ аппроксимировать центральными разностными, а интеграл от $f(x)$ — простейшей квадратурной формулой (на основе кусочно-постоянного представления функции по значениям в точках x_i):

$$
-\frac{p(x_{i+1/2})}{x_{i+1}-x_i}(v_{i+1}-v_i)+\frac{p(x_{i-1/2})}{x_i-x_{i-1}}(v_i-v_{i-1})=\frac{x_{i+1}-x_{i-1}}{2}f_i, \tag{5.110}
$$

имеющей при данной нормировке погрешность $O(h^2)$.

Останавливаясь для простоты на непрерывных коэффициентах диффузии, сразу замечаем, что уравнения (5.102) и (5.80) при равенстве шагов сетки совпадают (с точностью до масштабирования). В противном же случае левые части обоих разностных уравнений по-прежнему одинаковы, но коэффициент при f_i в "классическом" уравнении (5.80), в обозначениях элементно-ориентированной сетки на рис. 5.1, равен $(h_{i-1}+2\,h_i+h_{i+1})/4$ (вместо h_i в (5.102)).

Главное же отличие сравниваемых разностных схем заключается в погрешностях аппроксимаций. Чтобы в этом убедиться, перепишем ψ_i из (5.101) в более наглядном виде

$$\psi_i = -\frac{1}{4}\Big\{a_{i+1}\big[h_{i+1}^2 u''(\xi_i') - h_i^2 u''(\xi_i'')\big] - a_i\big[h_i^2 u''(\xi_i') - h_{i-1}^2 u''(\xi_{i-1}'')\big]\Big\}$$
$$+O(h^3). \tag{5.111}$$

Поскольку величины a_i имеют порядок $O(h^{-1})$, то нетрудно убедиться, что в общем случае $\psi_i = O(h)$ на неравномерной сетке, даже при гладких коэффициентах $p(x)$. А это означает, что если уравнение (5.101) поделить на h_i, т. е. отмасштабировать "обычным" образом, когда разностные коэффициенты становятся порядка $O(h^{-2})$, то оказывается, что полученная схема вообще не аппроксимирует исходное уравнение диффузии. Однако величина ψ_i/h_i при $h \to 0$ оказывается порядка $0(1)$, т. е. ограниченной. Это является очень существенным при анализе погрешности разностного решения (имеет значение и специальный вид погрешности ψ_i в (5.104)).

Остановимся теперь на особенностях аппроксимации краевых условий при использовании одномерных конечных объемов. Проще всего дело обстоит с условиями Неймана на границе. Если $u' = g$ в точках a и (или) b, то в балансных соотношениях (5.94) при $i = 0$ и (или) N надо соответственно значения $\frac{\mathrm{d}\,u}{\mathrm{d}\,x}\Big|_{x_0 - \frac{h_0}{2}}$, и (или) $\frac{\mathrm{d}\,u^-}{\mathrm{d}\,x}\Big|_{x_N + \frac{h_N}{2}}$ взять точно из краевых условий и перенести их в правую часть. В случае же условий 1-го или 3-го рода применяется подход с фиктивными точками: решение предполагается продолженным на полшага

за границы расчетного отрезка и вводятся вспомогательные узлы $x_{-1} = a - \frac{h_o}{2}$, $x_{N+1} = b + \frac{h_{N+1}}{2}$. Значения решения в граничных точках затем линейно аппроксимируются: $u_a = \frac{u_{-1}+u_o}{2} + O(h^2)$, $u_b = \frac{u_N+u_{N+1}}{2} + O(h^2)$, и "лишние" переменные u_{-1}, u_{N+1} с помощью этих равенств исключаются из нулевого и N-го разностных уравнений.

§ 5.4. Конечно-элементные методы и технологии

Целью данного параграфа является ознакомление именно с технологическими, а не теоретическими вопросами методов конечных элементов. Под МКЭ понимается реализация метода Ритца или Галеркина приближенного решения краевых задач, заключающаяся в разбиении расчетной области на конечное число элементов (ячеек) и аппроксимации вариационных уравнений на основе кусочно-полиномиальных интерполяций, которые представляются с помощью финитных базисных функций, отличных от нуля только в нескольких соседних элементах, что позволяет получать алгебраические системы с разреженными матрицами. Разнообразие МКЭ определяется видом решаемых краевых задач, конфигурацией используемых конечных элементов и способами аппроксимации, т. е. выбором базисных функций и искомых величин (значений функций или их производных в каких-то точках), относительно которых составляется алгебраическая система уравнений МКЭ. Отметим, что по различным аспектам методов конечных элементов опубликовано огромное количество работ, из которых мы отметим только приведенные в списке литера-

туры монографии [15], [29], [54].

5.4.1. Принцип поэлементного формирования алгебраической системы. Как известно, для успешного решения сложной математической проблемы надо в первую очередь ввести удачные обозначения и определения, с чего мы и начнем непосредственное изложение методов конечных элементов.

Пусть расчетная область Ω разбита на K непересекающихся конечных ячеек (элементов) Ω_k^h, $k = 1, \ldots, K$, объединение которых $\Omega^h = \bigcup_{k=1}^{K} \Omega_k^h$ составляет сеточную область, представляющую дискретный вариант (дискретизацию) исходной области Ω. В частности, может быть $\Omega^h = \Omega$, но допускается и приближенное равенство $\Omega = \Omega^h + R^h \approx \Omega^h$, где объем "погрешности аппроксимации области" R^h есть величина $O(h^{-d\gamma})$, d — размерность Ω, γ – порядок аппроксимации области, а h — характерный шаг сетки, равный по порядку "усредненному" диаметру элементарной ячейки. Под характерным шагом обычно понимается некоторая величина $h = O(K^{-d})$, и, строго говоря, она имеет смысл для сеток, которые можно назвать *псевдоравномерными*, характеризуемыми тем, что при $K \to \infty$ $(h \to 0)$ отношение максимального шага сетки к минимальному (*коэффициент неравномерности сетки*) остается величиной ограниченной.

Подчеркнем, что характер приближений в методах конечных элементов представляет интерес в двух отношениях. Во-первых, необходимо знать оценку погрешности для конкретной дискретизации области с конечной величиной шага сетки. Во-вторых, и это не менее важно, особенно в теоретическом

плане, исследовать асимптотическое поведение ошибки на последовательности сгущающихся сеток при $K \to \infty$ и $h \to 0$.

Пусть на выбранной сетке Ω^h приближенное решение u^h ищется в виде линейной комбинации *финитных функций* φ_i, $i = 1, \ldots, I$, отличных от нуля только на небольшом числе ячеек и составляющих *базис* конечномерного пространства V_h:

$$u^h = \sum_{i=1}^{I} v_i^h \varphi_i. \tag{5.112}$$

Тогда алгебраическая система конечно-элементных уравнений, получаемая в результате аппроксимации вариационной постановки (5.15) по *методу Ритца* (для симметричной билинейной формы $a(,))$ и по *методу Галеркина* — в общем случае, записывается в виде

$$A^h v^h = f^h, \tag{5.113}$$

где $v^h = \{v_i^h\}$ — искомый вектор порядка I, а квадратная матрица A^h и известный вектор f^h того же порядка определяются формулами

$$A^h = \left\{ a_{ij} = a\left(\sum_{j=1}^{I} \varphi_i, \varphi_j \right) \right\}, \qquad f^h = \left\{ f_i^h = (f, \varphi_i) \right\}. \tag{5.114}$$

Соотношение (5.106) формально можно рассматривать как операторное уравнение в конечномерном пространстве, $A^h \colon V_h \to V_h$ — как *сеточный оператор* дискретизированной краевой задачи, а искомый и заданный векторы v^h, f^h являются элементами пространства V_h, соответствующими выбранным базисным функциям φ_i.

Напомним, что здесь билинейная и линейная формы выражаются через интегралы по сеточной области $\Omega^h = \bigcup_{k=1}^{K} \Omega_k^h$ (для общего вида неоднородных краевых условий необходимо еще включить их соответствующие вклады, выражающиеся через поверхностные интегралы по границе, но мы ради простоты эти аспекты пока опускаем). Очевидно, что такие интегралы можно выразить как сумму интегралов по конечным элементам Ω_k^h, каждый из которых будет содержать конечное число членов, соответствующих тем базисным функциям, которые не равны нулю на соответствующем элементе. Отсюда матрица и вектор правой части из (4.62), (4.63) могут быть представлены в виде

$$A^h = \sum_{k=1}^{K} A_k^h, \quad f^h = \sum_{k=1}^{K} f_k^h, \tag{5.115}$$

где ненулевые элементы A_k^h и f_k^h записываются в форме

$$a_{ij}^{(k)} = a\left(\sum_{j=1}^{I} \varphi_i, \varphi_j \right)^{(k)}, \qquad f_i^{(k)} = (f, \varphi_i)^{(k)}. \tag{5.116}$$

Здесь верхний индекс (k) у билинейной или линейной формы означает номер соответствующего конечного элемента Ω_k^h, интеграл по которому берется в определении величин $a_{ij}^{(k)}$, $f_i^{(k)}$. Условно такие соотношения можно отобразить в следующей форме:

$$a(\cdot\,,\cdot)^{(k)} = \int\limits_{\Omega_k^h}, \qquad a(\cdot\,,\cdot) = \int\limits_{\Omega^h} = \sum_k a(\cdot\,,\cdot)^{(k)}. \tag{5.117}$$

Пусть теперь N_k означает число базисных функций, не равных тождественно нулю на элементе Ω_k^h, а I_k — совокупность соответствующих "глобальных" номеров базисных функций φ_i. Тогда аппроксимацию (5.105) на k-м элементе, т. е. для $x \in \Omega_k^h$, можно записать в виде

$$u^h|_{\Omega_k^h} = \sum_{i \in I_k} v_i^h \varphi_i = \sum_{i'=1}^{N_k} v_{i'}^h \varphi_{i'}. \tag{5.118}$$

Здесь i' означают "локальные" номера ненулевых на Ω_k^h базисных функций. Число N_k можно назвать *порядком k-го конечного элемента*, но исторически за ним закрепилось наименование *число степеней свободы*.

Обозначим далее через $\bar{A}_k$, $\bar{f}_k^h$ квадратную матрицу и вектор порядка N_k, состоящие только из определенных в (5.109) величин, которые равны соответствующим ненулевым элементам больших A_k^h и f_k^h. Очевидно, что между введенными матрицами и векторами различных размерностей можно установить такие взаимосвязи:

$$A_k^h = P_k \bar{A}_k^h P_k^t, \quad f_k^h = P_k \bar{f}_k^h. \tag{5.119}$$

Здесь P_k — прямоугольная матрица (I строк, N_k столбцов) перенумераций, имеющая в каждом столбце один ненулевой элемент, равный единице, а P_k^t означает транспонированную матрицу, имеющую N_k строк и I столбцов.

Определение 5.7. *Матрица $\bar{A}_k^h$ порядка N_k называется локальной матрицей метода конечных элементов (соответствующей элементу Ω_k^h), а матрица A^h — глобальной матрицей МКЭ.*

Если ввести дополнительно *локальные векторы неизвестных* порядка N_k

$$\bar{v}_k^h = P_k^t v^h = \{v_i^h,\ i \in I_k\}, \tag{5.120}$$

то связанные с алгебраической системой (5.106) билинейные формы можно записать как

$$\begin{aligned}(A^h v^h, w^h) &= \left(\sum_k A_k^h v^h, w^h\right) \\ &= \left(\sum_k P_k \bar{A}_k^h P_k^t v^h, w^h\right) = \sum_k (\bar{A}_k^h \bar{v}_k^h, \bar{w}_k^h)^{(k)},\end{aligned} \tag{5.121}$$

где $(\cdot\,,\cdot)^{(k)}$ означает скалярное произведение в N_k-мерном векторном евклидовом пространстве.

Одним из замечательных достижений метода конечных элементов является создание высокоэффективной в практическом отношении математической технологии формирования алгебраических систем уравнений на основе вычисления локальных матриц МКЭ и сборки (ассемблирования) глобальных матриц, что намного уменьшило вычислительную сложность алгоритмов и обеспечило унифицированный подход к реализации различного класса методов для самого широкого круга краевых задач, включая разные размерности, типы исходных уравнений и граничных условий, сложные конфигурации границ и функциональные зависимости коэффициентов.

Мы остановимся на этой важной проблеме подробнее, ради простоты вначале рассмотрев так называемые *лагранжевые конечные элементы*, когда в представлении (5.61) величины

v_i^h соответствуют значениям (приближенным) искомого решения в каких-то точках сеточной области $x_i \in \Omega^h$, а каждая базисная функция обладает свойством $\varphi_i(x_j) = \delta_{i,j}$ ($\delta_{i,j}$ — символ Кронекера). Мы еще более конкретизируем постановку задачи дискретизации, предполагая, что точки x_i (узлы сетки) суть вершины конечных элементов Ω^h, которые являются многоугольниками (треугольниками или прямоугольниками, например) в двумерном случае и полигонами (тетраэдрами, параллелепипедами) — в трехмерном. В этой ситуации каждому узлу сетки x_i можно сопоставить i-е уравнение системы уравнений МКЭ (5.106).

Пусть при каких-то выбранных упорядоченностях (не зависящих друг от друга) элементов Ω_k^h и узлов x_i через I_k обозначается совокупность узлов сетки, являющихся вершинами k-го элемента, а через K_i — совокупность номеров элементов, имеющих общую вершину — узел с номером i (*инцидентных* с узлом x_i). Предположим, опять же для упрощения изложения, что финитная базисная функция $\varphi_i(x)$ отлична от нуля только в элементах $\Omega_{k'}^h$ с номерами $k' \in K_i$.

В рассмотренных обозначениях и предпосылках существуют два альтернирующих алгоритма формирования алгебраической системы МКЭ: с поузловой и поэлементной обработкой данных.

Первый подход заключается в непосредственном определении всех элементов i-й строки матрицы A^h и i-го элемента вектора f^h элементов $\Omega_{k'}^h$, инцидентных с узлом x_i, и вычислении необходимых интегралов по этим элементам путем организации цикла по узлам сетки. Вторая стратегия — по-

элементная — состоит в нахождении локальной матрицы $\bar{A}_k^h$ и *локального вектора правой части* $\bar{f}_k^h$ порядка N_k с помощью расчета всех требуемых интегралов по элементу Ω_k^h:

$$\bar{A}_k^h = \left\{ a_{i',j'}^{(k)} = a\Big(\sum_{j\in I_k} \varphi_i, \varphi_j \Big)^{(k)}, \quad i', j' \in I_k \right\},$$
$$\bar{f}_k^h = \left\{ (f, \varphi_{i'})^{(k)}, \quad i' \in I_k \right\}. \tag{5.122}$$

При этом сборка глобальной матрицы сводится к организации цикла по элементам и последовательному суммированию (накоплению) элементов $a_{i',j'}^{(k)}$, $f_{i'}^{(k)}$ с соответствующими элементами $a_{i,j}$, f_i^h, что можно записать в форме псевдокода

$$f^h := 0, \qquad A^h := 0, \qquad k = 1, \ldots, K:$$
$$A^h := A^h + A_k^h, \qquad f^h := f^h + f_k^h,$$
$$a_{i,j} := a_{i,j} + a_{i',j'}^{(k)}, \quad f_i^h := f_i^h + f_{i'}^{(k)} \text{ для } i = i',\ j = j'. \tag{5.123}$$

Огромное преимущество последнего способа заключается в тривиальном устранении дублирующих вычислений и логической простоте алгоритма, что особенно ценно при его программировании. Другое важное достоинство, которое нельзя не упомянуть, — это возможность эффективного распараллеливания такого метода при его реализации на многопроцессорных вычислительных системах.

5.4.2. Построение уравнений МКЭ. Мы рассмотрим одномерную краевую задачу первого рода для дифференци-

ального уравнения второго порядка достаточно общего вида:

$$-(d\,u')' + b\,u' + cu = f(x), \qquad a < x < b,$$
$$u|_a = u_a, \qquad u|_b = u_b, \tag{5.124}$$

где требования на свойства функций d, b, c, f конкретизируем несколько позже.

Для дискретизации задачи введем *неравномерную сетку* Ω^h

$$\alpha = x_0 < x_1 < \ldots < x_N < x_{N+1} = \beta \tag{5.125}$$

с шагами $h_i = x_{i+1} - x_i$.

Для данной несамосопряженной задачи *метод Галеркина* нахождения приближенного обобщенного решения u^h в конечномерном подпространстве V_h определяется вариационным уравнением

$$a(u^h, v) = (f, v) \quad \forall v \in V_h, \;\; a(u^h, v) = \int\limits_\alpha^\beta \big[d(u^h)'v' + b(u^h)'v + c\,u\,v\big]\,d\,x, \tag{5.126}$$

Выберем сначала для простоты V_h как пространство *кусочно-линейных функций* на сетке Ω^h, линейных на каждом подинтервале (x_i, x_{i+1}) и обращающихся в нуль, естественно, на границах отрезка $[\alpha, \beta]$. Такое пространство имеет размерность N и *базисные финитные функции-"крышки"* $\varphi_i(x)$, каждая из которых отлична от нуля на интервале (x_{i-1}, x_{i+1}), обладает свойством $\varphi_i(x_j) = \delta_{i,j}$ и описывается формулами

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h_{i-1}, & x \in (x_{i-1}, x_i), \\ (x_{i+1} - x)/h_i, & x \in (x_i, x_{i+1}), \\ 0, & x \neq (x_{i-1}, x_{i+1}). \end{cases} \tag{5.127}$$

Графики таких функций изображены на рис. 5.2, а их первые производные, очевидно, равны соответственно $+1, -1$ и нулю.

Рис. 5.2. *Базисные кусочно-линейные функции — "крышки"*

Поскольку искомое решение удовлетворяет неоднородным условиям Дирихле, его приближение будем искать на множестве $\tilde{V}_h$, полученном из пространства V_h сдвигом на линейную функцию $w(x) = u_\alpha + \frac{x-\alpha}{\beta-\alpha}(u_\beta - u_\alpha)$. Это означает, что приближенное решение можно искать в виде разложения по базисным функциям

$$u^h(x) = u_\alpha \varphi_0 + \sum_{i=1}^{N} u_i \varphi_i + u_\beta \varphi_{N+1}, \tag{5.128}$$

где неизвестные $v_1, \ldots, v_N$ должны по смыслу аппроксимировать значения точного решения $u_i = u(x_i)$. Тогда конкретизация уравнений (5.119) метода Галеркина сводится к системе

$$a(u^h, \varphi_i) = (f, \varphi_i), \qquad i = 1, 2, \ldots, N, \tag{5.129}$$

которая в матричной записи имеет следующую форму:

$$A^h v^h = f^h. \tag{5.130}$$

Здесь $A^h = \{a_{i,j}\}$, $v^h = \{v_i\}$, $f^h = \{f_i^h\}$—квадратная (*глобальная*) матрица МКЭ и векторы порядка N, элементы которых равны

$$\begin{aligned} a_{i,j} &= a(\varphi_i, \varphi_j) = \int_\alpha^\beta (d\varphi_i'\varphi_j' + b\varphi_i\varphi_j' + c\varphi_i\varphi_j) d\,x, \\ f_i^h &:= \int_\alpha^\beta f\varphi_i\, d\,x - g_i, \qquad i,j = 1,\ldots,N, \end{aligned} \tag{5.131}$$

где величины $g_i = 0$ для $i = 2,\ldots,N-1$, а в околограничных узлах описываются формулами

$$g_1 = u_\alpha a_{1,0}, \qquad g_N = u_\beta a_{N,N+1}.$$

Здесь $a_{1,0}$, $a_{N,N+1}$ определяются соотношениями (5.124) при $j = 0$ и $j = N+1$ соответственно. Значения g_1, g_N обусловлены наличием неоднородных граничных условий Дирихле в краевой задаче (5.117) и образуются изначально в (5.122) при интегрировании по крайним сеточным интервалам, после чего они переносятся из левой части уравнений в правую.

Рассмотрим теперь подробнее структуру матрицы алгебраической системы, которую можно представить суммой

$$A^h = A_s^h + A_c^h + A_m^h, \tag{5.132}$$

где каждая из слагаемых матриц определяется соответствующими членами билинейной формы (5.124):

$$A_s^h = \left\{ a_{i,j}^{(s)} = \int_\alpha^\beta d\varphi_i' \varphi_j' \, d\, x \right\},$$
$$A_c^h = \left\{ a_{i,j}^{(c)} = \int_\alpha^\beta b\varphi_i \varphi_j' \, d\, x \right\}, \tag{5.133}$$
$$A_m^h = \left\{ a_{i,j}^{(m)} = \int_\alpha^\beta c\varphi_i \varphi_j \, d\, x \right\}.$$

За матрицами A_s^h и A_m^h исторически закрепились названия *матрица жесткости* и *матрица масс* (введенные, по-видимому, специалистами по приложению МКЭ к задачам теории упругости), а матрицу A_c^h мы назовем *матрицей конвекции*.

Перейдем теперь от рассмотрения глобальных матриц МКЭ к построению *локальных* матриц. Пусть элементарный сеточный отрезок $\Omega_i^h = (x_i, x_{i+1})$, $i = 0, 1, \ldots, N$, означает i-й конечный элемент. Поскольку на каждом из них отличны от нуля только две базисные функции из V_h (а на крайних — по одной), то матрица МКЭ может быть представлена следующей суммой:

$$A^h = A_0^h + \sum_{i=1}^{N-1} A_i^h + A_N^h, \tag{5.134}$$

где крайние слагаемые A_0^h, A_N^h содержат по одному ненулевому диагональному элементу: $a_{1,1}^{(0)} = a(\varphi_1, \varphi_1)^{(0)}$, $a_{N,N}^{(N)} = a(\varphi_N, \varphi_N)^{(N)}$ соответственно, а остальные — по четыре элемента вида $a_{i',i''}^{(i)} = a(\varphi_{i'}, \varphi_{i''})^{(i)}$ (здесь $i' = i, i+1$, $i'' = i, i+1$, а верхний индекс означает, что интегралы в билинейной форме берутся по интервалу (x_i, x_{i+1})). Матрицам A_i^h из (5.127)

можно сопоставить локальные матрицы $\bar{A}_i^h$ ($\bar{A}_0^h$ и $\bar{A}_N^h$ — первого порядка, а остальные — второго), каждая из которых связана с интегралами только по элементу Ω_i^h.

Естественно, что любая локальная матрица МКЭ разбивается на сумму локальных матриц жесткости, конвекции и масс:

$$\begin{aligned}
&\bar{A}_i^h = \bar{A}_{s,i}^h + \bar{A}_{c,i}^h + \bar{A}_{m,i}^h = \left\{ a_{i',i''}^{(i)} = a_{i',i''}^{s,i} + a_{i',i''}^{c,i} + a_{i',i''}^{m,i} \right\}, \\
&\bar{A}_{s,i}^h = \left\{ a_{i',i''}^{s,i} = \int\limits_{x_i}^{x_{i+1}} d\varphi_{i'}' \varphi_{i''}' \, dx \right\}, \\
&\bar{A}_{c,i}^h = \left\{ a_{i',i''}^{c,i} = \int\limits_{x_i}^{x_{i+1}} b\varphi_{i'} \varphi_{i''}' \, d\,x \right\}, \\
&\bar{A}_{m,i}^h = \left\{ a_{i',i''}^{m,i} = \int\limits_{x_i}^{x_{i+1}} c\varphi_{i'} \varphi_{i''} \, d\,x \right\}.
\end{aligned} \tag{5.135}$$

Здесь индексы i', i'' принимают значения или i, или $i+1$. Поскольку производные базисных функций кусочно-постоянны

$$\varphi_i' = \begin{cases} -h_i^{-1}, & x \in (x_i, x_{i+1}), \\ h_{i-1}^{-1}, & x \in (x_{i-1}, x_i), \\ 0, & x \neq (x_{i-1}, x_{i+1}), \end{cases} \tag{5.136}$$

то способ вычисления элементов матриц МКЭ определяется свойствами коэффициентов d, b, c — в некоторых ситуациях

интегралы из (5.128) могут быть найдены точно, но в общем случае для их вычисления необходимо привлекать приближенные квадратурные формулы.

Если эти коэффициенты постоянны на каждом элементе, т. е.

$$\begin{aligned} &d(x) = p_{i+1/2}, \qquad b(x) = b_{i+1/2}, \\ &c(x) = c_{i+1/2}, \qquad x \in (x_i, x_{i+1}) \end{aligned} \tag{5.137}$$

(при этом даже возможны их разрывы в узлах сетки), то элементы локальных матриц выражаются особенно просто:

$$\bar{A}^h_{s,i} = \frac{d_{i+1/2}}{h_i} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad \bar{A}^h_{c,i} = \frac{b_{i+1/2}}{2} \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix},$$

$$\bar{A}^h_{m,i} = c_{i+1/2} \frac{h_i}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad i = 1, 2, \ldots, N-1,$$

$$\bar{A}^h_{s,0} = d_{1/2} h_0^{-1}, \qquad \bar{A}^h_{c,0} = -b_{1/2}/2, \qquad \bar{A}^h_{m,0} = 2c_{1/2} h_0,$$

$$\bar{A}^h_{s,N} = d_{N+1/2} h_N^{-1}, \quad \bar{A}^h_{c,N} = b_{N+1/2}/2, \quad \bar{A}^h_{m,N} = 2c_{N+1/2} h_N. \tag{5.138}$$

Отсюда получаем, что для i-го узла алгебраическое уравнение метода Галеркина имеет следующий вид:

$$\frac{d_{i-1/2}}{h_{i-1}}(v_i - v_{i-1}) + \frac{d_{i+1/2}}{h_i}(v_i - v_{i+1}) + \frac{b_{i-1/2}}{2}(v_i + v_{i-1}) - \frac{b_{i+1/2}}{2}(v_i + v_{i+1}) + \frac{1}{6}\big[h_{i-1}c_{i-1/2}(2\,v_i + v_{i-1}) + h_i c_{i+1/2}(2\,v_i + v_{i+1})\big] = f_i^h, \qquad i = 1, 2, \ldots, N. \tag{5.139}$$

Учет краевых условий Дирихле для уравнений, соответствующих околограничным узлам, делается здесь тривиально — в первом уравнении надо положить $v_0 = u_a$ и перенести соответствующий член в правую часть, а в последнем уравнении аналогично полагаем $v_{N+1} = u_b$.

