Университетский учебник

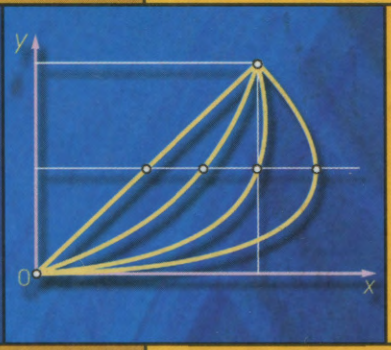
ЧИСЛЕННЫЕ МЕТОДЫ

В двух книгах

Книга 2

Н. Н. Калиткин
П. В. Корякин

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ФИЗИКИ



Прикладная математика и информатика

ACADEMIA

УНИВЕРСИТЕТСКИЙ УЧЕБНИК
Серия «Прикладная математика и информатика»

ЧИСЛЕННЫЕ МЕТОДЫ

В ДВУХ КНИГАХ

Книга 2

Н. Н. КАЛИТКИН, П. В. КОРЯКИН

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Допущено
Учебно-методическим объединением
по классическому университетскому образованию
в качестве учебника для студентов высших
учебных заведений, обучающихся по направлениям
«Прикладная математика и информатика»,
«Фундаментальная информатика и информационные технологии»

ACADEMIA
Москва
Издательский центр «Академия»
2013

УДК 51(075.8)
ББК 22.311я73
Ч-671

Рецензенты:
д-р физ.-мат. наук, проф. МГУ им. М. В. Ломоносова *А. В. Гулин*;
чл.-кор. РАН, зав. кафедрой вычислительной математики
Московского физико-технического института —
технического университета *А. С. Холодов*

Численные методы : в 2 кн. Кн. 2. Методы математической
Ч-671 физики : учебник для студ. учреждений высш. проф. образования / Н. Н. Калиткин, П. В. Корякин. — М. : Издательский центр «Академия», 2013. — 304 с. — (Университетский учебник. Сер. Прикладная математика и информатика).

ISBN 978-5-7695-5091-1

В учебнике излагаются основные численные методы решения широкого круга задач математической физики, возникающих при исследовании прикладных проблем. Это обыкновенные дифференциальные уравнения (включая жесткие задачи), уравнения в частных производных и интегральные уравнения.

В учебник включены только наиболее эффективные алгоритмы, пригодные как для расчетов на персональных компьютерах, так и для работы на многопроцессорных системах. Для каждого метода даны практические рекомендации по применению. Особое внимание уделено нахождению гарантированной оценки погрешности вычислений. Для лучшего понимания алгоритмов приведены численные расчеты.

Для студентов учреждений высшего профессионального образования.

УДК 51(075.8)
ББК 22.311я73

ISBN 978-5-7695-5091-1 (кн. 2)
ISBN 978-5-7695-5090-4

ПРЕДИСЛОВИЕ

Использование компьютеров позволило от простейших расчетов и оценок различных конструкций или процессов перейти к новой стадии работы — детальному математическому моделированию (вычислительному эксперименту), которое существенно сокращает потребность в дорогостоящих натурных экспериментах.

В основе вычислительного эксперимента лежит решение уравнений математической модели численными методами. Сложные вычислительные задачи, возникающие при исследовании различных физических и технических проблем, можно разделить на ряд элементарных. Традиционно эти задачи делят на две части. К первой части относят задачи математического анализа: решение алгебраических уравнений, нахождение интегралов, дифференцирование и т. п. Эти задачи считаются сравнительно простыми. Численное решение этих задач называют численным анализом. Проблемам численного анализа посвящено издание: Калиткин Н. Н. Численные методы: в 2 кн. Кн. 1. Численный анализ / Н. Н. Калиткин, Е. А. Альшина.

Более сложными и трудоемкими являются задачи, описывающиеся дифференциальными (обыкновенными и в частных производных) и интегральными уравнениями. Их называют задачами математической физики. Методам численного решения этих задач посвящена данная книга (кн. 2). Для задач численного анализа имеется немало методов и основанных на них стандартных программ. Однако в кн. 1 отмечалось, что бездумное пользование стандартными программами опасно: можно выйти за пределы применимости метода и программы (например, столкнуться с катастрофическим нарастанием компьютерных ошибок округления) и получить бессмысленный результат. Для задач математической физики эта проблема гораздо серьезнее. Такие задачи обычно настолько трудоемки, что их можно решать толь-

ко на компьютерах. Для различных задач разработаны методы и стандартные программы. Но насколько можно доверять этим программам?

В литературе описана задача Аренсторфа, относящаяся к небесной механике (рис. П.1). Луна обращается вокруг Земли. Начальное положение и скорость спутника выбирают так, чтобы при движении в поле тяготения Земли и Луны он описал за лунный месяц 4-витковую орбиту и вернулся в исходную точку с тем же значением вектора скорости (тонкая линия). Если рассчитывать данную орбиту по стандартной программе метода Рунге — Кутты 4-го порядка точности с числом шагов $\sim 10^4$ (жирная линия), то расчет резко отклоняется от истинной орбиты. Однако никаких указаний на катастрофическую ошибку программа обычно не дает. Только если пользователь догадается провести расчет с огромным числом шагов $\sim 10^6$, получится хорошее совпадение с истинной орбитой. Напомним, что для этого метода строго доказана сходимость численного решения к точному при стремлении величин шагов к нулю.

Этот пример показывает, что теоретического исследования сходимости метода недостаточно. Программа должна, одновременно с расчетом результата, находить фактическую оценку его погрешности. Поэтому в данной книге особенное внимание уделяется построению таких оценок.

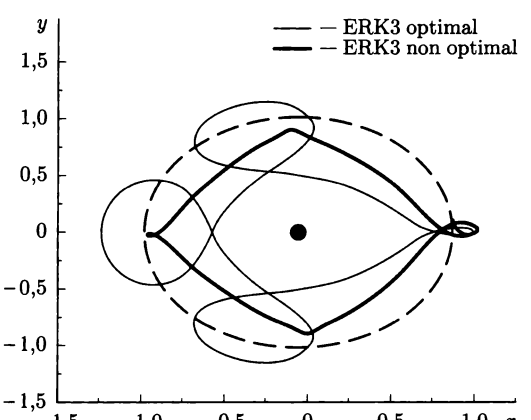


Рис. П.1. Задача Аренсторфа (тонкая линия — точное решение, жирная — расчет с недостаточно большим числом шагов, штриховая — орбита Луны)

Для одной и той же задачи в различных математических курсах нередко предлагают много различных алгоритмов решения. Например, для краевых задач (для обыкновенных дифференциальных уравнений, уравнений в частных производных и интегральных уравнений) применимы сеточные методы или методы Ритца и Галеркина. Однако для последних двух методов получить реальные оценки погрешности весьма сложно. Для сеточных же методов легко получить асимптотически точные апостериорные оценки погрешности, сгущая сетки и применяя методы Ричардсона или Эйткена. Поэтому в данной книге излагаются в основном сеточные методы.

Сеточных методов также очень много. Например, задачу Коши для нежестких обыкновенных дифференциальных уравнений можно решать явными методами Рунге — Кутты, Адамса — Башфорта, Милна, Нордсика, Лобатто, Мерсона и рядом других; в свою очередь, каждый из этих методов содержит немало конкретных схем, различающихся порядком точности и другими свойствами. Неспециалисту трудно разобраться в этом множестве методов и схем и выбрать наиболее пригодный для своих расчетов вариант.

Авторы попытались сделать такой выбор на основе собственного полувекового опыта численного решения прикладных задач.

Значительное внимание уделено одному нетривиальному классу схем — схемам с комплексными коэффициентами. Эти схемы мало известны даже в специальной литературе, а в учебники ранее никогда не включались. Такие схемы оказались исключительно эффективными для задач с диссипативными процессами (теплопроводностью, диффузией, вязким трением и др.), относящимся к так называемым жестким. Такие схемы обеспечивают не только хорошую точность, но и правильное качественное поведение численного решения (отсутствие или сильное сглаживание расчетной «ряби»).

Много изданий посвящено методу конечных элементов для решения краевых задач. Несколько десятилетий этот метод развивался как вполне самостоятельный. Однако постепенно стало ясно, что это лишь специфическая разновидность разностных схем; сам метод стали называть проекционно-сеточным. В настоящее время метод сохраняет свои позиции в основном потому, что на его основе написано много коммерческих и бесплатных программных пакетов для актуальных прикладных задач. По

мнению авторов, обычные разностные подходы более перспективны. Поэтому метод конечных элементов в данную книгу не включен.

По мере развития компьютеров приходится переоценивать пригодность тех или иных численных методов. Компьютеры первых поколений имели очень малую скорость и объем оперативной памяти. Поэтому тогда важнейшим требованием к методам была экономичность. По мере возрастания мощности компьютеров экономичность отходила на второй план, а на первый план выдвигалась надежность метода. Под надежностью подразумевается совокупность следующих свойств:

а) правильное качественное поведение численного решения, даже если шаги сетки или другие параметры расчета выбраны не слишком удачно;

б) возможность проведения расчетов большого объема без человеческого контроля;

в) получение оценки погрешности одновременно с результатом.

Быстродействие компьютеров позволило выполнять большие объемы вычислений. При этом погрешности округления возрастают в среднем как корень квадратный из числа операций. Они могут стать заметными, а в случае плохо обусловленных задач — очень большими. Частично от этого спасает большая разрядность чисел. Поэтому рекомендуется «не экономить на спичках» — использовать в расчетах всю максимально доступную разрядность компьютера.

Большинство существующих учебников по численным методам ориентировано на студентов и научных работников учреждений математического профиля, которые занимаются не столько использованием существующих, сколько разработкой новых численных методов. Данная же книга предназначена для широкого круга читателей — как учебник для студентов и аспирантов физических и технических специальностей вузов и как справочное пособие для научных сотрудников, инженеров и математиков-вычислителей. Авторы старались сочетать простоту изложения, разумную степень строгости, умеренный объем и широту охвата материала. Большое внимание в книге уделено рекомендациям по практическому применению алгоритмов; изложение пояснено рядом примеров. Для обоснования алгоритмов использован несложный математический аппарат, знакомый студентам физических и инженерных специальностей. Данная

книга является окончанием полного курса основ численных методов для физических и инженерных специальностей вузов; для математических специальностей вузов она может быть вводным курсом численных методов, после которого слушатели могут изучать углубленные спецкурсы.

Учебник написан на основе курса лекций, читавшихся сначала инженерам-конструкторам, а после переработки — студентам физического факультета МГУ им. М. В. Ломоносова и некоторых других вузов. Ранее по этому курсу была написана книга: Калиткин Н. Н. Численные методы. — М. : Физматлит, 1978. Данная книга существенно переработана и дополнена рядом актуальных разделов (решение жестких систем, задачи в неограниченных областях, бикомпактные разностные схемы для задач слоистых сред с разрывными коэффициентами и т. п.).

Книга разделена на главы, подразделы и пункты. Формулы имеют двойную нумерацию: номер главы и порядковый номер формул в главе; то же относится к нумерации рисунков, таблиц, теорем, следствий и определений. При ссылках на них указываются их номера. Например, гл. 2, рис. 2.5, п. 2.5.3, табл. 2.8 и т. п. Имеются также ссылки на кн. 1 данного курса (например, кн. 1 (3.18) означает формулу (3.18) из кн. 1). Конец доказательства теоремы отмечен знаком ■, а конец формулировки теоремы, леммы, определения или утверждения — знаком •.

В списке литературы приведены наиболее популярные учебники, которые ориентированы на широкий круг читателей; многие из них также рекомендуются для углубленного изучения отдельных разделов.

Общий подход к теории и практике вычислений, определивший стиль этой книги, сложился у авторов под влиянием многолетней совместной работы с А. А. Самарским и В. Я. Гольдиным. Ряд актуальных тем был включен по инициативе А. Г. Свешникова и В. Б. Гласко. Много ценных замечаний сделали В. Ф. Бутузов, А. В. Гулин, Б. Л. Рождественский, И. М. Соболь, И. В. Фрязинов, Е. В. Шикин. В оформлении рукописи большую помощь оказала Л. В. Кузьмина. Авторы искренне благодарны всем названным лицам, и особенно Александру Андреевичу Самарскому.

Глава 1

ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

1.1. ЗАДАЧА КОШИ

1.1.1. Элементы теории

Обыкновенными дифференциальными уравнениями (ОДУ) можно описать задачи движения системы взаимодействующих материальных точек (задачи небесной механики или баллистики), химической кинетики, электрических цепей с сосредоточенными параметрами и многие другие. Ряд важных задач для уравнений в частных производных также сводится к задачам для обыкновенных дифференциальных уравнений; примеры этого будут даны в следующих главах. Таким образом, решение обыкновенных дифференциальных уравнений занимает важное место среди прикладных задач физики, химии и техники. При этом конкретная прикладная задача может приводить к дифференциальному уравнению любого порядка, или к системе уравнений любого порядка. Например, химическая кинетика описывается системой уравнений первого порядка; число уравнений равно числу химических веществ. Трехмерное движение материальной точки описывается системой трех уравнений второго порядка.

Дифференциальное уравнение M-го порядка имеет общий вид

$$\frac{d^M u}{dt^M} = f(u, u', u'', \ldots, u^{(M-1)}, t); \tag{1.1}$$

аргумент обозначен через t, поскольку в задачах Коши он обозначает время. Припишем искомой функции индекс: $u(t) \equiv u_0(t)$. Все производные, кроме старшей, будем считать новыми функциями: $u^{(m)}(t) \equiv u_m(t)$, $m \leq M-1$. Тогда уравнение (1.1) можно переписать в виде системы M уравнений первого порядка:

$$\frac{du_{M-1}(t)}{dt} = f(u_0, u_1, \ldots, u_{M-1}, t);$$
$$\frac{du_m}{dt} = u_{m+1}(t), 0 \le m \le M-2. \tag{1.2}$$

Аналогично можно преобразовать к системе уравнений первого порядка произвольную систему уравнений любого порядка.

Поэтому далее будем рассматривать систему ОДУ первого порядка

$$\frac{du_m}{dt} = f_m(u_1, u_2, \ldots, u_M, t), 1 \le m \le M. \tag{1.3}$$

Систему (1.3) можно формально записать в векторных обозначениях:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}, t), \mathbf{u} = \{u_1, u_2, \ldots, u_M\}, \mathbf{f} = \{f_1, f_2, \ldots, f_M\}. \tag{1.4}$$

Далее нередко будем использовать форму (1.4) с нежирными буквами u, f, неявно подразумевая их векторами.

Известно, что система M-го порядка (1.4) имеет множество решений, которое в общем случае зависит от M параметров $\mathbf{c} = \{c_1, c_2, \ldots, c_M\}$ и может быть записано в форме $\mathbf{u} = \mathbf{u}(t, \mathbf{c})$. Для определения значений этих параметров, т. е. для выделения единственного (или нужного) решения, надо наложить M дополнительных условий на функции $u_m(t)$.

Задача Коши (задача с начальными условиями) имеет дополнительные условия, заданные в одной-единственной точке:

$$\mathbf{u}(t_0) = \mathbf{u}_0 \text{ или } u_m(t_0) = u_{m0}, 1 \le m \le M. \tag{1.5}$$

Это означает задание координат начальной точки интегральной кривой в $(M+1)$-мерном пространстве $\{t, u_1, \ldots, u_M\}$. Решение при этом обычно требуется найти на некотором отрезке $t_0 < t \le T$ или $T \le t < t_0$.

В курсах теории ОДУ доказаны следующие теоремы.

Теорема 1.1. Если правые части (1.3) — (1.4) непрерывны и ограничены в некоторой окрестности начальной точки $\{t_0, u_{10}, \ldots, u_{M0}\}$ интегральной кривой, то задача Коши (1.4) — (1.5) имеет решение. Однако это решение может быть не единственным. •

Теорема 1.2. Если правые части не только непрерывны, но и удовлетворяют условию Липшица по переменным u_m, то решение задачи Коши единственно и непрерывно зависит от координат начальной точки. Это означает корректность постановки задачи Коши. •

Теорема 1.3. Если вдобавок правые части имеют непрерывные производные вплоть до p-го порядка по всем аргументам (включая t), то решение $\mathbf{u}(t)$ имеет $(p+1)$-ю непрерывную производную по t. •

Эти теоремы следует учитывать при численных расчетах. Если выполнены только условия теоремы 1.1, то задача Коши некорректна, и численный расчет может «перескакивать» с одного допустимого решения на другое. При выполнении теоремы 1.2 численный расчет (по хорошей схеме) будет сходиться к точному решению, но оценку его точности невозможно получить. В условиях теоремы 1.3 целесообразно использовать схемы p-го порядка точности либо схемы меньшего порядка точности с уточнением по Ричардсону до p-го порядка; но использовать схемы порядка точности более p-го бессмысленно.

Методы решения можно условно разбить на точные, приближенные и численные. К точным относятся методы, позволяющие выразить решение через элементарные функции либо представить его в виде квадратур от элементарных функций. Эти методы изучают в курсах ОДУ. Нахождение точного решения всегда желательно. Однако это удается сделать лишь для отдельных классов уравнений. Например, несложное уравнение

$$\frac{du}{dt} = u^2 + t^2 \tag{1.6}$$

не решается в элементарных функциях.

Приближенными называют методы, в которых решение представляется как предел некоторой последовательности элементарных функций или их комбинаций $y_k(t)$ либо квадратур от таких комбинаций:

$$u(t) = \lim_{k\to\infty} y_k(t). \tag{1.7}$$

Конечная k-я итерация процесса (1.7) дает приближенное выражение $u(t)$ через элементарные функции или квадратуры. К приближенным методам относятся разложение решения в степенной или обобщенный степенной ряд, методы Пикара, Чаплыгина,

Пуанкаре (малого параметра) и др. Их довольно часто использовали до появления компьютеров; однако в настоящее время они почти полностью вытеснены численными методами и далее не рассматриваются.

В численных методах выбирается некоторая сетка значений аргумента $t_0 < t_1 < \ldots < t_N = T$ и вычисляются (приближенные) значения решения $u_n \approx u(t_n)$ в узлах этой сетки. Решение при этом имеет вид таблицы. Численные методы не позволяют найти общее решение системы (1.4). Они дают лишь какие-то частные решения, например решение задачи (1.4) — (1.5). Зато эти методы применимы к любым классам ОДУ и всем типам задач для них. Поэтому с появлением компьютеров численные методы стали основным инструментом решения прикладных задач.

Обусловленность. Численные методы можно применять только к корректно поставленным (или регуляризованным) задачам. Напомним (см. кн. 1), что корректность включает три требования: существование решения, его единственность и непрерывную зависимость от входных данных. Последнее означает, что если вариации входных данных $\delta\mathbf{u}_0 \to 0$, то соответствующие вариации решения $\delta\mathbf{u}(t) \to 0$. Это формально выполняется, например, если $\|\delta\mathbf{u}(t)\| \le c(t)\,\|\delta\mathbf{u}_0\|$, где $c(t)$ — ограниченная на $[t_0, T]$ функция.

Однако на практике возникают задачи, где $c(t)$ ограничена, но может принимать очень большие значения. Например, в задаче Аренсторфа (см. рис. П.1) ошибка начальной скорости спутника на 1 мм/с приводит к тому, что за месяц ошибка возрастает до 1 км/с! Это означает, что $c(t) \sim 10^6$ велико. Есть задачи с гораздо большими значениями $c(t)$.

Такие задачи называют ***плохо обусловленными***. Они формально корректны, но весьма трудны для численного расчета. В них ничтожные ошибки начальных данных, или эквивалентные им погрешности численного метода, и даже ошибки компьютерного округления могут сильно исказить решение.

Пример. Рассмотрим задачу

$$u'(t) = u - t, 0 \le t \le 100, \tag{1.8}$$

$$u(0) = 1\,. \tag{1.9}$$

Общее решение уравнения (1.8) содержит одну произвольную постоянную

$$u(t;c) = 1 + t + ce^t; \tag{1.10}$$

при начальном условии (1.9) константа $c = 0$, так что $u(100) = 101$. Однако небольшое изменение начального условия $\bar{u}(0) = 1{,}000001$ слегка меняет постоянную: $\bar{c} = 10^{-6}$; тогда $\bar{u}(100) \approx 2{,}7 \cdot 10^{37}$, т. е. решение изменилось очень сильно.

В подразделах 1.1—1.3 рассмотрены методы решения задачи Коши. Для простоты записи почти всюду ограничимся случаем одного уравнения первого порядка. Алгоритмы для систем M уравнений (1.3)—(1.4) легко получаются из алгоритмов для одного уравнения формальной заменой $u(t) \to \mathbf{u}(t)$, $f(u,t) \to \mathbf{f}(\mathbf{u},t)$.

Автономность. Если правые части системы (1.3—1.4) не зависят от неизвестных функций, т. е. имеют вид $\mathbf{f}(t)$, то решение задачи (1.3) тривиально записывается через квадратуры:

$$u_m(t) = \int_{t_0}^{t} f_m(\theta)d\theta + u_m(t_0). \tag{1.11}$$

Численное решение (1.11) находим по квадратурным формулам, приведенным в кн. 1. Этот случай не требует новых методов.

Интересен противоположный случай, когда правые части (1.3)—(1.4) не зависят явно от аргумента t:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}). \tag{1.12}$$

Системы (1.12) называют автономными, а системы общего вида (1.4) — неавтономными. Строить конкретные численные схемы для автономных систем проще, чем для неавтономных, поэтому полезен следующий прием, так называемая автономизация.

Автономизация. Введем дополнительную функцию $u_{M+1}(t) \equiv t$; очевидно, $du_{M+1}/dt = 1$. Заменим систему (1.3) с начальными условиями (1.5) следующей задачей Коши для системы $(M+1)$-го порядка:

$$\begin{gathered} \frac{du_m}{dt} = f_m(u_1, u_2, \ldots, u_M, u_{M+1}), 1 \le m \le M+1; \\ f_{M+1}(\mathbf{u}) \equiv 1; \quad u_m(t_0) = u_{m0}, 1 \le m \le M, u_{M+1,0} = t_0. \end{gathered} \tag{1.13}$$

Задача (1.13) автономна и эквивалентна исходной задаче (1.3), (1.5). В ней возникают два времени: время-аргумент и время-функция. Для точной задачи они совпадают, а при численном

решении могут отличаться в пределах погрешности алгоритма (хотя для большинства конкретных схем отличий не будет).

Прием автономизации позволяет преобразовать неавтономную систему к автономному виду и решать последнюю по схемам, построенным для автономных задач. Этот способ особенно выгоден для схем высокого порядка точности.

Длина дуги. Есть еще один способ автономизации, который выгодно применять даже для автономных систем. Рассмотрим интегральную кривую в $(M+1)$-мерном пространстве переменных $\{u_1, u_2, \ldots, u_M, t\}$. Длина дуги этой кривой определяется соотношением

$$(dl)^2 = (dt)^2 + \sum_{m=1}^{M} (du_m)^2. \tag{1.14}$$

Подставляя (1.3) в (1.14), получим

$$(dl)^2 = \left[1 + \sum_{m=1}^{M} f_m^2(u_1, u_2, \ldots, u_M, t)\right] (dt)^2. \tag{1.15}$$

Заменим в (1.13) величину dt на dl с помощью (1.15) и получим следующую систему $(M+1)$-го порядка:

$$\begin{gathered}
\frac{du_m}{dl} = F_m(u_1, \ldots, u_{M+1}), 1 \le m \le M+1, \\
F_m(u_1, \ldots, u_{M+1}) = f_m \Bigg/ \left(1 + \sum_{m=1}^{M} f_m^2\right)^{1/2}, \ 1 \le m \le M, \\
F_{M+1}(u_1, \ldots, u_{M+1}) = 1 \Bigg/ \left(1 + \sum_{m=1}^{M} f_m^2\right)^{1/2}, \ u_{M+1} \equiv t.
\end{gathered} \tag{1.16}$$

Правые части системы (1.16) не зависят явно от l, так что система автономна. Начальные данные для задачи Коши берутся из (1.13), а начальное значение длины дуги можно взять $l_0 = 0$. Правда, какие значения l соответствуют конечному моменту $t = T$, неизвестно.

Автономизация с помощью длины дуги имеет важное преимущество. При этом вектор правых частей в (1.16) имеет единичную длину:

$$\sum_{m=1}^{M+1} F_m^2(u_1, \ldots, u_{M+1}) = 1. \tag{1.17}$$

Это означает, что среди производных du_m/dt нет больших по величине, что облегчает численное решение задачи.

Заметим, что можно ввести масштабные множители по координатам интегральной кривой, т. е. перейти к величинам $u_m(t)/a_m$. Удачный подбор масштабов a_m (своих для каждой конкретной задачи) может облегчить численный расчет.

1.1.2. Методы Рунге — Кутты (РК)

Немного истории. Дифференциальные уравнения были введены в науку Ньютоном и Лейбницем в XVII в. Сначала изучались те задачи, для которых удавалось найти точное решение (например, движение планет вокруг Солнца). Серьезной потребности в численных методах долго не возникало.