Теперь коснемся вопроса о вычислении правых частей f_i^h. Поскольку приближенное решение мы строим в виде кусочно-линейной функции, то напрашивается такое же представление для функции $f(x)$ исходной задачи:

$$f(x) = \sum_{i=0}^{N+1} f_i \varphi_i(x) + \psi(x), \qquad f_i = f(x_i), \quad |\psi(x)| = O(h^2).$$

Тогда правая часть уравнения (5.132) имеет вид

$$f_i^h = \frac{1}{6}\big[h_{i-1}f_{i-1} + 2(h_{i-1} + h_i)f_i + h_i f_{i+1}\big], \tag{5.140}$$

а ее вычисление, очевидно, также может делаться поэлементно, при этом рассчитываются те же интегралы от произведений базисных функций, которые входят в элементы локальных матриц масс.

Система уравнений (5.132) после учета краевых условий Дирихле записывается в векторно-матричном виде

$$A^h v^h = f^h,$$

где векторы $v^h = (v_1, \ldots, v_N)^T$, $f^h = (f_1, \ldots, f_N)^T$ имеют размерность по числу внутренних узлов сетки, а матрица A^h является трехдиагональной:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & & & \\ a_{2,1} & a_{2,2} & & 0 & \\ \ddots & \ddots & \ddots & & \\ 0 & & & a_{N-1,N} & \\ & & a_{N,N-1} & a_{N,N} & \end{bmatrix}.$$

Ее элементы определяются из (5.132) следующим образом:

$$a_{i,i} = \frac{d_{i-1/2}}{h_{i-1}} + \frac{d_{i+1/2}}{h_i} + \frac{b_{i-1/2} - b_{i+1/2}}{2} + \frac{h_{i-1}c_{i-1/2} + h_i c_{i+1/2}}{3},$$
$$a_{i,i-1} = \frac{d_{i-1/2}}{h_{i-1}} + \frac{b_{i-1/2}}{2} + \frac{h_{i-1}c_{i+1/2}}{6},$$
$$a_{i,i+1} = \frac{d_{i+1/2}}{h_i} - \frac{b_{i+1/2}}{2} + \frac{h_i c_{i+1/2}}{6}.$$

Сравнение соотношений (5.132), (5.133) с конечно-разностными уравнениями, непосредственно аппроксимирующими диффузионно-конвективное уравнение (5.117) показывает, что левая часть (5.132), после его деления на "объем ячейки" $\frac{h_{i-1}+h_i}{2}$, будет та же, что у разностного уравнения, использующего аппроксимацию первой производной централь-

ной разностью (после перехода к равномерной сетке). Правая же часть f_i^h уравнений метода Галеркина содержит необъяснимое (с "разностной точки зрения") использование усредненных узловых значений функции $f(x)$. Отметим также, что в отсутствие конвективного члена ($b(x) = 0$) уравнение (5.132) является симметричным в отличие от разностного уравнения, даже на неравномерной сетке.

Естественным развитием кусочно-линейной аппроксимации является переход к *интерполяционным многочленам Лагранжа*, которые при использовании $n+1$ различного узла имеют порядок n и описываются формулой

$$L_n(x) = \sum_{l=0}^{n} f_l \psi_l(x), \tag{5.141}$$

где $f_l = f(x_l)$—узловые значения интерполируемой функции, а $\psi_l(x)$—базисные лагранжевые многочлены n-го порядка (см. [12], [30]):

$$\psi_l(x) = \frac{(x - x_0) \cdots (x - x_{l-1})(x - x_{l+1}) \cdots (x - x_n)}{(x_l - x_0) \cdots (x_l - x_{l-1})(x_l - x_{l+1}) \cdots (x_l - x_n)}, \tag{5.142}$$

обладающие свойством $\psi_l(x_m) = \delta_{lm}$. Аппроксимационные свойства многочлена (5.135) рассмотрим ниже.

Финитные базисные функции МКЭ на основе лагранжевых интерполянтов строятся следующим образом. На каждом сеточном интервале $[x_i, x_{i+1}]$ выбираются равномерно (для простоты) расположенные интерполяционные узлы

$$x_0^{(i)} = x_i, \quad x_n^{(i)} = x_{i+1}, \quad x_l^{(i)} = x_i + \frac{l h_i}{n}, \qquad l = 1, \ldots, n-1,$$

и каждому из них сопоставляется базисная лагранжевая функция $\psi_l^{(i)}$, получаемая из (5.135) приписыванием каждому интерполяционному узлу верхнего индекса i. С их помощью формируются финитные базисные функции двух типов: внутренние и узловые. Каждая внутренняя функция соответствует внутренней точке x_l сеточного отрезка и определяется как $\varphi_l^{(i)} = \psi_l^{(i)}$. Базисную функцию, соответствующую узлу x_i, обозначаем с помощью одного индекса — $\varphi^{(i)}$. Носителем каждой внутренней базисной функции $\varphi_l^{(i)}$, $l = 1, \ldots, n-1$, является интервал (x_i, x_{i+1}), а узловой — два смежных интервала:

$$\varphi_l^{(i)} = 0, \qquad x \notin (x_i, x_{i+1}), \quad l = 1, \ldots, n-1,$$

$$\varphi^{(i)} = \begin{cases} 0, & x \notin (x_{i-1}, x_{i+1}), \\ \varphi_0^{(i)}, & x \in (x_i, x_{i+1}), \\ \varphi_n^{(i-1)}, & x \in (x_{i-1}, x_i). \end{cases}$$

Для сетки (5.118) с $(n-1)$-й внутренней точкой на каждом из $N+1$ сеточных интервалов общее число базисных функций будет равно

$$\bar{N} = (n-1)(N+1) + N + 2 = n(N+1) + 1,$$

и их можно занумеровать одним индексом:

$$\varphi_k(x) = \varphi_l^{(i)}(x), \qquad k = i\,n + l, \quad k = 0, 1, \ldots, \bar{N} - 1.$$

Соответственно, можно упорядочить сквозным образом все узлы и внутренние интерполяционные точки:

$$x_k = x_l^{(i)}, \qquad k = 0, 1, \ldots, \bar{N} - 1.$$

Тогда приближенное решение для задачи Дирихле (5.117) по аналогии с (5.121) можно искать в виде

$$u^h(x) = u_\alpha \varphi_0(x) + \sum_{k=1}^{\bar{N}-2} v_k \varphi_k(x) + u_\beta \varphi_{\bar{N}-1}(x), \tag{5.143}$$

где величины v_k, по своему смыслу, должны аппроксимировать значения точного решения $u(x_k)$.

Алгебраическая система уравнений МКЭ в этом случае имеет также вид (5.123), (5.124), с точностью до определения базисных функций (формально надо заменить целые величины i, j, N на k, k', $\bar{N} - 2$ соответственно), а ее порядок равен $\bar{N} - 2$.

В данном случае локальные матрицы МКЭ имеют порядок $n+1$, по числу ненулевых базисных функций на одном интервале $[x_i, x_{i+1}]$. Приведем для примера *лагранжевые базисные функции* второго порядка, вид которых изображен на рис. 5.3:

$$\varphi^{(i)} = \begin{cases} 1 - 3\frac{x-x_i}{h_i} + 2\left(\frac{x-x_i}{h_i}\right)^2, & x \in (x_i, x_{i+1}), \\ 1 + 3\frac{x-x_i}{h_{i-1}} + 2\left(\frac{x-x_i}{h_{i-1}}\right)^2, & x \in (x_{i-1}, x_i), \end{cases} \tag{5.144}$$

$$\varphi_1^{(i)} = 4\left[\frac{x-x_i}{h_i} - \left(\frac{x-x_i}{h_i}\right)^2\right], \qquad x \in (x_i, x_{i+1}).$$

Рис. 5.3. Лагранжевые базисные функции второго порядка

Соответствующая сетка из “основных” и “вспомогательных” узлов приведена на рис. 5.4.

Рис. 5.4. Схема расположения основных и вспомогательных узлов для лагранжевых конечных элементов второго порядка

Локальная матрица жесткости (при постоянном коэффициенте $d(x)$ на интервале (x_i, x_{i+1})) в этом случае имеет следующий вид (соответствующий локальной упорядоченности неизвестных как v_i, $v_{i+1/2}$, v_{i+1}):

$$\bar{A}^h_{s,i} = \frac{p_{i+1/2}}{h_i} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}.$$

Если же коэффициент $d(x)$ зависит от x на интервале $[x_i, x_{i+1}]$, то данная матрица является простейшим приближением к точной матрице жесткости

$$\bar{A}^h_{s,i} = \left\{ a^{(i)}_{l,l'} = \int\limits_{x_i}^{x_{i+1}} d(x) \frac{d\varphi^{(i)}_l}{dx} \frac{d\varphi^{(i)}_{l'}}{d\,x} d\,x; \quad l, l' = 0, 1, 2 \right\}.$$

Как легко обнаружить, получаемые алгебраические уравнения имеют разную форму для узловых точек сетки и срединных точек (рассмотрим ради краткости исходную задачу с нулевыми коэффициентами $b(x)$ и $c(x)$):

$$\frac{d_{i+1/2}}{h_i}(-8v_i + 16v_{i+1/2} - 8v_{i+1}) = (f, \varphi_1^{(i)}), \qquad i = 0, \ldots, N-1,$$
$$\frac{d_{i-1/2}}{h_{i-1}}(v_{i-1} - 8v_{i-1/2} + 7v_i) + \frac{d_{i+1/2}}{h_i}(7v_i - 8v_{i+1/2} + v_{i+1}) = (f, \varphi_0^{(i)}),$$
$$i = 1, \ldots, N. \tag{5.145}$$

В этих уравнениях для $i = 0, 1, N$ вместо величин v_0 и v_{N+1} надо подставить известные из граничных условий значения u_α и u_β, а затем соответствующие члены перенести в правые части.

Получаемая матрица системы уравнений для $N = 4$ приведена ниже, где точками и звездочками изображены ненулевые элементы, соответствующие основным и вспомогательным узлам:

$$A^h = \begin{bmatrix} \star & \cdot & & & & & & & \\ \star & \cdot & \star & \cdot & & & & & \\ & \cdot & \star & \cdot & & & & & \\ & \cdot & \star & \cdot & \star & \cdot & & & \\ & & & \cdot & \star & \cdot & & & \\ & & & \cdot & \star & \cdot & \star & \cdot & \\ & & & & & \cdot & \star & \cdot & \\ & & & & & \cdot & \star & \cdot & \\ & & & & & & & \cdot & \star \end{bmatrix}.$$

Здесь полезно сделать одно замечание относительно системы уравнений (5.138): из трехчленных уравнений для срединных точек неизвестные $v_{i\pm1/2}$ могут быть легко выражены через узловые неизвестные и исключены из остальных уравнений, в результате чего мы редуцируем алгебраическую задачу к

системе N трехчленных уравнений относительно искомых v_i.

Этот прием получил название *конденсация* и имеет общую применимость для лагранжевых элементов любого порядка. Все "внутренние" неизвестные из одного сеточного интервала могут быть выражены через соответствующие узловые неизвестные (для этого надо обратить матрицы $(n-1)$-го порядка), и после их исключения мы приходим к трехчленной системе уравнения для v_i.

Остановимся на этом вопросе подробнее. Вводя по аналогии с обозначениями базисных функций "узловые" и "внутренние" неизвестные

$$v_i,\ v_l^{(i)}, \qquad l=1,\dots,n-1, \quad i=0,\dots,N,$$

выпишем соответственно "узловые" и "внутренние" уравнения:

$$a_{i,i-1}v_{i-1}+\sum_{l=1}^{n-1}a_{i,l}^{(i-1)}v_l^{(i-1)}+a_{i,i}v_i+\sum_{l=1}^{n-1}a_{i,l}^{(i)}v_l^{(i)}+a_{i,i+1}v_{i+1}=f_i^h,$$

$$i=1,\dots,N,$$

$$a_{l,0}^{(i)}v_i+\sum_{l'=1}^{n-1}a_{l,l'}^{(i)}v_l^{(i)}+a_{l,n}^{(i)}v_{i+1}=f_{l,i}^h, \quad i=0,\dots,N,\ \ l=1,\dots,n-1.$$

Здесь $a_{l,l'}^{(i)}$ — определяемые в соответствии с (5.128) элементы локальных матриц $\bar{A}_i^h$, а коэффициенты "узловых" уравнений равны (с учетом обозначений в (5.128), (5.129) и равенств $\varphi^{(i)}=\varphi_0^{(i)}=\varphi_n^{(i-1)}$)

$$a_{i,i-1} = a\big(\varphi^{(i-1)}, \varphi^{(i)}\big) = a_{0,n}^{(i-1)},$$

$$a_{i,i+1} = a\big(\varphi^{(i+1)}, \varphi^{(i)}\big) = a_{n,0}^{(i)},$$

$$a_{i,i} = a\big(\varphi^{(i)}, \varphi^{(i)}\big) = a_{n,n}^{(i-1)} + a_{0,0}^{(i)}.$$

Для небольших значений n "внутренние" неизвестные могут быть относительно легко выражены через "узловые" неизвестные с помощью "своих" $n-1$ уравнений:

$$v_l^{(i)} = p_l^{(i)} v_i + q_l^{(i)} v_{i+1} + g_l^{(i)}, \qquad l = 1, \ldots, n-1.$$

С помощью таких соотношений из "узловых" уравнений исключаются все $v_l^{(i-1)}, v_l^{(i)}$, $l = 1, \ldots, n-1$, что и составляет процедуру конденсации, после чего получаем систему трехточечных уравнений

$$\bar{a}_{i,i-1} v_{i-1} + \bar{a}_{i,i} v_i + \bar{a}_{i,i+1} v_{i+1} = \bar{f}_i, \qquad i = 1, \ldots, N.$$

Формулы для вычисляемых по ходу дела коэффициентов $p_l^{(i)}$, $q_l^{(i)}$, $g_l^{(i)}$ и $\bar{a}_{i,i\pm1}, \bar{a}_{i,i}, \bar{f}_i$ для краткости опускаем.

Отметим, что структура рассмотренных соотношений принципиально не меняется, если на каждом сеточном интервале $[x_i, x_{i+1}]$ брать разные порядки n_i интерполяционных многочленов.

Остановимся теперь на особенностях построения уравнений МКЭ при наличии граничных условий, содержащих производные от искомого решения. Пусть, например, в задаче (5.117) на правой границе вместо условия Дирихле задано краевое условие третьего рода

$$\varkappa u + \frac{\partial u}{\partial x} = \gamma, \qquad x = \beta.$$

В данном случае представление приближенного решения (5.136) вместо известного значения u_β содержит неизвестную величину $v_{\bar{N}-1}$, а порядок алгебраической системы (5.123) становится равным $\bar{N}-1$. При этом исходные уравнения (5.122) метода Галеркина, вследствие формулировки обобщенного решения (5.15), приобретают вид

$$a(u^h, \varphi_k) = (f, \varphi_k) - g_k, \qquad k = 0, 1, \ldots, \bar{N}-1,$$

где первые n значений g_k равны (мы полагаем для простоты $b = c = 0$):

$$g_k = u_\alpha \int_{x_0}^{x_1} d(x)\varphi_0'(x)\varphi_k'(x)\,dx = u_\alpha \int_{x_0}^{x_1} d(x)\varphi_0'(x)\big(\varphi_l^{(0)}(x)\big)'\,dx, k, l = 1, 2, \ldots, n$$

вследствие учета условия Дирихле на левой границе, "средние" значения g_k, $k = n+1, \ldots, \bar{N}-2$, — нулевые, а последнее, получаемое с учетом условия третьего рода на правой границе, имеет вид

$$g_{\bar{N}-1} = d(x_{N+1})\frac{du^h}{dx}(x_{N+1}) \cdot \varphi_{N+1}^{(0)}(x_{N+1}) = d_{N+1}\gamma - \varkappa v_{\bar{N}-1}.$$

Общим свойством лагранжевых конечных элементов любого порядка является непрерывность производных только нулевого порядка для определяемого на (α, β) приближенного решения, в силу самого способа его построения. Естественное развитие такого подхода заключается в переходе к *эрмитовой интерполяции*, использующей так называемые кратные

узлы, в которых задаются значения не только приближаемой функции, но и ее производных.

Очень простой и самой распространенной здесь является *кубическая эрмитовая интерполяция*, которая определяется по четырем задаваемым величинам v_i, v'_i, v_{i+1}, v'_{i+1} (см. [30]):

$$v^h(x) = v_i\varphi\Big(\frac{x-x_i}{h_i}\Big) + h_i v'_i\psi\Big(\frac{x-x_i}{h_i}\Big) + \\ v_{i+1}\varphi\Big(\frac{x_{i+1}-x}{h_i}\Big) + h_i v'_{i+1}\psi\Big(\frac{x_{i+1}-x}{h_i}\Big). \tag{5.146}$$

Здесь φ и ψ — финитные базисные функции с носителем (x_{i-1}, x_{i+1}), описываемые формулами

$$\varphi(\xi) = (|\xi|-1)^2(2|\xi|+1),$$

$$\psi(\xi) = \xi(|\xi|-1)^2, \qquad \xi = \frac{2x - x_{i-1} + x_{x+1}}{x_{i+1}-x_i},$$

и обладающие свойствами

$$\varphi(0) = \psi'(0) = 1, \;\; \varphi(\pm1) = \psi(\pm1) = \varphi'(\pm1) = \psi(0) = \psi'(\pm1) = 0.$$

Графики этих функций приведены на рис. 5.5, а локальная погрешность аппроксимации многочленом некоторой функции

$$w(x) \in C^4[x_i, x_{i+1}],$$

$$w(x_i) = v_i, \quad w'(x_i) = v'_{i+1}, \quad w(x_{i+1}) = v_{i+1}, \quad w'(x_{i+1}) = v'_{i+1}$$

равна

$$z(x) = w(x) - v^h(x) = \frac{w^{(4)}(\eta)}{4!}(x-x_i)^2(x-x_{i+1})^2, \quad \eta \in [x_i, x_{i+1}].$$

Локальные матрицы жесткости и масс (четвертого порядка) в данном случае, при кусочно-постоянных коэффициентах и упорядоченности неизвестных на элементе Ω_i^h в последовательности v_i, v_i', v_{i+1}, v_{i+1}', имеют следующий вид:

$$\bar{A}^h_{s,i} = \frac{d_{i+1/2}}{30h_i}\begin{bmatrix} 36 & 3h & -36 & 3h \\ \cdot & 4h^2 & -3h & -h \\ \cdot & \cdot & 36 & -3h \\ \cdot & \cdot & \cdot & 4h^2 \end{bmatrix},$$

$$\bar{A}^h_{m,i} = \frac{h_i c_{i+1/2}}{420}\begin{bmatrix} 156 & 22h & 54 & -13h \\ \cdot & 4h^2 & 13h & -3h^2 \\ \cdot & \cdot & 156 & -22h \\ \cdot & \cdot & \cdot & 4h^2 \end{bmatrix}. \qquad (5.147)$$

Рис. 5.5. Кубические эрмитовые базисные функции

Матрица $\bar{A}^h_{m,i}$ положительно определена, а локальная матрица жесткости $\bar{A}^h_{s,i}$ положительно полуопределена и имеет нулевое собственное значение, соответствующее функции $v^h(x) \equiv 1$, т. е. локальному собственному вектору $(1, 0, 1, 0)^T$.

В случае равномерной сетки с шагом $h_i = h$, постоянных коэффициентов d, c и нулевого конвективного члена ($b(x) = 0$) алгебраические уравнения получаемой системы имеют два характерных вида (их можно назвать уравнениями для искомой функции v_i и ее производной v'_i):

$$\begin{aligned}
&\frac{d}{30\,h}\big(-36\,v_{i-1} - 3\,h\,v'_{i-1} + 72\,v_i - 36\,v_{i+1} + 3\,h\,v'_{i+1}\big) + \\
&\frac{c\,h}{420}\big(54\,v_{i-1} + 13\,h\,v'_{i-1} + 312\,v_i + 54\,v_{i+1} - 13\,h\,v'_{i+1}\big) \\
&\qquad\qquad = f_i^h = \int\limits_\alpha^\beta f\varphi\Big(\frac{x}{h} - i\Big)\,d\,x, \\
&\frac{d}{30}\big(3\,v_{i-1} - h\,v'_{i-1} + 8\,h\,v'_i - 3\,v_{i+1} - h\,v'_{i+1}\big) + \\
&\frac{ch^2}{420}\big(-13\,v_{i-1} - 3\,h\,v'_{i-1} + 8\,h\,v'_i + 13\,v_{i+1} - 3\,h\,v'_{i+1}\big) \\
&\qquad\qquad = f'_i = h\int\limits_\alpha^\beta f\psi\Big(\frac{x}{h} - i\Big)\,d\,x.
\end{aligned} \tag{5.148}$$

Интересно рассмотреть эти уравнения с точки зрения разностной аппроксимации исходного дифференциального уравнения. Предположим, что $d = 1$, $b = c = 0$, $f = 1$, так что решаемое уравнение Пуассона есть $-u'' = 1$. Тогда из (5.141) мы имеем

$$\begin{aligned}
&-\frac{6}{5}\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} + \frac{1}{5}\frac{v'_{i+1} - v'_{i-1}}{2\,h} = 1, \\
&-\frac{1}{5}\frac{v_{i+1} - v_{i-1}}{2\,h} + \frac{1}{5}v'_i - \frac{1}{30}(v'_{i+1} - 2\,v'_i + v'_{i-1}) = 0.
\end{aligned} \tag{5.149}$$

Учет краевых условий в системе (5.142) делается просто. Если заданы условия первого рода, то в уравнениях для $i = 1$ и N вместо v_0 и v_{N+1} подставляются u_α и u_β соответственно. Если же, например, слева задано значение производной u'_α, то для $i = 1$ полагается $v'_0 = u'_\alpha$. Во всех возможных случаях мы получаем $2N+2$ уравнений для такого же числа неизвестных.

Отсюда с помощью тейлоровских разложений получаем, что первое сеточное уравнение аппроксимирует уравнение Пуассона с погрешностью $-\frac{h^2}{15}u^{(4)}$, а второе уравнение асимптотически дает ошибку $-\frac{h^4}{1230}u^{(5)}$.

Рассмотренный подход формально без труда переносится на эрмитовые базисные функции более высоких порядков, однако он связан с такой неприятностью, как повышение порядка локальных матриц МКЭ и усложнением структуры (с уменьшением степени разреженности) глобальной матрицы алгебраической системы уравнений.

Эффективные методы приближенных вычислений основаны, как правило, на последовательном уточнении искомых решений. При этом, если полученная на каком-то этапе погрешность представляется слишком большой, было бы желательно повысить точность за счет дополнительных незначительных вычислений.

Рассмотренные выше системы базисных функций высоких степеней такой возможностью, к сожалению, не обладают. Если, например, после использования лагранжевой интерполяции n-го порядка перейти к базисным функциям степени $n+1$, то все элементы матриц МКЭ приходится рассчитывать заново.

В этом плане более предпочтительными являются *иерархические базисы*, активно развиваемые в последние десятилетия И. Бабушкой и его последователями, см. [54]. Главные свойство и достоинство таких базисов заключаются в том, что при увеличении порядка аппроксимирующего пространства старые базисные функции остаются неизменными и к ним только добавляются новые.

Описание таких базисов проведем на *стандартном элементе*

$$\Omega_{st} = \{-1 < \xi < 1\},$$

который отображается в произвольный интервал (x_i, x_{i+1}) с помощью формулы

$$x = Q_i(\xi) \equiv \frac{1-\xi}{2} x_i + \frac{1+\xi}{2} x_{i+1}, \qquad \xi \in \Omega_{st}. \tag{5.150}$$

При этом обратное отображение имеет вид

$$\xi = Q_i^{-1}(x) = \frac{2x - x_i - x_{i+1}}{x_{i+1} - x_i}. \tag{5.151}$$

Базис p-го порядка включает $p+1$ функцию:

$$\varphi_1(\xi) = \frac{1-\xi}{2}, \qquad \varphi_2(\xi) = \frac{1+\xi}{2}, \qquad \varphi_l(\xi) = \psi_{l-1}(\xi),$$
$$l = 3, \ldots, p+1, \tag{5.152}$$

где вспомогательная функция $\psi_l(\xi)$ определяется через *многочлен Лежандра* P_{l-1}:

$$\psi_l(\xi) = \sqrt{\frac{2l-1}{2}} \int_1^{\xi} P_{l-1}(t)\,dt, \qquad l = 2, 3, \ldots \tag{5.153}$$

Напомним, что многочлены Лежандра ортогональны на рассматриваемом интервале

$$\int_{-1}^{1} P_i(x) P_j(x)\,dx = \frac{2}{2\,i+1}\delta_{i,j}$$

и могут быть вычислены из трехчленной рекуррентной формулы

$$(n+1)P_{n+1}(x) = (2\,n+1)xP_n(x) - n\,P_{n-1}(x), \qquad n = 1, 2, \ldots,$$

где первые многочлены имеют вид (см. [30])

$$P_0(x) = 1, \qquad P_1(x) = x, \qquad P_2(x) = \frac{1}{2}(3\,x^2 - 1).$$

Многочлены Лежандра также удовлетворяют следующим соотношениям:

$$(2n+1)P_n(x) = P'_{n+1}(x) - P'_{n-1}(x), \qquad n = 1, 2, \ldots$$

Все корни x_i многочлена Лежандра $P_n(x)$ находятся в интервале $(-1, 1)$ и являются абсциссами n-точечной гауссовской квадратурной формулы

$$\int_{-1}^{1} f\,dx \approx \sum_{i=1}^{n} w_i f(x_i).$$

Базисные функции φ_1, φ_2 называются узловыми (они обладают "лагранжевыми" свойствами $\varphi_1(-1) = 1$, $\varphi_1(1) = 0$, $\varphi_2(1) = 1$, $\varphi_2(-1) = 0$), а остальные — внутренними.

На основе свойств многочленов Лежандра показывается, что справедливо соотношение

$$\psi_l(\xi) = \frac{1}{\sqrt{2(2l-1)}}\big(P_l(\xi) - P_{l-2}(\xi)\big), \tag{5.154}$$

т. е. функция $\psi_l(\xi)$ является многочленом степени l. Легко также убедиться в том, что

$$\begin{aligned} &\psi_i(-1) = \psi_i(1) = 0, \qquad i = 2, 3, \ldots, \\ &\int\limits_{-1}^{+1} \frac{d\varphi_k}{d\xi}\frac{d\psi_i}{d\xi} = 0, \quad k = 1, 2; \qquad \int\limits_{-1}^{1} \frac{d\psi_i}{d\xi}\frac{d\psi_j}{d\xi}d\xi = \delta_{i,j}. \end{aligned} \tag{5.155}$$

В силу последних свойств локальные матрицы жесткости при постоянном коэффициенте $d(x)$ на интервале (x_i, x_{i+1}) вычисляются особенно легко, так как нахождение их диагональных коэффициентов не требует даже интегрирования:

$$\bar{A}^h_{s,i} = \frac{2d_{i+1/2}}{h_i} \begin{bmatrix} 1/2 & -1/2 & 0 & \cdot & 0 \\ -1/2 & 1/2 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdot & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}. \tag{5.156}$$

В таком случае система уравнений МКЭ для базиса сколь угодно высокого порядка записывается особенно просто, если

в (5.117) $b(x) = c(x) = 0$. Пусть на каждом сеточном отрезке $[x_i, x_{i-1}]$ используется $p_i + 1$ базисная функция вида (5.145), а приближенное решение записывается в виде

$$u^h(x) = v_i\varphi_1^{(i)}(x) + v_{i+1}\varphi_2^{(i)}(x) + \sum_{l=3}^{p_i+1} v_l^{(i)}\varphi_l^{(i)}(x).$$

В силу ортогональных свойств (5.148) производных от "внутренних" базисных функций $\varphi_l^{(i)}(x)$, $l \geq 3$, соответствующие неизвестные $v_l^{(i)}$ определяются независимо от остальных из соотношений

$$v_l^{(i)} = \big(f, \varphi_l^{(i)}\big)/a\big(\varphi_l^{(i)}, \varphi_l^{(i)}\big).$$

Каждая же пара узловых функций $\varphi_2^{(i-1)}(x)$, $\varphi_1^{(i)}(x)$ образует одну базисную функцию $\varphi_i(x)$ из (5.120) для кусочно-линейной интерполяции, и в итоге для неизвестных v_i получается трехточечная система уравнений вида (5.132) с учетом $b_{i\pm1/2} = c_{i\pm1/2} = 0$.

Если же функция $d(x)$ не постоянна, то локальная матрица жесткости $\bar{A}_{s,i}^h$ имеет уже не простой вид (5.149), а является плотной. Но поскольку "внутренние" базисные функции $\varphi_l^{(i)}(x)$ отличны от нуля только на одном сеточном интервале, как и в лагранжевых элементах высших порядков, то получаемые алгебраические системы имеют схожую алгебраическую структуру. К ним, в частности, также может быть эффективно применена процедура конденсации, т. е. исключения "внутренних" неизвестных $v_l^{(i)}$ и получение трехточечной системы для узловых переменных v_i.

5.4.3. Общие замечания о построении матриц МКЭ. Резюмируя проведенные выше рассмотрения, можно сказать,

что конкретный метод конечных элементов для решения поставленной краевой задачи полностью определяется способом разбиения расчетной области на конечные элементы Ω_k^h видом искомой аппроксимации искомого решения и совокупностью функционалов, характеризующих параметры выбранного приближения. Совокупность этих факторов и порождает конечномерное пространство функций V^h.

Очевидно, что все МКЭ можно разделить на однородные и неоднородные по степени "одинаковости" алгоритмов в области. К однородным отнесем методы с одинаковой формой конечных элементов (например, треугольники), с одинаковой на каждом элементе полиномиальной формой интерполянта и с однотипным набором функционалов (значений функций и/или их производных в заданном количестве точек). В принципе же представления приближенных решений на каждом элементе могут быть никак не связаны между собой, за исключением, возможно, условий непрерывности на смежных ребрах.

Последнее означает естественное требование, что определяемые на ребрах искомые величины должны быть общими для смежных элементарных ячеек, то же самое относится и к узловым неизвестным.

Остановимся теперь на некоторых конструктивных вопросах формирования локальных матриц МКЭ. Пусть $\mathcal{P}_k$ означает конечномерное подпространство (размерности N_k) функций, определенных на элементе Ω_k^h. Обозначим через $Q_k = \{q_1, \ldots, q_{N_k}\}$ базис в пространстве функционалов $\mathcal{P}_k'$, где N_k — число степеней свободы конечного элемента Ω_k^h. Пусть каж-

дая величина $q_l^{(k)}$ соответствует значению какой-то производной $D_j u(P_i)$ порядка $|D_j|$ в точке $P_i \in \bar{\Omega}_k^h$, где $u \in \mathbb{P}_k$, $i = 1, \ldots, M_k$, $j = 0, \ldots, \alpha_i^{(k)} - 1$, $\alpha_i^{(k)}$ — кратность каждого узла P_i, а M_k — их количество в $\bar{\Omega}_k^h$ (при этом каждый номер l соответствует паре индексов i, j, и мы можем записать $q_l^{(k)} = q_{i,j}^{(k)}$). Тогда совокупность линейно независимых функций $\Phi_k = \{\varphi_l^{(k)},\ i = 1, \ldots, N_k\}$, линейная оболочка которых составляет $\mathbb{P}_k$, является двойственным базисом к Q_k, если каждая базисная функция $\varphi_l^{(k)} = \varphi_{i,j}^{(k)}$ соответствует величине $q_l^{(k)} = q_{i,j}^{(k)}$ в смысле удовлетворения условиям $D_j \varphi_{i,j}^{(k)}(P_i) = \delta_{i,j}$. Поскольку элементы локальных матриц МКЭ определяются через скалярные произведения базисных функций, возникает не только теоретическая проблема их существования, но и алгоритмическая задача нахождения $\varphi_l^{(k)}$ по заданным $\mathbb{P}_k$ и Q_k.

В одномерном случае на эти вопросы находятся исчерпывающие ответы с помощью теории интерполирования. Пусть конечный элемент Ω_k^h есть отрезок $[x_k, x_{k+1}]$, на котором задано m_k различных точек $x_i^{(k)}$ и $N_k \geqslant m_k$ линейно независимых функций $\psi_j^{(k)}(x)$. Определим пространство $\mathbb{P}_k$ как совокупность всевозможных линейных комбинаций

$$\psi_k(x) = c_1^{(k)} \psi_1^{(k)} + \ldots + c_{N_k}^{(k)} \psi_{N_k}^{(k)} \tag{5.157}$$

с некоторыми коэффициентами $c_j^{(k)}$.

Теперь будем искать из всевозможных комбинаций (5.150) такую, которая удовлетворяет условиям обобщенной задачи интерполирования, т. е. в каждой из точек $x_i^{(k)}$, $i = 1, \ldots, m_k$,

функция $\Psi_k(x)$ и ее производные до $(\alpha_i^{(k)} - 1)$-го порядка включительно принимают заданные значения (число $\alpha_i^{(k)}$ называется *кратностью интерполяционного узла* $x_i^{(k)}$). Отсюда для N_k коэффициентов $c_j^{(k)}$ получаем систему линейных алгебраических уравнений

$$\psi_k\big(x_i^{(k)}\big) = q_j, \quad \psi_k'\big(x_i^{(k)}\big) = q_j^{(1)}, \ldots, \psi_k^{(\alpha_i^{(k)}-1)}\big(x_i^{(k)}\big) = q_j^{(\alpha_i^{(k)}-1)},$$

$$i = 1, \ldots, m_k, \tag{5.158}$$

где $q_i^{(l)}$ — заданные числа, а порядок системы равен

$$M_k = \alpha_1^{(k)} + \ldots + \alpha_m^{(k)}.$$

Для существования единственного решения системы (5.151) требуется, чтобы выполнялось равенство $N_k = M_k$ и квадратная матрица

$$
B = \begin{bmatrix}
\psi_1(x_1) & \psi_2(x_1) & \ldots & \psi_N(x_1(x_1) \\
\psi_1'(x_1) & \psi_2'(x_1) & \ldots & \psi_N'(x_1) \\
\multicolumn{4}{c}{\dotfill} \\
\psi_1^{(\alpha_1-1)}(x_1) & \psi_2^{(\alpha_1-1)}(x_1) & \ldots & \psi_N^{(\alpha_1-1)}(x_1) \\
\psi_1(x_2) & \psi_2(x_2) & \ldots & \psi_N(x_2) \\
\multicolumn{4}{c}{\dotfill} \\
\psi_1^{(\alpha_m-1)}(x_m) & \psi_2^{(\alpha_m-1)}(x_m) & \ldots & \psi_N^{(\alpha_m-1)}(x_m)
\end{bmatrix} \tag{5.159}
$$

была невырождена (здесь и ниже мы для краткости опускаем индекс k).

Обозначая через $c = \{c_j\}$, $q = \{q_i^{(l)}\}$ векторы порядка N, систему (5.151) запишем в матричном виде

$$
Bc = q. \tag{5.160}
$$

Установим некоторые факты о существовании и единственности решения системы (5.153), см. подробнее [12].

Определение 5.8. *Совокупность функций* $\psi_j(x)$, $j = 1, 2, \ldots, m$, *образует на отрезке* $[a, b]$ *систему Чебышева, если никакая их линейная комбинация*

$$
c_1\psi_1(x) + c_2\psi_2(x) + \ldots + c_m\psi_m(x)
$$

не обращается в ноль в точках $x_j \in [a,b]$, $j = 1,2,\ldots,m$, *т. е. не имеет на* $[a,b]$ m *различных корней.*

Теорема 5.3. *Если функции* $\psi_1(x), \psi_2(x), \ldots, \psi_m(x)$ m *раз дифференцируемы на отрезке* $[a,b]$ *и вронскиан*

$$W[\psi_1,\ldots,\psi_k] = \begin{vmatrix} \psi_1(x) & \psi_2(x) & \ldots & \psi_k(x) \\ \psi_1'(x) & \psi_2'(x) & \ldots & \psi_k'(x) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \psi_1^{(k)}(x) & \psi_2^{(k)}(x) & \ldots & \psi_k^{(k)}(x) \end{vmatrix}$$

не равен нулю на $[a,b]$ *при всех* $k = 1,\ldots,m$, *то эти функции образуют систему Чебышева.*

Подчеркнем , для того, чтобы функции $\psi_j(x)$ образовывали систему Чебышева, их линейная независимость является необходимой, но не достаточной. Простейшую систему Чебышева составляют мономы $\psi_j(x) = x^j$, $j = 0,1,\ldots$ Из теоремы 5.3 следует, в частности, существование и единственность интерполяционного многочлена Лагранжа, соответствующего случаю $\alpha_1 = \ldots = \alpha_m = 1$, когда все узлы x_j являются некратными, т. е. простыми.

В общем случае с кратными узлами и степенными функциями $\psi_j(x) = x^j$ система уравнений (5.151) определяет единственным образом *интерполяционный многочлен Эрмита* порядка $N_k - 1$. Для его построения необходимо найти вектор c из решения системы (5.153):

$$c = B^{-1}q.$$

Пусть теперь требуется найти коэффициенты $b_l^{(i)}$, $l = 1, \ldots, N$, многочленного представления базисной функции МКЭ:

$$\varphi_i(x) = b_1^{(i)} + b_2^{(i)}x + \ldots + b_N^{(i)}x^N, \tag{5.161}$$

соответствующей решению системы (5.153) с правой частью $q = e_i$ (e_i — вектор-орт с единственной ненулевой i-й компонентой, равной единице).

Очевидно, что вектор $b^{(i)} = \{b_l^{(i)}\}$ есть просто i-й столбец обратной матрицы $B^{-1} = \{b^{(i)},\ i = 1, \ldots, N\}$. Таким образом, однократное обращение матрицы B определяет все базисные функции подпространства $\mathcal{P}_k$, а локальные матрицы МКЭ находятся путем вычисления соответствующих билинейных форм:

$$a(\varphi_i, \varphi_j) = a(b_1^{(i)} + \ldots + b_N^{(i)}x^N, b_1^{(j)} + b_2^{(j)}x + \ldots + b_N^{(j)}x^N).$$

§ 5.5. Разрывные методы Галеркина

Рассматриваемые в § 4.15 вариационные принципы решения задач Коши для систем ОДУ, реализуемые с помощью разрывных методов Галеркина различных порядков точности, могут также успешно применяться и для нахождения численных решений краевых задач. Достаточно содержательный сравнительный анализ работ с различными подходами за 30 лет приведен в обзорной работе Д. Арнольда и Ф. Брецци с соавторами (D. Arnold, F. Brezzi, B. Cockburn, L. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal., v. 39, N 5, 2002, 1749-1779).

Однако мы в данном параграфе ограничимся изложением только одного аппроксимационного принципа: иммерсионных (от английского immersed — дословно “погруженный”) разрывных методов Галеркина РМГ(k) с базисными функциями высоких порядков k, которые строятся с учетом свойств решений дифференциальных уравнений, имеющих разрывные коэффициенты.

Такие алгоритмы, во-первых, могут строиться не на адаптивных сетках, привязанных жестко к геометрическим и материальным свойствам исходной задачи (точки разрыва коэффициентов не обязаны быть узлами сетки). Во-вторых, в данном случае строятся легко аппроксимации с переменными не только шагами, но и с меняющимися порядками на различных сеточных интервалах. Мы будем следовать в основном методике работы С. Аджерида и Т. Лина (S. Adjerid, T. Lin. High-order Immersed Discontinuous Galerkin Methods. Internat. J. of

information and system siences, N 3, 2007, 555-568).

Итак, пусть $u(x)$ есть решение краевой задачи Дирихле для уравнения диффузии

$$\begin{aligned} &(p\,u')' = f(x), \quad a < x < b, \\ &u(a) = u_a, \quad u(b) = u_b. \end{aligned} \tag{5.162}$$

Без ограничения общности будем считать, что расчетная область $\Omega = (a, b)$ разбита на две подобласти — (a, α) и (α, b) такие, что в точке $\alpha \in (a, b)$ диффузионный коэффициент $p(x)$ претерпевает скачок:

$$p(x) = \begin{cases} p^-, & x \in (a, \alpha), \\ p^+, & x \in (\alpha, b), \end{cases} \tag{5.163}$$

причем в интерфейсной точке α выполняются условия сопряжения

$$[u](\alpha) \equiv u(\alpha^+) - u(\alpha^-) = 0, \quad [p\,u'](\alpha) = 0. \tag{5.164}$$

Результаты, излагаемые ниже, в дальнейшем без труда могут быть перенесены на случай, когда в Ω имеется несколько точек со скачками коэффициента $p(x)$.

Применение РМГ к решению дифференциальных уравнений второго порядка заключается, прежде всего, в переходе к системе уравнений первого порядка, т. е. к так называемой смешанной постановке:

$$q' = f(x), \quad a < x < b,$$
$$p\,u' = q(x). \tag{5.165}$$

Введем теперь в расчетной области Ω сетку Ω^h, определяя на (a, b) множество узлов

$$a = x_0 < x_1 < ... < x_N < x_{N+1} = b,$$

без какой-либо их привязки, или локализации, по отношению к интерфейсной точке α. Далее, умножая первое и второе уравнения (5.158) на некоторые пробные функции $v(x)$ и $w(x)$ соответственно и применяя интегрирование по частям на интервалах

$$I_i = [x_i, x_{i+1}], \quad i = 0, 1, ..., N,$$

получаем так называемую локальную слабую постановку:

$$-p\,v\big|_{x_i}^{x_{i+1}} - (p, v')_{I_i} = (f, v)_{I_i} \equiv \int\limits_{x_i}^{x_{i+1}} f\,v\,d\,x,$$

$$p\,u\,w\big|_{x_i}^{x_{i+1}} - [p]u\,w(\alpha) - (p\,u, w')\big|_{I_i} = (q, w)\big|_{I_i}, \ \ \alpha \in I_i,$$

$$p\,u\,w\big|_{x_i}^{x_{i+1}} - (p\,u, w')\big|_{I_i} = (q, w)\big|_{I_i}, \ \ \alpha \notin I_i. \tag{5.166}$$

Отсюда получаем *иммерсионный локальный разрывный метод Галеркина k-го порядка* ИЛРМГ(к), если для аппроксимации $q(x)$ определим пространство разрывных кусочно-гладких полиномиальных функций

$$S^{N,k} = \{V | V|_{I_i} \in \mathcal{P}^{(k)}, \quad i = 0, 1, ..., N\},$$

где $\mathcal{P}^{(k)}$ есть пространство многочленов степени не выше k, а для аппроксимации $u(x)$ введем пространство *иммерсионных конечных элементов* $S_I^{N,k}$, которые представляются полиномами из $\mathcal{P}^{(k)}$ на каждом из интервалов I_i, не содержащем интерфейсную точку α, и кусочно-полиномиальными функциями из $\mathcal{P}^{(k)}$ на интервале, включающем точку α. Построение таких функций мы рассмотрим несколько позже.

Отсюда ИЛРМГ(к) заключается в нахождении пары функций

$$(u^h, q^h) \in X^{N,k} = S_I^{N,k} \times S^{N,k}$$

такой, что для всех $(V, W) \in X^{N,k}$ выполняются равенства

$$\hat{q}^h v_h \big|_{x_i}^{x_{i+1}} + (q^h v_h')\big|_{I_i} = (f, v_h)_{I_i},$$

$$p\,\hat{u}^h w_h \big|_{x_i}^{x_{i+1}} - (p\,u, w_h')\big|_{I_i} = (q^h, w_h)\big|_{I_i}, \quad \alpha \notin I_i,$$

$$p\,\hat{u}^h w_h \big|_{x_i}^{x_{i+1}} - [p](\alpha)(u^h w_h)(\alpha) - (p\,u^h, w_h')_{I_i} = (q^h, w_h)_{I_i}, \quad \alpha \in I_i. \tag{5.167}$$

Здесь величины $\hat{u}^h$ и $\hat{q}^h$ означают следы функций $u^h(x)$ и $q^h(x)$ в узлах сетки. В точках неоднозначности этих разрывных функций следы могут быть определены, в принципе, различным образом, и в рассматриваемых нами методах используется их следующая конкретизация:

$$\hat{u}^h(x_i) = u^h(x_i^-), \quad \hat{q}^h(x_i) = q^h(x_i^+), \quad i = 1, ..., N. \tag{5.168}$$

При этом на границах расчетной области для следов получаются уравнения

$$\hat{u}^h(a) = u(a), \quad \hat{q}^h(a) = q^h(a^+),$$
$$\hat{u}(b) = u(b), \quad \hat{q}(b) = q^h(b^-) - \frac{\gamma\, k}{h}(u(b) - u^h(b^-)) \tag{5.169}$$

Таким образом, получаем систему сеточных алгебраических уравнений порядка $2(N+2)$ для неизвестных $\hat{u}_i^h = \hat{u}^h(x_i)$, $\hat{q}_i^h = \hat{q}^h(x_i)$, $i = 0, 1, ..., N+1$. Отметим, что последний член в (5.162) с коэффициентом γ введен, с формальной точки зрения, несколько искусственно. В дальнейшем полагается $\gamma = 1$, а введение этого члена связано с идеологией множителей Лагранжа в методах штрафа, см. подробнее упомянутую выше обзорную работу Д. Арнольда с соавторами, а также приведенную в ней обширную специальную литературу. В данной статье дается сравнительный анализ и систематизация существующих подходов к построению разрывных методов Галеркина, основанных фактически на различных определениях следов функций $\hat{u}^h$ и $\hat{q}^h$, отличных от (5.161), (5.162).

Как видно из формулировки РМГ (5.160) – (5.162), данный алгоритм является обобщением методов конечных объемов с возможностями повышения порядков аппроксимации. С другой стороны, он является развитием смешанных методов конечных элементов.

Приведем теперь формулы, необходимые для вычисления интегралов в вариационных уравнениях (5.160), которые выводятся на основе представления базисных функций в “обыч-

ных" конечно-элементных пространствах $S^{N,k}$ и в ИКЭ – "интерфейсных" конечно-элементных простанствах $S_I^{N,k}$. В силу их разрывного характера, нам достаточно привести выражения на одном "стандартном" интервале $\hat{I} = (-1, 1)$, полученном после аффинных преобразований каждого из I_i. На интервале $\hat{I}$ иерархические "функции формы", как их иногда называют, имеют следующий вид:

$$\bar{\varphi}_0(t) = (1-t)/2, \quad \bar{\varphi}_1(t) = (1+t)/2, \quad t \in \bar{I},$$
$$\bar{\varphi}_j(t) = c_j(P_j(t) - P_{j-2}(t)), \quad j = 2, ..., k, \tag{5.170}$$

где коэффициенты c_j определяются по условиям нормировки. Таким образом мы имеем пространства

$$\bar{\mathcal{P}}^{(k)} = \text{span}\{\bar{\varphi}_j(t), \quad j = 0, 1, ..., k\}.$$

с помощью которых получаем локальные пространства $\bar{S}_I^{N,k}$ и $\bar{S}^{N,k}$ на интерфейсных и неинтерфейсных сеточных интервалах соответственно. Мы приведем базисные функции только для $\bar{S}_I^{N,k}$, которые точно удовлетворяют условиям сопряжения в точке $\bar{\alpha} \in (-1, 1)$, а гладкие функции из $\bar{S}^{N,k}$ различных порядков, называемые также иерархическим базисом Лобатто, получаются из кусочно-гладких как частный случай.

Две линейные функции ИКЭ-базиса имеют следующий вид ($i = 0, 1$):

$$\bar{\varphi}_i(t) = \begin{cases} a_{i,1}t + b_{i,1}, & -1 \leqslant t \leqslant \bar{\alpha}, \\ a_{i,2}t + b_{i,2}, & \bar{\alpha} \leqslant t \leqslant 1, \end{cases} \tag{5.171}$$

где восемь неизвестных коэффициентов определяются однозначно из такого же количества уравнений

$$\bar{\varphi}_0(-1) = 1, \quad \bar{\varphi}_0(1) = 0, \quad \bar{\varphi}_{-1}(1) = 0, \quad \bar{\varphi}_1(1) = 1,$$

$$[\bar{\varphi}_i](\bar{\alpha}) = 0, \quad [p\bar{\varphi}_i'](\bar{\alpha}) = 0, \quad i = 0, 1.$$

Квадратичная интерфейсная конечно-элементная базисная функция описывается формулой

$$\bar{\varphi}_2(t) = \begin{cases} (1+t)(a_{2,1}t + b_{2,1}), & -1 \leqslant t \leqslant \bar{\alpha}, \\ (t-1)(a_{2,2}t + b_{2,2}), & \bar{\alpha} \leqslant t \leqslant 1, \end{cases}$$

в которой четыре коэффициента находятся из трех условий сопряжения в интерфейсной точке $\bar{\alpha}$:

$$[\bar{\varphi}_2](\bar{\alpha}) = 0, \quad [p\bar{\varphi}_2'](\bar{\alpha}) = 0, \quad [p\bar{\varphi}_2''](\bar{\alpha}) = 0,$$

а также из условия нормировки $a_{2,2} = 1$. Заметим, что при этом выполняется условие ортогональности в смысле

$$< \bar{\varphi}_2, \bar{\varphi}_1 >_{\hat{I}} \equiv \int\limits_{-1}^{\bar{\alpha}} p^- \varphi_2' \varphi_1' \, d\,x + \int\limits_{\bar{\alpha}}^{1} p^+ \varphi_2' \varphi_1' \, d\,x = 0.$$

ИКЭ-функции формы k-х порядков для $k \geqslant 3$ имеют вид

$$\bar{\varphi}_k(t) = \begin{cases} (1+t) \sum\limits_{j=0}^{k-1} a_{k,j}^{(1)} t^j, & -1 \leqslant t \leqslant \bar{\alpha}, \\ (t-1) \sum\limits_{j=0}^{k-1} a_{k,j}^{(2)} t^j, & \bar{\alpha} \leqslant t \leqslant 1. \end{cases}$$

Здесь $2k$ неизвестных коэффициента определяются однозначно из $2k-1$ уравнения

$$[\varphi_k](\bar{\alpha})=0,\ \ [p\varphi_k^{(j)}](\bar{\alpha})=0,\ \ j=1,...,k+1,$$

$$<\bar{\varphi}_k,\bar{\varphi}_j>_{\hat{I}}=0,\ \ j=2,3,...,k-1,$$

а также из какого-либо условия нормировки, например, $a_{k,2}^{(2)}=1$.

Локальные пространства $S_I^{N,k}$ на интерфейсном элементе формируются с помощью пространств

$$\bar{\mathcal{P}}^{(k)}=\mathrm{span}\{\bar{\varphi}_i,\ \ i=0,1,...,k\}.$$

Если $\bar{\alpha}=\pm 1$ или если диффузионный коэффициент p непрерывен в интерфейсной точке $\bar{\alpha}$, то данные функции переходят в определяемые формулами (5.163) иерархические многочлены Лобатто см.§ 4.15. Отметим также, что при переходе от t к исходной переменной x иммерсионные базисные функции удовлетворяют условиям $[p\varphi^{(j)}](\alpha)=0$ для $j\geqslant 1$.