Однако уже в XVIII в. Л. Эйлер предложил первый численный метод. Обозначим шаг сетки через $\tau_n = t_{n+1} - t_n$ и грубо аппроксимируем приращение функции:

$$u(t_{n+1}) = u(t_n) + \tau_n (du/dt)_n = u(t_n) + \tau_n f\left(u(t_n),\ t_n\right). \quad (1.18)$$

Формулу (1.18) называют схемой Эйлера или схемой ломаных (так как на каждом шаге интегральная кривая заменяется прямолинейным звеном, наклон которого совпадает с касательной). Точность схемы (1.18) низка, так что на практике ее не используют.

Серьезная потребность в численных расчетах возникла в середине XIX в. Скорости пуль и снарядов возросли настолько, что потребовался аккуратный расчет их траекторий с учетом сопротивления воздуха. Английский математик Адамс разработал семейство схем высоких порядков точности, являющееся обобщением схемы Эйлера. Схемы Адамса (их называют также схемами Адамса — Башфорта) быстро стали популярными и широко использовались до середины XX в.

Достоинством схем Адамса является то, что в них легко строятся схемы очень высоких порядков точности $p > 10$. Основной недостаток состоит в том, что это многошаговые схемы. Чтобы выполнить один шаг $t_n \to t_{n+1}$, надо знать значения в p последних точках $t_n, t_{n-1}, \ldots, t_{n-p+1}$. В начальной точке t_0 эти значения неизвестны, и приходится разрабатывать специальные алгоритмы начала расчета (что непросто). Поэтому с появлением компьютеров эти схемы стали менее удобными.

Этого недостатка лишено принципиально другое, одношаговое обобщение схемы Эйлера. Первой такой схемой была двухстадийная схема Рунге (1895). Затем появились трехстадийные схемы и классическая четырехстадийная схема Кутты (1901). Бутчер разработал специальную технику построения схем типа Рунге—Кутты. Это позволило найти немало многостадийных схем высоких порядков точности. Все эти схемы оказались хорошо пригодными для компьютерных расчетов. Они постепенно вытеснили все прочие методы для большого класса прикладных задач (за исключением жестких, рассмотренных в подразделе 1.2).

Общий вид. Схемы Рунге—Кутты (РК) одношаговые, т. е. для перехода $t_n \to t_{n+1}$ надо знать только значение $u(t_n)$ в исходном узле сетки. Однако эти схемы многостадийны; эти стадии похожи на промежуточные шаги, хотя такая аналогия нестрога. Рассмотрим общий вид этих схем.

Будем различать точное решение дифференциальной задачи $u(t)$ и численное решение u_n (последнее определено только в узлах сетки t_n), хотя обычно будем обозначать их одной и той же буквой. Схема РК с s стадиями имеет следующий вид:

$$\begin{gathered} u_{n+1} = u_n + \tau_n \sum_{k=1}^{s} b_k w_k \,, \tau_n = t_{n+1} - t_n; \\ w_k = f\left(u_n + \tau_n \sum_{l=1}^{L} a_{kl} w_l \,,\ t_n + \tau_n a_k \right), 1 \le k \le s \,. \end{gathered} \tag{1.19}$$

Здесь w_k — правые части уравнения со сдвинутыми аргументами; сдвиги свои на каждой стадии. Коэффициенты схемы образуют два вектора (b_k), (a_l) и так называемую матрицу Бутчера (a_{kl}), табл. 1.1.

Очень важную роль играет предел суммирования L. Если $L = k - 1$, то каждое w_k выражается явно через w_l с меньшими

Таблица 1.1

Коэффициенты схем РК ($s = 4$)

a_1	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	b_1
a_2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	b_2
a_3	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	b_3
a_4	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$	b_4

индексами. Значит, формулы (1.19) являются явными, т. е. все вычисления выполняются за конечное, заранее известное число операций (даже если $\mathbf{u}, \mathbf{f}$ — векторы). В этом случае схемы РК называют ***явными***. Их матрица Бутчера поддиагональна: $a_{kl} = 0$ при $k \leq l$. В этом случае в табл. 1.1 остаются только те коэффициенты, которые набраны нормальным прямым шрифтом. Явные схемы РК наиболее часты в практике расчетов.

Схемы с $L = k$ называют ***диагонально-неявными***. В этом случае в табл. 1.1 добавляется диагональ матрицы Бутчера, набранная жирным курсивом. Величины w_k определяются последовательно, одна за другой. При этом для определения величины w_k надо решить нелинейное алгебраическое уравнение (а в случае системы M дифференциальных уравнений — систему M нелинейных алгебраических уравнений). Это существенно усложняет алгоритм по сравнению с явными схемами, хотя это много проще полностью неявных схем.

Схемы с $L = s$ называют ***полностью неявными***. Их матрица Бутчера полностью заполнена: все $a_{kl} \neq 0$ (добавленные элементы набраны нежирным курсивом). Величины w_k определяются одновременно для всех стадий из системы s нелинейных алгебраических уравнений (а для системы ОДУ M-го порядка — из системы sM уравнений). Это настолько сложно, что полностью неявные схемы практически не применяют.

Напомним, что метод (1.19) одношаговый, и при выполнении n-го шага не используется никакой информации о предыдущих шагах. Это позволяет немного упростить запись: записывать шаг τ без индекса (но считать шаг переменным). Введем обозначения

$$u_n \equiv u, u_{n+1} \equiv \hat{u},\ t_n \equiv t,\ t_{n+1} \equiv \hat{t} = t + \tau. \tag{1.20}$$

Последние обозначения введены для будущих формул.

Интерполяционность. Коэффициенты a_k в (1.19) показывают, на какую долю шага сдвигается аргумент правой части t на k-й стадии. Аналогичный сдвиг по аргументу u характеризуется величиной $\sum_l a_{kl}$. Естественно потребовать, чтобы эти сдвиги, а также частичные сдвиги a_{kl} не выводили расчет за границы отрезка $[t_n, t_{n+1}]$. Это приводит к так называемым условиям интерполяционности:

$$0 \le a_k \le 1, 0 \le b_k \le 1, 0 \le a_{kl},$$
$$\sum_l a_{kl} \le 1, \ 1 \le k \le s. \tag{1.21}$$

Условия (1.21) не являются необходимыми в задачах ОДУ, однако их соблюдение делает схему более надежной в расчетах. В гл. 3 будет показано, что для уравнений в частных производных нарушение условий интерполяционности приводит к неустойчивости.

Дадим наглядное сравнение. Один шаг можно рассматривать как подъем с одной лестничной площадки на следующую. Если мы поднимаемся по промежуточным ступеням, это интерполяция. Но если мы перескакиваем на следующий пролет лестницы, а потом соскакиваем обратно, то это нарушение условий (1.21). Ясно, что последнее неразумно.

Автономизация. Схема (1.19) написана для неавтономной системы ОДУ (1.4). Для автономных систем коэффициенты a_k отсутствовали бы. Преобразуем систему (1.4) к автономному виду (1.13) и возьмем для нее некоторую схему вида (1.19) без коэффициентов a_k.

Поскольку для функции $u_{+1} \equiv t$ величина $w_k \equiv 1$, то сдвиги аргумента будут равны $\tau \sum_l a_{kl}$. Они соответствуют сдвигам τa_k по переменной t. Отсюда вытекает важное следствие.

Следствие 1.1. Пусть имеется некоторая схема РК (1.19) для автономной системы (1.12), не содержащая коэффициентов a_k. Положив

$$a_k = \sum_l a_{kl}, \tag{1.22}$$

получаем схему РК для неавтономной системы (1.3). Прочие свойства обеих схем (точность, интерполяционность и т. п.) будут одинаковыми.

Это следствие позволяет ограничиться построением схем только для автономных систем, что значительно проще. Правда, при этом можно «прозевать» некоторые неавтономные схемы, не удовлетворяющие условиям (1.22). Но пропущенные схемы не дают существенных преимуществ.

1.1.3. Аппроксимация

Из каких соображений целесообразно выбирать коэффициенты схем РК? Очевидно, важнейшим должно быть требование, чтобы схема (1.19) как можно лучше аппроксимировала систему (1.3) — (1.4). Рассмотрим, как сформулировать такое требование. При этом ограничимся случаем автономных систем, что существенно упрощает выкладки, а уравнения будем записывать в векторных обозначениях.

Точный шаг. Для точного решения системы ОДУ (1.4) можно произвести переход от момента t к моменту $\hat{t} = t + \tau$ с помощью ряда Тейлора:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau\frac{d\mathbf{u}}{dt} + \frac{\tau^2}{2}\frac{d^2\mathbf{u}}{dt^2} + \frac{\tau^3}{6}\frac{d^3\mathbf{u}}{dt^3} + \frac{\tau^4}{24}\frac{d^4\mathbf{u}}{dt^4} + \dots . \tag{1.23}$$

Здесь и далее все производные берутся в момент t. Первая производная определяется непосредственно из системы (1.4): $d\mathbf{u}/dt = \mathbf{f}$. Вторую полную производную по времени определим через частные производные $\mathbf{f}(\mathbf{u})$ с учетом автономности:

$$\frac{d^2\mathbf{u}}{dt^2} = \frac{d}{dt}\left(\frac{d\mathbf{u}}{dt}\right) = \frac{d\mathbf{f}}{dt} = \frac{\partial\mathbf{f}}{\partial\mathbf{u}}\frac{d\mathbf{u}}{dt} = \mathbf{f}_\mathbf{u}\mathbf{f}\,. \tag{1.24}$$

Здесь $\mathbf{f}_\mathbf{u}$ есть матрица первых производных (тензор второго ранга), а $\mathbf{f}$ — вектор; они не коммутируют, так что в последнем произведении в (1.24) надо строго соблюдать указанный порядок сомножителей. Аналогично находим третью производную:

$$\frac{d^3\mathbf{u}}{dt^3} = \frac{d}{dt}\mathbf{f}_\mathbf{u}\mathbf{f} = \mathbf{f}_{\mathbf{uu}}\mathbf{ff} + \mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}\,, \tag{1.25}$$

где $\mathbf{f}_{\mathbf{uu}}$ — тензор третьего ранга. Подставляя все эти произведения в (1.23), получим следующее разложение точного решения:

$$\begin{gathered}
\hat{\mathbf{u}} = \mathbf{u} + \tau\mathbf{f} + \frac{\tau^2}{2}\mathbf{f}_\mathbf{u}\mathbf{f} + \frac{\tau^3}{6}\left(\mathbf{f}_{\mathbf{uu}}\mathbf{ff} + \mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}\right) + \\
+\frac{\tau^4}{24}\left(\mathbf{f}_{\mathbf{uuu}}\mathbf{fff} + 3\mathbf{f}_{\mathbf{uu}}\mathbf{f}_\mathbf{u}\mathbf{ff} + \mathbf{f}_\mathbf{u}\mathbf{f}_{\mathbf{uu}}\mathbf{ff} + \mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}\right) + \\
+\frac{1}{120}\tau^5\left(\mathbf{f}_{\mathbf{uuuu}}\mathbf{ffff} + 4\mathbf{f}_{\mathbf{uu}}\mathbf{f}_{\mathbf{uu}}\mathbf{fff} + 6\mathbf{f}_{\mathbf{uuu}}\mathbf{f}_\mathbf{u}\mathbf{fff} + \mathbf{f}_\mathbf{u}\mathbf{f}_{\mathbf{uuu}}\mathbf{fff} + \right. \\
\left. + 4\mathbf{f}_{\mathbf{uu}}\mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{ff} + 3\mathbf{f}_{\mathbf{uu}}\mathbf{f}_\mathbf{u}\mathbf{f}\mathbf{f}_\mathbf{u}\mathbf{f} + 3\mathbf{f}_\mathbf{u}\mathbf{f}_{\mathbf{uu}}\mathbf{f}_\mathbf{u}\mathbf{ff} + \mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}_{\mathbf{uu}}\mathbf{ff} + \mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}_\mathbf{u}\mathbf{f}\right) + \\
+ \dots + O(\tau^6);
\end{gathered} \tag{1.26}$$

здесь добавлены четвертая и пятая производные. Видно, что с ростом степени τ количество различных комбинаций производных стремительно увеличивается.

Если рассматривать неавтономные системы, то число комбинаций производных еще намного больше. В член $O(\tau^2)$ добавляется одна производная $\mathbf{f}_t$, в член $O(\tau^3)$ — уже три комбинации, а в член $O(\tau^4)$ — 14 комбинаций! Именно поэтому удобна автономизация.

Отметим существенную деталь вывода формулы (1.26). При почленном дифференцировании разных слагаемых очередной полной производной возникают одинаковые комбинации. Например, при получении члена $O(\tau^4)$ дифференцированием члена $O(\tau^3)$ дважды возникает комбинация $\mathbf{f_{uu}ff_uf}$; они тривиально складываются. Но есть еще нетривиальное совпадение этой комбинации с $\mathbf{f_{uu}f_uff}$.

В самом деле, тензор $\mathbf{f_{uu}}$ симметричен по нижним индексам. В этих комбинациях он свертывается с векторами $\mathbf{f_uf}$ и $\mathbf{f}$. Благодаря симметрии тензора эти векторы можно менять местами; это приводит к совпадению $\mathbf{f_{uu}f_uff} = \mathbf{f_{uu}ff_uf}$.

Численный шаг. Проделаем разложение по степеням для автомодельной схемы РК (1.19), аналогичное (1.26). При этом воспользуемся следующим разложением функции по приращению аргумента:

$$\mathbf{f}(\mathbf{u}+\tau\mathbf{w}) = \mathbf{f} + \tau\mathbf{f_u w} + \frac{1}{2}\tau^2\mathbf{f_{uu}ww} + \frac{1}{6}\tau^3\mathbf{f_{uuu}www} + \dots; \quad (1.27)$$

здесь под $\mathbf{f}, \mathbf{f_u}$ и т. д. без указания аргумента понимаются значения при аргументе $\mathbf{u}$. Подставим в (1.27) на первой стадии $\mathbf{w} = 0$, на второй $\mathbf{w} = a_{21}\mathbf{w}_1$, на третьей $\mathbf{w} = a_{31}\mathbf{w}_1 + a_{32}\mathbf{w}_2$, на четвертой $\mathbf{w} = a_{41}\mathbf{w}_1 + a_{42}\mathbf{w}_2 + a_{43}\mathbf{w}_3$. Введем для упрощения следующие обозначения:

$$a_2 = a_{21}, a_3 = a_{31} + a_{32}, a_4 = a_{41} + a_{42} + a_{43}, \dots; \quad (1.28)$$

они соответствуют тривиальной неавтономной схеме.

Будем учитывать в разложении $\mathbf{w}_k$ члены до $O(\tau^3)$ включительно. Выражение для $\mathbf{w}_1$ тривиально и не содержит степеней τ. Подставим его в $\mathbf{w}_2$ и выполним разложение; появятся уже все степени τ.

Затем подставим эти $\mathbf{w}_1$ и $\mathbf{w}_2$ в выражение $\mathbf{w}_3$ и т. д. В итоге получим следующие разложения:

$$\mathbf{w}_1 = \mathbf{f},$$

$$\mathbf{w}_2 = \mathbf{f} + \tau a_2 \mathbf{f_u f} + \frac{1}{2}\tau^2 a_2^2 \mathbf{f_{uu} ff} + \frac{\tau^3}{6} a_2^3 \mathbf{f_{uuu} fff} + \ldots;$$

$$\mathbf{w}_3 = \mathbf{f} + \tau a_3 \mathbf{f_u f} + \tau^2 \left(\frac{1}{2} a_3^2 \mathbf{f_{uu} ff} + a_2 a_{32} \mathbf{f_u f_u f} \right) +$$

$$+\tau^3 \left(\frac{1}{6} a_3^3 \mathbf{f_{uuu} fff} + a_2 a_3 a_{32} \mathbf{f_{uu} f_u ff} + \frac{1}{2} a_2^2 a_{32} \mathbf{f_u f_{uu} ff} \right) + \ldots;$$

$$\mathbf{w}_4 = \mathbf{f} + \tau a_4 \mathbf{f_u f} + \tau^2 \left[\frac{1}{2} a_4^2 \mathbf{f_{uu} ff} + (a_2 a_{42} + a_3 a_{43}) \mathbf{f_u f_u f} \right] +$$

$$+\tau^3 \left[\frac{1}{6} a_4^3 \mathbf{f_{uuu} fff} + a_4 (a_2 a_{42} + a_3 a_{43}) \mathbf{f_{uu} f_u ff} + \right.$$

$$\left. + \frac{1}{2} (a_2^2 a_{42} + a_3^2 a_{43}) \mathbf{f_u f_{uu} ff} + a_2 a_{32} a_{43} \mathbf{f_u f_u f_u f} \right] + \ldots. \tag{1.29}$$

Подставляя разложения (1.29) в схему РК (1.19), получим разложение численного решения по степеням шага τ:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau (b_1 + b_2 + b_3 + b_4) \mathbf{f} + \tau^2 (b_2 a_2 + b_3 a_3 + b_4 a_4) \mathbf{f_u f} +$$

$$+\tau^3 \left\{ \frac{1}{2} (b_2 a_2^2 + b_3 a_3^2 + b_4 a_4^2) \mathbf{f_{uu} ff} + \right.$$

$$\left. + [b_3 a_2 a_{32} + b_4 (a_2 a_{42} + a_3 a_{43})] \mathbf{f_u f_u f} \right\} +$$

$$+\tau^4 \left\{ \frac{1}{6} (b_2 a_2^3 + b_3 a_3^3 + b_4 a_4^3) \mathbf{f_{uuu} fff} + \right.$$

$$+ [b_3 a_3 a_2 a_{32} + b_4 a_4 (a_2 a_{42} + a_3 a_{43})] \mathbf{f_{uu} f_u ff} +$$

$$\left. + \frac{1}{2} [b_3 a_2^2 a_{32} + b_4 (a_2^2 a_{42} + a_3^2 a_{43})] \mathbf{f_u f_{uu} ff} + b_4 a_2 a_{32} a_{43} \mathbf{f_u f_u f_u f} \right\}. \tag{1.30}$$

В (1.29) и (1.30) оставлено столько членов разложения, сколько необходимо для 4-стадийных схем. При увеличении числа стадий число членов соответственно увеличивается, а громоздкость выражений (т. е. включенных комбинаций производных) стремительно возрастает.

Уравнения порядка. Сравним разложение схемы (1.30) и точного решения (1.26). При одинаковых степенях τ разложение (1.30) содержит такие же комбинации производных (хотя,

возможно, не все), как и разложение (1.26). Перед этими производными в (1.30) стоят некоторые комбинации коэффициентов схемы РК, а в (1.26) стоят числа. Требуя, чтобы соответствующие члены в (1.26) и (1.30) точно совпадали, получим так называемые уравнения порядка для 4-стадийных схем:

$$\tau\mathbf{f}:\ b_1+b_2+b_3+b_4=1\,; \tag{1.31}$$

$$\tau^2\mathbf{f_u f}:\ b_2a_2+b_3a_3+b_4a_4=\frac{1}{2}\,; \tag{1.32}$$

$$\tau^3\mathbf{f_{uu}ff}:\ b_2a_2^2+b_3a_3^2+b_4a_4^2=\frac{1}{3}; \tag{1.33}$$

$$\tau^3\mathbf{f_u f_u f}:\ b_3a_2a_{32}+b_4\left(a_2a_{42}+a_3a_{43}\right)=\frac{1}{6}; \tag{1.34}$$

$$\tau^4\mathbf{f_{uuu}fff}:\ b_2a_2^3+b_3a_3^3+b_4a_4^3=\frac{1}{4}; \tag{1.35}$$

$$\tau^4\mathbf{f_{uu}f_u ff}:\ b_3a_3a_2a_{32}+b_4a_4\left(a_2a_{42}+a_3a_{43}\right)=\frac{1}{8}; \tag{1.36}$$

$$\tau^4\mathbf{f_u f_{uu}ff}:\ b_3a_2^2a_{32}+b_4\left(a_2^2a_{42}+a_3^2a_{43}\right)=\frac{1}{12}; \tag{1.37}$$

$$\tau^4\mathbf{f_u f_u f_u f}:\ b_4a_2a_{32}a_{43}=\frac{1}{24}; \tag{1.38}$$

здесь указаны те степени τ и комбинации производных, которые согласованы. Уравнения порядка для меньшего числа стадий получаются из (1.31) — (1.38) отбрасыванием старших степеней τ и лишних коэффициентов.

Определение 1.1. Если коэффициенты b_k, a_{kl}, a_k схемы РК подобраны так, что все уравнения порядка для степеней τ^p и ниже удовлетворяются, то схема (1.19) *аппроксимирует* систему ОДУ (1.4) с p-м порядком. •

Далее для различного числа стадий s решаются уравнения порядка и находятся коэффициенты схем РК с максимально возможным порядком аппроксимации p при данном s. Напомним, что при этом получаются коэффициенты автономных и тривиальных неавтономных схем. Но нетривиальные неавтономные схемы известны лишь для одно- и двухстадийных схем, так что они не представляют существенного интереса.

1.1.4. Двухстадийная схема

Для полноты сначала рассмотрим одностадийную схему: $b_1 \neq \neq 0$, остальные b_k отсутствуют (т. е. формально равны нулю). Тогда первое уравнение порядка (1.31) удовлетворяется при $b_1 = 1$, что обеспечивает первый порядок аппроксимации: $p = 1$. Других коэффициентов автономная схема не имеет, так что она принимает следующий вид:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \mathbf{f}(\mathbf{u}) . \tag{1.39}$$

Второе уравнение порядка (1.32) не может удовлетворяться; это означает, что одностадийная схема не может иметь второй порядок аппроксимации.

Количественной мерой ошибки аппроксимации является главный член, на который различаются разложения точного решения (1.26) и схемы (1.30); чтобы он имел порядок $O(\tau^p)$, его надо дополнительно разделить на τ. Эту величину называют ***невязкой***. Видно, что невязка схемы (1.39) равна

$$\psi = \tau \mathbf{f}_{\mathbf{u}} \mathbf{f} = O(\tau) . \tag{1.40}$$

Тривиальная неавтономная схема для (1.39) имеет $a_1 = 0$, т. е. получается из (1.39) заменой $\mathbf{f}(\mathbf{u}) \to \mathbf{f}(\mathbf{u}, t)$. Она называется схемой Эйлера. Однако существует нетривиальная неавтономная схема с $a_1 \neq 0$; она также имеет первый порядок аппроксимации.

Схема (1.39) интерполяционна. Ее нетривиальное неавтономное обобщение интерполяционно при $0 \leq a_1 \leq 1$.

На практике схему Эйлера *не* употребляют из-за малой точности.

Две стадии. Автономная схема РК содержит только три коэффициента b_1, b_2, $a_2 \equiv a_{21}$ (старшие коэффициенты формально равны нулю). Она имеет следующий вид:

$$\begin{gathered} \hat{\mathbf{u}} = \mathbf{u} + \tau \left(b_1 \mathbf{w}_1 + b_2 \mathbf{w}_2 \right) , \\ \mathbf{w}_1 = \mathbf{f}(\mathbf{u}), \mathbf{w}_2 = \mathbf{f} \left(\mathbf{u} + \tau a_2 \mathbf{w}_1 \right) . \end{gathered} \tag{1.41}$$

С помощью имеющихся трех коэффициентов можно удовлетворить двум уравнениям порядка (1.31)–(1.32), в которых выброшены старшие коэффициенты. При этом один из коэффициентов можно выбрать свободным и выразить через него остальные

два; это дает однопараметрическое семейство решений. Выбирая a_2 в качестве параметра, получим

$$b_2 = 1/(2a_2),\ b_1 = 1 - b_2,\ (1/2 \le a_2 \le 1); \qquad (1.42)$$

в скобках приведены значения параметра a_2, обеспечивающие интерполяционность схемы (1.41).

Схема (1.41) — (1.42) имеет второй порядок аппроксимации. Удовлетворить одним свободным коэффициентом сразу обоим уравнениям порядка (1.33) — (1.34) и получить 3-й порядок аппроксимации невозможно.

Оптимум. Какое значение a_2 целесообразно выбрать? Обычно в учебниках рекомендуют два варианта: 1) схему «предиктор — корректор» с $a_2 = 1/2$ и 2) схему «с полусуммой» или $a_2 = 1$.

Однако при этом оба следующих уравнения порядка (1.33) — (1.34) не удовлетворяются. Если же положить $a_2 = 2/3$, то удовлетворяется первое из этих уравнений, т. е. комбинация производных $\tau^3 \mathbf{f_{uu}ff}$ точно согласуется (другую комбинацию $\mathbf{f_u f_u f}$ согласовать на двух стадиях невозможно). Поэтому оптимальным набором коэффициентов следует считать

$$a_2 \equiv a_{21} = 2/3, b_1 = 1/4, b_2 = 3/4. \qquad (1.43)$$

Невязка при этом равна

$$\psi = \frac{1}{2}\tau^2 \mathbf{f_u f_u f} = O(\tau^2). \qquad (1.44)$$

Заметим, что у двухстадийной схемы РК (1.41) есть нетривиальное неавтономное обобщение с $a_2 \neq a_{21}$. Это двухпараметрическое семейство решений. Однако и в нем оптимальным является набор (1.43).

Точность двухстадийной схемы РК удовлетворительна, так что эту схему нередко используют в прикладных расчетах.

Вложение. Рассмотрим первую стадию двухстадийной схемы (1.41) как самостоятельную схему; выбираем для нее $\bar{b}_1 = 1$ (очевидно $\bar{b}_2 = 0$). Полученная схема совпадает со схемой Эйлера (1.39), причем расчет шага по ней не требует дополнительного вычисления правых частей. Такая конструкция является частным случаем так называемых вложенных схем.