Примеры иммерсионных (для $\bar{\alpha}=-0.5$) и стандартных базисных функций первого — третьего порядков

Рис. 5.6. Иммерионные и стандартные базисные функции 1-го — 3-го порядков

В заключение данного параграфа отметим, что разрывные методы Галеркина в настоящее время представляют собой активно развиваемое направление, перспективное с точки

зрения параллельных вычислительных технологий. Для различных типов задач показано, как теоретически, так и экспериментально, высокая точность для численных решений и потоков:

$$||u - u^h|| = O(h^{k+2}), \;\; ||q - q^h|| = O(h^{k+1}).$$

Применение РМГ динамически расширяется для все новых классов задач, в том числе сингулярных, многомерных, нелинейных, нестационарных и т. д. Активное теоретическое и прикладное развитие разрывных методов Галеркина отражено в многочисленных журнальных статьях, но пока еще ждет своего обобщения и систематизации в монографической литературе.

§ 5.6. Алгебраические свойства сеточных уравнений и оценки погрешности.

Изучив построение сеточных уравнений и приступая к исследованию их свойств или методов решения, мы предварительно уделим внимание установлению алгебраической терминологии и ознакомимся с основными положениями теории матриц (см., например, [18], [28], [61]). В силу происхождения исследуемых систем уравнений все рассматриваемые матрицы и векторы предполагаем вещественными.

Понятие сеточного оператора мы уже применяли в предыдущих главах, не давая его формального определения. Очевидно, что под этим подразумевается преобразование одной сеточной функции в другую, причем области их определения,

т. е. множества узлов сетки, могут отличаться. Поскольку сеточная функция является совокупностью значений (чисел) в узлах сетки, то на алгебраическом языке она представляет собой вектор с размерностью, равной количеству узлов. При конечных величинах шагов сетки и ограниченных расчетных областях (а именно такие случаи мы и рассматриваем) множество определенных на сетке векторов образует конечномерное пространство. А так как любое линейное преобразование в конечномерном пространстве представляется матрицей, то понятия линейного разностного оператора и матрицы мы считаем эквивалентными (с точностью до упорядочения узлов и соответствующих значений сеточных функций). И поэтому, естественно, анализ сеточных уравнений будем проводить с помощью классического аппарата линейной алгебры, не затемняя существа дела новой терминологией и специальной техникой.

5.6.1. Структурные свойства матриц. Рассмотрим систему трехточечных разностных уравнений вида

$$\begin{aligned} &-a_i v_{i-1} + b_i v_i - c_i v_{i+1} = f_i, \;\; i = 1, ..., N, \\ &a_1 = c_N = 0, \end{aligned} \tag{5.172}$$

аппроксимирующих дифференциальное уравнение второго порядка на сетке с N внутренними узлами, дополненную граничными равенствами $v_0 = g_a, v_{N+1} = g_b$, соответствующими краевым условиям Дирихле. Вводя векторы-столбцы из $N+2$ компонент

$$\bar{v} = (v_o, v_1, \ldots, v_N, v_{N+1})^T, \qquad \bar{f} = (f_0,\ f_1, \ldots, f_N, f_{N+1})^T,$$

где в данном случае $f_o = u_a$, $f_{N+1} = u_b$, а символ "T" здесь и далее означает транспонирование, такую систему можно записать в форме

$$\bar{A}\bar{u} = \bar{f} \tag{5.173}$$

с трехдиагональной матрицей

$$\bar{A} = \begin{bmatrix} 1 & 0 & 0 & & & & \\ -a_1 & b_1 & -c_1 & & & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & -a_i & b_i & -c_i & & \\ & & & \ddots & \ddots & \ddots & \\ 0 & & & & -a_N & b_N & -c_N \\ & & & & 0 & 0 & 1 \end{bmatrix}. \tag{5.174}$$

В силу тривиальности первого и последнего уравнений переменные v_0, v_{N+1} из системы (5.166) можно исключить, после чего получаем новую систему:

$$
\begin{aligned}
Av = &\begin{bmatrix}
b_1 & -c_1 & & & \\
-a_2 & -b_2 & -c_2 & & \\
& \ddots & \ddots & \ddots & \\
0 & & -a_{N-1} & b_{N-1} & -c_{N-1} \\
& & & -a_N & b_N
\end{bmatrix}
\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{N-1} \\ v_N \end{bmatrix} \\
= f = &\begin{bmatrix}
f_1 & +a_1 u_a \\
f_2 & \\
\vdots & \\
f_{N-1} & \\
f_N & +c_N u_b
\end{bmatrix}
\end{aligned}
\tag{5.175}
$$

теперь уже N-го порядка, равного числу внутренних узлов. Матрица $\bar{A}$ в определенном смысле "хуже" A, так как она, например, заведомо несимметрична. И в дальнейшем мы всегда будем рассматривать системы, получаемые после исключения граничных условий 1-го рода.

Сейчас мы дадим несколько определений матричных свойств (неразложимость, диагональное преобладание, монотонность и др.), которые уже употреблялись нами при рас-

смотрении свойств сеточных СЛАУ в п. 5.2.2.

Определение 5.9. *Матрица A называется разложимой, если одноименными перестановками строк и столбцов она сводится к блочно-треугольному виду*

$$\tilde{A} = PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

с квадратными матрицами A_{11}, A_{22}; в противном случае она называется неразложимой.

Здесь P обозначает *матрицу перестановок*, имеющую на каждой строке и в каждом столбце только по одному ненулевому элементу, равному единице (напомним, что такая матрица является ортогональной, т е. $P^T = P^{-1}$).

Введенное понятие имеет тот простой смысл, что система уравнений с разложимой матрицей приводится к блочному виду

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2, \end{bmatrix},$$

что эквивалентно решению подсистем $A_{22}v_2 = f_2$, $A_{11}v_1 = f_1 - A_{12}v_2$, свойства которых могут изучаться отдельно. Таким образом, исследование неразложимых систем фактически не является ограничением общности, и именно на таких задачах мы и будем останавливаться в дальнейшем.

Определение 5.10. *Матрица $A = \{a_{k,l}\}$ обладает свойством диагонального преобладания, если*

$$\Delta_k = |a_{k,k}| - \sum_{\substack{l=1 \\ l\neq k}}^{N} |a_{k,l}| \geqslant 0, \qquad k = 1, 2, \ldots, N, \tag{5.176}$$

причем хотя бы для одного k неравенство является строгим.

Замечание 5.4. Данное свойство можно было бы назвать более конкретно — *диагональное преобладание по строкам*, так как аналогично можно определить *диагональное преобладание по столбцам*,, изменив только индекс суммирования в (5.169). Кроме введенного используется понятие *строгого диагонального преобладания* (гораздо реже встречающегося в разностных методах), когда модуль каждого диагонального элемента строго больше суммы модулей остальных элементов своей строки (или столбца).

Для трехдиагональной матрицы свойство диагонального преобладания выглядит особенно просто:

$$|b_i| \geqslant |a_i| + |c_i|, \qquad i = 1, 2, \ldots, N,$$

(здесь предполагается $a_1 = c_N = 0$, причем хотя бы для одного i неравенство должно быть строгим), а неразложимость означает отсутствие в ней нулевых коэффициентов a_i, c_i. Если они определяются в разностных уравнениях формулами из параграфа 5.2, то оба эти свойства выполняются при $p_{i\pm1/2}$, $r_i \geqslant 0$ и достаточно малом шаге сетки h (точнее, при $q_i < p_{i+1/2}/h$).

Две только что введенные матричные характеристики позволяют сформулировать результат, весьма актуальный для разностных уравнений вследствие его широкой и очень простой применимости.

Теорема 5.4 (Тауски). *Если матрица неразложима и обладает свойством диагонального преобладания, то она невырождена.*

Остановимся теперь на такой важной характеристике разностных уравнений, как симметричность. Понятие *симметричности системы уравнений* обычно связывается с ее матрицей. Однако ее можно ввести также и для пары любой совокупности уравнений как "равенство коэффициента при k-й переменной в i-м уравнении коэффициенту при i-й переменной в k-м уравнении" (аналогично можно определить и симметричность разностных выражений). Для совокупности трехточечных уравнений вида (5.155) это означает выполнение равенств $a_i = c_{i-1}$.

Очевидно, что для разностных уравнений Пуассона и даже уравнения диффузии с переменным коэффициентом условия симметрии выполняются, если только шаг сетки постоянный.

Если же величины h_i различны, то оказывается, что эти системы уравнений несимметричны, но относятся к другому важному классу — *симметризуемым системам уравнений* . Последнее означает, что они становятся симметричными после того, как каждое из них умножается на некоторое число (масштабирующий множитель). Как видно из (5.23), для разностного уравнения диффузии таким множителем является сумма $h_i + h_{i-1}$, поскольку

$$\tilde{a}_i = a_i(h_i + h_{i-1}) = \frac{2p_{i-1/2}}{h_{i-1}} = \tilde{c}_{i-1}(h_{i-1} + h_{i-2}).$$

Однако разностные диффузионно-конвективные уравнения (при любых рассмотренных выше или других способах

аппроксимации), в отличие от предыдущих, не являются симметричными даже в случае равномерной сетки и постоянства коэффициентов $p(x)$ и $q(x)$. Не симметризуются они и с помощью введения масштабирующего множителя $(h_i + h_{i-1})$.

Тем не менее системы трехчленных уравнений — и в этом их отличительное качество — могут быть симметризованы с помощью умножения каждого i-го уравнения на *масштабирующий множитель* $\varkappa_i$, если только разрешимы рекуррентные соотношения $\varkappa_i = \varkappa_{i-1}c_{i-1}/a_i$, которые следуют из условий симметрии отмасштабированных уравнений $\tilde{a}_i = \varkappa_i a_i = \tilde{c}_i = \varkappa_{i-1}c_{i-1}$.

Такая *симметризация* возможна (и только!) для *якобиевых систем трехчленных уравнений*, а также соответствующих матриц,, определяемых свойством $a_i c_{i-1} > 0$.

Для этих матриц существует такое преобразование подобия с диагональной матрицей D, что подобная матрица $\hat{A} = D^{-1}AD$ является симметричной. Элементы диагональной матрицы легко находятся из рекуррентных соотношений:

$$\begin{aligned} &\hat{A} = D^{-1}AD = \hat{A}^T, \qquad D = \operatorname{diag}\{d_i\}, \\ &d_{i+1} = d_i(a_{i+1}/c_i)^{1/2}, \qquad i = 1, \ldots, N-1, \end{aligned} \tag{5.177}$$

(величину d_1 можно положить равной единице). Так как подобные матрицы имеют одинаковые собственные числа, а симметричные обладают вещественным спектром, то последнее свойство относится и к якобиевым матрицам.

Аналогично введенному нами ранее определению монотонных систем уравнений вводится следующее понятие.

Определение 5.11. *Матрица A называется монотонной, если она невырождена, а обратная к ней матрица неотрицательна ($A^{-1} \geqslant 0$, где неравенство понимается поэлементно).*

Связь с определением монотонной СЛАУ легко усматривается из того, что если система $Av = f$ с монотонной матрицей имеет неотрицательный вектор правой части, то вследствие равенства $v = A^{-1}f$ решение v также является неотрицательным. Построение монотонных разностных схем (у которых матрицы систем монотонны) особенно естественно в тех случаях, когда исходная задача из физических соображений не может в принципе иметь отрицательных решений — например, функция плотности какой-либо субстанции: при неотрицательной правой части и численное решение будет неотрицательным.

Важным подклассом монотонных матриц является следующий.

Определение 5.12. *Матрица называется матрицей положительного типа, ,если она неразложима, обладает свойством диагонального преобладания и имеет положительные диагональные и неположительные внедиагональные элементы.*

Для трехдиагональных матриц из СЛАУ (5.165) эти условия вырождаются в неравенства $b_i, a_i, c_i > 0$, $b_i \geqslant a_i + c_i$, причем в последнем хотя бы для одного i неравенство строгое.

Доказательство монотонности матрицы положительного типа легко осуществимо с помощью представления $A = D - C$,

$D = \mathrm{diag}\{a_{i,i}\}$, где D и C — неотрицательные матрицы. Поскольку из теоремы Тауски следуют невырожденность A и неравенство для собственных чисел $|\lambda(D^{-1}C)| < 1$, то мы можем записать

$$A^{-1} = (I - D^{-1}C)^{-1}D^{-1} = [I + D^{-1}C + (D^{-1}C)^2 + \ldots]D^{-1}.$$

Так как степенной матричный ряд в квадратных скобках сходится, то неотрицательность матрицы A^{-1} следует из того, что она представима в виде произведения двух неотрицательных матриц.

Определение 5.13. *Матрица $A = \{a_{i,j}\}$ называется M-матрицей, если она монотонна и имеет неположительные внедиагональные элементы, т. е. $a_{i,j} \leqslant 0$ при $i \neq j$.*

Очевидно, что класс матриц положительного типа у́же, чем класс M-матриц, так как у последних необязательно диагональное преобладание. Если сумма матриц положительного типа есть матрица положительного типа, то сумма M-матриц не обязана быть M-матрицей.

Полезно также заметить, что если произведение монотонных матриц есть монотонная матрица, то при перемножении M-матриц или матриц положительного типа их изначальные свойства не сохраняются.

Определение 5.14. *Симметричная M-матрица называется матрицей Стилтъеса.*

Несложно показать, что у M-матрицы все диагональные элементы положительны (поскольку скалярное произведение k-й строки матрицы A на l-й столбец матрицы A^{-1} есть величина $\delta_{k,l}$).

Отметим, что у некоторых типов матриц неположительность внедиагональных элементов является необходимым условием монотонности. К таковым, например, относятся трехдиагональные матрицы с положительными диагональными элементами и свойством диагонального преобладания.

Очевидно, что симметричные матрицы положительного типа являются стилтьесовыми и положительно определенными.

Напомним, что последнее означает положительность для любого вектора скалярного произведения

$$(Av, v) \geqslant \delta(v, v), \qquad \delta > 0, \tag{5.178}$$

а его следствием является положительность собственных чисел матрицы A (которые у произвольной симметричной матрицы являются вещественными). Для указанного типа матриц свойство (5.155) проверяется непосредственно через оценку снизу скалярного произведения (Av, v) суммой (с весовыми множителями) квадратов разностей различных компонент вектора v, которая для матрицы A из (5.152) имеет простой вид:

$$(Av, v) > \sum_{i=2}^{N} a_i (v_i - v_{i-1})^2.$$

В общем случае данное утверждение доказывается с помощью следующего представления скалярного произведения для симметричной матрицы:

$$(Av, v) = \sum_{i=1}^{N} v_i^2 \sum_{j=1}^{N} a_{i,j} - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} a_{i,j} (v_i - v_j)^2.$$

Интересными особенностями обладает матрица системы разностных уравнений экспоненциального типа (5.71) (мы также предполагаем сейчас условия Дирихле). Обозначая ее элементы аналогично предыдущему

$$\begin{aligned} a_i &= \frac{s_i - s_{i-1}}{h^2(e^{s_i - s_{i-1}})}, \quad c_i = \frac{s_{i+1} - s_i}{h^2(1 - e^{s_i - s_{i+1}})} \\ b_i &= \frac{1}{h^2}\left(\frac{s_i - s_{i-1}}{1 - e^{s_{i-1} - s_i}} + \frac{s_{i+1} - s_i}{e^{s_{i+1} - s_i}}\right), \end{aligned} \tag{5.179}$$

мы легко обнаруживаем между ними связь

$$b_i \geqslant a_{i+1} + c_{i-1}, \tag{5.180}$$

означающую диагональное преобладание не по строкам, как ранее, а по столбцам. Если матрица симметрична, то различия в этом нет, но в данном случае она уже не является матрицей положительного типа. Однако для транспонированной матрицы A^T неравенство (5.168) приводит к диагональному преобладанию по строкам и в результате удовлетворяются все условия определения (5.12) о положительности типа. А так как матрицы A^{-1} и $(A^T)^{-1}$ могут быть неотрицательны только одновременно, то трехдиагональная матрица с элементами вида (5.172) является монотонной.

Отметим еще одно характерное свойство трехдиагональной матрицы A с элементами (5.167): будучи несимметричной, она симметризуется при умножении справа на диагональную матрицу:

$$\hat{A} = A\,D = (\hat{A})^T, \qquad D = \operatorname{diag}\{e^{-s_i}\}. \tag{5.181}$$

Симметризация соответствующей системы $A\,v = f$ в таком случае означает замену переменных $v = D^{-1}\hat{v}$, в результате чего приходим к $\hat{A}\hat{v} = f$. Напомним также, что матрица с элементами (5.167), как и любая трехдиагональная, может быть симметризована и с помощью преобразования подобия (5.154).

Среди рассмотренных трехдиагональных матриц большое методическое значение имеет наиболее простая из них, соответствующая аппроксимации задачи Дирихле для уравнения Пуассона на равномерной сетке:

$$A = \frac{1}{h^2}T_a = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{bmatrix}. \qquad (5.182)$$

Рассмотрим теперь задачу с краевыми условиями общего вида:

$$v_0 \cos\psi - v_1 \sin\psi = g_a,\ 0 \leqslant |\psi| < \frac{\pi}{2},$$

$$v_N \cos\varphi - v_N \sin\varphi = g_b,\ \ 0 \leqslant \varphi < \frac{\pi}{2},$$

$$(A_c v)_i \equiv \frac{1}{h^2}(T_c v)_i = \frac{1}{h^2}\begin{bmatrix} (2-tan\psi)v_1 - v_2 \\ \cdots\cdots\cdots \\ -v_{i-1} + 2\,v_i + v_{i+1} \\ \cdots\cdots\cdots \\ -v_{N-1} + (2 - \mathrm{tg}\varphi)v_N \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \cdots \\ v_i \\ \cdots \\ v_N \end{bmatrix} = \begin{bmatrix} f_0 + g_a/(\\ \cdots\cdots\cdots \\ f_i \\ \cdots\cdots\cdots \\ f_N + g_b/ \end{bmatrix} \tag{5.183}$$

Матрица A_c в (5.176), в отличие от A в (5.175), содержит в левом верхнем и правом нижнем углах элементы $2 - \mathrm{tg}\psi$ и $2 - \mathrm{tg}\varphi$ соответственно. В частности, если на левой и/или правой границах заданы условия Неймана, т. е. $\psi = \frac{\pi}{4}$ и/или $\varphi = \frac{\pi}{4}$, то угловые элементы равны 1. А при $\psi = \varphi = 0$ имеем $A_c = A$.

При $\psi = \pi/4$ и $\varphi = 0$, например, имеем матрицу

$$A = \frac{1}{h^2}T_b = \frac{1}{h^2}\begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{bmatrix}, \tag{5.184}$$

отличающуюся от (5.175) только левым верхним элементом. Она остается невырожденной, так как в последней строке имеется строгое диагональное преобладание.

Если же условие 2-го рода задается на обоих концах отрезка, то приходим уже к вырожденной матрице

$$T_c = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ 0 & & & -1 & 1 \end{bmatrix}, \qquad (5.185)$$

у которой вектор $e = (1, \ldots, 1)$ из единичных компонент является собственным, а соответствующее собственное число — нулевым.

Отметим еще одно полезное свойство для трехдиагональных матриц положительного типа с постоянными диагоналями, относящихся к классу *теплицевых матриц*:

$$A = \begin{bmatrix} b & -c & & & 0 \\ -a & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -c \\ 0 & & & -a & b \end{bmatrix}. \qquad (5.186)$$

Ее симметризация с помощью диагонального преобразования подобия (5.170) выглядит особенно просто:

$$D = \operatorname{diag}\left\{ \left(\frac{a}{c}\right)^{\frac{i-1}{2}} \right\}, \qquad A = D\tilde{A}D^{-1} = \{-\sqrt{ac},\ b, -\sqrt{ac}\}, \qquad (5.187)$$

т. е. само преобразование внешне заключается в делении поддиагональных элементов и в умножении наддиагональных на одно и то же число. Если при этом $a \neq c$, то подобная матрица $\hat{A}$ будет иметь большее диагональное преобладание, чем исходная A, так как $b - 2\sqrt{ac} > b - a - c$ вследствие того, что у двух разных чисел среднее геометрическое меньше среднего арифметического.

Перейдем далее к задаче с периодическими граничными условиями $v_{i+N} = v_i,\ i = 0, 1, \ldots$. Система сеточных уравнений при этом имеет вид

$$(A_d v)_i \equiv \frac{1}{h^2}(T_d v)_i = \frac{1}{h^2} \begin{bmatrix} 2\,v_1 - v_2 - v_N \\ \cdots\cdots\cdots \\ -v_{i-1} + 2\,v_i - v_{i+1} \\ \cdots\cdots\cdots \\ -v_{N-1} + 2\,v_N - v_1 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \cdots \\ v_i \\ \cdots \\ v_N \end{bmatrix} = \begin{bmatrix} f_1 \\ \cdots \\ f_i \\ \cdots \\ f_N \end{bmatrix}, \tag{5.188}$$

где матрица T_d записывается следующим образом

$$T_d = \begin{bmatrix} 2 & -1 & & & -1 \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & 2 & -1 \\ -1 & & & -1 & 2 \end{bmatrix}$$

и аналогично T_c имеет собственный вектор e с соответствующим нулевым собственным числом.

Теперь коснемся одномерных разностных уравнений повышенной точности, ограничившись одним примером — аппроксимацией (5.25) четвертого порядка точности. Матрица в данном случае является пятидиагональной:

$$A = -\Delta_h^{(5)} = \{\ldots 1, -16, 30, -16, 1, \ldots\}$$

и не является уже матрицей положительного типа. Надо при этом иметь в виду, что если в первом и предпоследнем узлах сетки около границы используются трехточечные разностные выражения, то соответствующие строки матрицы будут такие же, как у T_a из (5.175). Получаемые при этом матрицы относятся к широкому классу *ленточных матриц*, у которых ненулевые элементы расположены только около главной диагонали, т. е. $a_{i,j} = 0$ при $|i - j| > m$. Величина m называется *полушириной ленты*.

5.6.2. Спектральные характеристики матриц. Спектральный анализ является одним из основных в исследовании свойств систем сеточных уравнений, а также в обосновании и оптимизации методов их решения. Прежде чем перейти к конкретным задачам, приведем те минимальные данные из алгебраической теории матриц, которые нам потребуются.

Мы начнем рассмотрение с класса так называемых *нормальных матриц*, которые имеют N линейно-независимых собственных векторов и к которым относятся симметричные матрицы. Поскольку к ним, естественно, применимы процессы ортогонализации и нормировки, то можно сказать, что нормальная (и симметричная в том числе) матрица имеет ортонормальный базис из собственных векторов, т. е.

$$(z_q, z_{q'}) = \delta_{q,q'} = \left\{ \begin{array}{ll} 1, & q = q', \\ \\ 0, & q \neq q', \end{array} \right\},$$

где $\delta_{q,q'}$ есть символ Кронекера.

В силу определения собственных чисел λ_q и векторов z_q для нормальной матрицы справедливо представление (*спектральное разложение*)

$$A = Z\Lambda Z^T, \tag{5.189}$$

где Λ — диагональная матрица с элементами λ_q на главной диагонали, а Z — квадратная матрица, столбцами которой являются векторы z_q. Матрица Z относится к классу ортогональных матриц, обладающих свойством $Z^{-1} = Z^T$.

В силу представления (5.182) для симметричной положительно-полуопределенной матрицы, имеющей неотрицательные собственные числа, допустимо определение квадратного корня, являющегося матрицей с теми же собственными векторами и собственными числами $\sqrt{\lambda_g}$, т. е.

$$A^{1/2} = Z\Lambda^{1/2}Z^T, \tag{5.190}$$

и выполняется очевидное свойство $A^{1/2}A^{1/2} = A$ (читателю наверняка понятно, что $\Lambda^{1/2}$ означает диагональную матрицу с величинами $\sqrt{\lambda_q}$ на главной диагонали, т. е. $\Lambda^{1/2} = \mathrm{diag}\{\lambda_q^{1/2}\}$).

Аналогично могут быть введены и другие *функции от матрицы*. А именно, если для всех собственных чисел λ_q определена функция $F(\lambda_q)$, то

$$F(A) = Z^T \mathrm{diag}\{F(\lambda_q)\}Z. \tag{5.191}$$

Каждая симметричная положительно определенная матрица порождает некоторое новое *A-скалярное произведение и норму* :

$$(u,v)_A \equiv (A\,u,v) = (u, A\,v),$$
$$\|u\|_A = (A\,u,u)^{1/2} = \|A^{1/2}u_2\|. \tag{5.192}$$

Укажем сразу и на чрезвычайно полезное для нас в дальнейшем общее качество симметричных матриц — это *экстремальные свойства собственных чисел* λ:

$$\min_{v\neq 0}\left\{\frac{(Av,v)}{(v,v)}\right\} \leqslant \lambda \leqslant \max_{v\neq 0}\left\{\frac{(Av,v)}{(v,v)}\right\},$$

где записанная в фигурных скобках дробь называется *отношением Рэлея.*

Из предыдущего вытекает, в частности, что если в трехдиагональной матрице с коэффициентами вида (5.78) все значения r_i положительны, то справедливо неравенство $\lambda \geqslant \min\limits_i\{r_i\} > 0$.

Такая же оценка "верна"и для других рассмотренных выше аппроксимациях диффузионно-конвективного уравнения, например когда коэффициенты описываются приведенными в п. 5.2.2 формулами. А неравенства на соотношения между шагами сетки и коэффициентами $p_{i\pm 1/2}, q_i$ в алгебраической терминологии означают условия, при которых трехдиагональные матрицы с соответствующими коэффициентами будут матрицами положительного типа и, следовательно, монотонными (отсюда следует, что эти неравенства формально являются

только достаточными, но не необходимыми условиями монотонности).