Определение 1.2. Пусть s-стадийная схема с коэффициентами b_k и слагаемыми $\mathbf{w}_k$ имеет порядок аппроксимации p, а из тех же величин $\mathbf{w}_k$ с другими коэффициентами $\bar{b}_k$ составляется схема порядка аппроксимации $p-1$. Тогда вторую схему называют *вложенной* в первую. •

Очевидно, вторая схема не требует дополнительного вычисления правых частей, т. е. не увеличивает трудоемкости расчетов. Обычно вложенная схема имеет $s-1$ стадию, но иногда число ее стадий еще меньше. У схемы Эйлера не может быть вложенной схемы.

1.1.5. Три стадии

Коэффициенты. Автономная трехстадийная схема содержит коэффициенты, в качестве которых удобно выбрать a_2, a_3, a_{32}, b_1, b_2, b_3. Таким числом коэффициентов можно удовлетворить первым четырем уравнениям порядка (1.31) — (1.34), содержащим степени шага вплоть до τ^3, причем получится двупараметрическое семейство решений. Но удовлетворить следующим четырем уравнениям (1.35) — (1.38) со степенью τ^4 уже невозможно.

Таким образом, трехстадийная схема имеет двупараметрическое семейство решений, обеспечивающих порядок аппроксимации $p=3$, но не может иметь 4-й порядок аппроксимации. Найдем это семейство решений.

Для этого выберем a_2, a_3 в качестве свободных параметров. Тогда уравнения (1.32) и (1.33), в которых надо положить $b_4=0$, образуют линейную систему относительно b_2, b_3. Находим два последние коэффициента, затем определим b_1 из (1.31) и a_{32} из (1.34). Получим следующие выражения коэффициентов:

$$b_3=\frac{2-3a_2}{6a_3(a_3-a_2)}, b_2=\frac{3a_3-2}{6a_2(a_3-a_2)}, b_1=1-b_2-b_3,$$
$$a_{32}=\frac{1}{(6b_3a_2)}, a_{31}=a_3-a_{32}, a_{21}=a_2. \tag{1.45}$$

Для неавтономных систем существует лишь тривиальное решение (1.45).

Можно показать, что существует область значений параметров a_2, a_3, обеспечивающая интерполяционность всех коэффициентов (1.45). Она показана на рис. 1.1.

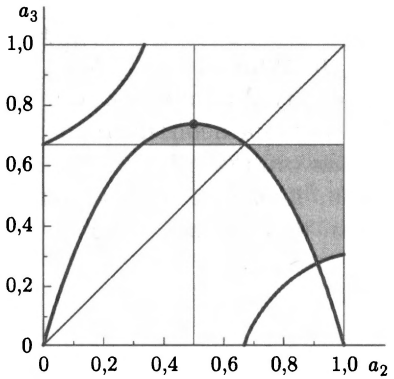


Рис. 1.1. Плоскость параметров a_2, a_3 для трехстадийных схем РК (заштрихованы области, обеспечивающие интерполяционность коэффициентов; точка — оптимальные параметры)

Оптимум. Двумя свободными параметрами a_2, a_3 можно удовлетворить двум из четырех оставшихся уравнений порядка. Последнему из них (1.38) удовлетворить нельзя, ибо трехстадийная схема не содержит коэффициента b_4. Особенно удобно удовлетворить уравнениям (1.36) — (1.37). Подставляя в них (1.45), получаем оптимальный набор коэффициентов:

$$\begin{gathered} a_2 = a_{21} = 1/2, a_3 = a_{32} = 3/4, a_{31} = 0, \\ b_1 = 2/9, b_2 = 3/9, b_3 = 4/9. \end{gathered} \tag{1.46}$$

Видно, что оптимальный набор интерполяционен. Если же попытаться удовлетворить уравнению (1.35) совместно с (1.36) либо (1.37), то получаются неинтерполяционные наборы, что хуже.

Невязка для оптимального набора (1.46) содержит две комбинации производных:

$$\psi = \frac{1}{6}\tau^3 \left(\frac{1}{48}\mathbf{f_{uuu}fff} + \mathbf{f_u f_u f_u f} \right) = O(\tau^3). \tag{1.47}$$

Однако первая комбинация входит с очень маленьким множителем, ибо уравнение (1.35) почти удовлетворяется для оптимального набора (1.46).

Вложение. Возьмем две первые стадии как самостоятельную двухстадийную схему. Ограничимся случаем оптимального

набора (1.46). Подставляя $a_2 = 1/2$ в общие формулы двухстадийной схемы (1.42), получим $\tilde{b}_2 = 1$, $\tilde{b}_1 = 0$. Это дает двухстадийную вложенную схему 2-го прядка аппроксимации. К сожалению, эта вложенная схема неоптимальна.

Аналогично, первая стадия с $\tilde{b}_1 = 1$ является схемой 1-го порядка аппроксимации, вложенной в найденную двухстадийную схему. Таким образом, получилась цепочка из трех рекуррентно вложенных схем.

1.1.6. Четыре стадии

Коэффициенты. Автономная 4-стадийная схема РК содержит 10 коэффициентов. Это позволяет удовлетворить 8 уравнениям порядка (1.31)–(1.38), причем два коэффициента остаются свободными параметрами. Получается двухпараметрическое семейство решений, обеспечивающее порядок аппроксимации $p = 4$.

Хотя нелинейная алгебраическая система (1.31) — (1.38) имеет весьма сложный вид, ее удается решить в радикалах. Приведем это решение без вывода. Примем за свободные параметры коэффициенты $a_2 \equiv a_{21}$ и a_{32}. Тогда остальные коэффициенты определяются следующими формулами:

$$
\begin{gathered}
a_3 = a_2/2 + \left[a_2^2/4 + 2a_2(1-2a_2)a_{32}\right]^{1/2}, a_{31} = a_3 - a_{32}; \\
a_4 = 1, a_{43} = \frac{(1-a_2)(1-a_3)}{2a_2a_{32}\left[3-4a_2-2(2-3a_2)a_{32}\right]}, \\
a_{42} = a_{43}\left[a_2(5-4a_2)a_{32} - (4a_2a_{32} - a_2 + 1)a_3\right] / \left[a_2(1-a_3)\right], \\
a_{41} = a_4 - a_{42} - a_{43}; \\
b_4 = 1/\left[24a_2a_{32}a_{43}\right], b_3 = 1/\left[24a_2a_{32}(1-a_3)\right], \\
b_2 = \frac{2a_2(3-2a_2)a_{32} - (4a_2a_{32} - a_2 + 1)a_3}{24a_2^2a_{32}(1-a_2)(1-a_3)}, b_1 = 1 - b_2 - b_3 - b_4.
\end{gathered}
\tag{1.48}
$$

Эти формулы записаны в таком порядке, что следующие коэффициенты выражаются через ранее вычисленные.

Следующий член разложения точного решения (1.26) порядка τ^5 содержит 9 различных комбинаций производных. Двумя свободными коэффициентами можно передать лишь две комби-

нации. Поэтому 4-стадийная явная схема РК не может иметь 5-й порядок аппроксимации.

Схема (1.48) имеет лишь тривиальное неавтономное обобщение, как в случае трех стадий.

Оптимум. Использовать два свободных параметра для точной передачи каких-то двух комбинаций производных в члене разложения $O(\tau^5)$ нецелесообразно: остается слишком много комбинаций, которые при этом не передаются. Более важен другой критерий. Есть лишь один набор коэффициентов, при котором 4-стадийная схема интерполяционна:

$$\begin{gathered} a_2 = a_{21} = a_3 = a_{32} = 1/2, a_4 = a_{43} = 1, \\ a_1 = a_{31} = a_{41} = a_{42} = 0, b_1 = b_4 = 1/6, b_2 = b_3 = 1/3. \end{gathered} \tag{1.49}$$

Этот набор построил Кутта (1901). Именно его следует считать оптимальным. Схема Кутты имеет хорошую точность и наиболее часто используется в инженерной практике. Для нее написаны стандартные программы.

Напомним, что если правая часть ОДУ не зависит от u, т. е. $f \equiv f(t)$, то решение задачи Коши для ОДУ сводится к вычислению интеграла. Легко видеть, что при этом 4-стадийная схема РК с оптимальными коэффициентами (1.49) принимает следующий вид:

$$\hat{u} = u + \tau \left[\frac{1}{6} f(t) + \frac{2}{3} f\left(t + \frac{\tau}{2}\right) + \frac{1}{6} f(t + \tau) \right].$$

Это квадратурная формула Симпсона, имеющая очень малый коэффициент в остаточном члене $O(\tau^4)$, как было показано в кн. 1. Данный предельный случай поясняет, почему схема Кутты оказалась столь точной.

Малый коэффициент в остаточном члене позволяет получать высокую точность при относительно крупном шаге. Это делает схему Кутты экономичной, несмотря на то, что на каждом шаге приходится 4 раза вычислять правую часть $\mathbf{f}$.

Вложение. Можно доказать, что у 4-стадийной схемы РК с аппроксимацией $O(\tau^4)$ отсутствует вложенная схема аппроксимации $O(\tau^3)$. Однако это обстоятельство не препятствует широкому практическому применению данной схемы.

Замечание. Из сказанного ранее видно, что у всех оптимальных схем с числом стадий $s \leq 4$ есть два свойства. Во-первых, порядок

аппроксимации равен числу стадий: $p = s$. Во-вторых, матрица Бутчера содержит только одну ненулевую линию — кодиагональ. Поэтому оптимальные схемы можно записать в упрощенной форме:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \sum_{k=1}^{s} b_k \mathbf{w}_k,$$
$$\mathbf{w}_k = f\left(\mathbf{u} + \tau a_k \mathbf{w}_{k-1},\ t + \tau a_k\right), 1 \le k \le s\ (s \le 4). \tag{1.50}$$

Запись дана для неавтономной задачи. Такая форма упрощает программирование. К сожалению, для многостадийных схем с $s \ge 5$ не найдено схем с такой простой формой записи.

Для $s = 4$ запись (1.50) нетривиальна: с помощью семи коэффициентов удается удовлетворить восьми уравнениям порядка. По аналогии Кутта предполагал, что и для $s \ge 5$ возможно удовлетворить уравнениям более высоких порядков с меньшим (чем число уравнений) числом коэффициентов. Но найти такие наборы коэффициентов пока не удалось.

1.1.7. Много стадий

Пороги Бутчера. Автономная явная схема РК с $s = 5$ стадиями имеет 15 коэффициентов: 5 коэффициентов b_k и 10 коэффициентов a_{kl}. Разложение точного решения (1.26) до члена $O(\tau^5)$ включительно содержит 17 различных комбинаций производных. Передать такое число комбинаций меньшим числом коэффициентов невозможно. Поэтому 5-стадийная явная схема РК не может иметь порядка аппроксимации $p = 5$. Таким образом, прибавление 5-й стадии, т. е. 5-го вычисления правой части $\mathbf{f}$ на шаге, не приводит к повышению порядка аппроксимации.

Это явление было подробно исследовано Бутчером. Оказалось, что только для явных схем РК с небольшим числом стадий $s \le 4$ возможен порядок аппроксимации $p = s$. При умеренном числе стадий $5 \le s \le 7$ максимально возможный порядок аппроксимации на 1 отстает от числа стадий: $p_{\max} = s - 1$. При большем числе стадий $s \ge 8$ отставание максимально возможного порядка аппроксимации от числа стадий снова увеличивается (табл. 1.2).

Границы, на которых увеличивается отставание, называют порогами Бутчера. Первый порог Бутчера есть переход от $s = 4$ к $s = 5$; отставание $s - p_{\max}$ увеличивается от 0 до 1. Второй порог

Таблица 1.2

Зависимость $p_{\max}(s)$ для явных схем РК (таблица поделена порогами Бутчера)

s	1	2	3	4	5	6	7	8	9	10	11	...
$p_{\max}$	1	2	3	4	4	5	6	6	7	7	8	...

есть переход от $s = 7$ к $s = 8$; здесь отставание увеличивается до 2. Начиная с многоточия отставание возрастает до 3 (хотя в литературе третий порог Бутчера трактуют немного иначе). При еще больших s пороги пока не изучены.

Схема Кутты (1.49)—(1.50) лежит непосредственно перед первым порогом Бутчера. Поэтому в ней эффективно используется каждое вычисление правой части.

Замечание. Все четыре оптимальные схемы с $s \leq 4$, лежащие ниже первого порога Бутчера, удовлетворяют требованиям интерполяционности. Пока не опубликовано ни одной схемы с $s < 4$, которая была бы интерполяционной. Быть может, существуют интерполяционные схемы с большим отставанием p от s. Но непосредственно перед вторым и следующим порогами Бутчера схема вряд ли может быть интерполяционной.

Схема Хаммуда. Непосредственно перед вторым порогом Бутчера лежат схемы с $s = 7$ стадиями. Бутчер показал, что есть 4-параметрическое семейство коэффициентов, обеспечивающее 6-й порядок аппроксимации для 7-стадийных схем РК, и описал алгоритм вычисления коэффициентов. Этот алгоритм громоздок и включает промежуточное решение алгебраических уравнений, так что явно выразить коэффициенты схем через свободные параметры не удается. Бутчер также нашел один конкретный набор коэффициентов в виде рациональных чисел; его коэффициенты не удовлетворяют условиям интерполяционности (1.21), причем максимально нарушает их $a_{74} \approx -3{,}0$.

Набор коэффициентов для $p = 6$ с существенно меньшей неинтерполяционностью нашел Хаммуд (2001); в нем наихудший коэффициент имеет те же индексы и равен $a_{74} \approx -1{,}27$. Все коэффициенты выражаются не через бесконечные десятичные дроби, а через радикалы, что намного удобнее для вычислений с максимальной разрядностью компьютера. В оригинальной работе имелись опечатки; приведем этот набор с исправленными опечатками:

$$
\begin{aligned}
a_{21} &= \frac{4}{7}; & a_{72} &= -\frac{425}{96} + \frac{51}{32}\sqrt{5};\\
a_{31} &= \frac{115}{112}; & a_{73} &= \frac{52}{15} - \frac{4}{5}\sqrt{5};\\
a_{32} &= -\frac{5}{16}; & a_{74} &= -\frac{27}{16} + \frac{3}{16}\sqrt{5};\\
a_{41} &= \frac{589}{630}; & a_{75} &= \frac{5}{4} - \frac{3}{4}\sqrt{5};\\
a_{42} &= \frac{5}{18}; & a_{76} &= \frac{5}{2} - \frac{1}{2}\sqrt{5};\\
a_{43} &= -\frac{16}{45}; & a_1 &= 0;\\
a_{51} &= \frac{229}{1200} - \frac{29}{6000}\sqrt{5}; & a_2 &= \frac{4}{7};\\
a_{52} &= \frac{119}{240} - \frac{187}{1200}\sqrt{5}; & a_3 &= \frac{5}{7};\\
a_{53} &= -\frac{14}{75} + \frac{34}{375}\sqrt{5}; & a_4 &= \frac{6}{7};\\
a_{54} &= -\frac{3}{100}\sqrt{5}; & a_5 &= \frac{5-\sqrt{5}}{10};\\
a_{61} &= \frac{71}{2400} - \frac{587}{12000}\sqrt{5}; & a_6 &= \frac{5+\sqrt{5}}{10};\\
a_{62} &= \frac{187}{480} - \frac{391}{2400}\sqrt{5}; & a_7 &= 1;\\
a_{63} &= -\frac{38}{75} + \frac{26}{375}\sqrt{5}; & b_1 &= \frac{1}{12};\\
a_{64} &= \frac{27}{80} - \frac{3}{400}\sqrt{5}; & b_2 &= b_3 = b_4 = 0;\\
a_{65} &= \frac{1+\sqrt{5}}{4}; & b_5 &= b_6 = \frac{5}{12};\\
a_{71} &= -\frac{49}{480} + \frac{43}{160}\sqrt{5}; & b_7 &= \frac{1}{12}.
\end{aligned}
\tag{1.51}
$$

Неавтономная схема (1.51) является тривиальным обобщением автономной. Вложенной схемы для (1.51) не построено, и неизвестно, возможно ли это.

Схема Хаммуда хорошо зарекомендовала себя в прикладных расчетах. Несмотря на сравнительную громоздкость, ее рекомендуется использовать в тех задачах, где точность 4-стадийной схемы Кутты (1.49) — (1.50) оказывается недостаточной.

Другие схемы. Отметим наиболее популярные многостадийные схемы РК, описанные в литературе. Укажем для каждой схемы число стадий, порядок точности, а также коэффициент, наиболее сильно нарушающий условия интерполяционности (1.21). Почти все эти схемы имеют вложенные схемы; в этом случае порядок вложенной схемы указан в скобках после порядка основной схемы.

Схема Фельберга: $s = 6, p = 5(4), a_{42} = -8$.

Схема Дормана—Принса: $s = 7, p = 5(4), a_{52} = -11,6$; по ней составлена известная программа с автоматическим выбором шага dopri5.

Схема Бутчера: $s = 7, p = 6, a_{74} = -3,0$; она упомянута ранее.

1.1.8. Общая характеристика

Явные схемы РК обладают рядом свойств, необходимых для компьютерных расчетов. 1) Они одношаговые, поэтому не требуют специальных формул начала счета; каждый шаг, включая первый, выполняется по одинаковым стандартным формулам. 2) Одношаговость позволяет легко менять шаг, так как во все формулы шага входит только величина τ с данного шага. 3) Многостадийные схемы имеют не только высокий порядок точности, но и малый численный коэффициент в остаточном члене. Это легко проверить следующим способом. Возьмем правую часть вида f(t). Тогда схема РК перейдет в некоторую квадратурную формулу. Например, наиболее распространенная четырехстадийная схема (1.48) примет при этом следующий вид:

$$u = \hat{u} + \tau/6\,[f(t) + 4f(t + \tau/2) + f(t + \tau)]\,.$$

Нетрудно видеть, что это квадратурная формула Симпсона, локальный остаточный член которой равен $\tau^5 f'''''(t)/2880$. В формуле Адамса того же порядка точности остаточный член имеет тот же вид, но с численным коэффициентом $1/3$; это в 960 раз больше! Даже если учесть, что схема (1.48) требует четырех вычислений $f(u,t)$ за один шаг, все равно она обеспечивает лучшую точность при одинаковом объеме вычислений.

Помимо схем РК и Адамса, в литературе описано много других явных схем. Однако пока среди них нет схем, превосходящих по сумме свойств схемы РК.

1.1.9. Сходимость

Ограничиваясь случаем равномерной сетки ($\tau = \text{const}$), дадим не вполне строгий вывод асимптотически точной оценки погрешности численного решения.

Для этого предположим, что $u(t)$ имеет $(p+1)$-у непрерывную производную; напомним, что для этого $f(u,t)$ должна иметь p непрерывных производных по всем своим аргументам. Разложим точное решение задачи Коши (1.4) — (1.5) в ряд Тейлора по степеням шага:

$$\hat{u} = u + \sum_{m=1}^{p+1} \tau^m A_m(u,t) + o(\tau^{p+1}), A_m(u,t) = d^m u(t)/dt^m. \quad (1.52)$$

Напомним, что $A_1(u,t) = f(u,t)$, а $A_m(u,t)$ выражаются через различные комбинации производных правой части (см. п. 1.1.3).

В этом пункте будем обозначать численное решение буквой $v(t)$, где t принимает только дискретные значения t_m. Для перехода сеточного решения на новый шаг справедливо соотношение вида (1.52). Пусть коэффициенты схемы РК выбраны так, что в разложении правильно передаются все комбинации производных до члена $O(\tau^p)$ включительно; следующую же комбинацию передать не удается. Тогда имеем

$$\hat{v} = v + \sum_{m=1}^{p} \tau^m A_m(v,t) + \tau^{p+1} [A_{p+1}(v,t) + \psi(v,t)] + o(\tau^{p+1}). \quad (1.53)$$

Здесь $\psi(v,t)$ есть та часть комбинации старших производных, которую схема не может передать; величину $\tau^p \psi(v,t)$ называют ***невязкой***.

Введем погрешность $z(t) = v(t) - u(t)$; она определена также лишь на дискретном наборе t_m. Будем считать, что шаг сетки достаточно мал, т. е. точность расчета хорошая, и $z(t)$ невелико. Подставим $z + u$ вместо v в (1.53) и вычтем (1.52) из (1.53). Получим

$$\hat{z} = z + \sum_{m=1}^{p+1} \tau^m [A_m(u+z,t) - A_m(u,t)] + \tau^{p+1} \psi(u+z,t) + o(\tau^{p+1}). \quad (1.54)$$

В (1.54) перенесем z в левую часть, поделим все на τ и разложим квадратные скобки по малому z:

$$(\hat{z} - z)/\tau = z \sum_{m=1}^{p+1} \tau^{m-1} \partial A_m(u,t)/\partial u + \tau^p \psi(u,t) + O(z^2) + o(\tau^p). \tag{1.55}$$

Главными членами в правой части (1.55) являются невязка и первый член суммы, который, с учетом вида A_1, равен f_u. Сделаем нестрогое предположение, что $z(t)$ можно считать дифференцируемой функцией непрерывного аргумента, и заменим левую часть (1.55) на производную. С учетом сказанного (1.55) перейдет в линейное неоднородное дифференциальное уравнение

$$dz/dt \approx f_u(u,t)z + \tau^p \psi(u,t). \tag{1.56}$$

На точном решении $u(t)$ коэффициенты уравнения (1.56) являются заданными функциями от t. Поэтому решение этого уравнения выражается через квадратуры:

$$\begin{gathered} z(t) \approx z_0 b_0(t_0,t) + \tau^p b_1(t_0,t), \\ b_0(\eta,t) = \exp \int_\eta^t f_u(\xi) d\xi, \\ b_1(t_0,t) = \int_{t_0}^t \psi(\eta) b_0(\eta,t) d\eta. \end{gathered} \tag{1.57}$$

Обсудим выражение (1.57).

Первое слагаемое в правой части показывает влияние начальной ошибки z_0. Если начальное значение в расчете задается точно (со всеми знаками компьютера), то это слагаемое выпадает. Если же начальная ошибка $z_0 \neq 0$, то ее влияние определяется множителем $b_0(t_0,t)$. Когда f_u невелико, этот множитель также невелик, и влияние начальной ошибки мало. В этом случае задача Коши является хорошо обусловленной и ее называют ***мягкой***. Если f_u принимает большие отрицательные значения, то $b_0(t_0,t)$ будет еще меньшей величиной, и задача Коши будет еще лучше обусловленной; такие задачи называют ***жесткими***. Когда же f_u принимает большие положительные значения, то множитель b_0 может становиться огромным, и задача Коши

является ***плохо обусловленной***. Плохо обусловленные задачи не всегда удается сосчитать на компьютере из-за накопления ошибок.

Второе слагаемое в (1.57) не исчезает и не зависит от начальной ошибки. Оно показывает вклад невязки в погрешность, и есть $O(\tau^p)$. Отсюда видно, что если с учетом этого оценить отброшенные ранее члены $O(z^2)$ и $o(\tau^p)$, то нетрудно убедиться, что ими действительно можно было пренебречь. Таким образом, второй член в (1.57) дает оценку погрешности, асимптотически точную при $\tau \to 0$. Эта погрешность есть $O(\tau^p)$.

Таким образом, если правая часть $f(u,t)$ имеет непрерывные p-е производные по всем своим аргументам, а разностная схема имеет аппроксимацию порядка p, то точное решение имеет погрешность $O(\tau^p)$, т. е. p-й порядок точности.

Описанный вывод относится не только к схемам РК, но и к любым другим схемам.

Замечание. Из (1.54) видно, что погрешность $z(t)$ разлагается в ряд, содержащий все степени τ. Это существенно отличается от погрешности квадратурных формул трапеций, средних, Симпсона и др., описанных в кн. 1: в них погрешность разлагалась в ряд только по четным степеням шага. Такая разница объясняется тем, что формулы РК (как и практически любые схемы решения задачи Коши) построены несимметрично, а названные квадратурные формулы симметричны. Асимметрия ухудшает сходимость ряда по степеням τ и замедляет выход погрешности на асимптотическую оценку (1.57).

1.1.10. Контроль точности

Единственным надежным методом, обеспечивающим гарантированную оценку погрешности расчета, является ***глобальное сгущение сетки***. Этот способ основан на методе Ричардсона, подробно описанном в кн. 1. Изложим алгоритм применения этого метода к дифференциальным уравнениям.

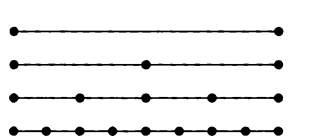


Рис. 1.2. Глобальные сгущения сеток в два раза

Построим последовательность равномерных сеток с шагами τ, $\tau/2$, $\tau/4$ и т. д. Узлы каждой сетки совпадают с четными узлами следующей сетки (рис. 1.2), т. е. соответствуют одному и тому же моменту t. Решим по выбранной схеме задачу Коши на каждой сетке. Рассмотрим численные ре-

шения на двух соседних сетках. Обозначим решение на первой сетке через $v_1(t)$ и на второй — $v_2(t)$. Согласно правилу Ричардсона (см. кн. 1), асимптотически точная оценка погрешности в данной точке определяется формулой

$$\Delta(t) = [v_2(t) - v_1(t)]\,/(r^p - 1); \tag{1.58}$$

здесь $r = 2$ есть коэффициент сгущения сетки, а p — теоретический порядок точности численного метода. Эта оценка погрешности относится к численному решению $v_2(t)$, т. е. к более подробной сетке.