Приведем без доказательства некоторые нужные нам для дальнейшего спектральные свойства матриц.

Теорема 5.5. *Нормальные матрицы A, B перестановочны ($A\,B = B\,A$) тогда и только тогда, когда они имеют одинаковый базис из собственных векторов. Собственные числа суммы $A + B$ и произведения $A\,B$ равны сумме и произведению соответствующих собственных чисел матриц A, B.*

Для произвольной вещественной квадратной матрицы обобщением соотношения (5.183) является следующее утверждение.

Теорема 5.6 (Шура). *Для любой вещественной квадратной матрицы существует вещественная ортогональная матрица Z такая, что*

$$Z^T A Z = \begin{bmatrix} A_1 & & & * \\ & \ddots & & \\ & & A_i & \\ & & & \ddots \\ 0 & & & A_N \end{bmatrix}, \tag{5.193}$$

где для каждого i матрица A_i имеет размер 1×1 или 2×2, отвечая соответственно вещественному собственному значению или паре комплексно-сопряженных собственных зна-

чений матрицы A. Блоки A_i можно расположить в произвольном порядке. Если две матрицы коммутативны, то их приведение к форме (5.186) *осуществляется одной и той же матрицей Z.*

Подчеркнем, что матрица (5.186) является *верхней хессенберговой* (т. е. элементы $a_{i,j}$ равны 0 при $i > j+1$) и становится верхней треугольной, если ее собственные числа вещественны.

Для оценки границ спектра матриц важное значение имеет следующий результат.

Теорема 5.7 (Гершгорина). *Каждое собственное число матрицы всегда расположено в одном из кругов комплексной плоскости*

$$|a_{i,i} - \lambda| \leqslant \sum_{\substack{j=1 \\ j\neq 1}}^{N} |a_{i,j}|, \quad i = 1, 2, ..., N. \tag{5.194}$$

Отсюда следует, например, что если матрица имеет по всем строкам строгое диагональное преобладание

$$\delta = \min_i \left\{ |a_{i,j}| \} - \sum_{\substack{j=1 \\ j\neq 1}}^{N} |a_{i,j}| \right\} > 0 \tag{5.195}$$

и все $a_{i,i}$ положительны, то для любого собственного числа λ его вещественная часть удовлетворяет оценке

$$\delta \leqslant \mathrm{Re}\lambda \leqslant ||A||_\infty. \tag{5.196}$$

Определение 5.15. *Матрицы A, B называются подобными, если существует невырожденная матрица T такая, что справедливо равенство $B = T^{-1} A\, T$.*

Очевидно, что подобные матрицы A, B имеют одинаковые собственные числа, а их соответствующие собственные векторы z_q, y_q связаны соотношением $z_q = Ty_q$.

Определение 5.16. *Если существует невырожденная матрица S такая, что*

$$B = SAS', \tag{5.197}$$

то матрица B называется конгруэнтной к A, а само выражение (5.190) *— преобразованием конгруэнтности. Такие матрицы A и B при этом называются конгруэнтными.*

Очевидно, что если матрица A симметрична (и, возможно, положительно определена), то такими же свойствами обладает конгруэнтная к ней матрица B. Более глубокий результат заключается в следующем так называемом законе инерции Сильвестра.

Теорема 5.8 (закон инерции). *Симметричные матрицы A, B конгруэнтны тогда и только тогда, когда у них одинаковое количество положительных, отрицательных и нулевых собственных чисел.*

Естественно, что для спектрального анализа определяющее значение имеет нахождение собственных чисел или их оценок для сумм и произведений матриц. Алгебраический аппарат, если не погружаться в технические дебри, ограничивается в данной области немногими результатами. Естественно, самый простой из них заключается в теореме 5.5 о свойствах спектра перестановочных матриц. Ненамного сложнее следующие утверждения.

Теорема 5.9. *Собственные числа многочлена от нормальной матрицы $P_n(A)$ суть значения многочлена от соответствующих собственных чисел, т. е.*

$$\lambda_q(P_n(A)) = P_n(\lambda_q). \tag{5.198}$$

Теорема 5.10. *Если матрица A симметрична, а B симметрична и положительно определена, то матричное произведение BA подобно матрице $B^{1/2}AB^{1/2}$, конгруэнтной с A, и, следовательно, имеет одинаковое с ней количество положительных, отрицательных и нулевых собственных чисел. В частности, если $\lambda_q(A)$, $\lambda_q(B) > 0$ для всех q, то $\lambda_q(AB) > 0$.*

Теорема 5.11. *Для собственных чисел λ_q суммы симметричных матриц $A + B$ справедливы следующие оценки через собственные числа μ_q и ν_q матриц A, B:*

$$\min_q\{\mu_q(A)\}+\min_q\{\nu_q(B)\}\leqslant\lambda_q(A+B)\leqslant\max_q\{\mu_q(A)\}+\max_q\{\nu_q(B)\}. \tag{5.199}$$

Последний результат непосредственно вытекает из экстремального свойства собственных чисел.

Теорема 5.12. *Если матрицы A, B симметричны и положительно определены, то для собственных чисел их произведения A выполняют неравенства*

$$\min_q\{\mu_q(A)\}\min_q\{\nu_q(B)\}\leqslant\lambda_q(AB)\leqslant\max_q\{\mu_q(A)\}\max_q\{\nu_q(B)\}. \tag{5.200}$$

Для определенности в дальнейшем будем предполагать, что если матрица симметрична, то ее собственные числа λ_q пронумерованы в порядке возрастания: $\lambda_1 \leqslant \lambda_2 \leqslant \ldots < \lambda_N$.

Далее для некоторых матриц с постоянными коэффициентами мы приводим (без вывода) явные выражения собственных чисел и векторов, а когда это невозможно, даем оценки для границ спектра.

Надо сказать, что формально все спектральные свойства одномерных разностных операторов можно вывести из теории уравнений в конечных разностях (см. [16], [47]), по аналогии с обыкновенными дифференциальными уравнениями, однако это технически достаточно сложно и мы эти вопросы опускаем для сокращения изложения.

Начнем с простейшей трехдиагональной теплицевой матрицы $T_a = \{\ldots -1, 2, -1, \ldots\}$ из (5.175), соответствующей аппроксимации задачи Дирихле для уравнения Пуассона на равномерной сетке. В данном случае легко проверяется следующий результат:

$$T_a z_q^a = \lambda_q^a z_q^a, \qquad q = 1, 2, \ldots, N, \qquad \lambda_q^a = 4 \sin^2 \frac{q\pi}{2(N+1)},$$
$$z_q^a = \left\{ z_{q,i}^a = \sqrt{\tfrac{2}{N+1}} \sin q \frac{i\pi}{N+1} \right\}, \tag{5.201}$$

где множитель в собственном векторе введен по условию нормировки $(z_q^a, z_{q'}) = \delta_{q,q'}$. При этом полезно отметить неравенства

$$\lambda_1^a \geqslant \Big(\frac{\pi}{N+1}\Big)^2, \qquad \lambda_N^a \leqslant 4\Big[1 - \Big(\frac{\pi}{N+1}\Big)^2\Big], \tag{5.202}$$

которые в асимптотическом случае $N \gg 1$ переходят в приближенные равенства, а также вытекающую отсюда оценку числа обусловленности

$$\operatorname{cond}_2(T_a) \leqslant \Big[\frac{2(N+1)}{\pi}\Big]^2.$$

Для матрицы T_b из (5.177), соответствующей смешанной краевой задаче (условие Дирихле на одном конце расчетного отрезка и условие Неймана — на другом), имеем

$$\begin{gathered} T_b\, z_q^b = \lambda_q^b z_q^b, \qquad q = 1, 2, \ldots, N, \\ \lambda_q^b = 4\sin^2\frac{2q-1}{2(2N+1)}\pi, \qquad z_q^b = \Big\{\sqrt{\tfrac{2}{2N+1}}\sin\frac{2q-1}{2N+1}\Big\}, \\ \frac{\pi^2}{(2N+1)^2} \approx \lambda_1 \leqslant \lambda_q^b \leqslant \lambda_N \approx \Big[1 - \frac{\pi^2}{(2N+1)^2}\Big]. \end{gathered} \tag{5.203}$$

Отсюда следует, в частности, что число обусловленности

$$\operatorname{cond}_2(T_b) \leqslant \Big[\frac{2(2N+1)}{\pi}\Big]^2$$

асимптотически, т. е. при $N \gg 1$, в четыре раза больше, чем у матрицы T_a.

Для задачи Неймана, порождающей матрицу T_c из (5.178), получаем

$$T_c z_q^c = \lambda_q^c z_q^c, \qquad q = 1, 2, \ldots, N, \lambda_q^c = 4 \sin^2 \frac{(q-1)\pi}{2N},$$
$$z_q^c = \Big\{ \frac{2}{\sqrt{2N+1}} \sin(N+1-i)\frac{q\pi}{N} \Big\}, 0 \leqslant \lambda_q^c \leqslant 4. \tag{5.204}$$

Рассмотрим также трехдиагональную циклическую матрицу T_d из (5.171), соответствующую периодической краевой задаче. Ее спектральные свойства описываются следующими выражениями:

$$T_d z_q^d = \lambda_q^d z_q^d, \qquad q = 1, 2, \ldots, N+1, \qquad \lambda_q^d = 4 \sin^2 \frac{q\pi}{N+1},$$
$$z_q^d = \Big\{ \frac{1}{\sqrt{N+1}} \Big(\cos \frac{iq\pi}{N+1} + \sin \frac{iq\pi}{N+1} \Big), \ \ i = 1, \ldots, N+1 \Big\}. \tag{5.205}$$

Первым шагом к переходу на матрицы более общего вида является рассмотрение трехдиагональной симметричной теплицевой матрицы, для которой ввиду элементарной связи с T_a

$$A = \{\ldots - a, b, -a, \ldots\} = aT_a + (b - 2a)I$$

собственные числа определяются простым выражением

$$\lambda_q(A) = 4\, a \sin^2 \frac{q\pi}{2(N+1)} - 2\, a + b. \tag{5.206}$$

Если же трехдиагональная теплицева матрица несимметрична, то поскольку диагональным преобразованием подобия (5.180) она сводится к симметричной, для нее также формулируется результат, аналогичный предыдущему:

$$A = \{\dots - a, b, -c, \dots\} = D\big[\sqrt{ac}\,T_a + \big(b - 2\sqrt{ac}\big)I\big]D^{-1},$$

$$\lambda_q(A) = 4\sqrt{ac}\ \sin^2\frac{q\pi}{2(N+1)} - 2\sqrt{ac} + b. \tag{5.207}$$

Здесь надо иметь в виду, что если соответствующие собственным числам $\lambda_q(A)$ из (5.200) являются описанные в (5.194) собственные векторы z_q^a, то собственными векторами несимметричной матрицы из (5.200) будут

$$z_q(A) = D^{-1}z_q^a = \left\{\left(\frac{a}{c}\right)^{\frac{i-1}{2}} \sin\left(iq\frac{\pi}{N+1}\right)\right\}. \tag{5.208}$$

Следует указать, что при $a \neq c$ и $b \geq a + c$ полученная после симметризации матрица $D^{-1}AD$ при $N \gg 1$ будет иметь тем лучше обусловленность, чем сильнее разница в значениях a, c, так как при этом растет величина $a + c - 2\sqrt{ac}$. Заметим, что говорить о числе обусловленности исходной матрицы A надо с осторожностью, так как в силу ее несимметричности величина $\mathrm{cond}_2(A)$ выражается не через значения $\lambda_q(A)$ из (5.200), а через сингулярные числа.

Что касается представлений или оценок для матриц с переменными элементами, то здесь круг результатов ограничен. Основные инструменты исследования в данном случае — использование теоремы Гершгорина, выражений для спектра матричных сумм и произведений или специальных технических приемов.

Например, для трехдиагональной стилтьесовой матрицы, соответствующей аппроксимации задачи Дирихле для уравнения Пуассона на неравномерной сетке, можно выписать ад-

дитивное представление

$$A = \{\ldots - a_i, b_i, -a_{i+1}..\} = A_0 + (A - A_0),$$

$$A_0 = aT_a, \qquad a = \min\{a_i\},$$

такое, что разность $A - A_0$ будет положительно полуопределенной матрицей. Отсюда (с применением кругов Гершгорина для верхней оценки) получаем $a\lambda_1^a \leqslant \lambda_q(A) \leqslant \max\limits_i\{a_i + b_i + c_i\}$.

В данном случае полезным примером является аппроксимация задачи Дирихле для уравнения диффузии на неравномерной сетке с разностными коэффициентами, определяемыми из обобщения формул (5.43), когда матрица

$$A = \left\{ \cdots - \frac{2p_{i-1/2}}{h_{i-1}(h_i + h_{i-1})},\ \frac{2(h_i p_{i-1/2} + h_{i-1} p_{i+1/2})}{h_{i-1} h_i (h_i + h_{i-1})} - \frac{2p_{i+1/2}}{h_i(h_i + h_{i-1})}, \ldots \right\}$$

является матрицей положительного типа, но несимметричной. Для нее на основе описанных подходов получаются двусторонние оценки

$$\frac{\pi^2 \check{p}}{(N+1)^2 \hat{h}^2} \leqslant \frac{\check{p}}{\hat{h}^2} 4 \sin^2 \frac{\pi}{2(N+1)} \leqslant \lambda_q(A) \leqslant \frac{4\hat{p}}{\check{h}^2}, \qquad (5.209)$$

где введены обозначения

$$\check{p} = \min_i\{p_{i\pm 1/2}\}, \quad \check{h} = \min_i\{h_i\}, \quad \hat{p} = \max_i\{p_{i\pm 1/2}\}, \quad \hat{h} = \max_i\{h_i\}.$$

Отсюда следует оценка для спектрального числа обусловленности

$$\mathrm{cond}_2(A) \leqslant \left[\frac{2(N+1)}{\pi}\right]^2 \frac{\hat{p}}{\check{p}} \left(\frac{\hat{h}}{\check{h}}\right)^2,$$

которая указывает на ухудшение обусловленности матрицы как при усилении неравномерности шагов сетки, так и при разбросе значений коэффициента диффузии.

Универсальный подход к решению проблемы собственных значений $Az_q - \lambda_q z_q = 0$ для трехдиагональной матрицы $A = 3\text{-diag}\{-a_i, b_i, -c_i\}$ заключается в следующем. Требуется найти общее решение однородной системы

$$(A\,v)_i \equiv -a_i v_{i-1} + (b_i + \lambda) v_i - c_i v_{i+1} = 0, \;\; i = 1, ..., N,$$

$$l_a v|_{i=0} = 0, \;\; l_b v|_{N+1} = 0 \tag{5.210}$$

с однородными же "сеточными" граничными условиями, конкретный вид которых определяется способом аппроксимации краевых условий исходной дифференциальной задачи. Общее решение (5.203) можно искать в виде

$$v_i = c_1 v_i^{(1)} + c_2 v_i^{(2)},$$

где $v^{(q)} = \{v_i^{(q)}\}$ — некоторые линейно-независимые решения, а c_q — константы, $q = 1, 2$, которые можно определить из двух последних (граничных) уравнений.

Решение данной задачи достаточно сложно, и мы его продемонстрируем для случая постоянных коэффициентов (теплицева матрица A):

$$v_{i+1} - 2\,b\,v_i + c\,v_{i-1} = 0.$$

Искомые величины ищем в виде $v_i^{(q)} = \lambda_q^i$, после подстановки которого в систему и сокращения на λ_q^{i-1} (в предположении его отличия от нуля) получаем квадратное уравнение $\lambda^2 - 2\,b\lambda + c = 0$, а оно при $b^2 - c > 0$ имеет разные вещественные корни $\lambda_{1,2} = b \pm \sqrt{b^2 - c}$. Если же детерминант $b^2 - c$ равен нулю, то $\lambda_1 = \lambda_2 = b$ и линейно-независимые решения ищутся в виде $v_i^{(1)} = b^i, v_i^{(2)} = ib^i$. При $b^2 - c = 0$ и $b > 0$ имеем $v_i^{(1)} = \cos i\theta, v_i^{(2)} = \rho^i \sin i\theta$, где $\theta = \text{arc } \cos(a/\sqrt{b}),\ \rho = \sqrt{b}$.

В случае аппроксимаций повышенной точности, например (5.33), или при дискретизации одномерных краевых задач для дифференциальных уравнений более высокого порядка возникают ленточные матрицы с шириной полуполосы $p > 1$, для которых зачастую существуют аддитивно-мультипликативные представления через трехдиагональные матрицы и анализ спектра удается осуществить на основе предыдущих результатов.

Знание минимального собственного числа с.п.о. матрицы позволяет легко получить оценку погрешности сеточного уравнения в евклидовой норме.

Переписывая уравнение (5.49),приведем связь между вектором погрешности аппроксимации ψ и вектором ошибки в виде

$$A\,z = \psi.$$

Далее получаем, что если матрица A невырожденна, то справедливо неравенство

$$||z|| \leqslant ||A^{-1}|| \cdot ||\psi||.$$

В частности, если A есть с.п.о. — матрица, т. е. $A = A^T$ и $(Ax, x) \geqslant \delta(x, x), \delta > 0$, то для евклидовой нормы имеем $||A^{-1}||_2 \leqslant \delta^{-1}$. Отсюда при условиях

$$||A^{-1}||_2 \leqslant C_1 h^{-\alpha},\ ||\psi||_2 \leqslant C_2 h^{\gamma+\alpha}$$

получаем оценку ошибки

$$||z||_2 \leqslant C_1 C_2 h^{\gamma},$$

которая означает сходимость сеточного решения с порядком γ. Отметим, что в приведенных неравенствах введен показатель степени α, который в окончательной оценке сокращается и больше не участвует. Он был введен для того, чтобы подчеркнуть, что возможные различные нормировки сеточных уравнений влияют на оценки $||A^{-1}||$ и $||\psi||$, но не на итоговый результат.

Для рассматриваемых нами сеточных СЛАУ, аппроксимирующих дифференциальные уравнения второго порядка, норма $||A^{-1}||_2$ (при используемой естественной нормировке, когда коэффициенты уравнений имеют порядок $O(h^{-2})$) есть величина $O(h^0)$, т. е. в (5.77) $\alpha = 0$, а типичное значение γ — один, два или четыре.

5.6.3. Свойства монотонных сеточных схем. Напомним, что система линейных алгебраических уравнений называется монотонной, если ее матрица монотонна, т. е. обратная к ней матрица существует и неотрицательна ($A^{-1} \geqslant 0$). Частным случаем монотонных матриц являются матрицы положительного типа, наиболее легко идентифицируемые (см. определение 5.12).

Чтобы обогатить понятие монотонных матриц, нам потребуется рассмотреть еще несколько определений и теорем.

Определение 5.17. *Матрица положительного типа со строгим диагональным преобладанием называется матрицей Минковского.* .

Имеется одна существенная связь между матрицами Минковского и M-матрицами.

Теорема 5.13. (Островского). *Если A есть M-матрица, то существует положительно определенная диагональная матрица D такая, что $D^{-1}AD$ есть матрица Минковского.*

Этот результат потенциально важен для нижней оценки собственных чисел, однако вопрос заключается в конструктивном нахождении необходимых диагональных матриц.

Для характеризации M-матриц полезны следующие два утверждения.

Теорема 5.14. *Вещественная матрица A с неположительными внедиагональными элементами является M-матрицей тогда и только тогда, когда ее диагональные элементы положительны и справедливо неравенство $\rho(B) < 1$ для спектрального радиуса матрицы $B = I - D^{-1}A$, где $D = \operatorname{diag}\{A\}$.*

Доказательство данного утверждения следует из того известного факта, что неравенство $\rho(B) < 1$ является необходимым и достаточным для сходимости матричного ряда

$$(D^{-1}A)^{-1} = (I - B)^{-1} = I + B + B^2 \ldots \geqslant 0.$$

Напомним, что *спектральным радиусом матрицы* назы-

вается максимальный из модулей ее собственных чисел

$$\rho(B) = \{\max_q |\lambda_q(B)|\}.$$

Тесно связанным с теоремой 5.14 является следующее утверждение: если матрица B неотрицательна, то матрица $I - B$ является монотонной тогда и только тогда, когда $\rho(B) < 1$.

Теорема 5.15. *Для того чтобы вещественная неразложимая матрица A с неположительными внедиагональными элементами была M-матрицей, необходимо и достаточно наличие у нее свойства обобщенного диагонального преобладания, т. е. существование положительного вектора $y > 0$ такого, что $Ay \geqslant 0$, причем хотя бы для одной компоненты вектора $A\,y$ неравенство строгое ($(A\,y)_i \geqslant 0$).*

Нетрудно проверить, что свойство обобщенного диагонального преобладания для матриц с неположительными внедиагональными элементами эквивалентно свойству "классического" диагонального преобладания для матрицы $\tilde{A} = A\,Y$, где $Y = \mathrm{diag}\{y_i\}$ – диагональная матрица в силу очевидного равенства $A\,y = A\,Y\,e$.

Что касается достаточности "обычного" диагонального преобладания для монотонности неразложимой матрицы с неположительными внедиагональными элементами, то это следует из свойств матриц положительного типа.

Важное прикладное значение может иметь следующее утверждение, которое в определенном смысле можно назвать теоремой сравнения для M-матриц.

Теорема 5.16. *Пусть $A_1 = D_1 - C_1$ есть M-матрица*

($D_1 = \mathrm{diag}\{A_1\}$, $C_1 \geqslant 0$ — неотрицательная матрица с нулевой главной диагональю). Пусть также $D_2 \geqslant D_1$ — некоторая положительно определенная диагональная матрица и C_2 — неотрицательная матрица, удовлетворяющая неравенству $C_2 \leqslant C_1$. Тогда $A_2 = D_2 - C_2$ является M -матрицей и выполняется условие $A_2^{-1} \leqslant A_1^{-1}$.

Следствием данного результата является тот факт, что если A есть M-матрица, D — неотрицательная диагональная матрица, то $A + D$ — тоже M-матрица, причем $(A + D)^{-1} \leqslant A^{-1}$.

Поскольку у монотонной матрицы обратная матрица имеет неотрицательные элементы, важное значение имеют следующие спектральные свойства неотрицательных матриц.

Теорема 5.17 (Фробениуса). *Неразложимая неотрицательная матрица A имеет простое положительное собственное число ρ с положительным соответствующим собственным вектором. Абсолютные значения остальных собственных чисел не превосходят ρ, а матрица A не имеет других положительных собственных векторов.*

Доказательство этой теоремы включает следующую важную характеристику, которую можно назвать минимаксным свойством спектрального радиуса неотрицательной матрицы:

$$\rho = \max_{x>0} \min_i \left\{ \frac{(A\,x)_i}{x_i} \right\} = \min_{x>0} \max_i \left\{ \frac{(A\,x)_i}{x_i} \right\}.$$

Отсюда вытекает, в частности, такой результат: если неразложимая неотрицательная матрица A представима в форме $A = B + \alpha C$, где B и C – неотрицательные матрицы ($C \neq 0$),

то

$$d\rho/d\alpha > 0.$$

Для монотонных матриц приведем еще несколько полезных для многих приложений результатов.

Теорема 5.18(о норме матрицы, обратной к монотонной). *Пусть A — монотонная матрица, а v — некоторый вектор такой, что справедливы соотношения $Av \geqslant \alpha e$, $\alpha > 0$, $\|v\|_\infty = 1$. Тогда имеет место неравенство $\|A^{-1}\|_\infty \leqslant \alpha^{-1}$.*

Доказательство данного факта приведем ввиду его краткости и изящности. Действительно, из условия теоремы следует векторное неравенство $A^{-1}e \leqslant \alpha^{-1}v$, из которого сразу следует искомый результат для нормы.

Совершенно аналогично может быть доказано и более общее утверждение.

Теорема 5.19. *Если A и B — монотонная и неотрицательная матрицы такие, что*

$$Av \geqslant \alpha Be, \qquad \alpha > 0, \qquad \|v\|_\infty = 1,$$

то справедливо неравенство $\|A^{-1}B\|_\infty \leqslant \alpha^{-1}$.

Определение 5.18. *Представление матрицы A в виде $A = B - C$ называется регулярным разложением, если матрицы B, C являются соответственно монотонной и неотрицательной, т. е. $B^{-1} \geqslant 0$, $C \geqslant 0$. Если вместо последнего условия выполняется — $B^{-1}C \geqslant 0$, то данное представление называется слабо регулярным разложением. Очевидно, что регулярное разложение является слабо регулярным разложением, но обратное неверно.*

Теорема 5.20 (Варги). *Если равенство $A = B - C$ есть слабо регулярное разложение монотонной матрицы A, то для спектрального радиуса матрицы $B^{-1}C$ справедливы соотношения*

$$\rho(B^{-1}C) = \frac{\rho(A^{-1}C)}{1 + \rho(A^{-1}C)} < 1. \tag{5.211}$$

Кроме того, если для слабо регулярного разложения выполняется неравенство (5.204), то A — монотонная матрица.

Отсюда вытекает следующий результат. Пусть для монотонной матрицы имеются два регулярных разложения $A = B_k - C_k, \quad k = 1, 2$, причем матрицы C_1, C_2 связаны неравенством $C_1 \leqslant C_2$. Тогда для спектральных радиусов матриц $B_k^{-1}C_k$ имеет место неравенство

$$\rho(B_1^{-1}C_1) \leqslant \rho(B_2^{-1}C_2). \tag{5.212}$$

Определение 5.19. *Представление монотонной матрицы A в виде $A = Q^{-1}(B-C)P^{-1}$, где P, B и Q — монотонные матрицы, а C — неотрицательная, называется квазирегулярным разложением.*

Теорема 5.21(о квазирегулярном разложении). *Если равенство $A = Q^{-1}(B - C)P^{-1}$ есть квазирегулярное разложение, то для спектрального радиуса матрицы $B^{-1}C$ справедливы соотношения*

$$\rho(B^{-1}C) = \frac{\rho(P^{-1}A^{-1}Q^{-1}C)}{1 + \rho(P^{-1}A^{-1}Q^{-1}C)} < 1.$$

Отсюда аналогично (5.19) получаем следствие: пусть матрица A имеет два квазирегулярных разложения

$$A = Q_k^{-1}(B_k - C_k)P_k^{-1}, \qquad k = 1, 2,$$

причем дополнительно выполняется неравенство

$$Q_1^{-1}CP_1^{-1} \leqslant Q_2^{-1}CP_2^1 \quad \Big(Q_1^{-1}CP_1^{-1} < Q_2^{-1}CP_2^{-1},$$

тогда для спектральных радиусов матриц $B_k^{-1}C_k$ выполняется соотношение

$$\rho(B_1^{-1}C_1) \leqslant \rho(B_2^{-1}C_2) \quad \Big(\rho(B_1^{-1}C_1) < \rho(B_2^{-1}C_2)\Big).$$

Этот факт вытекает из соотношения (5.205) и того, что неотрицательная матрица $P_k^{-1}A^{-1}Q_k^{-1}C_k$ подобна $A^{-1}Q_k^{-1}C_kP_k^{-1}$.

Приведем одно достаточное условие монотонности матрицы, являющееся вполне конструктивным и хорошо дополняющее теорему Варги о регулярном разложении.

Теорема 5.22 (О. Аксельсон, Л. Ю. Колотилина). *Пусть для матрицы A существует слабо регулярное разложение, т. е.*

$$A = B - C, \qquad B^{-1} \geqslant 0, \qquad B^{-1}C \geqslant 0.$$

Пусть также существует положительный вектор $v > 0$ такой, что справедливо неравенство $B^{-1}A\,v > 0$. Тогда матрица A является монотонной. □

Далее мы рассмотрим результат, который при всей своей очевидности имеет фундаментальное значение и может быть назван как теорема сравнения для решений монотонных систем уравнений.

Теорема 5.23 (сравнения). *Пусть A — монотонная матрица, а $u^{(1)}$, $u^{(2)}$ — два решения систем с разными правыми частями f_1, f_2, связанными соотношениями упорядочения:*

$$A\,u^{(k)} = f^{(k)}, \qquad k = 1, 2, \qquad f^{(1)} \geqslant f^{(2)} \qquad (f^{(1)} > f^{(2)}).$$

Тогда выполняются неравенства $u^{(1)} \geqslant u^{(2)}$ $(u^{(1)} > u^{(2)})$.

Практическое значение чаще имеет теорема (ее можно назвать второй теоремой сравнения), являющаяся следствием предыдущей.

Теорема 5.24. *Пусть векторы $u^{(1)}$, $u^{(2)}$ являются решениями систем уравнений $A\,u^{(k)} = f^{(k)}$, $k = 1, 2$, с монотонной матрицей A и правыми частями, удовлетворяющими условиям*

$$f^{(1)} = \bigl\{f_i^{(1)} \geqslant 0\bigr\}, \qquad f^{(2)} = \bigl\{f_i^{(2)} : |f_i^{(2)}| \leqslant f_i^{(1)}\bigr\}.$$

Тогда компоненты решений связаны неравенствами

$$|u_i^{(2)}| \leqslant u_i^{(1)}, \qquad i = 1, \ldots, N. \tag{5.213}$$

Методическое значение описанных результатов будет очевидно далее, но смысл их заключается в следующем. Конечной целью в данном случае является оценка равномерной нормы решения системы $A\,u = f$ через такую же норму вектора правой части. Она дается очевидным неравенством

$$\|u\|_\infty \leqslant \|A^{-1}\|_\infty \|f\|_\infty, \tag{5.214}$$

которое, однако, не может помочь в тех случаях, когда не удается оценить норму обратной матрицы (например, с помощью теоремы 5.24). Зачастую же удается конструктивно построить некоторую мажорирующую функцию $u^{(1)}$ и через нее выписать искомую оценку для $u^{(2)}$.

Следующим нашим шагом будет переход к рассмотрению систем уравнений, которые характерны для сеточных методов и имеют пересечение с классом монотонных. Их матрицы являются, с одной стороны, сужением монотонных, а с другой —их расширением. Сложившегося наименования для них нет, и мы для определенности назовем их псевдомонотонными.

Определение 5.20. *Пусть $I_0(N)$ означает совокупность индексов i_k, $1 \leqslant i_k \leqslant N$, $k = 1, \ldots, m$, причем $m < N$; пусть также $I_1(N) = \{1 \leqslant i_k \leqslant N,\, i_k \notin I_0(N)\}$ — совокупность остальных индексов, а матрица $A = \{a_{i,j}\}$ N-го порядка обладает следующими свойствами:*

а) *для $i \in I_0(N)$ строки матрицы являются строками неотрицательного типа —*

$$a_{i,i} > 0, \quad a_{i,j} \leqslant 0, \quad a_{i,i} - \sum_{j=1}^{N} a_{i,j} \geqslant 0, \quad i \in I_0(N), \quad j \neq i;$$

б) *для $i \in I_1(N)$ строки матрицы A имеют строгое диагональное преобладание —*

$$\delta_i = \frac{1}{|a_{i,i}|}\left(|a_{i,i}| - \sum_{\substack{j=1 \\ j\neq 1}}^{N} |a_{i,j}|\right) \geqslant \delta > 0, \qquad i \in I_1(N).$$

Тогда матрица A называется псевдомонотонной.

Очевидно, что если в $I_1(N)$ строки, кроме свойства (б), обладают еще и качеством неотрицательности, то псевдомонотонная матрица A является матрицей положительного типа. Однако в общем случае она не монотонная.

Следующая теорема является алгебраической интерпретацией *принципа максимума.*.

Теорема 5.25 (принцип максимума). *Если вектор правой части системы уравнений $A\,u = f$ с псевдомонотонной матрицей удовлетворяет условиям $f_i \leqslant 0$, $i \in I_0(N)$, то для компонент его решения справедливо неравенство*

$$\max_{i\in I_0}\{u_i\} \leqslant \max_{i\in I_1}\{u_i\},$$

т. е. их значения не достигают максимума в I_0.

Теорема 5.26. *Пусть $\bar{u}$ — неотрицательный вектор такой, что для псевдомонотонной матрицы A выполняются неравенства $(A\bar{u})_i = \bar{f}_i \geqslant \Delta > 0$, $i \in I_0$, а правая часть уравнения $A\,u = f$ удовлетворяет условию $|f_i| \leqslant \varepsilon$, $i \in I_0$. Тогда имеет место оценка*

$$\max_{i\in I_0}\{|u_i|\} \leqslant \max_{i\in I_1}\{|u_i|\} + \frac{\varepsilon}{\Delta}\max_{i\in I_0}\{\bar{u}_i\}. \tag{5.215}$$

Рассмотренные две теоремы являлись, по сути, подготовительными для обоснования следующего утверждения.

Теорема 5.27. *Пусть выполняются условия предыдущей теоремы и неравенства $|f_i/a_{i,i}| \leqslant \varepsilon_1$, $i \in I_1(N)$. Тогда для решения системы $Au = f$ с псевдомонотонной матрицей справедлива оценка*

$$|u_i| \leqslant \frac{1}{\delta}\Big(\frac{\varepsilon}{\Delta}\max_{i\in I_0}\{\bar{u}_i\} + \varepsilon_1\Big). \tag{5.216}$$

Если ранее объектом внимания при исследовании погрешностей численных решений были симметричные или симметризуемые сеточные уравнения, для которых явно находятся или оцениваются собственные числа, то ниже мы будем рассматривать только монотонные системы, в которых для обратных матриц удается оценить кубические, или равномерные, нормы. Эти две разные алгебраические техники применимы одновременно, естественно, только к стилтьесовым матрицам, обладающим свойствами как монотонности, так и положительной определенности.

Мы начнем анализ с простейших задач, для которых удается применить теорему 5.18 о норме матрицы, обратной к монотонной. Вопрос здесь заключается в конструктивном нахождении положительного вектора y, для которого выполняется условие $Ay \geqslant \alpha e$, $\alpha > 0$.

Рассмотрим одномерную разностную задачу Дирихле для уравнения Пуассона на неравномерной сетке $a = x_0 < x_1 < \ldots < x_{N+1} = b$:

$$-\frac{2v_{i-1}}{h_{i-1}(h_i + h_{i-1})} + \frac{2v_i}{h_i h_{i-1}} - \frac{2v_{i+1}}{h_i(h_i + h_{i-1})} = f_i, \quad i = 1, \ldots, N, \tag{5.217}$$

трехдиагональная матрица A которой после исключения v_0, v_{N+1} принимает вид (5.155) с элементами $a_i = \frac{1}{h_{i-1}(h_i+h_{i-1})}$,

$b_i = \dfrac{1}{h_i h_{i-1}}$, $c_i = \dfrac{1}{h_i(h_i + h_{i-1})}$ и погрешность аппроксимации удовлетворяет неравенству

$$|\psi_i| \leqslant \psi_m = \max_i \left\{ \frac{|h_i - h_{i-1}|}{3} M_3 + \frac{h_i^2 - h_i h_{i-1} + h_{i-1}^2}{12} M_4 \right\}.$$

В качестве “пробного"вектора берем сеточную функцию с компонентами

$$y_i = 1 - \frac{4(x_i - \frac{a+b}{2})^2}{(b-a)^2}, \qquad i = 1, \ldots, N, \tag{5.218}$$

обеспечивающими выполнение условий

$$\|y\|_\infty = 1, \qquad Ay \geqslant \alpha e, \qquad \alpha = \frac{8}{(b-a)^2}.$$

Отсюда из равенства $z = A^{-1}\psi$ и вытекающего из теоремы 5.18 неравенства $\|A^{-1}\|_\infty \leqslant (b-a)^2/8$ получаем сразу равномерную оценку погрешности для задачи (5.211):

$$\|z\|_\infty \leq (b-a)^2 \psi_m \tag{5.219}$$

с первым порядком для неравномерной сетки и вторым для равномерной.

Если в краевой задаче (5.211) заменить на правом конце отрезка граничное условие Дирихле на Неймана ($v_{N+1} - v_N = h_N g$), то необходимый пробный вектор также легко строится:

$$y_i = 1 - \frac{(x_i - c)^2}{(a-c)^2}, \qquad i = 1, \ldots, N, \qquad c = a + 1 + \sqrt{1 - 2(b-a)},$$

и он удовлетворяет условиям $Ay \geqslant \alpha e$, $\alpha = \frac{1}{(a-c)^2}$. Отсюда для нормы ошибки получаем оценку

$$\|z\|_\infty \leqslant (a-c)^2 \|\psi\|_\infty,$$

имеющую тот же характер, что и (5.212).

§ 5.7. Методы решения СЛАУ с ленточными матрицами

В этом параграфе главное внимание уделяется методам решения трехдиагональных СЛАУ, которые представим в покомпонентной форме

$$\begin{aligned} &(Av)_i \equiv -a_i v_{i-1} + b_i v_i - c_i v_{i+1} = f_i, \\ &a_1 = c_N = 0, \qquad i = 1, \ldots, N. \end{aligned} \tag{5.220}$$

Для решения системы (5.213) рассмотрим несколько различных алгоритмов не столько в силу их практической значимости, сколько в методических целях, так как заложенные в них принципы переносятся и на более сложные задачи.

Практически все рассматриваемые в данном параграфе методы базируются на описанных ранее факторизациях матриц, однако для простоты изложения зачастую удобно привлекать их описание в покомпонентном представлении.

5.7.1. Стандартный метод прогонок. Мы начнем с простейшего, или стандартного, алгоритма *прогонок*, для которого в методическом плане полезно дать следующий краткий вывод формул.

Решение системы (5.214) ищем в виде рекуррентной последовательности

$$v_i = \beta_i v_{i+1} + z_i, \qquad i = N, N-1, \ldots, 1, \tag{5.221}$$

где β_i и z_i — пока неизвестные величины. Выражая с помощью (5.204) величину v_{i-1} из (5.214), приходим к соотношению $(b_i - a_i\beta_{i-1})v_i = c_i v_{i+1} + f_i + a_i z_{i-1}$. Поделив обе его части на скобку и приравнивая однотипные члены в получаемом равенстве и (5.204), мы можем утверждать, что решение в форме выписанной рекурсии действительно будет удовлетворять исходной системе, если

$$\beta_i = \frac{c_i}{b_i - a_i\beta_{i-1}}, \qquad z_i = \frac{f_i + a_i z_{i-1}}{b_i - a_i\beta_{i-1}}, \qquad i = 1, 2, \ldots, N. \tag{5.222}$$

Формулы (5.214), (5.215) и определяют знаменитый метод прогонки, причем последние составляют первый этап (прямой ход), а предыдущие – второй этап (обратный ход). Выписанные рекурсии “стартуют” автоматически, так как в силу особенностей коэффициентов системы фактически для $i = 1, N$ формулы имеют вид $\beta_1 = c_1/b_1, \quad z_1 = f_1/b_1, \quad v_N = z_N$.

Для трехдиагональных систем уравнений общего вида метод прогонки является самым экономичным алгоритмом и требует выполнения $8N$ арифметических действий. При многократном решении систем с одинаковыми матрицами и разными правыми частями общий объем операций можно сократить путем запоминания вспомогательных величин. При этом один раз рассчитываются значения

$$\beta_i = c_i d_i, \qquad d_i = (b_i - a_i\beta_{i-1})^{-1}, \qquad i = 1, \ldots, N, \tag{5.223}$$

а затем для каждой правой части прямой ход состоит только в вычислении

$$z_i = d_i(f_i + a_i z_{i-1}), \qquad i = 1, \ldots, N. \tag{5.224}$$

Формулы (5.216) и (5.217), как легко подсчитать, требуют выполнения только $5N$ действий.

Нетрудно убедиться, что если система уравнений (5.214) обладает свойством диагонального преобладания ($|b_i| \geqslant |a_i| + |c_i|$, причем хотя бы для одного $i = i_0$ неравенство строгое), то метод прогонки корректен в том смысле, что вычисления реализуемы (проходят без деления на нуль). Действительно, в этом случае по индукции легко показываются неравенства $|\beta_i| \leqslant 1$ (для $i = 1$ это очевидно) и $|d_i^{-1}| \geqslant |c_i|$, причем при $i > i_0$ неравенства строгие. Необходимо отметить, что сделанное утверждение справедливо только при абсолютно точной арифметике. Чтобы в этом убедиться, достаточно рассмотреть симметричную матрицу с элементами $b_i = 2$, $c_i = 1$ для $i = 1, \ldots, N-1$ и $b_N = a_N = 1$. При вычислениях на машине со сколь угодно малой, но конечной, погрешностью округлений найдутся такие достаточно большие N,что величина β_{N-1} будет с точностью до всех разрядов равна единице, и это приведет при расчете z_N к авосту (делению на нуль).

Если равенства (5.215) и (5.218) переписать в виде $g_i{=}d_i^{-1}$, $v_i{-}\beta_i v_{i+1}{=}z_i$, $g_i z_i - a_i z_{i-1} = f_i$, то очевидно, что они реализуют решение систем с треугольными матрицами L и U:

$$Lz = f, \qquad Uv = z, \tag{5.225}$$

которые имеют вид

$$L = \begin{bmatrix} g_1 & & 0 \\ -a_2 & \ddots & \\ & \ddots & \ddots \\ 0 & -a_N & g_N \end{bmatrix}, \quad U = \begin{bmatrix} 1 & -\beta_1 & 0 \\ & \ddots & \ddots \\ & 1 & -\beta_{N-1} \\ 0 & & 1 \end{bmatrix}$$

и представляют собой треугольные множители факторизации $A = LU$.

Если ввести некоторую диагональную матрицу $G = \operatorname{diag}\{g_i\}$, то разложение можно записать в виде соотношений $A = \bar{L}G\overline{U}$, $\bar{L} = LG^{-1}$, $\overline{U} = U$, которые в случае симметричности A дают $\bar{L}' = \overline{U}$. При этом для положительных d_i можно определить $G^{1/2} = \operatorname{diag}\{g_i^{1/2}\}$ и ввести новую факторизацию $A = \tilde{L}\tilde{L}^T$, $\tilde{L} = \bar{L}G^{1/2} = (G^{1/2}\overline{U})^T = (\widetilde{U}^T)$, являющуюся разложением Холесского. Применение его к решению систем уравнений с трехдиагональной матрицей приводит к подсистемам $\tilde{L}w = f$, $\widetilde{U}v = w$, которые в покомпонентной форме реализуются с помощью рекуррентных соотношений

$$g_1 = b_1^{1/2}, \quad g_i = (b_i - a_i^2/g_{i-1}^2)^{1/2},$$

$$w_1 = f_1/g_1, \quad w_i = (f_i + a_i g_{i-1} w_{i-1})/g_i, \quad i = 2, \ldots, N,$$

$$v_N = w_N/g_N, \quad v_i = (w_i + \beta_i v_{i+1})/g_i, \quad i = N-1, \ldots, 1. \tag{5.226}$$

Отсюда видно, что метод Холесского не дает по числу опе-

раций преимущества в сравнении с алгоритмом прогонок, хотя последний не использует свойство симметрии матрицы.

5.7.2. Метод встречных прогонок. Совершенно очевидно, что вместо рассмотренных формул прогонки (5.215), (5.218) можно взять альтернативный к ним вариант, основанный на противоположном направлении вычислений в прямом и обратном ходе. Такой метод определяется как

$$\begin{aligned} &\hat{\beta}_i = a_i\hat{d}_i, \quad \hat{z}_i = \hat{d}_i(f_i + c_i\hat{z}_{i+1}), \\ &\hat{d}_i = (b_i - c_i\hat{\beta}_{i+1})^{-1}, \quad i = N, N-1, \ldots, 1, \\ &v_i = \hat{\beta}_i v_{i-1} + \hat{z}_i, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{5.227}$$

Это совершенно эквивалентный алгоритм, не хуже и не лучше предыдущего. Основанием к выбору одного из них может быть повышение устойчивости численного решения к погрешности машинных округлений или другие соображения, но мы на этих вопросах не останавливаемся. В отличие от (5.220) его реализация сводится к другой последовательности систем с треугольными матрицами: $\widehat{U}\hat{z} = f$, $\hat{L}v = \hat{z}$, где первой решается верхняя треугольная. Отсюда видно, что такая прогонка базируется на новой факторизации, в которой левый множитель есть правая треугольная матрица, а правый — левая треугольная

$$A = \widehat{U}\hat{L},\ \widehat{U} = \begin{bmatrix} \hat{d}_1^{-1} & -c_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & -c_{N-1} \\ 0 & & & \hat{d}_N^{-1} \end{bmatrix}, \quad \hat{L} = \begin{bmatrix} 1 & & & 0 \\ -\hat{\beta}_2 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & -\hat{\beta}_N & 1 \end{bmatrix}.$$

На основе двух различных прогонок может быть построен третий алгоритм – так называемый *метод встречных прогонок* (на самом деле это семейство, отличающееся "местом встречи" альтернативных рекурсий). Сущность подхода заключается в следующем.

Пусть задано некоторое целое $1 < i_0 < N$ (его значение — свободный параметр семейства алгоритмов). Выполним прямой ход для $1 \leqslant i \leqslant i_0 - 1$ по формулам (5.217) и для $i_0 + 1 \leqslant i \leqslant N$ по формулам (5.211). Далее с помощью выражений

$$v_{i_0-1} = \beta_{i_0-1} v_{i_0} + z_{i_0-1}, \qquad v_{i_0+1} = \hat{\beta}_{i_0+1} v_{i_0} + \hat{z}_{i_0+1}$$

исключим из i_0-го уравнения неизвестные v_{i_0+1}, v_{i_0+1} и вычислим

$$v_{i_0} = \frac{f_{i_0} + a_{i_0} z_{i_0-1} + c_{i_0} \hat{z}_{i_0-1}}{b_{i_0} - a_{i_0}\beta_{i_0} - c_{i_0}\beta_{i_0+1}}. \tag{5.228}$$

После этого значения v_i для $1 \leqslant i < i_0$ находятся по формулам обратного хода (5.215), а для $i_0 < i \leqslant N$ — по формулам (5.221).

Здесь на первых шагах рекурсий формально можно положить $\beta_{i_0-1} = \hat{\beta}_{i_0+1} = 0$ и откорректировать величины z_{i_0-1} и $\hat{z}_{i_0+1}$, пересчитав их новые значения по формулам

$$z_{i_0-1} = v_{i_0-1} = d_{i_0-1}(f_{i_0-1} + c_{i_0-1}v_{i_0} + a_{i_0-1}z_{i_0-2}),$$

$$\hat{z}_{i_0+1} = v_{i_0+1} = \hat{d}_{i_0+1}(f_{i_0+1} + a_{i_0+1}v_{i_0} + c_{i_0+1}\hat{z}_{i_0+2}).$$

При этом мы фактически переходим к решению сеточных подзадач, у которых в точке i_0 уже задано условие 1-го рода и его требуется учесть в правых частях уравнений для прилегающих узлов.

Сформулированный таким образом метод встречных прогонок практически не уступает по числу операций обычному алгоритму (требуется только на 6 арифметических действий больше для расчета v_{i_0}). Целесообразность его применения может быть оправдана как повышением численной устойчивости (за счет сокращения длины рекурсии или особых свойств разностных коэффициентов), так и потенциальной возможностью распараллеливания вычислений, которые при наличии двух процессоров могут выполняться одновременно в подобластях слева и справа от точки i_0.

С методической точки зрения интересно представить матричную интерпретацию алгоритма встречных прогонок. Для этого перенумеруем последовательно компоненты векторов v, f в следующем порядке: сначала от 1-й до (i_0-1)-й, далее — от (i_0+1)-й до N-й и последнюю с номером i_0 (здесь указываются их "старые" номера). Затем векторы представим в блочной форме $v = (v_a, v_b, \bar{v}_0)$, $f = (f_a, f_b, \bar{f}_0)$, где подвекторы v_a, f_a

имеют размерность $i_0 - 1$, v_b и f_b — размерности $(N - i_0)$, а $\bar{v}_0 = v_{i_0}$, $\bar{f}_0 = f_{i_0}$ — фактически скаляры.

При этом система приводится к виду

$$\begin{bmatrix} A_a & 0 & -U_a \\ 0 & A_b & -U_b \\ -L_a & -L_b & A_0 \end{bmatrix} \begin{bmatrix} v_a \\ v_b \\ v_0 \end{bmatrix} = \begin{bmatrix} f_a \\ f_b \\ f_0 \end{bmatrix}, \tag{5.229}$$

в котором A_a, A_b — квадратные трехдиагональные подматрицы порядка $i_0 - 1$ и $N - i_0$, $L_a = (0, \ldots, 0, a_{i_0})$ и $L_b = (c_{i_0}, 0, \ldots, 0)$ — строки тех же размерностей, $U_a = (0, \ldots, 0, c_{i_0-1})^T$ и $U_b = (a_{i_0+1}, 0, \ldots, 0)^T$ — вектор-столбцы ,а $A_0 = b_{i_0}$ — "матрица первого порядка".

Для системы (5.10) проведем еще дальнейшее "укрупнение", объединив подвекторы v_a, v_b в один вектор $\bar{v} = (v_a, v_b)$ порядка $N - 1$ и соответственно определив $\bar{f} = (f_a, f_b)$. После этого получим новое блочное представление:

$$\begin{bmatrix} \bar{A} & -\overline{U} \\ -\bar{L} & A_0 \end{bmatrix} \begin{bmatrix} \bar{v} \\ \bar{v}_0 \end{bmatrix} = \begin{bmatrix} \bar{f} \\ \bar{f}_0 \end{bmatrix}, \tag{5.230}$$

где $\bar{L} = (L_a, L_b)$ и $\overline{U} = (U_a', U_b')'$ — вектор-строка и вектор-столбец размерности $N - 1$, а $\bar{A}$ — блочно-диагональная матрица с блоками A_a, A_b на диагонали.

Применим к матрице блочного второго порядка из (5.223) разложение на блочно-треугольные множители:

$$\begin{bmatrix} \bar{A} & \overline{U} \\ -\bar{L} & A_0 \end{bmatrix} = \begin{bmatrix} \bar{A} & 0 \\ -\bar{L} & G_0 \end{bmatrix} \begin{bmatrix} I & -\bar{A}^{-1}\overline{U} \\ 0 & I_0 \end{bmatrix}, \qquad (5.231)$$

$$G_0 = A_0 - \bar{L}\bar{A}^{-1}\overline{U},$$

где I, I_0 — единичные матрицы соответствующих порядков. После этого исходную систему разобьем на подсистемы

$$\begin{bmatrix} \bar{A} & 0 \\ -\bar{L} & G_0 \end{bmatrix} \begin{bmatrix} \bar{z} \\ \bar{z}_0 \end{bmatrix} = \begin{bmatrix} \bar{f} \\ \bar{f}_0 \end{bmatrix}, \quad \begin{bmatrix} E & -\bar{A}^{-1}\overline{U} \\ 0 & E_0 \end{bmatrix} \begin{bmatrix} \bar{v} \\ \bar{v}_0 \end{bmatrix} = \begin{bmatrix} \bar{z} \\ \bar{z}_0 \end{bmatrix}.$$

Как отсюда видно, для нахождения подвектора $\bar{z} = (z'_a, z'_b)'$ в силу блочно-диагональной структуры $\bar{A}$ необходимо решить подсистемы уравнений с матрицами A_a и A_b. Для первой из них применим разложение вида (5.219), а для второй – факторизацию с правой треугольной матрицей в качестве первого множителя и левой треугольной — в качестве второго:

$$\check{L}\check{U}z_a = f_a, \qquad \widehat{U}\hat{L}z_b = f_b. \qquad (5.232)$$

Введенные двухдиагональные матрицы $\check{L} = \{-a_i, d_i^{-1}\}$, $\widehat{U} = \{\hat{d}_i^{-1}, -c_i, \}$ имеют порядки $i_0 - 1$, $N - i_0$ и содержат диагональные элементы d_i^{-1}, $i = 1, \ldots, i_0 - 1$ и $\hat{d}_i^{-1}$, $i = i_0 + 1, \ldots, N$, вычисляемые с помощью формул (5.206) и (5.211) соответственно.