Напомним, что оценка погрешности называется асимптотически точной, если при $\tau \to 0$ она становится сколь угодно близкой к фактической погрешности. Поэтому ее можно прибавить к решению на подробной сетке и получить уточненное решение:

$$\widetilde{v}_2(t) = v_2(t) + \Delta(t). \tag{1.59}$$

В этом уточненном решении исключена погрешность $O(\tau^p)$. Поскольку в схемах РК погрешность разлагается в ряд по всем степеням τ, погрешность $\widetilde{v}_2(t)$ будет $O(\tau^{p+1})$; порядок точности повысится лишь на 1.

Уточнение (1.59) вычисляется только в четных узлах подробной сетки t_{2n}, поскольку погрешность (1.58) непосредственно вычисляется только в этих узлах. Однако можно интерполировать эту погрешность на нечетные узлы. Ограничимся простейшей линейной интерполяцией:

$$\Delta(t_{2n+1}) = [\Delta(t_{2n}) + \Delta(t_{2n+2})]\,/2. \tag{1.60}$$

Сама погрешность Δ в четных узлах есть $O(\tau^p)$. Линейная интерполяция (1.60) вносит в нечетных узлах дополнительную относительную погрешность $O(\tau^2)$; значит, дополнительная абсолютная погрешность есть $O(\tau^{p+2})$. Она на один порядок меньше, чем погрешность формулы уточнения (1.59). Поэтому ей можно пренебречь и прибавлять величину интерполированной погрешности (1.60) к значениям $v_2(t_{2n+1})$. Это позволяет получить уточненные значения $\widetilde{v}_2(t)$ в нечетных узлах подробной сетки. Таким образом, мы получаем уточненное решение с погрешностью $O(\tau^{p+1})$ во всех узлах подробной сетки, почти не увеличивая общей трудоемкости расчета.

Анализ погрешности в каждом узле сетки нецелесообразен. Обычно достаточно рассмотреть две наиболее употребительные нормы погрешности:

$$\|\Delta\|_c = \max_{1\le n\le N} |\Delta(t_n)|\,, \|\Delta\|_{l_2} = \left[1/N \sum_{n=1}^{N} \Delta^2(t_n)\right]^{1/2}; \qquad (1.61)$$

здесь N — полное число точек в той сетке, по которой вычисляются данные нормы.

Опишем процедуру достижения заданной точности ϵ. Проведем расчеты на первой и второй сетках и вычислим нормы погрешностей (1.61). Если они меньше ϵ, то расчет прекращается: заданная точность достигнута. В противном случае сгущаем вторую сетку вдвое и аналогично находим погрешность и уточненное решение для новой пары сеток: второй и третьей. Повторяем эту процедуру до тех пор, пока нормы погрешностей не станут меньше ϵ.

Заметим, что итоговая погрешность относится к **не**уточненному решению на последней сетке. Погрешность уточненного решения на последней сетке мы не знаем, но интуитивно предполагаем, что она еще на один порядок по τ меньше. Поэтому обычно выдают последнее уточненное решение в качестве ответа, а последнюю погрешность — в качестве заведомо завышенной оценки точности ответа.

Первую сетку обычно выбирают с небольшим числом узлов N. Для окончательного ответа нужны лишь последняя и предпоследняя сетки. Может показаться, что при этом тратится много времени на расчеты на первых, довольно грубых сетках. Однако на самом деле суммарный объем расчетов на всех предшествующих сетках по трудоемкости равен одной трети расчета на двух последних сетках. Так что общее увеличение трудоемкости незначительно. Наличие же расчетов на многих сетках, как будет показано далее, повышает надежность процедуры контроля.

Визуализация. Выведем на экран компьютера зависимость нормы погрешности от шага τ (или числа узлов $N \sim 1/\tau$). Поскольку асимптотика погрешности имеет степенной закон, целесообразен двойной логарифмический масштаб: $\log_{10} N$ по оси абсцисс и $\log_{10} \|\Delta\|$ по оси ординат. Типичный пример такого графика приведен на рис. 1.3. На первых грубых сетках погрешность еще не успела выйти на свою асимптотику и линии имеют заметную кривизну (на трудных задачах они могут иметь и более сложный характер). При увеличении N происходит выход на асимптотику и линии приближаются к прямым с наклоном

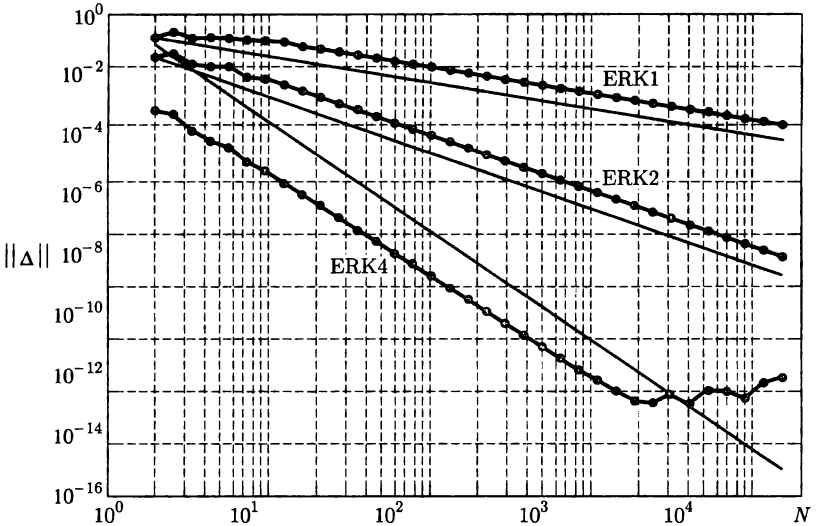


Рис. 1.3. Типичное поведение погрешности для схем РК (точки — численные расчеты, они соединены жирными линиями, нормы указаны около кривых; тонкие линии — теоретический наклон)

$\mathrm{tg}(\alpha) = -p$, где p — теоретический порядок точности (чтобы легче заметить это, полезно на графике провести прямую с указанным наклоном). Если мы видим на графике четкий выход на теоретический наклон, то можно останавливать расчет по достижению заданной точности.

Заметим, что для схем РК выход на асимптотику происходит позднее, чем для симметричных квадратурных формул, так как погрешность разлагается в ряд по степеням τ, а не τ^2. Кроме того, для нормы l_2 выход на теоретический наклон происходит обычно заметно раньше, чем для нормы c, и сама точность ϵ достигается также раньше. Вычислитель должен решить, по какой из норм останавливать расчет, исходя из прикладного смысла задачи.

На очень подробных сетках погрешность перестает убывать. Это означает, что расчет вышел на ошибки округления. Уровень ошибок округления зависит от трудности задачи. Он может оказаться бóльшим, чем заданная вычислителем точность ϵ. В таком случае заданная точность не может быть достигнута. Такая ситуация легко обнаруживается при визуальном контроле.

Визуальный контроль наиболее надежен, но у вычислителя не всегда есть время для работы в интерактивном режиме. Для

автоматического контроля целесообразно сравнивать отношения норм погрешностей на соседних сетках. Если эти отношения близки к r^p (где $r = 2$ — коэффициент сгущения сетки), то поведение погрешности близко к асимптотическому. Удобно ввести эффективный порядок точности

$$p_{eff} = \left[\log(\|\Delta_\tau\| \,/\, \|\Delta_{\tau/2}\|)\right] / \log r. \tag{1.62}$$

Если $|p_{eff} - p| < 0{,}05$, то результат можно считать удовлетворительным. Вдобавок, при дальнейшем сгущении сеток p_{eff} должно монотонно стремиться к p. В этом случае можно применять описанную процедуру определения погрешности. Нарушение монотонного хода $p_{eff}(N)$ обычно указывает на приближение к ошибкам округления.

Автоматический выбор шага. В литературе описано много алгоритмов автоматического выбора шага τ. Их идея состоит в том, чтобы уменьшать шаг там, где решение быстро меняется, и увеличивать шаг на участках слабого изменения решения. Для определения τ в них исследуется локальная погрешность, вносимая на шаге. Обычно используют два способа.

Первый способ основан на локальном сгущении сетки. Задается допустимая на шаге погрешность δ. Если предыдущий шаг имел величину τ, то новый шаг рассчитывают дважды, с шагами τ и $\tau/2$. Сравнивают эти два результата между собой по правилу Ричардсона. Если внесенная на шаге погрешность близка к δ, то результат принимают. Если эта погрешность заметно отличается от δ, то соответственно увеличивают или уменьшают величину τ, чтобы сделать ее равной δ.

Второй способ использует расчет по основной схеме порядка p и вложенной в нее схеме порядка $p - 1$. Разность этих результатов можно рассматривать как локальную погрешность вложенной схемы. Если она близка к δ, то результат принимается. В противном случае шаг корректируется.

Оба способа не учитывают того, что внесенная на шаге локальная погрешность затем домножается на экспоненциальный множитель из (1.57). В зависимости от знака f_u этот множитель гасит или усиливает локальную ошибку. Поэтому окончательная точность может существенно отличаться от заданной точности δ. Тем самым эти методы не гарантируют декларированной точности, и к ним нужно относиться осторожно. В прикладных расчетах фактическая погрешность нередко оказывается в 10—100 раз больше или меньше заявленной.

1.2. ЖЕСТКИЕ СИСТЕМЫ

1.2.1. Классификация систем

Далеко не всегда задачу Коши (1.4) удается решить явными схемами. Например, есть такие важные прикладные задачи, как цепочки химических реакций или процессы в сложных радиоконтурах. Для получения разумных результатов в них требовался неприемлемо малый шаг. Ситуацию можно иллюстрировать на несложном примере — так называемом тесте Далквиста (который интерпретируется как радиоактивный распад):

$$du/dt = \lambda u, 0 \le t \le T, u(0) = u_0, \lambda T \ll -1. \qquad (1.63)$$

Здесь λ — большая отрицательная величина, а точное решение задачи (1.63) есть $u(t) = u_0 \exp(\lambda t)$. Оно сохраняет знак и быстро стремится к 0.

Попробуем решить задачу (1.63) явными схемами РК. Простейшая явная схема Эйлера дает

$$\hat{u} = u + \tau\lambda u \equiv R_1(z)u, R_1(z) = 1 + z, z = \tau\lambda; \qquad (1.64)$$

множитель $R_1(z)$ называют ***функцией устойчивости***. Чтобы численное решение при $\lambda < 0$ затухало, надо взять очень малый шаг $\tau < 2/|\lambda|$. Полное число шагов при этом будет $N = T/\tau > T|\lambda| \gg 1$ согласно условиям задачи (1.63). Если же потребовать, чтобы решение затухало с сохранением своего знака, то надо взять τ еще вдвое меньше.

Аналогичные результаты получаются для многостадийных явных схем РК, приведенных в подразделе 1.1. Нетрудно проверить, что для них формулы перехода на новый шаг аналогичны (1.64), но с другими функциями устойчивости. Например, для s-стадийных явных схем РК с порядком аппроксимации $p = s$ получим следующие функции устойчивости:

$$\begin{aligned} R_2(z) &= 1 + z + z^2/2, \\ R_3(z) &= 1 + z + z^2/2 + z^3/6, \\ R_4(z) &= 1 + z + z^2/2 + z^3/6 + z^4/24. \end{aligned} \qquad (1.65)$$

Напомним, что $p = s$ возможно только при $s \le 4$. Нетрудно заметить, что функции устойчивости (1.64) — (1.65) являются

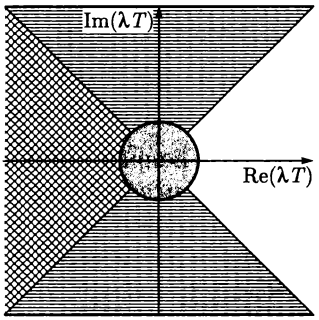


Рис. 1.4. Классификация задач Коши (фон — мягкие, крестообразная штриховка — жесткие, горизонтальная — быстро осциллирующие, отсутствие штриховки — плохо обусловленные)

отрезками ряда Тейлора для $\exp(z)$. Для всех этих схем ограничения на шаг имеют вид $\tau = O(|\lambda|^{-1})$, но с различными константами.

В задачах радиоактивного распада λ может быть большой отрицательной величиной, а точное решение — быстро затухающей функцией. Такие задачи называют ***жесткими***. Заметим, что такие задачи хорошо обусловлены: при небольшом изменении u_0 точное решение все равно быстро затухает. В задачах теплового взрыва возникают большие положительные λ, а точное решение быстро возрастает. Эти задачи являются ***плохо обусловленными***, так как при небольшом изменении u_0 интегральные кривые сильно расходятся. В электрических контурах возможны комплексные λ с большой мнимой частью, когда точное решение быстро осциллирует. Поэтому для наглядной классификации таких задач рассматривают величину λT на комплексной плоскости на рис. 1.4 (разумеется границы областей на этом рисунке достаточно условны).

Для мягких задач (к которым относится очень много инженерных приложений) хорошо применимы явные методы РК. Для остальных типов задач явные методы требуют неприемлемо малого шага τ и нужно рассматривать другие классы схем.

1.2.2. Устойчивость

Тест Далквиста (1.63) содержит линейное однородное дифференциальное уравнение. Если численный алгоритм также описывается линейной формулой (как, например, все явные схемы РК), то переход на новый шаг приводится к форме $\hat{u} = R(z)u$, аналогично (1.64). Разумеется, вид функции устойчивости зависит от конкретной схемы.

Если λ комплексна, то $z = \tau\lambda$ также комплексно. Наиболее важен для нас случай, когда $\mathrm{Re}(\lambda) < 0$; при этом точное решение затухает. Будем требовать от схемы, чтобы численное решение при этом тоже затухало. Поскольку точное решение затухает экспоненциально, т. е. достаточно быстро, желательно достаточно быстрое затухание численного решения. Поэтому введем следующие определения.

Определение 1.3. Схема называется *A-устойчивой*, если $|R(z)| \leq 1$ при $\mathrm{Re}(z) \leq 0$.

Определение 1.4. Схема называется *L-устойчивой*, если она A–устойчива, и $R(z) \to 0$ при $z \to \infty$.

Определение 1.5. Схема называется *Lp-устойчивой*, если она A–устойчива, и $R(z) = O(z^{-p})$ при $z \to \infty$.

Определение 1.6. Схема называется *t-монотонной*, если она A-устойчива, и $R(z) > 0$ при вещественном отрицательном z.

Если схема удовлетворяет определению 1.3, то для жестких задач численное решение будет убывать по модулю при переходе на новый шаг. Очевидно, это свойство необходимо для того, чтобы схема позволяла решать жесткие задачи при не слишком маленьком шаге τ. Если это свойство отсутствует, то расчет задач большой жесткости потребует неприемлемо малого шага τ, т. е. будет практически невозможным.

A-устойчивость еще не гарантирует быстрого убывания численного решения за один шаг (примеры этого увидим далее). Для задач большой жесткости, когда $|z| \gg 1$, L-устойчивость обеспечивает достаточно быстрое убывание численного решения. Еще более конструктивна Lp-устойчивость: она уточняет скорость этого убывания. Это убывание является степенным, т. е. оно медленнее экспоненциального убывания точного решения. Чем больше p, тем ближе убывание численного решения к экспоненциальному.

Наконец, t-монотонность обеспечивает монотонное убывание численного решения, если λ вещественная отрицательная. При этом поведение численного решения будет качественно похоже на поведение точного решения. Особенно хорошим будет качественное соответствие, если схема одновременно Lp-устойчива и t-монотонна.

Есть неплохие схемы, которые не вполне удовлетворяют приведенным определениям. Поэтому рассматривают некоторые

обобщения. Например, называют схему $A(\alpha)$-устойчивой, если $|R(z)| \leq 1$ не во всей левой полуплоскости комплексного z, а в лишь в области, отрезанной углами α от мнимой оси. Аналогично определяются $L(\alpha)$-устойчивость и $Lp(\alpha)$-устойчивость. Если угол α мал, а $|R(z)|$ в левой полуплоскости не сильно превышает 1 (максимальное превышение обычно лежит на мнимой оси), то можно надеяться на удовлетворительные результаты. В противном случае схемы обычно ненадежны.

Явные схемы РК низших порядков имеют функции устойчивости (1.64)–(1.65). Это многочлены от z, и в левой полуплоскости они неограниченно растут при $z \to \infty$. Нетрудно заметить, что у явных схем РК с числом стадий $s > 4$ функции устойчивости также будут многочленами. Поэтому все явные схемы РК не являются A-устойчивыми или $A(\alpha)$-устойчивыми. Тем самым они непригодны для задач большой жесткости.

Доказано, что не только схемы РК, но и любые другие явные схемы не могут быть A-устойчивыми.

1.2.3. Одностадийные схемы Розенброка

Поскольку явные схемы малопригодны для жестких систем, рассмотрим построение неявных схем. Например, запишем простейшую одностадийную схему РК, которую называют чисто неявной, или обратной схемой Эйлера:

$$\hat{u} = u + \tau f(\hat{u}). \tag{1.66}$$

Для линейного теста Далквиста (1.63) схема (1.66) приводится к каноническому виду:

$$\hat{u} = R(z)u, R(z) = 1/(1-z), z = \tau\lambda. \tag{1.67}$$

Если λ вещественная отрицательная, т. е. $-\infty < z < 0$, то $0 < R(z) < 1$ при любом шаге τ. Численное решение теста Далквиста становится монотонно убывающим, как и точное решение $\exp(z)$.

Таким образом, для линейных жестких задач неявная схема (1.66) — (1.67) дает разумные результаты. Однако для нелинейных задач значение $\hat{u}$ находится не из (1.67), а из решения нелинейного уравнений (1.66). Решать нелинейное уравнение можно только каким-либо итерационным процессом (см. кн. 1). Для жестких систем такие итерационные процессы далеко не всегда

сходятся. Еще труднее добиться сходимости, если решается не одно дифференциальное уравнение, а система дифференциальных уравнений, т. е. u, f являются векторами, а (1.66) становится системой нелинейных алгебраических уравнений относительно вектора $\hat{u}$.

Поэтому не будем добиваться сходимости итерационного процесса, а просто возьмем первую итерацию метода Ньютона как самостоятельную схему. Для этого разложим в ряд Тейлора $f(\hat{u}) \approx f(u) + f_u(u)(\hat{u} - u)$. Для системы уравнений f_u есть матрица производных. Подставляя это приближение в (1.66), приведем ее к следующему виду:

$$[E - \tau f_u(u)]\,(\hat{u} - u) = \tau f(u); \tag{1.68}$$

здесь E-единичная матрица. Полученная схема является неявным алгебраическим выражением относительно $\hat{u}$. Однако нахождение $\hat{u} - u$ свелось к решению системы линейных уравнений, а это можно сделать прямыми методами (например, методом Гаусса) за конечное число действий. Поэтому схемы типа (1.68) называют ***явно-неявными*** или ***полуявными***.

Для теста Далквиста схема (1.68) переходит в (1.67). Поэтому она удовлетворительна для жестких систем.

Семейство ROS1. Обобщим схему (1.68), введя свободные скалярные параметры, которые можно варьировать для получения тех или иных свойств. Схему запишем для неавтономной задачи:

$$\hat{u} = u + \tau b w, [E - a\tau f_u(u,t)]\, w = f(u, t + c\tau). \tag{1.69}$$

Это одностадийная схема; она является первой стадией семейства схем, предложенных Розенброком (1963). Проведем исследование ее свойств.

Аппроксимация. Подставим в (1.69) разложение в ряд Тейлора: $f(u, t+c\tau) = f + c\tau f_t + O(\tau^2)$; здесь отсутствие аргумента у функции или производных означает, что они берутся на исходном шаге. Обратим оператор в квадратных скобках, учитывая малость τ:

$$\begin{aligned} w &= [E - a\tau f_u]^{-1} f(u, t + c\tau) = \\ &= \left[E + a\tau f_u + O(\tau^2)\right](f + c\tau f_t + O(\tau^2)) = \\ &= f + \tau(a f_u f + c f_t) + O(\tau^2). \end{aligned} \tag{1.70}$$

Подставляя это разложение в (1.69), найдем разложение численного решения в ряд Тейлора:

$$\hat{u} = u + \tau bf + \tau^2(baf_u f + bcf_t) + O(\tau^3). \tag{1.71}$$

Сравним его с разложением точного решения в ряд Тейлора (1.26). Видно, что для согласования членов $O(\tau)$, т. е. для 1-го порядка аппроксимации, необходимо положить $b = 1$. Можно добиться 2-го порядка аппроксимации, если дополнительно положить $a = c = 1/2$. Таким образом, одностадийная схема Розенброка может иметь второй порядок аппроксимации (напомним, что одностадийная явная схема РК могла иметь только первый порядок аппроксимации).

A-устойчивость. Напишем функцию устойчивости схемы (1.69) на тесте Далквиста. Для этого произведем в схеме следующие замены: $E \to 1, f \to \lambda u, f_u \to \lambda$. Поскольку схемы хуже первого порядка аппроксимации бессмысленны, сразу полагаем $b = 1$. Тогда получим

$$R(z) = \frac{1 + (1-a)z}{1 - az}. \tag{1.72}$$

Найдем область комплексного переменного z, в которой модуль $|R(z)| \leq 1$. Последнее условие можно переписать как $|R(z)|^2 \equiv \equiv R(z)R(z^*) \leq 1$, где $*$ означает комплексное сопряжение. Последнее условие можно переписать в следующем виде:

$$[1 + (1-a)z]\,[1 + (1-a)z^*] \leq (1 - az)(1 - az_*). \tag{1.73}$$

Введем вещественную и мнимую части: $z = r + \mathtt{i}j$, где $\mathtt{i} = \sqrt{-1}$. Тогда последнее неравенство принимает форму

$$2r \leq (2a - 1)(r^2 + j^2). \tag{1.74}$$

В левой полуплоскости $r \leq 0$, и тогда в левой части (1.74) стоит отрицательная величина. Видно, что если $a \geq 1/2$, то в правой части стоит неотрицательная величина и неравенство выполняется. Таким образом, при $a \geq 1/2$ схема (1.69) является A-устойчивой.

Если $a < 1/2$, то первая круглая скобка в (1.74) отрицательна. Тогда, какой бы ни было взято $r < 0$, всегда найдется настолько большое j, что правая часть (1.74) будет меньше левой и неравенство не выполнится. Поэтому при $a < 1/2$ схема (1.69) не является A-устойчивой. Таким образом, доказана теорема.

Теорема 1.4. Условие $a \geq 1/2$ необходимо и достаточно для A–устойчивости схемы (1.69). •

Параметр c на A-устойчивость не влияет.

L-устойчивость. Асимптотическое значение функции устойчивости (1.72) есть $R(\infty) = (a-1)/a$. При $a = 1$ выполняется $R(\infty) = 0$, т. е. схема становится L-устойчивой. При других значениях параметра a схема (1.69) не является L-устойчивой.

Результат при $a = 1$ можно усилить. Функция устойчивости принимает вид $R(z) = 1/(1-z)$, так что $R(z) = O(1/z)$ при $z \to \infty$. Следовательно, при $a = 1$ схема (1.69) является $L1$-устойчивой. При этом она переходит в обратную схему Эйлера.

Монотонность. Она связана с поведением $R(z)$ на вещественной отрицательной полуоси. Из вида функции устойчивости ясно, что при $a < 1$ на этой полуоси $R(z)$ положительно при небольших значениях z, и становится отрицательным при $z < -1/(1-a)$; в этом случае схема немонотонна. Немонотонность становится особенно сильной при $a = 1/2$, когда $R(\infty) = -1$; при этом на задачах большой жесткости численное решение почти не затухает.

Если $a \geq 1$, то на отрицательной вещественной полуоси $R(z) > 0$ и схема становится t-монотонной.

Рекомендации. При практических вычислениях используют только одно из двух вещественных значений параметра a. Если задача является очень жесткой, то полагают $a = 1$. Получают обратную схему Эйлера, которая $L1$-устойчива и t-монотонна, т. е. обеспечивает хорошее качественное поведение численного решения. Однако она имеет лишь первый порядок точности, что является платой за высокую надежность.

Во втором случае берут $a = 1/2$, получая схему «с полусуммой». Это схема второго порядка точности. Зато она лишь A-устойчива; ни L-устойчивости, ни t-монотонности у нее нет, так что надежность ее невелика. Поэтому ее применяют лишь на задачах умеренной жесткости. На задачах большой жесткости в этой схеме решение может становиться знакопеременным в тех ситуациях, когда точное решение обязано быть знакопостоянным.

Для второго порядка необходимо $c = 1/2$. Поскольку c не влияет на устойчивость, то во всех случаях полагают $c = 1/2$. Даже на обратной схеме Эйлера это аннулирует производную

f_t в остаточном члене, что несколько улучшает фактическую точность.