Вводя теперь в (5.225) векторы $\check{z} = \check{U}z_a$ и $\hat{z} = \hat{L}z_b$, легко обнаруживаем, что их компоненты $\check{z}_i$ для $i = 1, \ldots, i_0 - 1$ и $\hat{z}_i$

для $i = i_0+1, \ldots, N$ также вычисляются по формулам прямых этапов прогонок — соответственно (5.215) и (5.221).

Далее несложный анализ векторных равенств (5.215) показывает, что величина v_0 равна z_0 и в силу структуры строки $\bar{L}$ удовлетворяет соотношению

$$G_0\bar{v}_0 = f_0 + L_a z_a + L_b z_b. \tag{5.233}$$

Теперь надо вспомнить, что строки L_a и L_b содержат только по одному ненулевому элементу, так что правая часть (5.226) равна числителю дроби в (5.222).

Для вычисления $v_0 = v_{i_0}$ надо еще учесть структуру вектор-столбца $\overline{U}$, содержащего только два ненулевых элемента, а также блочно-диагональный характер матрицы $\bar{A}$ и равенство $A_0 = b_{i_0}$. Отсюда, используя факторизованные представления обратных матриц $A_a^{-1} = \check{U}^{-1}\check{L}^{-1}$, $A_b^{-1} = \widehat{U}^{-1}\hat{L}^{-1}$, можно установить, что в формуле для G_0 из (5.226) последний член представляет собой сумму двух слагаемых $\bar{L}\bar{A}_a^{-1}\overline{U} = a_{i_0}\beta_{i_0-1} + c_{i_0}\hat{\beta}_{i_0+1}$, причем множитель β_{i_0-1} есть правый нижний элемент обратной матрицы A_a^{-1}, а β_{i_0+1} — левый верхний элемент матрицы A_b^{-1} (это действительно так, поскольку диагональные элементы матриц $\check{L}^{-1}$ и $\hat{U}^{-1}$ равны соответственно β_i/c_i и $\hat{\beta}_i/a_i$, а матрицы $\check{U}^{-1}$ и $\hat{L}^{-1}$ имеют на главных диагоналях единицы). Таким образом, формулы (5.227) и (5.222) являются эквивалентными, т. е. прямой ход метода встречных прогонок описывается векторно-матричной формой (5.226).

После этого надо выполнить обратный ход метода, алгебраическая интерпретация которого состоит в нахождении подвектора $\bar{v}$ из (5.216):

$$\bar{v} = \bar{z} + \bar{A}^{-1}\overline{U}\bar{v}_0 = \bar{A}^{-1}(f + \overline{U}\bar{v}_0).$$

Подробнее это можно представить в виде

$$A_a v_b = f_a + v_0 U_a, \qquad A_b v_b = f_b + v_0 U_b, \tag{5.234}$$

откуда с учетом треугольных разложений матриц A_a и A_b нетрудно усмотреть, что оставшиеся вычисления действительно реализуются формулами обратного хода "обычных" прогонок, с предварительной корректировкой значений z_{i_0} и $\hat{z}_{i_0+1}$, так как именно для этих значений индексов векторы $v_0 U_a$ и $v_0 U_b$ имеют единственные ненулевые компоненты.

Аддитивные представления правых частей в (5.227) позволяют дать следующую интерпретацию решений подсистем на языке разностных краевых задач для $1 \leqslant i \leqslant i_0 - 1$ и $i_0 + 1 \leqslant i \leqslant N$: каждое из них есть сумма двух слагаемых $v_a = v_a^{(1)} + v_a^{(2)}$, $v_b = v_b^{(1)} + v_b^{(2)}$, где индекс 1 означает решения с правыми частями f_i и однородными (нулевыми) граничными условиями Дирихле в i_0-й точке, а индекс 2 относится к решениям с нулевыми правыми частями и значением $v_0 = v_{i_0}$ в точке раздела исходного расчетного отрезка.

Такой взгляд дает право считать рассмотренный алгоритм простейшим из методов декомпозиции области на подобласти, составляющим актуальное направление в численном решении многомерных краевых задач. А на алгебраическом языке метод встречных прогонок можно назвать методом редукции, означающим сведение решения исходной системы уравнений к решению меньших подсистем.

5.7.3. Методы редукции трехдиагональных систем.

Совершенно естественно напрашивается обобщение метода встречных прогонок путем перехода к редукции исходной системы уравнений не на две, а на произвольное число $m < N$ подсистем.

Пусть заданы номера $1 < i_1 < \ldots < i_m < N$ m точек, разделяющих сеточную область Ω_h на подобласти Ω_k, $k = 1, \ldots, m+1$, каждая из которых содержит узлы с номерами от i_{k-1} до $i_k - 1$ (полагаем $i_0 = 1$ и $i_{m+1} = N$). Обозначим через $\bar{v}_0$ подвектор m-го порядка со значениями, соответствующими узлам-разделителям, и через $\bar{v}_k$, $k = 1, \ldots, m+1$, — подвекторы с размерностями $N_k = i_k - 1 - i_{k-1}$ и компонентами из Ω_k. Вводя еще "объединенный" вектор $\bar{v} = (\bar{v}_1, \ldots, \bar{v}_{m+1})$ размерности $\overline{N} = N - m$, систему уравнений опять можно представить в блочном виде (5.224), но теперь блочно-диагональная матрица $\bar{A} = \mathrm{diag}\{A_k\}$ будет иметь блочный порядок $m+1$, а $A_0 = \mathrm{diag}\{b_{i_0}\}$ представляет собой диагональную матрицу m-го порядка. Далее, $\bar{L}$ является уже "блочной" строкой, т. е. совокупность m строк длины N, каждая из которых имеет по два ненулевых элемента. Разбивая эти строки на $m+1$ частей в соответствии с порядком матриц A, запишем (аналогично поступаем и с блочным столбцом $\overline{U}$)

$$\bar{L} = (L_1, \ldots, L_m), \qquad \overline{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_m \end{bmatrix}.$$

Каждая прямоугольная матрица L_k содержит только два

ненулевых элемента, причем в k-й строке она имеет правый ненулевой элемент a_{i_k}, а в $(k+1)$-й строке — левый ненулевой элемент c_{i_k}. Прямоугольные матрицы $\overline{U}_k$ содержат столько же ненулевых элементов, расположенных симметрично относительно ненулевых элементов L_k.

Разложение матрицы системы на блочно-треугольные множители и в данном случае имеет вид (5.225), где теперь матрица m-го порядка G_0 является трехдиагональной, что следует из выражения

$$G_0 = A_0 - \sum_{k=1}^{M} L_k A_k^{-1} U_k. \tag{5.235}$$

Как и в методе встречных прогонок, элементы G_0 будут выражаться через коэффициенты исходных уравнений и угловые элементы обратных матриц A_k^{-1}, а разделительный вектор $\bar{v}_0 = \bar{z}_0$ (см. (5.227)) находится из решения подсистемы

$$G_0 \bar{v}_0 = \bar{f}_0 + \bar{L}\bar{z}. \tag{5.236}$$

Предварительно блочные компоненты вектора $\bar{z} = (\bar{z}_1', \ldots, \bar{z}_{m+1}')'$ определяются как решение вспомогательных подзадач $A_k \bar{z}_k = \bar{f}_k$, а искомые решения в подобластях вычисляются на последнем этапе (обратный ход) по формуле

$$\bar{v}_k = A_k^{-1}(U_k \bar{v}_0 + \bar{f}_k). \tag{5.237}$$

Более наглядно вычислительный процесс можно описать в терминах прогоночных коэффициентов для решения подсистем в сформированных подобластях.

Рассмотрим совокупность трехточечных уравнений на отрезке Ω_k для $i = i_{k-1}+1, \ldots, i_k-1$, предполагая известными значениями решения $v_{i_{k-1}}$, v_{i_k} в его концевых точках. В силу линейности подсистемы ее решение $\bar{v}_k$ можно представить как сумму трех слагаемых

$$\bar{v}_k = \tilde{v}_k + v_{i_k}\check{v}_k + v_{i_{k-1}}\hat{v}_k. \tag{5.238}$$

Первое из них — решение уравнений с заданными правыми частями и нулевыми граничными условиями 1-го рода в i_{k-1}-й и i_k-й точках, а $\check{v}_k$, $\hat{v}_k$ — решения той же подсистемы с нулевыми правыми частями и следующими граничными условиями: для $\hat{v}_k$ решение равно единице на левом конце и нулю — на правом, для $\check{v}_k$ — наоборот. Естественно, что их значения находятся очень легко:

$$\check{v}_{i_k-1} = c_{i_k-1}\Big/\Big(b_{i_k-1} - a_{i_k-1}\cdot\beta_{i_k-2}\Big),\ \check{v}_i = \beta_i\check{v}_{i+1},$$

$$i = i_k-2, \ldots, i_{k-1}+1;\ \ \hat{v}_{i_{k-1}+1} = a_{i_{k-1}+1}\Big/\Big(b_{i_{k-1}+1} - c_{i_{k-1}+1}\cdot\beta_{i_{k-1}+2}\Big),$$

$$\hat{v}_i = \hat{\beta}_i\hat{v}_{i-1},\ i = i_{k-1}+2, \ldots, i_k-1,$$

где β_i, $\hat{\beta}_i$ определяются формулами (5.215), (5.221), но только для подсистемы с матрицей A_k.

Выписывая аналогичное аддитивное представление подвектора решения $\bar{v}_{k+1}$, подставим его вместе с $\bar{v}_k$ в i_k-е уравнение в точке на стыке подобластей Ω_k и Ω_{k+1}. В итоге получаем равенство

$$-\tilde{a}_k v_{i_{k-1}} + \tilde{b}_k v_{i_k} - \tilde{c}_k v_{i_{k+1}} = \tilde{f}_k \tag{5.239}$$

такого же трехчленного вида, как и в исходной системе, но связывающее на этот раз значения решения в соседних разделительных точках. Коэффициенты и правые части новой редуцированной подсистемы имеют вид

$$\tilde{b}_k = b_{i_k} - a_{i_k} v_{i_k-1} - c_{i_k} \hat{v}_{i_k+1},$$

$$\tilde{a}_k = a_{i_k} \hat{v}_{i_k-1}, \qquad \tilde{c}_k = c_{i_k} \check{v}_{i_k+1},$$

$$\tilde{f}_k = f_{i_k} + a_{i_k} \tilde{v}_{i_k-1} + c_{i_k} \tilde{v}_{i_k+1}.$$

Детальный анализ элементов матрицы G_0 и правой части уравнения (5.231) подтверждает идентичность последнего с системой (5.232).

После нахождения вектора $\bar{v}_0$ с помощью аддитивного представления (5.232) можно вычислить остальные искомые компоненты решения, формально не прибегая к выполнению обратного хода с помощью выражения (5.231). Однако фактически обратный ход все равно делается при расчете векторов как $\bar{v}_k$, так и $\check{v}_k$, $\hat{v}_k$.

Хотя в основе описанного *метода редукции* лежит все тот же алгоритм прогонки, общее количество операций в данном случае увеличивается хотя бы из-за необходимости вычисления в подобластях Ω_k дополнительных прогоночных коэффициентов, а также реализации аддитивного представления решения (5.232) (исключение составляет случай $m = 1$, но это как раз рассмотренный выше метод встречных прогонок).

Изложенный метод редукции, или декомпозиции области, может оказаться полезным при необходимости вычисления решения только в отдельных точках, а особенно – при многократном повторении этой процедуры с какими-то изменениями исходных данных. Алгоритм может применяться также для повышения устойчивости вычислений в задачах при наличии подобластей с резко меняющимися свойствами и специальными особенностями решений. И конечно, главное потенциальное качество метода — возможность гибкого распараллеливания на вычислительных системах с различным числом процессоров. За все эти качества, как мы видели, приходится платить дополнительной сложностью, и вопросы реального повышения эффективности требуют более углубленного изучения.

5.7.4. Метод циклической редукции. При осмысливании методов редукции может возникнуть интерес ко второму крайнему случаю (кроме $m = 1$), когда система разбивается на максимально возможное число подсистем. Такой случай один, и реализуется он при исключении каждой второй неизвестной.

Пусть для простоты в исходной системе (5.214) число уравнений $N = 2^k - 1$, где k — целое число. Выражая неизвестные для нечетных i с помощью соотношений

$$v_i = (f_i + a_i v_{i-1} + c_i v_{i+1})/b_i, \tag{5.240}$$

после их исключения получим подсистемы порядка $N_1 = 2^{k-1} - 1$ для четных неизвестных

$$-a_i^{(1)}v_{i-2}+b_i^{(1)}v_i-c_i^{(1)}v_{i+2}=f_i^{(1)},\qquad i=2,4,\ldots,N-1,$$

в которой новые коэффициенты выражаются формулами

$$b_i^{(1)}=b_i-a_i\frac{c_{i-1}}{b_{i-1}}-c_i\frac{a_{i+1}}{b_{i+1}},\quad a_i^{(1)}=a_i\frac{a_{i-1}}{b_{i-1}},$$
$$f_i^{(1)}=f_i+a_i\frac{f_{i-1}}{b_{i-1}}+c_i\frac{f_{i+1}}{b_{i+1}},\quad c_i^{(1)}=c_i\frac{c_{i+1}}{b_{i+1}}.$$

В этой подсистеме исключим опять каждую вторую неизвестную, получая новую редуцированную подсистему порядка $N_2=2^{k-2}-1$. Продолжая этот процесс далее, на каждом j-м этапе редукции будем иметь систему из $N_j=2^{k-j}-1$ уравнений

$$-a_{i_j}^{(j)}v_{i_j-2^j}+b_{i_j}^{(j)}v_{i_j}-c_{i_j}^{(j)}v_{i_j+2^j}=f_{i_j}^{(j)},\qquad i_j=2^j,\ldots,N_j,\tag{5.241}$$

коэффициенты которой выражаются рекуррентно по j через коэффициенты предыдущего этапа редукции:

$$\begin{aligned}
b_{i_j}^{(j)}&=b_{i_j}^{(j-1)}-a_{i_j}^{(j-1)}\frac{c_{i_j-2^{j-1}}^{(j-1)}}{b_{i_j-2^{j-1}}^{(j-1)}}-c_{i_j}^{(j-1)}\frac{a_{i_j+2^{j-1}}^{(j-1)}}{b_{i_j+2^{j-1}}^{(j-1)}},\\
a_{i_j}^{(j)}&=a_{i_j}^{(j-1)}\frac{a_{i_j-2^{j-1}}^{(j-1)}}{b_{i_j-2^{j-1}}^{(j-1)}},\qquad c_{i_j}^{(j)}=c_{i_j}^{(j-1)}\frac{c_{i_j+2^{j-1}}^{(j-1)}}{b_{i_j+2^{j-1}}^{(j-1)}},\\
f_{i_j}^{(j)}&=f_{i_j}^{(j-1)}+a_{i_j}^{(j-1)}\frac{f_{i_j-2^{j-1}}^{(j-1)}}{b_{i_j-2^{j-1}}^{(j-1)}}+c_{i_j}^{(j-1)}\frac{f_{i_j+2^{j-1}}^{(j-1)}}{b_{i_j+2^{j-1}}^{(j-1)}}.
\end{aligned}\tag{5.242}$$

На последнем этапе редукции получим одно уравнение относительно неизвестной с номером $(N+1)/2$. После его вычисления начинается обратный ход редукции с поэтапным

определением остальных неизвестных по аналогичным (5.234) формулам, полученным из уравнений (5.235):

$$v_{i_j} = \left(f_{i_j}^{(j-1)} + a_{i_j}^{(j)} v_{i_j-2^j} + c_{i_j}^{(j)} v_{i_j+2^j}\right)/b_{i_j}^{(j)}. \tag{5.243}$$

В блочной интерпретации *метод циклической редукции* также представляется формулами (5.224), (5.225). В данном случае подвектор $\bar{v}$ состоит из нечетных компонент вектора v, а $\bar{v}_0$ — из четных. Матрицы $\bar{A}$ и A_0 — диагональные, а $\overline{U}$ и $\bar{L}$ содержат максимум по два ненулевых элемента в каждой строке. Матрица G_0 оказывается трехдиагональной порядка N_1, для решения системы (5.227) аналогично предыдущему проводится разбиение вектора $\bar{v}_0$ на подвекторы с четными и нечетными компонентами, и далее процесс редукции продолжается “в глубину”.

Простой подсчет числа арифметических операций в формулах (5.236), (5.237) показывает, что данный вычислительный процесс значительно проигрывает методу прогонки. С точки зрения распараллеливания этот алгоритм также обладает существенным недостатком: если имеется в наличии, например, N_1 процессоров, то на первом этапе редукции они могут быть эффективно использованы для пересчета коэффициентов вида (5.225), однако на последующих этапах остается все меньше и меньше уравнений и значительное количество процессоров будет простаивать.

Для устранения этого эффекта можно предложить *метод циклической редукции без обратного хода*, который заключается в следующем. На первом этапе поочередно исключаются

четные и нечетные неизвестные, в результате чего получаются две независимые подсистемы. На втором каждая из них, в свою очередь, редуцируется аналогичным образом, что приводит к четырем независимым подсистемам. На k-м этапе такого процесса формируется N уравнений с одним неизвестным, и на их вычислениях алгоритм заканчивается.

В таком методе при наличии N процессоров на каждом этапе все они будут задействованы, а данная задача служит примером того, как с точки зрения быстрейшего ее решения на многопроцессорной вычислительной системе может оказаться целесообразным выбирать алгоритм, увеличивающий общее число арифметических действий.

5.7.5. Метод циклической прогонки. Под таким названием в литературе закрепился метод решения системы трехточечных уравнений типа (5.171) с циклической матрицей A_d, которая возникает при аппроксимации уравнения Пуассона с периодическими краевыми условиями. По сути, данный алгоритм основан на редукции СЛАУ, который использовался и в предыдущих пунктах 5.7.2 — 5.7.4.

Мы рассмотрим задачу в несколько более общей постановке, предполагая, что некоторое дифференциальное уравнение вида (5.1) задано на интервале $x \in (a, b)$ с периодическими граничными условиями $u(a) = u(b)$ или $u(x) = u(x + b - a)$. При этом аппроксимация осуществляется на сетке (не обязательно равномерной)

$$a = x_0 < x_1 < ... < x_N = b,$$

а решаемая невырожденная СЛАУ с диагональным преобла-

данием записывается в виде

$$\begin{gathered} -a_i u_{i-1} + b_i u_i - c_i u_{i+1} = f_i, \\ |b_i| \geqslant |a_i| + |c|_i, \;\; i = 1, ..., N; \;\; u_0 = u_N, \; u_1 = u_{N+1}, \end{gathered} \tag{5.244}$$

причем в последних неравенствах хотя бы для одного i неравенство является строгим.

Сделаем временное предположение, что величина u_N нам известна (тоже можно было бы сделать и относительно любой другой компоненты решения u_i), и для "соседних" уравнений с номерами $N-1$ и 1 соответствующие члены перенесем в правую часть:

$$\begin{gathered} -a_i u_{i-1} + b_i u_i - c_i u_{i+1} = f_i, \;\; i = 2, ..., N-2, \\ b_1 u_1 - c_1 u_2 = f_1 + a_1 u_N, \\ -a_{N-1} u_{N-2} + b_{N-1} u_{N-1} = f_{N-1} + c_{N-1} u_N. \end{gathered} \tag{5.245}$$

В силу линейности системы (5.238) к ее решению применим следующий принцип суперпозиции. Введем две вспомогательные СЛАУ с решениями $u_i^{(0)}$ и $u_i^{(1)}$ следующего вида:

$$\begin{gathered} -a_i u_{i-1}^{(0)} + b_i u_i^{(0)} - c_i u_{i+1}^{(0)} = 0, \;\; i = 2, ..., N-2, \\ b_1 u_1^{(0)} - c_1 u_2^{(0)} = a_1, \;\; -a_{N-1} u_{N-2}^{(0)} + b_{N-1} u_{N-1}^{(0)} = c_{N-1} \end{gathered} \tag{5.246}$$

$$-a_i u_{i-1}^{(1)} + b_i u_i^{(1)} - c_i u_{i+1}^{(1)} = f_i, \;\; i = 2, ..., N-2,$$

$$b_1 u_1^{(1)} - c_1 u_2^{(1)} = f_1, \;\; -a_{N-1} u_{N-2}^{(1)} + b_{N-1} u_{N-1}^{(1)} = f_{N-1}. \tag{5.247}$$

Очевидно, что искомое решение исходной системы представимо в форме

$$u_i = u_i^{(1)} + u_N \cdot u_i^{(0)}, \;\; i = 1, ..., N-1. \tag{5.248}$$

Чтобы в этом убедиться, достаточно все уравнения (5.239) умножить на u_N и сложить с соответствующими уравнениями (5.240), в результате чего приходим (с учетом (5.241)) к уравнениям (5.238).

Вспомогательные решения $u_i^{(0)}$ и $u_i^{(1)}$ легко находятся с помощью "стандартных" формул прогонки вида (5.214), (5.215) из решения СЛАУ (5.239), (5.240) $N-1$-го порядка. Можно только отметить, что поскольку у этих систем матрицы одинаковые, то одинаковыми будут и прогоночные коэффициенты β_i.

Для нахождения пока неизвестного u_N подставим представление (5.241) в N-е уравнение исходной системы:

$$-a_N(u_{N-1}^{(1)} + u_N u_{N-1}^{(0)}) + b_N u_N - c_N(u_1^{(1)} + u_N u_1^{(0)}) = f_N,$$

откуда получаем необходимое значение

$$u_N = (f_N + a_N u_{N-1}^{(1)} + c_N u_1^{(1)})/(b_N - a_N u_{N-1}^{(0)} - c_N u_1^{(0)}), \tag{5.249}$$

с помощью которого искомое решение находится из формулы (5.241).

Можно показать, что если матрица системы (5.238) невырожденная, то знаменатель дроби в (5.242) не обращается в нуль.

Легко также видеть, что нахождение вспомогательных величин $u_i^{(0)}, u_i^{(1)}$ без труда распараллеливаются на двух вычислительных потоках (или процессорах).

5.7.6. Метод прогонки для решения ленточных систем. Рассмотрим систему алгебраических уравнений с ленточной матрицей, имеющей ширину полуполосы m и записанной в виде

$$a_{i,i-m}u_{i-m} + ... + a_{i,i-1}u_{i-1} + a_{i,i}u_i + a_{i,i+1}u_{i+1} + ... + a_{i,i+m}u_{i+m} = f_i,$$

$$i = 1, ..., N;\ a_{i,j} = 0 \text{ при } \max(1, i-m) > j > \min(N, i+m).$$

Такая СЛАУ может рассматриваться как система $(2m+1)$-точечных сеточных уравнений с матрицей

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} & 0 & \cdots \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,m+1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ a_{m,1} & & & & & a_{N-m,N} \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ & & & & & a_{N-1,N} \\ 0 & & a_{N,N-m} & \cdots & a_{N,N-1} & a_{N,N} \end{bmatrix}.$$

В частности, при $m = 2$ отсюда имеем пятиточечные уравнения Лапласа (5.25), если положить

$$a_{i,i} = 30, \quad a_{i,i-1} = a_{i,i+1} = -16, \quad a_{i,i-2} = a_{i,i+2} = 1.$$

Решение $(2m+1)$-диагональной СЛАУ ищем, после введения вспомогательных величин z_i с помощью двух рекуррентных соотношений

$$u_i = z_i - \beta_{i,1}u_{i+1} + ... + \beta_{i,m}u_{i+m}, \;\; i = N, N-1, ..., 1,$$

$$z_i = f_i - \alpha_{i,1}z_{i-1} - ... - \alpha_{i,m}z_{i-m}/\alpha_{i,0}, \;\; i = 1, 2, ..., N,$$

первое из которых называется обратной прогонкой, а второе — прямой.

Для нахождения неизвестных пока прогоночных коэффициентов $\beta_{i,j}$ и $\alpha_{i,j}$ выражаем z_i из первых соотношений и подставляем их во вторые, в результате чего получаем

$$\alpha_{i,0}(u_i + \beta_{i,1}u_{i+1} + ... + \beta_{i,m}u_{i+m}) =$$

$$= \alpha_{i,1}(u_{i-1} + \beta_{i-1,1}u_i + ... + \beta_{i-1,m}u_{i+m-1}) + ...$$

$$\alpha_{i,m}(u_{i-m} + \beta_{i-m,1}u_{i-m+1} + ... + \beta_{i-m,m}u_i) + f_i.$$

Требуя эквивалентности полученного уравнения с исходным, после приравнивания коэфициентов при соответствующих членах, приходим к рекуррентным формулам, легко реализуемым для $i = 1, 2, ..., N$, в указанной ниже последовательности:

$$\alpha_{i,m} = a_{i,i-m}, \;\; \alpha_{i,m-1} = a_{i,i-m+1} - \alpha_{i,m}\beta_{i-m,1}...,$$

$$\alpha_{i,1} = a_{i,i-1} - \alpha_{i,2}\beta_{i-2,1} - ... - \alpha_{i,m}\beta_{i-m,m-1},$$

$$\alpha_{i,0} = (a_{i,i} - \alpha_{i,1}\beta_{i-1,1} - ... - \alpha_{i,m}\beta_{i-m,m})^{-1},$$

$$\beta_{i,m} = a_{i,i+m}/\alpha_{i,0}, \;\; \beta_{i,m-1} = (a_{i,i+m-1} - \alpha_{i,1}\beta_{i-1,m})/\alpha_{i,0}, ...,$$

$$\beta_{i,1} = (a_{i,i+1} - \alpha_{i,1}\beta_{i-1,2} - ... - \alpha_{i,m}\beta_{i-m,m-1})/\alpha_{i,0}.$$

Естественно, здесь предполагается, что в указанных арифметических действиях отсутствует деление на нуль. Отметим ту особенность данных рекурсий, что в них формально неопре-

деленные величины умножаются на нуль (но только при реализации в представленном порядке: сначала для $i = 1$ вычисляются $\alpha_{i,m}, ..., \alpha_{i,1}, \alpha_{i,0}, \beta_{i,m}, ..., \beta_{i,1}$, а затем все повторяется для $i = 2$ и т. д.).