1.2.4. Комплексная схема Розенброка

Можно взять коэффициент a схемы (1.69) комплексным. Тогда потребуется решать систему линейных уравнений с комплексной матрицей, и величина w также станет комплексной. Поэтому для получения вещественного $\hat{u}$ придется брать $\mathtt{Re}(w)$. Было показано, что оптимальной в классе комплексных коэффициентов является схема с $a = (1+\mathtt{i})/2$. Она имеет следующий вид:

$$\hat{u} = u + \tau \mathtt{Re}(w), \left[E - \frac{1+\mathtt{i}}{2}\tau f_u(u,t)\right] w = f(u, t+\tau/2). \quad (1.75)$$

Докажем, что эта схема является $L2$-устойчивой, t-монотонной, и имеет точность $O(\tau^2)$.

Точность. Явно вычислим w с точностью $O(\tau^2)$, обращая оператор в квадратных скобках и разлагая правую часть в ряд Тейлора:

$$\begin{aligned} w &= \left[E - \frac{1+\mathtt{i}}{2}\tau f_u(u,t)\right]^{-1} f(u, t+\tau/2) = \\ &= \left[E + \frac{1+\mathtt{i}}{2}\tau f_u + O(\tau^2)\right] \left[f + \tau/2 f_t + O(\tau^2)\right] = \\ &= f + \tau\left(\frac{1+\mathtt{i}}{2} f_u f + \frac{1}{2} f_t\right) + O(\tau^2); \end{aligned} \quad (1.76)$$

здесь функции и производные без аргумента считаются взятыми в исходной точке. Беря вещественную часть, находим

$$\hat{u} = u + \tau f + \frac{\tau^2}{2}(f_u f + f_t) + O(\tau^3). \quad (1.77)$$

В полученном выражении выписанные комбинации производных совпадают с аналогичными членами разложения точного решения (1.26). Тем самым, комплексная схема Розенброка (CROS) имеет второй порядок аппроксимации.

Устойчивость. На тесте Далквиста надо полагать $E \to 1$, $f \to \lambda u, f_u \to \lambda$. Проводя соответствующие замены в (1.75), получим

$$\left(1-\frac{1+\mathtt{i}}{2}\tau\lambda\right)w=\lambda u. \tag{1.78}$$

Умножим это равенство на комплексно сопряженную скобку $\left(1-\frac{1-\mathtt{i}}{2}\tau\lambda\right)$:

$$\left[1-\tau\lambda+\frac{1}{2}(\tau\lambda)^2\right]w=\left(1-\frac{1-\mathtt{i}}{2}\tau\lambda\right)\lambda u. \tag{1.79}$$

Поскольку в квадратных скобках стоит вещественное выражение, то для взятия $\mathtt{Re}(w)$ достаточно отбросить в правой части $\mathtt{i}$: $\mathtt{Re}(w)=\left(1-\frac{\tau\lambda}{2}\right)\lambda u/\left[1-\tau\lambda+\frac{1}{2}(\tau\lambda)^2\right]$. Подставляя последнее выражение в схему (1.75), получим ее функцию устойчивости:

$$\hat{u}=R(z)u,\ R(z)=1/\left(1-z+\frac{z^2}{2}\right),z=\tau\lambda. \tag{1.80}$$

Функция устойчивости оказалась так называемой Паде-аппроксимацией экспоненты: вместо ряда Тейлора для $\exp(z)$ берется дробь, знаменатель которой есть ряд Тейлора для $\exp(-z)$.

Найдем область, в которой $|R(z)|^2\equiv R(z)R(z^*)\le 1$. С учетом (1.80) это неравенство переписывается в следующем виде:

$$\left(1-z+\frac{z^2}{2}\right)\left(1-z^*+\frac{(z^*)^2}{2}\right)\ge 1. \tag{1.81}$$

Перемножая скобки, получим

$$\frac{1}{4}(zz^*)^2-\frac{1}{2}(z+z^*)zz^*+\frac{1}{2}(z+z^*)^2-(z+z^*)\ge 0. \tag{1.82}$$

Подставим сюда $z=r+\mathtt{i}j, z^*=r-\mathtt{i}j$ и получим

$$\left(r^2+j^2\right)^2-4r\left(r^2+j^2\right)+8r^2-8r\ge 0, \tag{1.83}$$

или

$$\left[\left(r^2-2r\right)+j^2\right]^2+4\left(r^2-2r\right)\ge 0. \tag{1.84}$$

Это неравенство выполняется ***вне*** области, ограниченной замкнутой кривой четвертого порядка:

$$j^2=\sqrt{r(2-r)}\left[2-\sqrt{r(2-r)}\right],0\le r\le 2. \tag{1.85}$$

Кривая (1.85) лежит в правой полуплоскости, касаясь мнимой оси в начале координат. Тем самым в левой полуплоскости неравенство (1.84) выполняется и схема CROS является A-устойчивой.

Функция устойчивости (1.80) есть $O(z^{-2})$ при $z \to \infty$. Тем самым, схема CROS является $L2$-устойчивой. Это наилучший показатель устойчивости среди всех известных одностадийных схем.

Монотонность. Видно, что для вещественных $z < 0$ функция устойчивости $R(z) > 0$ и монотонно убывает при $z \to -\infty$. Тем самым, схема CROS является t-монотонной. Это свойство повышает надежность схемы.

На рис. 1.5 приведены профили $R(z)$ при вещественных $z < 0$ для точного решения и рассмотренных здесь схем. Видно, что для схемы CROS функция устойчивости наиболее близка к точной и качественно и количественно. Для обратной схемы Эйлера функция устойчивости качественно правильная, но количественно заметно хуже. Для схемы «с полусуммой» функция устойчивости переходит в отрицательную область, т. е. является качественно неправильной.

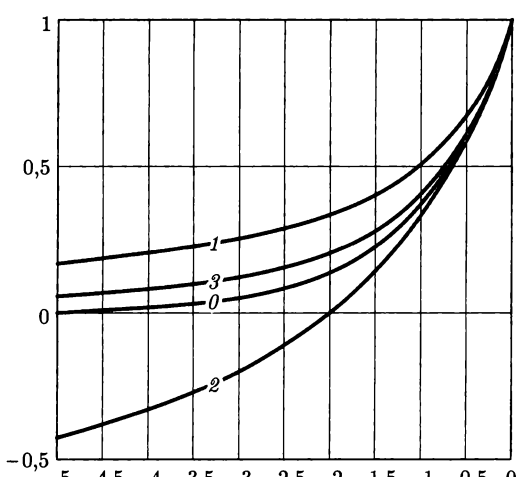


Рис. 1.5. Функция устойчивости при вещественном $z < 0$:

0 — точное решение; *1* — обратная схема Эйлера; *2* — схема «с полусуммой»; *3* — схема CROS

Вывод. Среди всех одностадийных безытерационных схем наиболее точной и одновременно наиболее надежной является схема CROS.

Схема CROS требует решения системы уравнений с комплексной арифметикой. Поэтому она более трудоемка, чем вещественные схемы. Но превышение трудоемкости невелико, так как значительный объем уходит на вычисление вещественной матрицы f_u, а он одинаков у всех схем.

Вычисление матрицы. Во всех схемах Розенброка требуется вычислять матрицу Якоби f_u. Если требуется исключительно высокая надежность расчета, то следует вычислять эту матрицу в явном виде (вручную или используя программы символьных вычислений). Однако в большинстве практических приложений можно ограничиться численным дифференцированием (см. кн. 1). Напомним соответствующую формулу.

Для системы M уравнений правая часть есть векторная функция от векторного аргумента с компонентами $f_m(u_1, \ldots, u_k, \ldots, u_M)$. Выберем некоторый шаг численного дифференцирования h и возьмем симметричную формулу

$$f_u = \left[\frac{\partial f_m}{\partial u_k}\right], \quad \frac{\partial f_m}{\partial u_k} \approx$$

$$\approx \frac{f_m(u_1, \ldots, u_k(1+h), \ldots, u_M) - f_m(u_1, \ldots, u_k(1-h), \ldots, u_M)}{2u_k h}. \tag{1.86}$$

Фактически разностный шаг равен не h, а $u_m h$; так лучше учитываются масштабы величин при записи чисел с плавающей точкой. Аппроксимация (1.86) выполняется для первой производной и имеет точность $O\left(h^2\right)$. Оптимален такой разностный шаг h, при котором ошибка аппроксимации формулы примерно равна ошибке округления при вычислении разности в числителе. Для 64-разрядных чисел оптимум составляет $h \approx 10^{-5}$. Для перестраховки можно увеличить h, но уменьшать недопустимо.

1.2.5. Многостадийные схемы Розенброка

Для лучшей точности расчета строят схемы более высокого порядка аппроксимации. С этой целью используют многостадийные схемы. Розенброк написал общее выражение семейства явно-неявных многостадийных схем, являющееся обобщением одно-

стадийной схемы (1.69). Оно содержит много свободных параметров, аналогично схемам РК. Схема не итерационная, но на каждой стадии вычисляется матрица Якоби и решается система линейных уравнений.

Однако схемы Розенброка с вещественными коэффициентами оказались недостаточно эффективными. Например, для вещественных двухстадийных схем было проведено полное исследование семейства возможных решений и доказано следующее:

- A-устойчивой схемы точности $O\left(\tau^4\right)$ не существует;
- существует A-устойчивая схема точности $O\left(\tau^3\right)$, но она не является L-устойчивой или монотонной, так что она недостаточно надежна для жестких задач;
- есть $L1$-устойчивые схемы точности $O\left(\tau^2\right)$, но $L2$-устойчивости у них нет.

Тем самым, эти двухстадийные схемы уступают по точности и качественным свойствам схеме CROS, поэтому использовать их в расчетах нецелесообразно.

Для s-стадийных вещественных схем доказано, что они не могут иметь одновременно точность $O\left(\tau^s\right)$ и $L2$-устойчивость.

Перспективными остаются лишь многостадийные схемы с комплексными коэффициентами. Общий вид двухстадийных схем для автономных (или автономизированных) систем таков:

$$\begin{gathered}\hat{u} = u + \tau \mathtt{Re}\left(b_1 w_1 + b_2 w_2\right), \\ \left(E - a_{11}\tau F_u(u)\right) w_1 = F(u), \\ \left(E - a_{22}\tau F_u\left(u + \tau \mathtt{Re}\left(a_{21} w_1\right)\right)\right) w_2 = F\left(u + \tau \mathtt{Re}\left(c_{21} w_1\right)\right).\end{gathered} \tag{1.87}$$

Исследовать такие схемы и находить их коэффициенты очень трудно. В настоящее время построено лишь несколько двухстадийных схем с комплексными коэффициентами. Из них приведем одну схему со следующим набором коэффициентов:

$$\begin{gathered}a_{11} = \frac{1}{10} + \frac{\sqrt{11}}{30}\mathtt{i};\ a_{22} = \frac{2}{10} + \frac{1}{10}\mathtt{i}; \\ b_1 = 0.1941430241155180 - 0.2246898944678803\mathtt{i}; \\ b_2 = 0.8058569758844820 - 0.8870089521907592\mathtt{i}; \\ c_{21} = 0.2554708972958462 - 0.2026195833570109\mathtt{i}; \\ a_{21} = 0.5617645150714754 - 1.148223341045841\mathtt{i}.\end{gathered} \tag{1.88}$$

Схема (1.87) — (1.88) имеет точность $O\left(\tau^4\right)$ и является $L2$-устойчивой. Число знаков коэффициентов рассчитано на 64-разрядный компьютер.

1.2.6. О других схемах

Схемы Ваннера являются модификацией многостадийных вещественных схем Розенброка. В них на всех стадиях сохраняется матрица f_u, вычисленная на первой стадии. Коэффициент перед ней в линейной системе также одинаков на всех стадиях, поэтому линейные системы на всех стадиях имеют одну и ту же матрицу. Это уменьшает трудоемкость; вычисление всех стадий при этом лишь незначительно превышает трудоемкость одной стадии. Поэтому такие схемы получили достаточно широкое распространение. Для них написано немало программных реализаций, в том числе с автоматическим выбором шага.

Однако среди известных схем Ваннера нет схем с устойчивостью лучше, чем $L1$, а большинство известных схем лишь A-устойчивы. Поэтому их надежность на задачах высокой жесткости может оказаться недостаточной. Косвенно об этом свидетельствуют «срывы» автоматического выбора шага: нередко на участках даже слабого изменения решения программа вдруг резко уменьшает шаг τ в $10^3 - 10^4$ раз и затем на протяжении многих шагов постепенно увеличивает шаг τ до первоначального значения, после чего может последовать новый срыв и так далее.

Кроме того, в п. 1.1.2 отмечалось, что автоматический выбор шага не гарантирует заданной пользователем точности. Если для мягких задач отличие фактической и заявленной точности в хороших программах не очень велико, то для жестких задач оно может превышать три-четыре порядка. Выигрыш же в трудоемкости при современном повышении быстродействия компьютеров становится мало значимым.

Методы Гира были первыми работоспособными методами для жестких систем. Они основаны на схемах дифференцирования назад, предложенных Гиршфельдером и Кертиссом. Они являются многошаговыми, т. е. для перехода от момента t_n к t_{n+1} надо использовать значения решения с нескольких предыдущих шагов (тем самым, непосредственный расчет с нулевого шага невозможен, и несколько первых шагов приходится вычис-

лять по явным схемам или другим способом). Запишем первые схемы этого семейства на равномерной сетке, т. е. с постоянным шагом τ:

$$u_{n+1} - u_n = \tau f(u_{n+1}); \tag{1.89}$$

$$\frac{3}{2}u_{n+1} - 2u_n + \frac{1}{2}u_{n-1} = \tau f(u_{n+1}); \tag{1.90}$$

$$\frac{11}{6}u_{n+1} - 3u_n + \frac{3}{2}u_{n-1} - \frac{1}{3}u_{n-2} = \tau f(u_{n+1}). \tag{1.91}$$

Схема (1.89) является одношаговой, имеет точность $O(\tau)$ и совпадает с обратной схемой Эйлера (1.66). Схема (1.90) является двухшаговой, а ее точность есть $O(\tau^2)$. Схема (1.91) трехшаговая точности $O(\tau^3)$ и т. д. Доказано, что первые шесть схем этого семейства являются A-устойчивыми, а последующие — нет.

Схемы выглядят очень просто, однако лишь при постоянном шаге τ; если шаг переменный, то левые части принимают достаточно сложный вид. Кроме того, эти схемы являются алгебраическими уравнениями, нелинейными относительно u_{n+1}. Поэтому они требуют итераций для нахождения u_{n+1}. При крупных шагах τ любой итерационный процесс может плохо сходиться или вообще не сходиться. Однако для этих методов были созданы хорошие вспомогательные алгоритмы, в том числе алгоритмы автоматического выбора шага, и написаны программные реализации.

Методы Гира являются L-устойчивыми. При этом одностадийная схема (1.89) имеет функцию устойчивости (1.67) и является $L1$-устойчивой. Исследуем двухстадийную схему (1.90) на тесте Далквиста. Для этого положим $u_{n+1} = R(z)\,u_n$, $u_n = R(z)\,u_{n-1}$, $f(u_{n+1}) = \lambda u_{n+1}$. Подстановка этих выражений в (1.90) дает квадратное уравнение относительно $R(z)$: $\left(\frac{3}{2} - z\right)R^2(z) - 2R(z) + \frac{1}{2} = 0$, так что функция устойчивости имеет следующий вид:

$$R(z) = 1/\left(2 \pm \sqrt{1+2z}\right). \tag{1.92}$$

Для этой схемы доказана A-устойчивость. Из (1.92) видно, что $R(z) = O(z - 1/2)$ при $z \to \infty$. Значит, схема $L_{1/2}$-устойчивая, т. е. показатель устойчивости оказался дробным.

Аналогично можно показать, что схемы Гиршфельдера — Кертисса p-го порядка точности являются $L_{1/p}$-устойчивыми.

Чем выше порядок точности, тем хуже затухание $R(z)$ при $z \to \infty$. Поэтому такие схемы, особенно высокого порядка, теряют надежность на задачах большой жесткости.

Химическая кинетика. Для частных, но практически важных задач нередко пишут нестандартные схемы, учитывающие специфику этих задач. Например, рассмотрим задачу химической кинетики. Неизвестными функциями $u_m(x)$ являются концентрации отдельных химических компонент — молекул, атомов и различных радикалов. Эти концентрации неотрицательны, причем только в начальный момент времени они могут быть нулевыми. Правые части являются суммами скоростей отдельных реакций. Скорости отдельных реакций имеют специфический вид: это произведения концентраций (или целых степеней концентраций) отдельных компонент. Если в данной реакции вещество сгорает, то его концентрация обязательно входит в это произведение. Поэтому формально систему уравнений можно записать так:

$$du_m/dt = \phi_m(\mathbf{u}, t) - u_m\psi_m(\mathbf{u}, t), \phi_m \geq 0, \psi_m \geq 0,\ 1 \leq m \leq M. \tag{1.93}$$

Здесь m — индекс отдельной компоненты, M — полное число компонент, $\mathbf{u} = \{u_1, \ldots, u_M\}$ обозначает совокупность всех компонент. Знаки в правых частях следуют из описанных свойств химических реакций: ϕ_m отвечает за синтез данного вещества, ψ_m — за его уничтожение.

Нетрудно построить схему 1-го порядка точности. Для этого производную заменяем разностью, в правой части аргумент функций $\mathbf{u}$ берем с исходного шага, а отдельный множитель u_m — с нового шага (как $\hat{u}_m$). После преобразований получаем явную одностадийную схему

$$\hat{u}_m = \left[u_m + \tau\phi_m\left(\mathbf{u}\right)\right] / \left[1 + \tau\psi_m\left(\mathbf{u}\right)\right], 1 \leq m \leq M. \tag{1.94}$$

Видно, что численное решение $\hat{u}_m > 0$, причем реакции синтеза данного вещества ведут к увеличению его концентрации, а реакции его уничтожения — к уменьшению. Это обеспечивает качественно правильное поведение решения.

Можно построить схему второго порядка точности. Для этого все концентрации в правой части (1.93) заменим полусуммами с исходного и нового шагов. Приведем выражение к следующему виду:

$$\frac{\hat{u}_m - u_m}{\tau} = \phi_m\left(\frac{\hat{\mathbf{u}} + \mathbf{u}}{2}\right) - \frac{1}{2}\hat{u}_m\left(1 + \frac{u_m}{\hat{u}_m}\right)\psi_m\left(\frac{\hat{\mathbf{u}} + \mathbf{u}}{2}\right).$$

Это выражение можно преобразовать так:

$$\hat{u}_m = \left[u_m + \tau\phi_m\left(\frac{\hat{\mathbf{u}} + \mathbf{u}}{2}\right)\right] / \left[1 + \frac{\tau}{2}\left(1 + \frac{u_m}{\hat{u}_m}\right)\psi_m\left(\frac{\hat{\mathbf{u}} + \mathbf{u}}{2}\right)\right]. \tag{1.95}$$

Выражение (1.95) есть нелинейное уравнение относительно концентраций на новом шаге. Если в правой части отбросить «крышку», т. е. взять все концентрации на исходном слое, то (1.95) перейдет в схему (1.94). Рассмотрим схему (1.94) как предиктор, т. е. вычислим по ней предварительное решение на новом шаге. Подставим это решение в правую часть (1.95), т. е. произведем коррекцию. Тем самым, схема становится двухстадийной. Разложением в ряд Тейлора можно показать, что скорректированное решение имеет второй порядок точности. При этом качественное поведение численного решения будет таким же, как у предыдущей схемы.

Скорректированная схема (1.95) обеспечивает и хорошую точность, и хорошее качественное поведение численного решения. Она успешно применяется в задачах химической кинетики, включая плохо обусловленные задачи теплового взрыва.

Система (1.93) может быть неавтономной (например, если температура среды зависит от времени). В этом случае в схему-предиктор (1.94) можно поставить значение t с исходного шага или с нового. Но в схему-корректор (1.95) надо подставлять $t + \tau/2$, чтобы обеспечить второй порядок точности.

Обратные схемы РК. Полностью неявные схемы Рунге — Кутты (1.19) с $L = s$ практически не используются, так как в них требуется решать систему нелинейных алгебраических уравнений порядка sM, где s — число стадий, M — порядок системы. Однако среди них можно выделить один интересный класс схем, которые будем называть обратными.

Для одностадийных схем $s = 1$ известны явная схема Эйлера (1.18) и неявная (обратная) схема Эйлера (1.66). Напомним их вид:

$$\hat{u} - u = \tau f(u), \quad \hat{u} - u = \tau f(\hat{u}).$$

Аналогичные построения возможны для многостадийных схем с $s \leq 4$ стадиями. Ограничимся явными оптимальными схемами

Таблица 1.3

Коэффициенты формулы (1.96)

k	s							
	1		2		3		4	
	b_k	a_k	b_k	a_k	b_k	a_k	b_k	a_k
1	1	0	1/4	0	2/9	0	1/6	0
2	—	—	3/4	2/3	3/9	1/2	2/6	1/2
3	—	—	—	—	4/9	3/4	2/6	1/2
4	—	—	—	—	—	—	1/6	1

РК. Их матрица Бутчера содержит только одну нижнюю кодиагональ коэффициентов. Поэтому такие схемы можно записать в следующем виде:

$$\begin{gathered}\hat{\mathbf{u}} = \mathbf{u} + \tau \sum_{k=1}^{s} b_k \mathbf{w}_k, \ s \le 4; \\ \mathbf{w}_k = f\left(\mathbf{u} + \tau a_k \mathbf{w}_{k-1}\right), \ 1 \le k \le s.\end{gathered} \tag{1.96}$$

Коэффициенты этих формул приведены в табл. 1.3.

Обратные схемы построим следующим образом. Рассмотрим движение в обратном направлении. Напишем явную схему для перехода $\hat{t} \to t$. Для этого в (1.96) нужно поменять местами $t \leftrightarrow \hat{t}$, $\mathbf{u} \leftrightarrow \hat{\mathbf{u}}$ и изменить знак τ. Получим следующие оптимальные обратные схемы:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau \sum_{k=1}^{s} b_k \hat{\mathbf{w}}_k, \ \hat{\mathbf{w}}_k = f\left(\mathbf{u} - \tau a_k \hat{\mathbf{w}}_{k-1}\right). \tag{1.97}$$

Поскольку явные формулы (1.96) с $s \le 4$ имеют точность $O(\tau^s)$, обратные схемы (1.97) также имеют точность $O(\tau^s)$.

Обратные схемы можно записать не только для оптимальных явных схем, но и для произвольных. Однако практический интерес из них представляет лишь схема точности $O(\tau^2)$, обратная к схеме «с полусуммой». Она записывается через рекурсивную функцию:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau f\left(\hat{\mathbf{u}} - \frac{1}{2}\tau f\left(\hat{\mathbf{u}}\right)\right). \tag{1.98}$$

Благодаря особенно простой форме, схема (1.98) оказалась полезной при решении квазилинейных уравнений в частных производных.

Нетрудно проверить, что функции устойчивости явных схем (1.96) являются полиномиальными аппроксимациями $\exp(z)$ степени s:

$$R_s(z) = \sum_{k=0}^{s} z^k/k!.$$

Для обратных схем (1.97) — (1.98) функции устойчивости будут Паде-аппроксимациями $\exp(z)$:

$$R_s(z) = \left[\sum_{k=0}^{s} (-z)^k/k!\right]^{-1}.$$

Поэтому обратные схемы (1.97) — (1.98) являются Ls-устойчивыми. В частности, у схемы с $s = 2$ функция устойчивости оказалась такой же, как для комплексной одностадийной схемы Розенброка (1.75).

Уравнения (1.97) — (1.98) являются нелинейными алгебраическими относительно $\hat{u}$; для системы дифференциальных уравнений они будут системой нелинейных алгебраических уравнений. Ее можно решать методом Ньютона или простых итераций (см. кн. 1). При этом s итераций каждого метода достаточно, чтобы обеспечить s-й порядок точности. Однако для сохранения Ls-устойчивости необходимо доводить итерации до сходимости. Заметим, что трудоемкость каждой итерации такова же, как для одностадийной обратной схемы Эйлера. Для жестких систем простые итерации обычно плохо сходятся и могут вообще не сходиться. Поэтому целесообразно использовать ньютоновские итерации.

При использовании ньютоновских итераций требуется матрица производных. Находить ее явный вид для рекурсивных функций достаточно сложно, даже если использовать программу символьных вычислений. Целесообразнее заменять производные разностными отношениями.

Заметим, что для $s > 4$ обратные схемы строить нецелесообразно: из-за наличия порогов Бутчера они не могут иметь точность $O(\tau^s)$ и Ls-устойчивость. Но для практики вполне достаточно приведенных схем.

1.2.7. Точность расчетов

Длина дуги. Прикладные задачи редко бывают чисто жесткими. Обычно в них присутствуют как быстро затухающие компоненты (радиоактивный распад), так и быстро нарастающие (задачи о тепловом взрыве). Графики отдельных компонент $u_m(t)$ содержат участки резкого изменения — почти вертикальные скачки. Для расчета таких участков с хорошей точностью необходим очень малый шаг. В промежутках же между скачками можно брать довольно крупный шаг.

Положение улучшается, если провести автономизацию системы, вводя в качестве параметра длину дуги в $(M+1)$-мерном пространстве $t, u_1, u_2, \ldots, u_M$. Она проводится по (1.16). Как указано в п. 1.1.1, в новых переменных вместо почти скачков $u_m(t)$ появляются почти изломы кривых $u_m(l)$, и скачки заменяются участками новых кривых с наклоном около 45 град. Это позволяет брать гораздо более крупный шаг на таких участках. Поэтому для жестких или вообще трудных задач всегда рекомендуется переходить от времени к длине дуги.

Сгущение сеток. Если произведена автономизация с помощью длины дуги, то расчет даже жестких задач удовлетворительно можно производить при постоянном шаге Δl. Разумеется, локальная погрешность будет возрастать в районе изломов, но не чрезмерно сильно. Поэтому можно применять процедуру глобального сгущения сетки, описанную в п. 1.1.10, и оценивать точность расчета методом Ричардсона. Кривые зависимости погрешности от шага здесь не столь четко выходят на асимптотику, как для мягких задач. Поэтому здесь целесообразна визуализация расчетов; автоматический же контроль точности работает гораздо менее надежно. Ричардсоновское рекуррентное уточнение по нескольким сеткам здесь также хуже работает. Особенно плохо при этом воспроизводятся участки резкого (почти разрывного) изменения функций. Положения скачков могут смещаться на несколько шагов сетки относительно точных положений. На разных сетках расчетные положения скачков при этом не совпадают.

Автоматический выбор шага в жестких задачах зачастую работает неудовлетворительно. На задачах большой жесткости фактическая точность таких расчетов может на три-четыре порядка отличаться от точности, заданной вычислителем. При этом у пользователя таких программ отсутствует возмож-

ность проконтролировать фактическую точность расчета. Эти программы нужно применять осторожно.

1.3. ДИФФЕРЕНЦИАЛЬНО-АЛГЕБРАИЧЕСКИЕ СИСТЕМЫ

1.3.1. Постановки задачи

Есть много прикладных задач, которые описываются одновременно дифференциальными и алгебраическими уравнениями. Один пример — кинематика механизмов. Отдельная деталь механизма есть материальное тело, и ее движение описывается дифференциальными уравнениями ньютоновской механики. Детали соединены друг с другом, так что координаты определенных точек у них совпадают. Такие связи являются алгебраическими уравнениями. Другой пример — разветвленные электрические цепи. Участок между двумя узлами описывается дифференциальными уравнениями для тока и напряжения. Алгебраическая сумма токов в каждом узле равна нулю, а падение напряжения на всех участках, соединяющих одну пару узлов, одинаково.

Напишем простейший вид дифференциально-алгебраической системы:

$$\begin{gathered} du_m/dt = f_m\left(\mathbf{u},\mathbf{v},t\right), 1 \le m \le M, \\ \mathbf{u} = \{u_1, u_2, \ldots, u_M\}, \mathbf{u}(0) = \mathbf{u}^0; \\ 0 = g_k\left(\mathbf{u},\mathbf{v},t\right), 1 \le k \le K, \\ \mathbf{v} = \{v_1, v_2, \ldots, v_K\}, g_k(\mathbf{u}^0, \mathbf{v}^0, 0) = 0. \end{gathered} \tag{1.99}$$

Первая подсистема (1.99) является дифференциальной, вторая — алгебраической. Без нарушения общности считаем, что начальный момент времени $t = 0$. Начальные данные для дифференциальной компоненты $\mathbf{u}(t)$ могут быть заданы произвольно, как для задачи Коши. Начальные данные для алгебраической компоненты $\mathbf{v}(t)$ задают так, чтобы в начальный момент алгебраическая подсистема (1.99) удовлетворялась; тем самым, они согласуются с начальными данными задачи Коши.

Для согласованного задания $\mathbf{v}^0$ надо решить алгебраическую подсистему (1.99). Это возможно, если якобиан $\det\left(\partial\mathbf{g}/\partial\mathbf{v}\right) \neq 0$. Таким дифференциально-алгебраическим системам приписывают ***индекс 1***. Здесь будем рассматривать только такие системы:

они наиболее распространены в приложениях и достаточно просто решаются. Системам с вырожденной матрицей Якоби приписывают более высокие индексы.

Для систем индекса 1 в принципе можно решить алгебраическую подсистему и выразить алгебраические компоненты через дифференциальные, т. е. найти $\mathbf{v}(\mathbf{u}, t)$. Подставляя эту зависимость в дифференциальную подсистему (1.99), получим обычную задачу Коши для дифференциальной системы порядка M:

$$d\mathbf{u}/dt = \mathbf{F}(\mathbf{u}, t) \equiv \mathbf{f}(\mathbf{u}, \mathbf{v}(\mathbf{u}, t), t), \mathbf{u}(0) = \mathbf{u}^0.$$

Напомним, что решение последней задачи существует, единственно и $p+1$ раз непрерывно дифференцируемо, если $\mathbf{F}(\mathbf{u}, t)$ имеет непрерывные p-е производные по всем аргументам. Для этого достаточно, чтобы исходная система (1.99) имела индекс 1, а ее правые части $\mathbf{f}, \mathbf{g}$ имели p-е непрерывные производные по всем аргументам.

Жесткость. Поставим в левой части алгебраической подсистемы (1.99) не нули, а производные с малым коэффициентом ϵ. Модифицированная подсистема примет следующий вид:

$$\epsilon \frac{d\mathbf{v}}{dt} = \mathbf{g}(\mathbf{u}, \mathbf{v}, t), |\epsilon| \ll 1. \tag{1.100}$$

Система (1.99) с видоизменением (1.100) становится чисто дифференциальной. Приведем подсистему (1.100) к каноническому виду, поделив обе части на ϵ. Ввиду малости ϵ новые правые части станут очень большими по модулю. Следовательно, подсистема (1.100) является жесткой или плохо обусловленной. При $\epsilon \to 0$ жесткость или плохая обусловленность неограниченно усиливаются, а дифференциальная подсистема (1.100) переходит в алгебраическую.

Поэтому можно рассматривать алгебраические системы как предельный случай дифференциальных, а дифференциально-алгебраические системы трактовать как системы огромной жесткости.

Общий вид. Систему (1.99) можно формально переписать в матричной форме:

$$\begin{pmatrix} E_M & O \\ O & O_K \end{pmatrix} \begin{pmatrix} d\mathbf{u}/dt \\ d\mathbf{v}/dt \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{1.101}$$

Здесь E_M — квадратная единичная матрица порядка M, O_K — квадратная нулевая матрица порядка K, остальные две матрицы O — нулевые прямоугольные. Полная матрица в левой части является вырожденной; ее ранг (который равен порядку наибольшего ненулевого минора матрицы) есть $r = \mathtt{rank}\, G = M$.

В системе (1.99) уравнения для дифференциальных и алгебраических компонент четко разделялись. Однако в практических задачах они нередко «перепутываются», и трудно сказать, какая компонента является дифференциальной, а какая — алгебраической. Поэтому общий вид дифференциально-алгебраической системы таков:

$$G\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{U}, t). \tag{1.102}$$

Здесь G — вырожденная матрица порядка N. Она имеет ранг $r < N$. Количество дифференциальных компонент решения равно r, а количество алгебраических есть $N - r$.

Общая форма записи приводит к некоторым трудностям при практических расчетах. Нахождение ранга матрицы — довольно трудоемкая процедура, так что ранг чаще всего не вычисляют. Но тогда неизвестно, сколько следует задавать начальных условий, а сколько выбирать согласованно с этими начальными условиями. Обычно применяют следующий прием. Берут столько начальных условий, сколько поставлено в прикладной задаче (очевидно, их число должно быть не более порядка всей системы). Недостающие условия ставят более или менее правдоподобно. Уже после небольшого числа шагов алгебраические компоненты, благодаря своей сверхжесткости, сильно затухают, и численное решение выходит на правильный предел. Первым шагам такого расчета нельзя доверять, но следующие шаги являются удовлетворительными.

1.3.2. Метод ϵ-вложений

Схемы Розенброка. Рассмотрим систему общего вида (1.102). Потребуем, чтобы матрица G не зависела от решения U; от времени t она может зависеть. Временно забудем, что матрица G — вырожденная. Тогда систему (1.99) можно переписать, разрешив относительно производной:

$$d\mathbf{U}/dt = G^{-1}\mathbf{F}(\mathbf{U}, t). \tag{1.103}$$

Будем рассматривать эту систему как жесткую и применим к ее решению одностадийную схему Розенброка. Коэффициент a этой схемы можно брать как вещественным, так и комплексным. Поэтому схема примет следующий вид:

$$\hat{\mathbf{U}} = \mathbf{U} + \tau \mathbf{Rew}, \left[E - a\tau \frac{\partial \left(G^{-1}\mathbf{F} \right)}{\partial \mathbf{U}} \right] \mathbf{w} = G^{-1}\mathbf{F}.$$

Поскольку G не зависит от решения, то G^{-1} выносится за знак производной. Умножим последнее уравнение слева на G и получим окончательную схему:

$$\hat{\mathbf{U}} = \mathbf{U} + \tau \mathbf{Rew}, \left(G - a\tau \mathbf{F}_U \right) \mathbf{w} = \mathbf{F}. \tag{1.104}$$

То, что матрица G особенная, не препятствует расчету: в круглых скобках стоит неособенная матрица, и линейная система имеет решение. Очевидно, входящий в $\mathbf{F}_U$ и $\mathbf{F}$ аргумент U надо брать на исходном шаге.

Но для неавтономных задач есть еще аргумент t; его выбор не тривиален и обсуждается далее.

Поскольку задача (1.99) характеризуется огромной жесткостью, схема (1.104) должна быть L-устойчивой. Поэтому возможны лишь два значения параметра: $a = 1$ или $a = (1 + \mathrm{i})/2$.

Этот метод получил название метода ϵ-вложений из-за аналогии с переходом к чисто дифференциальным системам (1.100) с помощью параметра $\epsilon \to 0$.

Обратная схема Эйлера. Она получается при $a = 1$ и является $L1$-устойчивой. Для пояснения ее смысла напишем схему (1.104) для простейшей системы (1.99). Проводя разбиение матриц и векторов на клетки аналогично (1.101) и учитывая, что $\mathbf{w} = \left(\hat{\mathbf{U}} - \mathbf{U} \right)/\tau$, приведем систему к следующему виду:

$$\begin{pmatrix} E - \tau \mathbf{f_u} & -\tau \mathbf{f_v} \\ -\tau \mathbf{g_u} & -\tau \mathbf{g_v} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}} - \mathbf{u} \\ \hat{\mathbf{v}} - \mathbf{v} \end{pmatrix} = \begin{pmatrix} \tau \mathbf{f} \\ \tau \mathbf{g} \end{pmatrix}. \tag{1.105}$$

Проводя поклеточное умножение, запишем эту систему в виде двух подсистем:

$$\begin{gathered} (E - \tau f_{\mathbf{u}}) (\hat{\mathbf{u}} - \mathbf{u}) - \tau \mathbf{f_v} \cdot (\hat{\mathbf{v}} - \mathbf{v}) = \tau \mathbf{f}, \\ \mathbf{g_u} \cdot (\hat{\mathbf{u}} - \mathbf{u}) + \mathbf{g_v} \cdot (\hat{\mathbf{v}} - \mathbf{v}) = -\mathbf{g}. \end{gathered} \tag{1.106}$$

Нетрудно заметить, что верхняя строка (1.106) есть обратная схема Эйлера для дифференциальной подсистемы (1.99). Вто-

рая строка (1.106) не содержит шага τ и является первой ньютоновской итерацией для алгебраической подсистемы (1.99).

Алгебраические уравнения должны удовлетворяться на новом шаге $\hat{t}$ насколько возможно точнее. Для этого следовало бы в правой части (1.106) взять $\mathbf{g}(\mathbf{u}, \mathbf{v}, \hat{t})$. Но схема Эйлера имеет точность лишь $O(\tau)$. Поэтому для неавтономных задач безразлично, какой момент времени является аргументом в правых частях (1.106): t или $\hat{t}$.

Обратная схема Эйлера (1.106) удобна своей $L1$-устойчивостью, монотонностью, надежностью и возможностью применения к неавтономным задачам. Однако ее точность $O(\tau)$ невелика.

Комплексная схема Розенброка. Для схемы надо пользоваться записью (1.104), полагая $a = (1 + \mathrm{i})/2$. Записать ее в клеточном виде можно, но привести к просто трактуемой форме типа (1.106) не удается. Однако наглядный смысл этой схемы остается таким же. Для дифференциальных уравнений системы шаг выполняется по комплексной схеме Розенброка, а для алгебраической подсистемы делается первая ньютоновская итерация.

Для чисто дифференциальных систем схема CROS имеет точность $O(\tau^2)$. Однако для неавтономных дифференциально-алгебраических систем при этом возникают противоречивые требования. Чтобы дифференциальные уравнения обеспечивали точность $O(\tau^2)$, в правой части (1.104) надо брать $\mathbf{F}(\mathbf{U}, t + \tau/2)$. Но чтобы ньютоновская итерация для алгебраической подсистемы давала такую же точность, надо брать $\mathbf{F}(\mathbf{U}, t + \tau)$. Поэтому для неавтономных дифференциально-алгебраических систем схема CROS дает точность лишь $O(\tau)$.

Чтобы обеспечить точность $O(\tau^2)$, надо предварительно автономизировать исходную систему (см. п. 1.1.10). Напомним, что при этом время t вводится как дополнительная функция; тем самым, в правых частях явная зависимость от t исчезает. За аргумент при этом рекомендуется брать не t, а длину дуги; это не повышает порядка точности, но улучшает фактическую точность.

В этом случае схема CROS будет $L2$-устойчивой, монотонной и очень надежной в расчетах, а ее точность будет $O(\tau^2)$. Она точнее обратной схемы Эйлера и не уступает ей по надежности. Ее можно рекомендовать для большинства практических расчетов.

Схемы высоких порядков. Для получения больших порядков точности целесообразно использовать многостадийные схемы Розенброка с комплексными коэффициентами. Как и для схемы CROS, необходимо предварительно автономизировать исходную систему. Следует использовать двухстадийные схемы общего вида, где вместо матрицы E надо подставлять матрицу G.

Однако большинство наборов коэффициентов, дающих точность $O\left(\tau^4\right)$ для чисто дифференциальных систем, на дифференциально-алгебраических системах имеют точность лишь $O\left(\tau^2\right)$. Набор (1.88) дает для дифференциально-алгебраических систем точность $O\left(\tau^3\right)$.

Обратные схемы РК (см. п. 1.2.6) позволяют получить точность до $O(\tau^4)$. Они также требуют предварительной автономизации системы. Эти схемы итерационны, но обеспечивают хорошее качество решения лишь в том случае, когда итерации сходятся с высокой точностью. Поэтому один шаг таких схем более трудоемок, чем для комплексных схем.

1.4. КРАЕВЫЕ ЗАДАЧИ

1.4.1. Постановки задач

Для системы дифференциальных уравнений порядка M полная постановка задачи содержит M дополнительных условий (см. п. 1.1.1). Если $M = 1$, то дополнительное условие единственное, а точку его задания можно считать начальной. Если же $M \geq 2$, то дополнительные условия могут быть заданы в разных точках. Если таких точек две, их считают границами отрезка, а сами задачи называют краевыми.

В прикладных задачах Коши аргумент часто интерпретировался как время и обозначался через t. В краевых задачах аргумент обычно интерпретируется как пространственная переменная и обозначается через x. При этом нередко более естественной оказывается запись дифференциальных уравнений не как системы M уравнений первого порядка, а как одного уравнения порядка M. При этом для производных наряду с обозначениями $du/dx, d^2u/dx^2$ будем использовать обозначения u_x, u_{xx} и т. д. Приведем некоторые примеры краевых задач.

Струна. Горизонтально натянутая струна прогибается под действием внешних нагрузок и собственного веса. Этот прогиб

описывается линейным дифференциальным уравнением второго порядка. Запишем общий вид такого уравнения:

$$u_{xx} + q(x)\, u_x - r(x)\, u = f(x), a < x < b. \tag{1.107}$$

Пусть каждый конец струны закреплен на определенной высоте. Тогда граничные условия имеют следующий вид:

$$u(a) = \alpha, u(b) = \beta. \tag{1.108}$$

Такие граничные условия называют условиями первого рода; возможны и другие граничные условия.

Из теории известно, что если $r(x), q(x), f(x)$ имеют p непрерывных производных, то $u(x)$ имеет $p+2$ непрерывные производные. Знание гладкости решения нам потребуется для выбора численного метода и применения метода Ричардсона для контроля точности путем сгущения сеток.

Обобщение. Напишем общий вид нелинейного уравнения второго порядка, разрешенного относительно старшей производной:

$$u_{xx} = f(x, u, u_x), a < x < b. \tag{1.109}$$

Краевые условия могут быть взяты согласно (1.108).

Если функция f имеет p-е непрерывные производные по всем аргументам, то $u(x)$ имеет $p+2$ непрерывные производные.

Брусок. Прогиб упругого бруска описывается уравнением четвертого порядка. Задача требует четырех краевых условий. Запишем простейший пример такой задачи:

$$\begin{gathered} u_{xxxx} = f(x), a < x < b; \\ u(a) = 0, u_x(a) = 0, u(b) = 0, u_x(b) = 0. \end{gathered} \tag{1.110}$$

Здесь брусок прогибается только под действием внешней силы, его концы расположены на одинаковой высоте (принятой за нулевую) и вмурованы в стенки, так что первые производные на границах сохраняют горизонтальное направление. Если функция f имеет p непрерывных производных по всем аргументам, то $u(x)$ имеет $p+4$ непрерывные производные.

Нелокальные краевые условия связывают между собой значения функции в разных точках. Задачу с нелокальным краевым условием можно поставить даже для дифференциального

уравнения первого порядка! Приведем простейший пример (он взят абстрактно, а не из прикладной задачи):

$$\frac{du}{dx} = u,\ 0 \le x \le a;\ u(a) - u(0) = c. \tag{1.111}$$

Нетрудно записать точное решение этой задачи:

$$u(x) = c\mathrm{e}^x / \left(\mathrm{e}^a - 1\right);$$

оно может служить тестом для численного расчета.

В квантовой механике возникает нелокальное условие нормировки, имеющее интегральный вид $\int u^2(x)dx = 1$.

1.4.2. Сеточный метод

В настоящее время наиболее распространен сеточный метод решения краевых задач. Этот метод заключается в следующем. 1) Выбираем некоторую сетку $a = x_0 < x_1 < x_2 < \cdots < x_N = b$. 2) Заменяем все производные в дифференциальном уравнении некоторыми разностными соотношениями, используя значения решения в узлах сетки: $u_n = u\left(x_n\right)$. При этом дифференциальное уравнение превращается в разностное, т. е. в алгебраическое уравнение относительно узловых значений u_n. Это уравнение называют ***разностной схемой***, а совокупность всех входящих в это уравнение узлов — ***шаблоном*** разностной схемы. 3) Доказываем существование решения полученной системы алгебраических уравнений и строим алгоритм вычисления этого решения. 4) Доказываем, что полученное сеточное решение стремится к точному при сгущении сетки; по возможности находим теоретическую мажорантную или асимптотически точную оценку погрешности. 5) На основе метода сгущения сеток строим процедуру нахождения сеточного решения с заданной точностью.

Рассмотрим реализацию этой программы на различных примерах.

Струна. Простейшим случаем является задача (1.107) — (1.108). Выберем на отрезке $[a, b]$ некоторую сетку. Для простоты возьмем равномерную сетку с N интервалами и $N + 1$ узлами:

$$x_n = a + nh, 0 \le n \le N, h = (b - a)/N; \tag{1.112}$$

величина h есть шаг сетки. Аппроксимируем уравнение (1.107) разностной схемой. Для аппроксимации второй производной

простейшая схема должна содержать три соседних узла: $n-1$, n, $n+1$. Заменяя производные симметричными разностными выражениями (см. кн. 1), получим следующую схему для внутренних узлов:

$$\frac{1}{h^2}\left(u_{n+1}-2u_n+u_{n-1}\right)+\frac{q_n}{2h}\left(u_{n+1}-u_{n-1}\right)-r_n u_n=f_n, \qquad 1\le n\le N-1. \tag{1.113}$$

Здесь $r_n=r\left(x_n\right), q_n=q\left(x_n\right), f_n=f\left(x_n\right)$, а величины u_n являются ***приближенным*** решением в узлах x_n.

Система (1.113) содержит $N-1$ уравнение с $N+1$ неизвестными, т. е. является недоопределенной. В граничных узлах $n=0$, $n=N$ разностные уравнения (1.113) писать нельзя, иначе индексы узлов выйдут за допустимые пределы. Для доопределения системы используем граничные условия (1.108). Их запись очевидна:

$$u_0=\alpha, \quad u_N=\beta. \tag{1.114}$$

Алгебраическая система (1.113), (1.114) является полностью определенной.

Эта алгебраическая система оказалась линейной благодаря линейности исходной задачи. Матрица этой системы трехдиагональна. Поэтому ее нетрудно решить методом Гаусса для ленточной матрицы или прогонкой (см. кн. 1). Это прямые методы. Они позволяют найти решение, затратив около девяти арифметических операций на каждый узел.

Остановимся на существовании разностного решения и его сходимости к точному. Существует широкий класс коэффициентов, для которых нетрудно провести необходимые доказательства. Справедлива следующая теорема.

Теорема 1.5. Пусть $q(x), r(x), f(x)$ дважды непрерывно дифференцируемы на $[a,b]$, а $r(x)\ge m$, где константа $m>0$. Пусть также шаг h достаточно мал, так что $h\max|q(x)|\le 2$. Тогда разностное решение существует и отличается от точного в норме c на величину $O(h^2)$. •

Доказательство. В пределах данного доказательства будем обозначать точное решение через $v(x)$, чтобы отличать его от сеточного. Из двукратной дифференцируемости коэффициентов следует, что $v(x)$ четырежды непрерывно дифференцируема. Разложим $v(x)$ в ряд Тейлора — Маклорена с центром в точке x_n:

$$v_{n\pm1} = v \pm hv_x + \frac{1}{2}h^2v_{xx} \pm \frac{1}{6}h^3v_{xxx} + \frac{1}{24}h^4v_{xxxx} + o(h^4); \quad (1.115)$$

здесь функции без аргументов считаются взятыми к точке x_n. Отсюда получаем

$$\frac{1}{h^2}\left(v_{n+1} - 2v_n + v_{n-1}\right) = v_{xx} + \frac{1}{12}h^2v_{xxxx} + o\left(h^2\right),$$

$$\frac{1}{2h}\left(v_{n+1} - v_{n-1}\right) = v_x + \frac{1}{6}h^2v_{xxx} + o\left(h^2\right)$$

Выражая отсюда v_{xx} и v_x через разности и подставляя их в дифференциальное уравнение (1.107), получим

$$\begin{aligned}&\frac{1}{h^2}\left(v_{n+1} - 2v_n + v_{n-1}\right) + \frac{q_n}{2h}\left(v_{n-1} - v_{n-1}\right) - r_nv_n = \\ &= f_n - \frac{1}{12}h^2v_{xxxx} + \frac{1}{6}h^2q_nv_{xxx} + o(h^2), 1 \le n \le N-1;\end{aligned} \quad (1.116)$$

это разностное уравнение, которому удовлетворяет точное решение.

Введем погрешность сеточного решения $z_n = u_n - v_n, 0 \le n \le N$. Для краевых условий первого рода (1.108) на границах сеточное решение (1.114) совпадает с точным. Значит, $z_0 = 0$, $z_N = 0$. Вычитая (1.116) из (1.113), умножая на h^2 и перегруппировывая члены, получим разностные уравнения для погрешности:

$$\begin{aligned}&\left(1 + \frac{1}{2}hq_n\right)z_{n+1} - \left(2 + h^2r_n\right)z_n + \left(1 - \frac{1}{2}hq_n\right)z_{n-1} = \\ &= \frac{1}{12}h^4v_{xxxx} - \frac{1}{6}h^4q_nv_{xxx} + o(h^4), 1 \le n \le N-1.\end{aligned} \quad (1.117)$$

По условию малости шага h все выражения в скобках не отрицательны.

Оценим из (1.117) норму погрешности $\|z\|_c = \max|z_n|$. Поскольку на границах $z_n = 0$, этот максимум достигается в одном из внутренних узлов k. Напишем (1.117) для k-го узла, перенесем слагаемые с $z_{k\pm1}$ в правую часть и применим неравенство многоугольника; это дает

$$\begin{aligned}\left(2 + h^2r_k\right)|z_k| \le &\left(1 + \frac{1}{2}hq_k\right)|z_{k+1}| + \left(1 - \frac{1}{2}hq_k\right)|z_{k-1}| + \\ &+ \frac{1}{12}h^4|v_{xxxx}| + \frac{1}{6}h^4|q_kv_{xxx}| + o(h^4).\end{aligned} \quad (1.118)$$

Здесь в левой части стоит $|z_k| = \|z\|_c$. Подставляя в правую часть норму $\|z\|_c$ вместо меньших величин $|z_{k\pm1}|$, а вместо v_{xxxx} и v_{xxx} их нормы C, мы только усилим неравенство. Сокращая при этом совпадающие члены в левой и правой частях, получим

$$h^2 r_k \|z\|_c \le \frac{1}{12} h^4 \|v_{xxxx}\|_C + \frac{1}{6} h^4 \|q v_{xxx}\|_C + o(h^4).$$

Сокращая на h^2 и учитывая соотношение $r_k \ge m > 0$, получим окончательную оценку погрешности:

$$\|z\|_c \le \frac{h^2}{m} \left(\frac{1}{12} \|v_{xxxx}\|_C + \frac{1}{6} \|q v_{xxx}\|_C \right) = O(h^2). \qquad (1.119)$$

Доказана сходимость разностного решения с вторым порядком точности.