Приведенные формулы определяют *метод* $(2m + 1)$-*точечных прогонок*. Легко проверить, что при $m = 1$ они переходят в алгоритм трехточечных прогонок.

Как нетрудно увидеть, рассмотренные формулы прямой и обратной прогонок можно записать в матричной форме

$$L\,z = f,\ \ U\,u = z,$$

где L и U — нижняя и верхняя треугольные ленточные матрицы, которые реализуют разложение исходной матрицы на треугольные множители ($A = L\,U$) и имеют следующий вид:

$$L = \begin{bmatrix} \alpha_{1,0} & 0 & & & \\ \alpha_{2,1} & \alpha_{2,0} & & & 0 \\ \vdots & \ddots & \ddots & & \\ \alpha_{i+m,m} & & & & \\ 0 & \alpha_{N,m} & \cdots & \alpha_{N,1} & \alpha_{N,0} \end{bmatrix},$$

$$U = \begin{bmatrix} 1 & \beta_{1,1} & \cdots & \beta_{1,m} & & 0 \\ & \ddots & & \ddots & & \\ & & & & \beta_{N-m,m} & \\ & & & & \vdots & \\ 0 & & & & \beta_{N-1,1} & \\ & & & & 1 & \end{bmatrix}.$$

5.7.7. Анализ устойчивости метода прогонки. Рассмотрим невырожденную трехдиагональную СЛАУ с диагональным преобладанием

$$-a_i u_{i-1} + b_i u_i - c_i u_{i+1},\ i = 1, ..., N,$$

$$a_1 = c_N = 0,\ |b_i| \geqslant |a_i| + |c_i| \geqslant 0,$$

где для определенности будем считать, что в последнем неравенстве при $i = 1$ имеет место строгое неравенство.

Пусть при точных арифметических вычислениях решение вычисляется последовательно по формулам прямой и обратной "стандартной" прогонки

$$\beta_i = \frac{c_i}{b_i - a_i \beta_{i-1}},\ \ z_i = \frac{f_i + a_i z_{i-1}}{b_i - a_i \beta_{i-1}},\ \ i = 1, ..., N,$$

$$u_i = \beta_i u_{i+1} + z_i,\ \ i = N, N-1, ..., 1.$$

Отметим прежде всего, что в силу свойства диагонального преобладания прогоночные коэффициенты β_i удовлетворяют неравенствам

$$|\beta_1| < 1, \;\; |\beta_i| \leqslant \frac{|c_i|}{|b_i| - |a_i| \cdot |\beta_{i-1}|} < 1, \;\; i = 1, 2, ..., N-1, \;\; \beta_N = 0.$$

Предположим теперь, что вследствие выполнения арифметических операций с конечным числом знаков вместо точных величин β_i вычисляются их приближенные значения $\bar{\beta}_i = \beta_i + \delta\beta_i$, в соответствии с формулами

$$\bar{\beta}_{i-1} = \beta_{i-1} + \delta\beta_{i-1}, \;\; \bar{\beta}_{i-1} = \frac{c_i}{b_i - a_i\bar{\beta}_{i-1}} + \varepsilon_i,$$

где ε_i — величина погрешности, обусловленная наличием округлений при реализации дроби. Отсюда для $\delta\beta_i$ следует рекуррентное соотношение

$$\delta\beta_i = \beta_i\gamma_i\frac{\delta\beta_{i-1}}{1 - \gamma_i\delta\beta_{i-1}} + \varepsilon_i, \;\; \gamma_i = \beta_i a_i/c_i.$$

Делая далее предположения

$$\varepsilon = \max_i\{|\varepsilon_i|\}, \;\; |\delta\beta_1| \leqslant \varepsilon,$$

для модулей ошибок $|\delta\beta_i|$ получаем неравенства

$$|\delta\beta_i| \leqslant \frac{|\gamma_i||\delta\beta_{i-1}|}{(1 - |\gamma_i| \cdot |\delta\beta_{i-1}|} + \varepsilon, \;\; i = 2, 3, ...$$

Допуская теперь выполнение естественного при $\varepsilon \ll 1$ условия

$$\varkappa = \max_k \{\varkappa_k = \prod_{i=2}^{k} |\gamma_i|\},$$

$$\varepsilon N(N-1)\varkappa/2 \le 1 - 1/C, \;\; C > 1,$$

получаем итоговую оценку для ошибок вычисления прогоночных коэффициентов:

$$|\delta\beta_i| \leqslant C\varepsilon\varkappa i \leqslant C\varepsilon\varkappa N, \;\; i = 2, ..., N-1.$$

Таким образом, погрешность расчета прогоночных коэффициентов β_i пропорциональна величине ошибки машинных операций ε и пропорциональна размерности N решаемой СЛАУ.

Проведем теперь аналогичным образом прямой анализ погрешностей приближенных вычислений величин $\bar{z}_i = z_i + \delta z_i$:

$$\bar{z}_i = z_i + \delta z_i = \frac{f_i + a_i \bar{z}_{i-1}}{b_i - a_i \bar{\beta}_{i-1}} + \hat{\varepsilon}_i,$$

где $\hat{\varepsilon}_i$ — ошибки округлений при реализации дробей с использованием значений $\bar{z}_{i-1}$ и $\bar{\beta}_{i-1}$.

Вводя обозначения $\hat{\varepsilon} = \max\limits_i \{|\hat{\varepsilon}_i|\}$, для ошибок δz_i получаем рекурсии

$$\delta z_1 = \hat{\varepsilon}_1, \;\; \delta z_i = \gamma_i \delta z_{i-1} + \alpha_i \delta\beta_i + \hat{\varepsilon}_i, \;\; i = 2, ..., N-1,$$

$$\alpha_i = (f_i + a_i \bar{z}_{i-1})/c_i,$$

из которых следуют оценки

$$|\delta z_i| \leqslant \gamma_i|\delta z_{i-1}| + C_1\varkappa_i\varepsilon i,\ \ C_1 = \max_i\{|\alpha_i|\} + 1,$$

$$|\delta z_i| \leqslant \varkappa\frac{i(i+1)}{2}C_1\hat{\varepsilon} \leqslant \varkappa C_1 N(N+1)\hat{\varepsilon}/2.$$

В соответствии с данными неравенствами машинные округления в расчетах z_i накапливаются уже квадратичным образом по N.

И наконец, итоговая ошибка реализации обратной прогонки определяется соотношениями

$$\bar{u}_i = u_i + \delta u_i = \bar{\beta}_i\bar{u}_{i+1} + \bar{z}_i + \check{\varepsilon}_i,\ i = N-1, ..., 1,\ \bar{u}_N = \bar{z}_N,$$

где $\check{\varepsilon}_i$ есть фактически ошибка выполнения одного умножения и одного сложения. Используя здесь полученные выше представления для $\bar{\beta}_i$ и $\bar{z}_i$, получаем следующие равенства и неравенства:

$$\delta u_N = \delta z_N,\ \delta u_i = \beta_i\delta u_{i+1} + \delta z_i + \delta\beta_i\bar{u}_{i-1} + \check{\varepsilon}_i,\ i = N-1, ..., 1,$$

$$|\delta u_i| \leqslant |\beta_i||\delta u_{i+1}| + \bar{\varepsilon},\ \bar{\varepsilon} = \max_i\{|\delta z_i| + |\delta\beta_i|\cdot|\bar{u}_{i+1}| + |\check{\varepsilon}_i|\},$$

$$|\delta u_i|_i \leqslant C_2N^3\tilde{\varepsilon},\ \tilde{\varepsilon} = \max_i\{|\varepsilon_i|, |\hat{\varepsilon}_i|, |\check{\varepsilon}_i|\}.$$

Полученный конечный результат свидетельствует, что оценка погрешности численной реализации метода прогонки пропорциональна N^3 и величине $\tilde{\varepsilon}$, определяемой ошибкой арифметических действий.

Данный факт свидетельствует об относительном характере понятия устойчивости. С одной стороны, для

фиксированной величине $\tilde{\varepsilon}$ найдется достаточно большое N, когда ошибка решения будет заведомо неприемлемой. И наоборот, для заданного сколь угодно большого порядка СЛАУ N можно подобрать точность машинной арифметики такую, что погрешность численного решения будет меньше любой требуемой величины.

§ 5.8. Задачи к главе 5

5.8.1. Доказать, что подобные матрицы имеют одинаковые след и определитель.

5.8.2. Пусть матрицы A и $B = P^{-1}AP$ подобны. Однозначно ли определена трансформирующая матрица P?

5.8.3. Показать, что скалярная матрица αI подобна лишь самой себе.

5.8.4. Доказать, что нильпотентная матрица не имеет отличных от нуля собственных значений.

5.8.5. Показать, что матрица проектирования имеет простую структуру (т. е. подобна диагональной матрице).

5.8.6. Доказать, что всякий многочлен $P(A)$ от оператора простой структуры сам имеет простую структуру.

5.8.7. Доказать, что ранг оператора проектирования равен его следу.

5.8.8. Найти собственные числа и векторы матриц

$$A = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 3+i & -1 \\ 2i & 1-i \end{bmatrix}.$$

5.8.9. Доказать, что если матрицы A и B перестановочны, то перестановочны и транспонированные матрицы A^T, B^T.

5.8.10. Показать, что оператор, сопряженный к оператору проектирования, сам является оператором проектирования.

5.8.11. Доказать, что всякий оператор ортогонального проектирования является отрицательно полуопределенным.

5.8.12. Показать, что в положительно определенной матрице максимальный элемент стоит на главной диагонали.

5.8.13. Найти квадратный корень матрицы

$$A = \begin{bmatrix} 4 & -3 \\ -3 & 4 \end{bmatrix}.$$

5.8.14. Доказать, что для минимального по модулю собственного числа λ_N и минимального сингулярного числа α_N матрицы $A \in \mathbb{R}^{N,N}$ выполняется соотношение $|\lambda_N| \geqslant \alpha_N$.

5.8.15. Пусть A_k — произвольная главная подматрица с.п.о. матрицы A. Доказать, что

$$\mathrm{cond}_2(A_k) \leqslant \mathrm{cond}(A).$$

5.8.16. Доказать, что для спектрального радиуса $\rho(A)$ матрицы A выполняется неравенство

$$\rho(A) \leqslant ||A||,$$

какова бы ни была матричная норма $||A||$.

5.8.17. Доказать, что ранг кососимметричной матрицы есть четное число.

5.8.18. Показать, что сумма и произведение M-матриц не обязаны быть M-матрицей.

5.8.19. Показать, что произведение матриц положительного типа не обязано быть матрицей положительного типа.

5.8.20. Показать, что неположительность внедиагональных элементов трехдиагональной матрицы является необходимым условием ее монотонности.

5.8.21. Доказать, что всякая норма матрицы согласована с некоторой нормой вектора.

5.8.22. Показать, что спектральный радиус нормальной матрицы равен $\rho(A) = ||A||_2$.

5.8.23. Доказать, что у вещественной ортогональной матрицы все собственные числа по модулю равны единице.

5.8.24. Показать, что если A и B — с.п.о. матрицы, то собственные числа матрицы AB положительны.

5.7.25. Доказать, что для произвольных матриц A, B собственные числа $A\,B$ и $B\,A$ совпадают.

5.8.26. Показать, что для с.п.о. — матрицы A величина $\mathrm{cond}_2(A + \alpha I)$ есть монотонно убывающая функция от α при $\alpha > 0$.

Предметный указатель

B-сходящимся, 246
B-согласованным, 246
-скалярное произведение, 595
k-шаговый метод, 111
Алгоритмы оптимальные по порядку, 118
Аппроксимация
  монотонная, 475
  Паде, 221
  Шортли–Уэллера, 470
Асимптотически устойчивым, 68
Автономные ОДУ, 19
Базис собственный, 76
Блочный метод Якоби—Ньютона, 187
Частное решение, 20
Число Куранта, 383
Диагональное преобладание
  по столбцам, 578
  строгое, 578
Дифференциальная прогонка, 453
Дифференциальное уравнение в полных дифференцилах, 24
Дискретные уравнения теплопроводности, 29
Двухшаговая схема Симпсона, 146
Экстремальные свойства собственных чисел, 595
Фазовая плоскость, 42
Фазовая траектория, 42
Фазовый портрет, 42
Фокус, 46
Форма
  жордановая, 47
Формула
  Эйлера, 60
Формула
  Лобатто, 216
  Радо, 214
Формула Тейлора, 109
Фундаментальная матрица системы ОДУ, 54
Функции от матрицы, 594
Функционал действия, 31
Функционал гамильтониан, 35

Функция
абсолютно монотонная, 237
контрактивно в евклидовой норме, 236
Функция A—допустимой, 141
Функция Лагранжа, 30
Функция Ляпунова, 70
Функция диссипативная, 90
Функция лагранжианом, 30
Функция относительной устойчивости, 140
Функция устойчивости, 140
Гамильтоновая система, 35
Гамильтоновы системы, 34
Идекс дифференциально-алгебраического уравнения, 42
Индекс нильпотентности пучка матриц , 42
Интеграл
по правилу трапеций, 213
Интеграл абсолютной устойчивости, 139
Интеграл уравнения Эйлера, 32
Интеграл вариационной (функциональной) производной, 32
Интерполяция
кубическая эрмитовая, 547
Инвариант, 35
Канонические переменные, 36
Канонические преобразования, 35
Коэффициент
неравномерности сетки, 521
Коэффициенты
узлами и весами МРК, 193
Коэффициенты уравнения с разделяющимися переменными, 23
Коллокационные методы, 212
Комбинация
трехточечная, 458
Константа
погрешности ММ, 331
Корневое условие, 136
Кривая локуса корней, 344
Круги Гершгорина, 605
Квадратичная интерполяция, 469
Квадратичная сходимость, 177
Линейная интерполяция, 469
Логарифмическая норма матрицы, 90
Масштабирующий множитель, 580

Матрица
Минковского, 610
положительного типа, 582
разложимая, 158
Стилтьеса, 583
трехдиагональная, 574
Матрица
баланса, 489
глобальная, 489
ленточная, 592
локальная, 489
Минковского, 610
нормальная, 593
нулевого порядка, 469
Островского, 610
перестановок, 577
строго, или вполне, невырожденная, 158
теплицевая, 590
верхняя хессенберговая, 597
Якоби, 39
Матрица резольвентная, 54
Матрица Грина системы ОДУ, 54
Матрица Якоби, 66
Матрица диагонализируемая, 76
Матрица нормальная, 60
Матрица перехода, 132, 143
Матрица плохо обусловленная, 155
Матрица положительно определенная, 153
Матрица симметричная положительно определенная, 153
Матрица симплектическая, 36
Матричная экспонента, 59
Метод
циклической редукции, 642
циклической редукции без обратного хода, 643
редукции, 639
встречных прогонок, 628
интегро-балансный, 505
интегро-интерполяционный, 502
исключения Гаусса, 158
конечных объемов МКО, 488
Метод
одноопорный, 324
$(2m+1)$-точечных прогонок, 650
A-контрактивный, 368
G-устойчивый, 360

бисекции, 171
Галеркина, 522, 529
Кенига, 184
конечных элементов, 449
Милна, 320
Ньютона, 175
Нюстрема, 294, 319
одновременных смещений, 185
последовательных смещений, 185
прогноза и коррекции, 318
простой итерации, 173
пространства состояний, 298
ПВР—Ньютона, 186
Ритца, 522
Метод A-устойчивый, 139
Метод Рунге—Кутты, 119
Метод линеаризации, 74
Метод неявный, 115
Метод секущих, 180
Метод трапеций, 317
Метод устойчивый, 136
Метод вариации постоянных, 26
Метод явный, 116
Методы Адамса–Башфорта, 312
Многошаговые методы, 111
Множитель интегрирующий, 25
Неустойчивый фокус, 46
Неустойчивый узел, 44
Неявные методы Адамса, 316
Норма
-норма, 595
Норма
подчиненная матричная, 153
Норма евклидова, 74
Норма сферическая, вектора, 74
ОДУ с постоянными коэффициентами, 20
Обыкновенное дифференциальное уравнение, 17
Область
абсолютной устойчивости, 343
Область абсолютной устойчивости, 140
Область относительной устойчивости, 140
Обобщенная задача Маркова, 386
Общее решение, 20
Одноопорный метод, 356
Одношаговые методы, 111
Одношаговый метод L-

устойчивый, 141
Оператор
разностный или конечно-разностный, 465
Оператор диагональный, 76
Определение
Гир, 351
Видлунд, 351
Отношение Рэлея, 595
Отображение
сжимающее, 172
Параметр бифуркации, 97
Переходные процессы, 79
Переменная обобщенный импульс, 34
Первая теорема Батчера, 210
Пикара., 58
Пограничный слой, 82
Погрешность аппроксимации, 198
Погрешность локальная, 198
Погрешность локальной ошибки, 112
Полином Чебышева 1-го рода, 386
Полином второй характеристический многочлен, 145
Порядок
с неявным правиом средней точки, 213
стадийный, 243
уравнения, 17
Постоянная Липшица, 50
Предельный цикл, 94
Представление монотонной матрицы, 615
Преобразование Лежандра, 35
Преобразование конгруэнтности, 598
Преобразование подобия, 47
Принцип Гамильтона, 34
Принцип наименьшего действия, 34
Принцип суперпозиции, 21
Приведенная система, 38
Производящая функция, 35
Пространство интегральной кривой, 20
Прямая интерполяция, 180
Разложение
Холесского, 160
Разность
односторонняя, 476
Решение
финитных функций, 521
неподвижной точки, 172
системы ОДУ, 20
Решение устойчивое, 68
Решение задачи Коши, 20
Семейство

неявных ФДН, 323
неявных методов Штермера, 376
явных k-шаговых методов Штермера, 375
Сетка
элементно-ориентированные, 512
точечно-ориентированные, 512
Сетка квазиравномерная, 107
Сетка неравномерная, 107
Сетка равномерная, 107
Сетка регулярная, 107
Сетка сходящаяся, 107
Сетка вложенная, 107
Схемы
балансными, 448
компактные разностные , 481
консервативными, конечно-разностными, 448
Симметризация, 581
Система уравнений
симметричная, 579
положительного типа, 473
симметризуемая, 580
якобиевая, 581
Система жесткая, 81
Соотношения
методами конечных объемов, 448
Спектральный радиус матрицы, 611
Стандартный метод прогонок., 623
Существенным или главным, 452
Свойство контрактивности, 89
Таблица
Батчера, 196
Теорема
эквивалентности коллокационных МРК, 286
принцип максимума, 619
сравнения, 616
Теорема
абсолютной устойчивости, 344
алгебраической устойчивости, 237
Далквист., 358
Дж. фон Нейман, 236
эквивалентности A- и G-устойчивости., 360

эквивалентности ООММ и ММ, Далквист., 357
эквивалентности разрывных коллокационных МРК, 290
Фробениуса, 612
Гершгорина, 597
Хайрер, 234
Като., 367
Кенига, 184
Крайсс., 367
необходимых условий сходимости ММ, 341
о B-сходимости, 247
о B-согласованности НМРК, 245
о достаточном условии B-устойчивости, 239
о единственности НМРК, 242
о контрактивности МРК, 237
о квазирегулярном разложении , 615
о норме матрицы, обратной к монотонной, 613
о порядке сходимости и оценках погрешностей коллокационных МКР, 290
о возмущениях НМРК, 241
О. Аксельсон, Л. Ю. Колотилина, 616
об интегральном представлении локальной погрешности ММ, 332
об оценки устойчивости, 246
первый барьер Далквиста, 339
погрешности производных в коллокационных МРК, 288
Пуанкаре, 94, 274
сходимости коллокационных МРК, 288
симметричности ММ, 340
сравнения, 616
существования НМРК, 240
Штермера—Верле, 278
Шура, 596
Тауски, 579
условия порядка ММ, 327
устойчивости ФДН-

методов, 338
Варги, 614
Вогеларе, 276
второй барьер Далквиста, 350
закона инерции, 599
Теорема Далквиста., 92
Теорема Грёбнера., 65
Теорема Лиувилля., 54
Теорема Ляпунова., 71
Теорема дифференцируемости решений по параметру., 64
Теорема эквивалентности Лакса, 138
Теорема линейной зависимости или независимости системы векторных функций., 53
Теорема о гладкости решений ОДУ., 48
Теорема существование и единственность решений линейных ОДУ., 49
Теорема существования и единственности, 48
Теорема устойчивости системы с постоянными коэффициентами., 71
Точка покая, 43
Точка покоя, 72
Точка положение равновесия системы, 43
Третья теорема Батчера, 234
Уравнение трехточечное, 472
Уравнение Деккера–Вервера, 85
Ван-дер-Поля, 93
Уравнения Далквиста, 145
линейные, 19
Лотки—Вольтерра, 100
Уравнения Кеплера, 36
Уравнения Лагранжа, 30
Уравнения Ван-дер-Поля, 37
Уравнения движения, 27
Уравнения колебаний, 28
Уравнения лагранжевой механики, 30
Уравнения сингулярно возмущенные, 83
Уравнения сингулярно-возмущенные и дифференциально-алгебраические, 37
Условие

I-устойчивости, 221
3-го рода, или Робена, 442
первого рода или Дирихле, 441
Условия
нераспадающиеся, 439
распадающиеся, 439
согласованности, 328
Условия Липшица, 49
Устойчивый фокус, 46
Устойчивый узел, 44
Утверждение
B-устойчивость $\Longrightarrow$ AN-устойчивость $\Longrightarrow$ A-устойчивость, 239
Вектор
сеточного решения, 480
Вектор обобщенных координат, 31
Вектор обобщенных скоростей, 31
Векторное поле, 21
Величина
псевдоравномерная, 521
Величина локальной ошибки, 111
Вронскианский определитель, 52
Вторая теорема Батчера, 211
Ядро
Пеано, 333
Явные методы Адамса, 312
Задача
двухточечная, 440
изоклан, 22
Коши, 18
многоточечная, 440
Задача Коши асимптотически устойчива для системы ОДУ, 69
Задача абсолютно устойчивая, 67
Задачи
некорректные, 443
плохо обусловленные, 443
сингулярно-возмущенные, 446
условно корректные, 443
Жесткая задача, 79
абсолютно устойчивым с шагом h, 135
алгебраически устойчивый, 234
аппроксимация экспоненциального типа, 511
аппроксимация первой производной, 458
аттрактора, 95

бифуркационными параметрами, 97
бифуркация Хопфа, 97
билинейную форму, 450
блочный метод ПВР, 186
блочный метод Якоби, 187
центральная разность, 457
числа Фибоначчи, 182
число стадий, 119
диагональная аппроксимация Паде, 221
двусторонние приближения, 128
двусторонняя разность первого порядка, 456
естественным, 452
экстраполяцией Ричардсона, 127
экстраполяция, 126
эрмитовой интерполяции, 547
характеристическим полиномом, 145
иммерсионные конечные элементы, 566
интервал интегрирования, 19
коэффициент неравномерности, 107
коллокационные точеки, 285
коллокационный многочлен, 285
корректор, 123
квазиравномерными, 455
линейные многошаговые методы, 116, 309
метод ε-вложения, 297
методы переменных направлений, 187
многочлен Тейлора, 109
многошаговые методы, 371
модельная задача Протеро–Робинсона, 242
неприводимыми, 337
неявные дифференциальные уравнения, 226
неявные формулы дифференцирования назад, 323
неявный метод Эйлера, 110
норма Гельдера, 153
норма кубическая (равномерная), 153
норма октаэдрическая, 153
норма сферическая, 153
область абсолютной устойчивости, 136
обратная интерполяция, 180
общие линейные методы или многошаговые методы Рунге—Кутты, 123
общими линейными методами, 369

одностороннее условие Липшица, 87
односторонние разности, 456
онечных разностей первого порядка, 456
ориентированная площадь, 270
первый характеристический полином, 136
пограничный слой, 446
погрешность аппроксимации, 111
пороговым коэффициентом, 237
порядок точности метода, 115
последовательная верхняя релаксация, 185
последовательности сгущающихся сеток, 107
правила Рунге, 128
правило средней точки, 319
предиктор, 123
предиктор-корректор, 318
прогноза и коррекции, 123
пространством пробных функций, 451
размерность системы, 19
разность назад, 456
разность вперед, 456
разность второго порядка, 459
разрывным коллокационным, 289
регулярным аттрактором, 96
релаксационный параметр, 185
сетка, 106
сеточным шаблоном, 463
скалярные уравнения, 20
скалярное произведение, 450
слабой, постановкой, 450
спектральное число обусловленности, 155
спектральное разложение, 593
странный аттрактор, 96
точками бифуркации, 97
уравнение в пространстве состояний, 297
уравнения Протеро—Робинсона, 151
условия коллокации, 118
вариационной, 450
явные ФДН, 322
явные одношаговые методы типа Розенброка, 120
явный метод Эйлера, 110
жестко точные НМРК, 299

численный поток, 275
иммерсионный локальный

разрывный метод Галеркина k-го порядка, 566
методам конечных разностей, 448
регулярными, 455