В разностной схеме (1.113) также можно сгруппировать члены и привести ее к такому же виду, как левая часть (1.117). При этом для нахождения значений u_n получается линейная система с трехдиагональной матрицей, элементами которой являются скобки в (1.117). Диагональный элемент по модулю больше суммы внедиагональных в силу условий теоремы. Известно, что при этом решение линейной системы существует и единственно (см. кн. 1). Вычисление решения методом Гаусса или прогонкой при этом устойчиво, т. е. происходит без накопления ошибок округления. Теорема полностью доказана. ■

Оценка (1.119) является мажорантной. При дополнительных предположениях можно построить и асимптотически точную оценку погрешности; это будет сделано в гл. 2. Тогда можно применять сгущение сетки и метод Ричардсона для нахождения апостериорной оценки погрешности и расчетов с контролем точности, как это сделано для задачи Коши в п. 1.1.10.

Условия теоремы являются достаточными, но не необходимыми. Даже если условия не соблюдены, то в большинстве случаев разностное решение существует и сходится к точному. Это можно проверить, сгущая сетку и применяя метод Ричардсона.

Разностная схема (1.113) симметрична относительно центрального узла шаблона. Отсюда следует, что при разложении решения в ряд Тейлора сокращаются все члены с нечетными степенями шага и погрешность разлагается в ряд по степеням h^2. В таких условиях метод Ричардсона гораздо эффективнее, чем при решении задачи Коши, где ошибка разлагается в ряд по всем

степеням h. При нерекуррентном применении метода отброшенный член погрешности есть не $O(h)$ по отношению к главному, а $O(h^2)$, т. е. его вклад намного меньше. Кривые зависимости погрешности от h при этом гораздо быстрее выходят на асимптотику. Если же метод Ричардсона применяется рекуррентно, то при каждом сгущении сетки порядок точности повышается не на 1, а на 2.

Бикомпактные схемы. Схема (1.113) была написана для равномерной сетки и гладких коэффициентов уравнения. Однако в приложениях важную роль занимают задачи для слоистых сред. Например, это может быть задача прогиба струны, сваренной из двух разных материалов, или задача о стационарном профиле тепла в многослойной оболочке. Для таких задач дифференциальное уравнение принимает несколько иной вид:

$$\frac{d}{dx}\left[q(x)\frac{du}{dx}\right] - r(x)u = f(x), u(a) = \alpha, u(b) = \beta. \qquad (1.120)$$

Для задач теплопроводности $f(x)$ имеет смысл внешних источников тепла, а $q(x)$ является коэффициентом теплопроводности. Пусть границей двух слоев является внутренняя точка $x_* \in (a, b)$. Будем предполагать, что внутри слоев коэффициенты $q(x)$, $r(x)$, $f(x)$ дважды непрерывно дифференцируемы, а на границе слоев x_* коэффициенты претерпевают разрыв.

Из курса математической физики известно, что задача (1.120) имеет множество решений, т. е. является недоопределенной. Чтобы выделить единственное решение из множества формально допустимых, необходимо поставить два внутренних краевых условия в точке x_*. Для задачи теплопроводности такими внутренними краевыми условиями являются требования непрерывности температуры $u(x)$ и теплового потока $w(x) = -q(x)u_x$ на границе слоев:

$$u(x_* - 0) = u(x_* + 0),\ (qu_x)_{x_*-0} = (qu_x)_{x_*+0}. \qquad (1.121)$$

Доказано, что решение задачи (1.120) — (1.121) единственно, а если все коэффициенты имеют p непрерывных производных внутри каждого слоя, то $u(x)$ имеет $p + 2$ непрерывных производных внутри каждого слоя.

Из второго условия (1.121) видно, что поскольку в точке x_* коэффициент $q(x)$ разрывен, то u_x также должно иметь разрыв для сохранения непрерывности произведения. Следовательно, u_{xx} и более высокие производные в точке x_* не существуют.

Поэтому в тех узлах, для которых точка x_* попадает внутрь шаблона (т. е. $x_{n-1} < x_* < x_{n+1}$), схему типа (1.113) записывать нельзя.

Как записать схему, пригодную для любых интервалов? Дадим следующее определение.

Определение 1.7. Схема называется *бикомпактной*, если она построена по следующим правилам:

1) если имеется особая точка решения x_*, в нее надо поставить узел сетки;

2) дифференциальное уравнение высокого порядка надо заменить системой уравнений первого порядка;

3) следует использовать двухточечные схемы, беря неизвестные функции только в узлах сетки x_n (брать неизвестные функции в полуцелых точках $x_{n+1/2}$ нельзя);

4) при написании разностной схемы надо дифференциальные уравнения проинтегрировать и аппроксимировать полученные квадратуры.

Бикомпактные схемы позволяют успешно решать задачи для слоистых сред с переменными и даже разрывными коэффициентами, причем на произвольных неравномерных сетках.

Построим бикомпактную схему для задачи (1.120) — (1.121). Для этого заменим уравнение второго порядка (1.120) системой двух уравнений первого порядка:

$$w = -q(x)\frac{du}{dx},\ \frac{dw}{dx} = -r(x)u - f(x).$$

Поделим обе части первого уравнения на $q(x)$ и проинтегрируем оба уравнения на отрезке $[x_{n-1}, x_n]$. Поскольку внутри интервала все функции многократно дифференцируемы, интегралы от производных возьмем точно и получим следующее соотношение:

$$\begin{gathered} \int\limits_{x_{n-1}}^{x_n} \frac{w(x)}{q(x)}dx = u_{n-1} - u_n, \\ w_n - w_{n-1} = -\int\limits_{x_{n-1}}^{x_n} [r(x)u(x) + f(x)]\,dx, \\ 1 \le n \le N. \end{gathered} \tag{1.122}$$

Эти соотношения справедливы во всех интервалах, в том числе примыкающих к границе слоев. В первом интеграле применим гибридную формулу: неизвестную функцию w возьмем в узлах сетки, что соответствует формуле трапеций, а известную функцию $q(x)$ — в полуцелой точке (в узлах ее брать не следует, поскольку один из узлов является точкой x_*, т. е. разрывом коэффициента). Во втором интеграле u берем в целых узлах, а коэффициенты — в полуцелых точках. Все квадратуры имеют погрешность $O(h^2)$. В результате получаем следующую разностную схему:

$$\frac{h_n}{2q_{n-1/2}}(w_{n-1}+w_n) = u_{n-1}-u_n,$$
$$w_{n-1}-w_n = \frac{h_n}{2}r_{n-1/2}(u_{n-1}+u_n) + h_n f_{n-1/2}, \tag{1.123}$$
$$1 \le n \le N;\ u_0 = \alpha,\ u_N = \beta.$$

Здесь дописаны условия на внешних границах, после чего количество неизвестных значений u_n, w_n совпадает с числом уравнений. Внутренние граничные условия (1.121) не входят явно в схему (1.123); они автоматически удовлетворяются, поскольку правые и левые пределы неизвестных функций $u(x)$, $w(x)$ во всех узлах x_n считаются одними и теми же значениями u_n, w_n.

Схема (1.123) является системой линейных уравнений относительно неизвестных u_n, w_n. Ее можно решить прямым методом Гаусса. Перед этим целесообразно упорядочить вектор неизвестных значений следующим образом: $u_0, w_0, u_1, w_1, \cdots, u_N, w_N$. Тогда матрица линейной системы будет ленточной с узкой лентой и число операций в методе Гаусса станет небольшим. Однако можно упростить алгоритм, если исключить w с помощью следующего приема.

В первом из уравнений (1.123) умножим обе части на $2q_{n-1/2}/h_n$; в левой части будет стоять $w_{n-1}+w_n$. В полученном уравнении увеличим индекс на единицу; в левой части будет стоять w_n+w_{n+1}. Вычитая одно такое уравнение из другого, получим в левой части $w_{n+1}-w_{n-1}$. Затем напишем второе из уравнений (1.123) для индекса на единицу больше и сложим с исходным; в левой части опять будет стоять $w_{n+1}-w_{n-1}$. Приравниваем правые части, тем самым исключая w_n, и получаем следующую трехточечную схему:

$$\frac{q_{n+1/2}}{h_{n+1}}(u_{n+1}-u_n)-\frac{q_{n-1/2}}{h_n}(u_n-u_{n-1})-$$

$$-\frac{1}{4}\left[r_{n+1/2}h_{n+1}(u_{n+1}+u_n)+r_{n-1/2}h_n(u_n+u_{n-1})\right]= \tag{1.124}$$

$$=\frac{1}{2}\left(h_n f_{n-1/2}+h_{n+1}f_{n+1/2}\right),$$

$$1\le n\le N-1; u_0=\alpha, u_N=\beta.$$

Схема (1.124) содержит только значения u_n и формально является трехточечной. Однако она является строгим следствием схемы (1.123). Следовательно, она бикомпактная и применима к задачам слоистых сред, причем на произвольной неравномерной сетке.

Если $q(x)>0$ и $r(x)>0$, то в трехдиагональной матрице схемы выполнено условие преобладания диагонального элемента. Поэтому разностное решение легко находится методом Гаусса для ленты или прогонкой. Это заметно проще, чем решать полную систему (1.123).

Повышенная точность. Если для вычисления интегралов использовать формулы Эйлера—Маклорена, то можно построить разностные схемы точности $O(h^4)$ и выше. При этом полученные выражения аналогичны системе (1.123) с числом неизвестных $2N+2$. Исключить из них w_n и привести к форме типа (1.124) не удается.

Неограниченная область. Рассмотрим задачу (1.120) в неограниченной области: $a=0, b=+\infty$. Построим в этой области квазиравномерную сетку (см. кн. 1) с подходящей производящей функцией, например:

$$x(\xi)=c\,\xi/\left(1-\xi^2\right)^s,\ 0\le\xi\le 1,\ c>0,\ s>0. \tag{1.125}$$

По вспомогательной переменной ξ возьмем равномерную сетку: $\xi_n=\dfrac{n}{N}$, $0\le n\le N$. Ей соответствует квазиравномерная сетка $x_n=x(\xi_n)$, причем $x_0=0$, $x_N=+\infty$; последний узел квазиравномерной сетки является бесконечно удаленной точкой. Середины интервалов надо понимать в смысле преобразования (1.125): $x_{n+1/2}=x(\xi_{n+1/2})\ne(x_{n+1}+x_n)/2$. Поэтому середина последнего неограниченного интервала $x_{N-1/2}$ оказывается конечной точкой. В неограниченной области шаг сетки необ-

ходимо переопределить с помощью следующего соотношения: $h_n = \frac{1}{N}\left(\frac{dx}{d\xi}\right)_{n-1/2} \neq x_n - x_{n-1}$; при этом шаг будет конечен всюду, включая последний неограниченный интервал.

Подставляя квазиравномерную сетку и переопределенный шаг в уравнение (1.123) или (1.124), получим разностные схемы в неограниченной области. В этих схемах краевое условие на бесконечности ставится естественным образом. При этом схемы сохраняют второй порядок точности, только вместо $O(h^2)$ надо брать $O(N^{-2})$. Как и на равномерных сетках, можно сгущать квазиравномерные сетки последовательно вдвое и применять метод Ричардсона с поточечным контролем сходимости.

Обычно для преобразования (1.125) берут $s = 1/2$. Параметр c рекомендуется выбирать так, чтобы в области существенного изменения решения лежало около 75 % узлов сетки.

Нелинейные задачи. Существенным усложнением является нелинейная задача (1.109). Нетрудно взять разностную сетку (1.112) и заменить производные разностями, как сделано выше. Подставляя эти разности в дифференциальное уравнение (1.109), получим следующую разностную схему:

$$u_{n+1} - 2u_n + u_{n-1} = h^2 f\left(x_n, u_n, \frac{u_{n+1} - u_{n-1}}{2h}\right), \quad 1 \leq n \leq N-1;\ u_0 = \alpha, u_N = \beta. \tag{1.126}$$

Мы получили систему нелинейных алгебраических уравнений для определения сеточного решения. Самым сложным является вопрос — существует ли у этой системы вещественное решение, и каким алгоритмом его можно вычислить. Интуитивно представляется, что если у дифференциального уравнения существует решение, то при достаточно подробной сетке ($N \gg 1$) сеточное решение должно существовать. Однако нетрудно построить примеры, когда на грубой сетке вещественного сеточного решения не существует.

Будем считать,, что задача для исходного дифференциального уравнения корректно поставлена, а разностная сетка достаточно подробна. Тогда сеточное решение должно существовать и быть единственным в определенном смысле: формально может существовать другое решение сеточных уравнений, но оно будет хорошо отличимо от «правильного». Тогда для решения системы (1.126) целесообразно применять итерационный метод Ньютона

(см. кн. 1). Для этого формально перепишем схему (1.126) в виде $\mathbf{F}(\mathbf{u}) = 0$, где компоненты вектора $\mathbf{F}$ имеют следующий вид:

$$F_n = u_{n+1} - 2u_n + u_{n-1} - h^2 f\left(x_n, u_n, \frac{u_{n+1} - u_{n-1}}{2h}\right), \quad 1 \le n \le N-1; \qquad F_0 = u_0 - \alpha, F_N = u_N - \beta. \tag{1.127}$$

Матрица Якоби для этой системы трехдиагональна. Ее ненулевые элементы имеют следующий вид:

$$\frac{\partial F_n}{\partial u_{n-1}} = \left(1 + \frac{h}{2}\frac{\partial f}{\partial u_x}\right)_n, \quad \frac{\partial F_n}{\partial u_n} = -\left(2 + h^2 \frac{\partial f}{\partial u}\right)_n, \quad \frac{\partial F_n}{\partial u_{n+1}} = \left(1 - \frac{h}{2}\frac{\partial f}{\partial u_x}\right)_n, \quad 1 \le n \le N-1; \quad \frac{\partial F_0}{\partial u_0} = 1, \frac{\partial F_N}{\partial u_N} = 1. \tag{1.128}$$

Здесь индекс n около скобок означает, что в производных от f аргументы берутся в n-м узле, как в (1.113).

Напомним, что в методе Ньютона выбирается некоторое нулевое приближение для вектора $\mathbf{u} = \{u_n, 0 \le n \le N\}$. Далее выполняют итерационный процесс $s = 0, 1, \ldots$ по следующим формулам:

$$\mathbf{u}^{(s+1)} = \mathbf{u}^{(s)} + \mathbf{y}^{(s)}, \frac{\partial \mathbf{F}\left(\mathbf{u}^{(s)}\right)}{\partial \mathbf{u}} \mathbf{y}^{(s)} = -\mathbf{F}\left(\mathbf{u}^{(s)}\right). \tag{1.129}$$

Для вычисления вектора приращений $\mathbf{y}^{(s)}$ надо решить линейную систему с матрицей Якоби (1.128). Справедливо следующее утверждение.

Утверждение. Если во всех узлах сетки $\dfrac{\partial f}{\partial u} > 0$, а шаг h достаточно мал, так что $h\left|\dfrac{\partial f}{\partial u_x}\right| \le 2$, то решение линейной системы (1.129) существует и единственно. •

Утверждение легко доказывается: при сделанных предположениях диагональные элементы матрицы преобладают. Сформулированные условия являются достаточными, но не необходимыми. Обычно решение линейной системы возможно и при их несоблюдении, но ошибки округления при этом могут оказаться заметно больше.

Усечение. Если на каждой итерации линейная система (1.129) разрешима, а решение сеточных уравнений (1.126) существует, это еще не гарантирует сходимости итерационного процесса к сеточному решению. Метод Ньютона хорошо сходится лишь в достаточно малой окрестности корня, а при неудачном выборе нулевого приближения итерационный процесс может «болтаться», не сходясь ни к какому пределу. Это обычно опознается по поведению невязки линейной системы

$$\left|\mathbf{F}\left(\mathbf{u}^{(s)}\right)\right| = \sum_{n=0}^{N}\left[F_n\left(\mathbf{u}^{(s)}\right)\right]^2. \qquad (1.130)$$

В окрестности корня ньютоновские итерации сходятся квадратично, т. е. невязка при этом монотонно и очень быстро убывает. Немонотонное поведение невязки служит признаком «разболтки».

В этих случаях нередко удается добиться сходимости следующим приемом. Вычислив приращение $\mathbf{y}^{(s)}$ из линейной системы (1.129), найдем следующую итерацию по формуле

$$\mathbf{u}^{(s+1)} = \mathbf{u}^{(s)} + \tau_s \mathbf{y}^{(s)}, 0 < \tau_s \leq 1. \qquad (1.131)$$

Значение $\tau_s = 1$ соответствует методу Ньютона. При выполнении итераций сначала полагают $\tau_s = 1$. Если при этом невязка уменьшилась, т. е. $\left|\mathbf{F}\left(\mathbf{u}^{(s+1)}\right)\right| < \left|\mathbf{F}\left(\mathbf{u}^{(s)}\right)\right|$, то результат засчитывается. Если невязка возросла, то вектор $\mathbf{y}^{(s)}$ не пересчитывается, но берется меньшее τ_s; целесообразно полагать $\tau_s = 0,5$. Снова находятся $\mathbf{u}^{(s+1)}$ и невязка. Если невязка по-прежнему возросла, то шаг τ_s опять уменьшается вдвое. И так до тех пор, пока не удастся добиться уменьшения невязки.

Не рекомендуется использовать какие-либо варианты метода простых итераций. Во-первых, нетрудно построить такие примеры, когда сеточное решение существует, но простые итерации расходятся в его окрестности; в этом случае ни при каком нулевом приближении получить сходимость нельзя. Во-вторых, даже

если простые итерации сходятся, то их сходимость в окрестности корня линейна и нередко оказывается очень медленной, в то время как ньютоновские итерации вблизи простого корня сходятся квадратично.

Существование разностного решения априори не очевидно. Однако, если итерации сходятся к некоторому пределу, то в силу непрерывности f этот предел является решением алгебраической системы, т. е. разностное решение существует.

Сходимость разностного решения к точному при $h \to 0$ исследуется аналогично случаю линейного уравнения (1.113). При этом аналогами коэффициентов $q(x)$ и $r(x)$ в нелинейном уравнении (1.126) являются соответственно f_{u_x} и f_u. Справедлива теорема.

Теорема 1.6. Пусть $f(x, u, u_x)$ имеет вторые непрерывные производные по всем своим аргументам, $f_u \geq m > 0$, а шаг настолько мал, что $h|f_{u_x}| \leq 2$. Тогда разностное решение (если оно существует) сходится к точному в норме c с погрешностью $O(h^2)$. •

Обратим внимание, что термин «сходимость» традиционно употребляется в двух смыслах: как сходимость итерационного процесса к сеточному решению при $s \to \infty$ или как сходимость сеточного решения к точному при $h \to 0$. Всегда надо четко понимать, о какой сходимости идет речь.

Многосеточный метод. Любой численный расчет должен выполняться с контролем точности; это является одним из важных требований практики. Для разностных методов это означает, что расчет должен выполняться на последовательности сгущающихся сеток, когда точность сеточного решения можно оценивать методом Ричардсона.

В нелинейных задачах сгущение сеток позволяет добиваться быстрой сходимости итерационного процесса нахождения сеточного решения. В самом деле, пусть на некоторой сетке с числом интервалов N уже вычислено сеточное решение $\bar{u}_n, 0 \leq n \leq N$. Мы вдвое сгущаем сетку и хотим найти на ней решение $\tilde{u}_k, 0 \leq \leq k \leq 2N$. Если N не слишком мало, то оба сеточных решения близки к точному, а следовательно, и друг к другу. Возьмем в качестве нулевого приближения на сгущенной сетке сеточное решение с предыдущей сетки. В совпадающих узлах, т. е. при $k = 2n$, это делается простым переносом: $\tilde{u}_{2n}^{(0)} = \bar{u}_n$. В нечетных

узлах подробной сетки можно использовать линейную интерполяцию: $\tilde{u}^{(0)}_{2n+1} = (\bar{u}_{n+1} + \bar{u}_n)/2$. Такое нулевое приближение оказывается близким к $\tilde{u}_n$. Оно обычно попадает в область квадратичной сходимости ньютоновских итераций, так что не требуется усечения, и уже две-четыре итерации обеспечивают высокую точность. Только на первой сетке надо позаботиться о хорошем выборе нулевого приближения, а число итераций может быть значительным. Такая многосеточная процедура обеспечивает высокую надежность и малую трудоемкость расчетов.

Дадим рекомендации по критерию окончания ньютоновских итераций для 64-разрядных вычислений. Бессмысленно требовать, чтобы невязка была близка к 10^{-16}; из-за ошибок округления этот критерий может никогда не выполняться. Выбирать невязку на уровне $\sim 10^{-5}$ нежелательно: это не позволит эффективно повышать точность методом Ричардсона, так как сеточные решения будут найдены довольно грубо. Целесообразно оканчивать итерации по невязке $10^{-10} - 10^{-12}$. С учетом квадратичной сходимости метода Ньютона отличие достигнутого решения от точного будет лежать на уровне ошибок округления.

Брусок. Сеточный метод одинаково легко применим как к уравнениям второго порядка (наименьший возможный порядок для краевых задач), так и к уравнениям высоких порядков. Рассмотрим последний случай на примере задачи для упругого бруска (1.110). Это уравнение имеет четвертый порядок и дополняется четырьмя краевыми условиями. Построим для него схему, имеющую аппроксимацию $O(h^2)$. Два краевых условия первого рода аппроксимируются точно, аналогично случаю струны. Однако остаются еще два краевых условия, содержащих граничные производные. Если аппроксимировать производную u'_0 по двум точкам x_0, x_1, получим только первый порядок $O(h)$, что нас не устраивает.

Краевые условия. Удобно аппроксимировать граничные производные, вводя ***фиктивные точки*** за границей: $x_{-1} = x_0 - h, x_{N+1} = x_N + h$. Тогда граничные производные заменяются симметричными разностями, что дает следующие разностные уравнения:

$$x_0 = 0,\ x_N = 0,\ (x_1 - x_{-1})/2h = 0,\ (x_{N+1} - x_{N-1})/2h = 0. \tag{1.132}$$

Они обеспечивают аппроксимацию $O(h^2)$.

Можно не вводить фиктивные точки, аппроксимируя граничные производные не симметричной разностью по трем сеточным точкам с точностью $O(h^2)$:

$$\begin{aligned} u_0' &\approx \left(-\frac{3}{2}u_0 + 2u_1 - \frac{1}{2}u_2\right)/h = 0, \\ u_N' &\approx \left(\frac{3}{2}u_N - 2u_{N-1} + \frac{1}{2}u_{N-2}\right)/h = 0. \end{aligned} \tag{1.133}$$

Аппроксимация легко проверяется разложением u_n в ряды Тейлора. Такой способ немножко сложнее. Он применяется, если по каким-то соображениям фиктивную точку вводить нежелательно (например, исходное уравнение содержит коэффициенты, имеющие особенности на границах). Разумеется, краевые условия $x_0 = 0$, $x_N = 0$ берутся по-прежнему.

Регулярные узлы. Во внутренних узлах отрезка можно аппроксимировать четвертую производную симметричной пятиточечной разностью; это дает сеточное уравнение

$$(u_{n-2} - 4u_{n-1} + 6u_n - 4u_{n+1} + u_{n+2})/h^4 = f_n. \tag{1.134}$$

Мы имеем право менять индекс n в границах $1 \le n \le N-1$, если введено по одному фиктивному узлу за каждой границей и взяты краевые условия (1.132): при этом индексы $n-2$ и $n+2$ не выходят за пределы дополненной сетки. Если же фиктивные узлы не вводились и взяты краевые условия (1.133), то индекс меняется в пределах $2 \le n \le N-2$.

Решение. Разностные уравнения (1.134) с разностными краевыми условиями (1.132) или (1.133) образуют систему линейных уравнений относительно вектора u_n. Число уравнений в обоих случаях равно числу неизвестных. Матрица системы пятидиагональная. Поэтому найти численное решение можно методом Гаусса для ленточной матрицы. Число арифметических операций при этом пропорционально числу узлов сетки, т. е. алгоритм экономичен. Для матрицы с элементами из (1.134) преобладания диагонального элемента нет. Поэтому простое условие существования разностного решения, в отличие от уравнения струны, здесь получить не удается. Однако в реальных численных расчетах это решение легко получается, так что оно существует. Для более сложных дифференциальных уравнений возможно отсутствие разностного решения (если случайно одно из собственных

значений полученной матрицы точно равно нулю); но это бывает исключительно редко.

Однако обусловленность линейной системы для дифференциальных уравнений высокого порядка будет существенно хуже, чем для уравнения второго порядка. Например, для уравнения четвертого порядка (1.134) число десятичных знаков, теряемых из-за ошибок округления компьютера, $k \approx \frac{7}{2} \lg N$; если для хорошей точности мы возьмем $N \approx 100 - 1000$, то потеряем на округлениях $k \approx 7 - 10{,}5$ десятичных знаков! Такая потеря существенна даже для 64-разрядного компьютера, а при 32-разрядных вычислениях в ответе не будет ни одного верного знака.

Поэтому задачи для уравнений высоких порядков можно считать только при достаточно большой разрядности компьютера. Чем выше порядок дифференциального уравнения и большее число узлов сетки берется для получения хорошей точности, тем бóльшая разрядность чисел необходима.

За исключением ухудшения обусловленности, в остальном сеточный метод позволяет единообразно и просто решать любые краевые задачи для дифференциальных уравнений высокого порядка с разнообразными граничными условиями, включая нелокальные краевые условия (когда связаны между собой значения функции на правой и левой границах или включены интегральные условия типа нормировки функции).

Нелокальные краевые условия также не представляют трудности для сеточного метода. Например, рассмотрим задачу (1.111). На интервале $(x_{n-1},\ x_n)$ аппроксимируем производную простейшей разностью, а правую часть уравнения — полусуммой узловых значений; это дает в регулярных узлах схему аппроксимации $O(h^2)$. Краевое условие аппроксимируем точно и запишем после регулярных уравнений.

Таблица 1.4

Матрица системы (1.135)

$$\left|\begin{array}{ccccccc}
\bullet & \bullet & \circ & \circ & \circ & \circ & \circ \\
\circ & \bullet & \bullet & \circ & \circ & \circ & \circ \\
\circ & \circ & \bullet & \bullet & \circ & \circ & \circ \\
\circ & \circ & \circ & \bullet & \bullet & \circ & \circ \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \circ \\
\circ & \circ & \circ & \circ & \circ & \bullet & \bullet \\
\bullet & \circ & \circ & \circ & \circ & \circ & \bullet
\end{array}\right|$$

Получим следующую схему:

$$(u_n - u_{n-1})/h = (u_n + u_{n-1})/2,\ 1 \le n \le N;\ u_N - u_0 = c. \quad (1.135)$$

Получилась система линейных уравнений, матрица которой показана в табл. 1.4. Систему нетрудно решить методом Гаусса. Вся матрица, за исключением последней строки, уже имеет верхнюю треугольную форму. Остается привести к такой форме только нижнюю строку. Заметим, что в этой схеме возникает ограничение на шаг сетки: надо требовать $h < 2$, иначе разностное решение будет знакопеременным (точное решение этой задачи является знакопостоянным и монотонным). Это ограничение вызвано тем, что мы использовали разностную схему «с полусуммой»; в п. 1.2.3 показано, что эта схема немонотонна.

1.4.3. Другие методы

Метод Ритца применяют к линейным самосопряженным дифференциальным уравнениям. Такие уравнения являются уравнениями Эйлера для вариационной задачи на минимум квадратичного функционала. Например, для уравнения (1.120) с нулевыми краевыми условиями $u(0) = 0$, $u(a) = 0$ это будет функционал

$$F[u] \equiv \int_0^a \left[q(x) \left(\frac{du}{dx} \right)^2 + r(x)u^2(x) + 2u(x)f(x) \right] dx = \min. \quad (1.136)$$

Выберем некоторую систему линейно независимых базисных функций $\phi_m(x)$, удовлетворяющую тем же самым краевым условиям: $\phi_m(0) = 0$, $\phi_m(a) = 0$, $m = 1, 2, \ldots$. Будем искать аппроксимацию решения в виде конечной суммы базисных функций:

$$u(x) \approx \Phi_M(x) \equiv \sum_{m=1}^{M} c_m \phi_m(x). \quad (1.137)$$

Подставим приближение (1.137) в функционал (1.136) и потребуем

$$\int_a^b \left[q(x) \left(\frac{d\Phi_M}{dx} \right)^2 + r(x)\Phi_M^2(x) + 2\Phi_M(x)f(x) \right] dx = \min. \quad (1.138)$$

Поскольку вид базисных функций известен, интеграл (1.138) есть квадратичная функция неизвестных коэффициентов c_m. Минимум (1.138) находим, приравнивая нулю все производные по c_m. Учитывая, что $\partial\Phi_M(x)/\partial c_m = \phi_m(x)$, получим следующую систему уравнений:

$$\sum_{k=1}^{M} A_{mk}c_k = B_m,\ 1 \le m \le M;$$

$$A_{mk} = \int_0^a \left[q(x)\frac{d\phi_m(x)}{dx}\frac{d\phi_k(x)}{dx} + r(x)\phi_m(x)\phi_k(x)\right]dx, \quad (1.139)$$

$$B_m = -\int_0^a \phi_m(x)f(x)dx.$$

Уравнения (1.139) являются линейной системой относительно коэффициентов c_m. Коэффициенты этой системы являются интегралами от известных функций и могут быть найдены либо точным интегрированием, либо численным. Из теории известно, что при $q(x) > 0$, $r(x) > 0$ определитель системы $\det A_{mk} > 0$, так что система имеет единственное решение. Это решение может быть найдено методом Гаусса; при этом выбор главного элемента не нужен, так как главный элемент автоматически находится на диагонали.

В общем случае матрица (A_{mk}) оказывается плотно заполненной. Такие матрицы зачастую оказываются плохо обусловленными; тогда при решении системы (1.139) ошибки округления могут стать значительными. Поэтому стараются выбрать такую систему базисных функций, при которой матрица (1.139) будет диагональной или почти диагональной. От удачного выбора базисных функций зависит успех вычислений.

Например, пусть $q(x) \equiv \text{const}$ и $r(x) \equiv \text{const}$. Выберем тригонометрическую подсистему синусов в качестве базиса:

$$\phi_m(x) = \sin(\pi m x/a), m = 1, 2, \ldots.$$

Подставляя эту систему в (1.139), получим

$$A_{mk} = a\left(\frac{\pi^2 q m^2}{a^2} + r\right)\delta_{mk}/2,$$

где δ_{mk} — символы Кронекера. Матрица оказалась диагональной, так что решение линейной системы сразу записывается явно.

Однако такое упрощение получается не часто. Уже при $q^{(x)} \neq$ $\neq$ const или $r(x) \neq$ const матрица системы станет недиагональной.

Достоинством метода Ритца является то, что приближенное решение получается в виде явной функции от непрерывного аргумента x, а не только в отдельных узлах сетки. Однако успех сильно зависит от удачного выбора базиса $\phi_m(x)$, т. е. от опыта вычислителя и предварительного знания поведения точного решения. Заметим, что этим методом пользуются не только для линейных самосопряженных уравнений, но и для нелинейных уравнений. Но тогда для определения коэффициентов c_m получается система нелинейных уравнений; это сильно осложняет нахождение приближенного решения.

Заметным недостатком метода Ритца является то, что в нем в отличие от сеточного метода Ричардсона не построены апостериорные асимптотически точные оценки погрешности. Это не позволяет надежно контролировать точность выполненных расчетов.

Метод Галеркина применим к любым уравнениям, в том числе к несамосопряженным. Пусть имеется произвольное дифференциальное уравнение, которое символически записано в виде $\Psi(u(x)) = 0$, $0 \leq x \leq a$; все производные включены в оператор Ψ. Аналогично методу Ритца, выберем базис $\phi_m(x)$ и аппроксимацию обобщенным многочленом (1.137). Подставим эту аппроксимацию в оператор Ψ; тогда этот оператор станет некоторой функцией, зависящей от x и неизвестных коэффициентов c_m. Потребуем, чтобы эта функция была ортогональна всем использованным базисным функциям; ортогональность понимаем в смысле интегрального скалярного произведения:

$$\int_0^a \Psi\left(\sum_{k=1}^{M} c_k \phi_k(x)\right) \phi_m(x) dx = 0,\ m = 1, 2, \ldots, M. \qquad (1.140)$$

Выполним в (1.140) все интегрирования по x, поскольку все входящие туда функции известны. Получится система уравнений для определения коэффициентов c_m, в общем случае нелинейная; тогда решать систему достаточно трудно. Если оператор Ψ линеен, то система (1.140) будет линейна относительно коэффициентов c_m; ее нетрудно решить методом Гаусса. Матрица системы обычно будет плотно заполненной; в исключительных случа-

ях она может оказаться диагональной или почти диагональной. Справедлива следующая теорема.

Теорема 1.7. Если функции $\phi_m(x)$ линейно независимы, а их система полна в классе достаточно гладких функций с заданными граничными условиями, то при $M \to \infty$ обобщенный многочлен (1.137) сходится к точному решению $u(x)$ в норме L_2. •

В качестве примера рассмотрим задачу упругого бруска (1.110) с четырьмя краевыми условиями. Выберем линейно независимые базисные функции, удовлетворяющие всем четырем краевым условиям:

$$\phi_m(x) = \sin^2(\pi m x/a).$$

Подставим эти функции и оператор (1.110) в уравнение (1.140); получим линейную систему для c_m, причем интегрирование матричных элементов выполняется точно (квадраты синусов удобно заменить через косинус двойного угла):

$$\sum_{k=1}^{M} A_{mk} c_k = B_m,$$

$$A_{mk} = \frac{2\pi^4 k^4}{a^3} \delta_{mk},$$

$$B_m = \int_0^a f(x) \sin^2(\pi m x/a)\, dx,$$

где δ_{nm} — символ Кронекера.

Матрица системы оказалась диагональной. Поэтому решение линейной системы тривиально: $C_m = B_m/A_{mm}$. Обусловленность вычислений идеальная, и расчеты можно проводить при больших значениях M, даже с 32-разрядными числами. Напомним, что сеточный метод для этой задачи требовал 64-разрядных чисел. Это показывает, что метод Галеркина может иметь важное преимущество.

Конечные элементы — это базисные функции, имеющие конечный носитель: каждая функция отлична от нуля на своем интервале, малом по сравнению со всем отрезком, на котором решается задача. Несовпадение носителей разных базисных функций обеспечивает их линейную независимость. Приведем

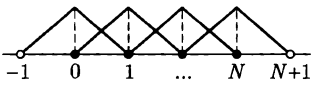


Рис. 1.6. Конечные элементы — линейные B-сплайны

пример такого базиса, часто используемого в прикладных расчетах. Это так называемые линейные B-сплайны. Вводят сетку x_n с шагами $h_n = x_{n+1} - x_n$ и полагают

$$\begin{cases} \phi_n(x) = (x - x_n)/h_n, \ \dfrac{d\phi_n}{dx} = \dfrac{1}{h_n}, \ x \in (x_n, x_{n+1}); \\ \phi_n(x) = (x_{n+2} - x)/h_{n+1}, \ \dfrac{d\phi_n}{dx} = \dfrac{-1}{h_{n+1}}, \ x \in (x_{n+1}, x_{n+2}); \\ \phi_n(x) = 0, \ \dfrac{d\phi_n}{dx} = 0, \ x \notin (x_n, x_{n+2}). \end{cases} \tag{1.141}$$

Функции (1.141) являются непрерывными кусочно линейными, а их первые производные —разрывными кусочно постоянными. Их носителем является отрезок $[x_n, x_{n+2}]$. Носители соседних функций перекрываются (рис. 1.6), а носители более далеких функций — нет.

Выражения (1.141) относятся к регулярным узлам сетки; для этого примера такими узлами будут $n = 0, 1, 2, \ldots, N-2$. При этом треугольники целиком принадлежат исходному отрезку $[a, b]$. Но для получения полной системы функций необходимо добавить и те треугольники, которые частично выходят за пределы отрезка (левый треугольник на рис. 1.6). Конечный элемент на левой границе начинается в фиктивном узле $n = -1$ и кончается в узле $n = 1$. Правый граничный элемент начинается в узле $n = N - 1$ и заканчивается в фиктивном узле $n = N + 1$.

Функции типа (1.141) можно использовать в методах Ритца или Галеркина. Поскольку перекрываются только носители близких функций, в матрицах (A_{mk}) будут отличны от нуля только элементы главной диагонали и нескольких соседних, т. е. матрица будет ленточной с неширокой лентой. Для системы (1.141), где перекрываются только соседние функции, ненулевыми будут лишь элементы A_{mm} и $A_{m,m\pm1}$; матрица окажется трехдиагональной. Это существенно облегчает задачу решения алгебраической системы для коэффициентов c_m.

Таким образом, метод конечных элементов является некоторым сочетанием сеточных методов с методами Ритца и Галеркина. Поэтому его нередко называют проекционно-сеточным методом. Окончательную систему алгебраических уравнений для c_m

в методе конечных элементов можно рассматривать как некоторую специфическую разностную схему.

Отметим основной недостаток метода конечных элементов. Простейшие конечные элементы типа (1.141) приводят к достаточно простым алгоритмам; но эти элементы имеют невысокую гладкость: у них отсутствует вторая производная. Поэтому их можно применять только к задачам, которые можно преобразовать к виду, содержащему лишь первые производные. Вдобавок, точность выше чем $O(h^2)$ получить не удается. Если же взять более гладкие конечные элементы (например, параболические или кубические B-сплайны), то алгоритмы становятся очень громоздкими.

Стрельба. Ранее этот метод широко использовался. Поясним его суть на примере простейшей краевой задачи (1.107) — (1.108). Вместо краевых условий на разных границах (1.108) поставим два условия на левой границе. Одно условие будет соответствовать точной постановке: $u(a) = \alpha$. Второе условие будет $\dfrac{du(a)}{dx} = \gamma$, где значение γ выбрано «с потолка». Уравнение второго порядка (1.107) вместе с новыми краевыми условиями образуют задачу Коши (только в подразделе 1.1 аргумент мы обозначали через t, а не через x).

Для решения задачи Коши можно использовать любые схемы из подраздела 1.1, например схемы РК высокого порядка точности. Однако когда расчет дойдет до значения $x = b$, полученное значение $u(b)$ может сильно отличаться от требуемой величины β («выстрел» мимо цели). Тогда меняют величину γ и снова производят расчет. Так поступают до тех пор, пока не получат $u(b) \approx \beta$ с требуемой точностью.

Для изменения γ конструируют различные алгоритмы. Простейший алгоритм похож на пристрелку цели артиллеристами: если один расчет дает $u(b) > \beta$, а другой — $u(b) < \beta$, то делят «вилку» пропорционально. Отсюда и название метода.

Основным преимуществом метода стрельбы считалась возможность использовать схемы РК и другие высокого порядка точности. Однако у этого метода много серьезных недостатков. 1. Метод трудно применять к уравнениям высоких порядков, когда на каждой границе поставлено два или более условий; приходится «пристреливать» сразу несколько параметров, и эффективный алгоритм их подбора трудно сконструировать. А раз-

ностные схемы для таких краевых задач записываются естественно, как мы видели на примере упругого бруска. 2. Краевая задача может быть хорошо обусловленной, а сопутствующая задача Коши — очень плохо обусловленной. Тогда при решении задачи Коши очень сильно нарастает погрешность и расчет может даже выйти за пределы представимых на компьютере чисел. Однако в разностных схемах для таких задач никаких осложнений не возникает.

Поэтому в настоящее время метод стрельбы почти вытеснен из употребления разностными схемами.

1.5. ЗАДАЧИ НА СОБСТВЕННЫЕ ЗНАЧЕНИЯ

1.5.1. Постановки задач

Существует много краевых задач, в уравнение которых входит один или несколько неизвестных параметров λ. При решении таких задач надо одновременно определить как неизвестную функцию $u(x)$, так и параметр λ. Такие задачи называют задачами на собственные значения, а удовлетворяющие этой задаче $u(x)$ и λ — собственной функцией и собственным значением.

В подразделе 1.3 отмечалось, что краевая задача для уравнения p-го порядка требует задания p краевых условий. В задаче на собственные значения помимо $u(x)$ требуется еще определить параметр λ (в общем случае s неизвестных параметров). Очевидно, для корректности постановки задачи надо брать дополнительные краевые условия; в общем случае число дополнительных условий будет равно $p + s$.

Приведем некоторые примеры таких задач.

Струна. Рассмотрим неоднородную струну, натянутую горизонтально, с концами закрепленными в точках $x = a$ и $x = b$. Пусть возбуждены свободные колебания этой струны (ударом рояльного молоточка или движением скрипичного смычка). В курсах математической физики показано, что отклонение струны от положения равновесия имеет вид линейной комбинации так называемых гармоник; первую гармонику называют основным тоном струны, следующие — обертонами.

При этом k-й обертон дает отклонение $u_k(x)\sin(\lambda_k t)$; здесь x — пространственная координата; t — время. Функцию $u_k(x)$

называют амплитудой колебания; она удовлетворяет дифференциальному уравнению со следующими краевыми условиями:

$$\frac{d}{dx}\left[q(x)\frac{du}{dx}\right] - r(x)u(x) + \lambda u(x) = 0,\ 0 \le x \le a; u(0) = 0, u(a) = 0. \tag{1.142}$$

Здесь записано уравнение второго порядка с неизвестным параметром. Для такой задачи требуется три краевых условия, а в (1.142) имеется только два. На самом деле здесь есть третье условие, заданное в скрытой форме. В самом деле уравнение (1.142) является линейным однородным относительно $u(x)$; следовательно, умножение $u(x)$ на постоянный множитель также является решением задачи. Фиксация этого множителя есть третье условие (в жизни этот множитель определяется силой удара молоточка).

Найдем точное решение задачи (1.142) в простейшем случае $q(x) = \text{const} = 1$, $r(x) = \text{const} = 0$. Нетрудно проверить, что тогда задаче удовлетворяет следующий набор решений:

$$u_k(x) = \sin(\pi k x/a),\ \lambda_k = \pi^2 k^2/a^2,\ k = 1, 2, \ldots. \tag{1.143}$$

Число гармоник не ограничено, а их частоты возрастают пропорционально k^2. Совокупность всех собственных значений называют спектром задачи. Спектр вида (1.143) называют дискретным, так как он состоит из набора отдельных чисел.

Задача (1.142) возникает и в других прикладных областях. Например, она возникает при решении основного уравнения квантовой механики — уравнения Шредингера для движения частицы в силовом поле, имеющем плоскую цилиндрическую или сферическую симметрию.

Брусок. Камертон является упругим однородным бруском, концы которого свободны (он изогнут и закреплен в средней точке, но это практически не влияет на описывающее его уравнение). Поэтому задачу на собственные колебания камертона можно записать в следующем виде (мысленно разворачивая камертон в прямой брусок):

$$\begin{gathered} u_{xxxx} = \lambda u, \\ 0 \le x \le a;\ u(0) = 0,\ u_{xx}(0) = 0,\ u(a) = 0,\ u_{xx}(a) = 0. \end{gathered} \tag{1.144}$$

Обозначим k-ю собственную функцию через $v_k(x)$. Нетрудно проверить, что точное решение задачи (1.144) имеет следующий

дискретный набор собственных значений и собственных функций:

$$v_k(x) = \sin(\pi k x/a)\,,\ \lambda_k = (\pi k/a)^4\,,\ k = 1, 2, 3, \ldots\,. \qquad (1.145)$$

Заметим, что собственные значения камертона при увеличении номера k возрастают гораздо быстрее, чем для струны. Поэтому у камертона слышен практически только основной тон.

Другие задачи. Рассмотренные задачи были линейными однородными, содержали только одно собственное значение и имели дискретный спектр. Такие задачи относят к простейшим. Существует много важных задач, имеющих гораздо более сложные спектры. Приведем примеры.

Уравнение Шредингера для одной частицы в силовом поле, имеющем форму ямы, в общем случае имеет дискретный спектр (в квантовой механике спектральные значения называют уровнями энергии). Если эта яма неглубокая, то задача может вообще не иметь решения, т. е. в яме не окажется ни одного уровня. В более глубокой яме появляется дискретный спектр с конечным числом уровней. Чем глубже и шире яма, тем большее число уровней в ней возможно. В достаточно большой яме число уровней становится бесконечным.

Силовое поле кристаллической решетки является периодической функцией — набором потенциальных ям. При этом собственные значения оказываются комплексными и образуют не дискретный спектр, а непрерывный. При этом спектр состоит из так называемых полос; в пределах полосы λ меняются непрерывно, а между полосами уровни отсутствуют.

В многоэлектронных атомах уравнение Шредингера ни точно, ни численно решить не удается. Поэтому уравнение заменяют приближенным уравнением Хартри, Хартри — Фока или другими; такое уравнение записано не для одного электрона, а для всех электронов атома. Каждый электрон имеет свою собственную функцию и свое собственное значение. В этой задаче есть не один неизвестный параметр λ, а множество; их число равно числу электронов атома. Заметим, что само уравнение при этом оказывается нелинейным относительно $u(x)$. Для того чтобы численно решать подобные задачи, необходимо предварительно исследовать поведение точных решений и определить характер ожидаемого спектра. Не зная этого, трудно построить хороший численный метод.

1.5.2. Сеточный метод

Сеточный метод является наиболее распространенным способом численного решения задач на собственные значения. Аналогично краевым задачам, вводят сетку x_n и для сеточной функции $u_n \approx u(x_n)$ записывают разностную схему. Однако теперь неизвестным будет еще значение λ. Рассмотрим на простейших примерах, к чему это приводит.

Струна. Возьмем для простоты равномерную сетку с шагом $h = a/N$. Возьмем упрощенную задачу (1.142) с $q(x) = 1$, $r(x) \equiv \equiv 0$. Заменяя вторую производную разностным отношением и умножая на h^2, получим следующую схему:

$$u_{n-1} - \left(2 - h^2\lambda\right) u_n + u_{n+1} = 0,\ 1 \le n \le N-1;\ u_0 = 0,\ u_N = 0. \tag{1.146}$$

В подразделе 1.3 сеточная задача оказывалась системой алгебраических уравнений для сеточной функции u_n. Теперь же, благодаря наличию неизвестного параметра λ, получается другой класс математической задачи: задача на собственные значения вида

$$Au = \lambda u. \tag{1.147}$$

Матрица A в (1.147) является трехдиагональной; если принять за собственное значение величину $h^2\lambda$, то ненулевыми элементами матрицы будут $a_{nn} = -2$, $a_{n,n\pm1} = 1$.

Покажем, как с помощью простейших вычислений можно удовлетворительно получить по крайней мере первое собственное значение. Как известно, спектр матрицы A определяется из решения характеристического уравнения; для данной задачи оно принимает вид $\det\left(A + h^2\lambda E\right) = 0$. Порядок матрицы A равен числу регулярных узлов сетки. В данном случае это $N-1$. Именно столько собственных значений будет у матрицы. Очевидно, они должны быть приближениями к первым $N-1$ собственным значениям точной задачи (1.147). Решим задачу на самых грубых сетках. Для упрощения всех выражений положим $a = 1$.

Сначала положим $N = 2$, $h = 1/2$, порядок матрицы равен 1, и характеристическое уравнение принимает вид $-2 + h^2\lambda = 0$, т. е. $\lambda = 8$. Это приближение к значению $\lambda_1 = \pi^2 \approx 9{,}87$. Видно, что ошибка составляет около 20 %, что неплохо для столь грубой сетки.

Затем положим $N = 3$, $h = 1/3$. Порядок матрицы равен 2, а характеристическое уравнение легко приводится к виду

$\left(h^2\lambda - 2\right)^2 - 1 = 0$. Это квадратное уравнение; оно имеет два корня $\lambda = 9$ и $\lambda = 27$. Они являются приближениями к точным значениям $\lambda_1 = \pi^2$ и $\lambda_2 = 4\pi^2$. Теперь первое собственное значение получается с точностью около 9 %. Ошибка второго собственного значения составляет около 30 %, что существенно хуже.

Уточним первое собственное значение по методу Ричардсона, используя расчеты на этих двух сетках. Для аппроксимации использовалась симметричная разность, обеспечивающая порядок точности $p = 2$. Коэффициент сгущения сетки составлял 3/2. Используя стандартную технику метода Ричардсона (см. п. 1.1.10), получим уточненное значение

$$\widetilde{\lambda} = 9 + \frac{9-8}{(3/2)^2 - 1} = 9,80.$$

Отличие от точного значения составляет всего 0,7%! Этот пример показывает, как без компьютера («на пальцах») можно получать хорошие оценки.

Разностная задача (1.146) часто возникает в разностных схемах для уравнений в частных производных. Поэтому приведем ее точное решение, которое нетрудно проверить прямой подстановкой:

$$\begin{gathered} u_n^k = \sin\left(\pi k x_n / a\right), \ 0 \le n \le N; \\ \lambda_k = (2N/a)^2 \sin^2\left(\frac{\pi k}{2N}\right), \ 1 \le k \le N-1. \end{gathered} \tag{1.148}$$

Нетрудно убедиться, что при $N = 2$ или $N = 3$ получаем вычисленные выше собственные значения. Видно также, что первые собственные значения с $k \ll N$ близки к соответствующим точным собственным значениям. Для высоких гармоник точность численного решения быстро ухудшается.

Брусок. Составим разностную схему для задачи камертона (1.144). Возьмем равномерную сетку. Чтобы аппроксимировать краевые условия с второй производной, введем фиктивные точки x_{-1} и x_{N+1}. Тогда нетрудно записать следующую систему уравнений:

$$\begin{gathered} u_0 = 0, \ u_{-1} - 2u_0 + u_1 = 0; \\ \left(u_{n-2} - 4u_{n-1} + 6u_n - 4u_{n+1} + u_{n+2}\right)/h^4 = \lambda u_n, \ 1 \le n \le N-1; \\ u_{N-1} - 2u_N + u_{N+1} = 0, \ u_N = 0. \end{gathered} \tag{1.149}$$

Первые два уравнения (1.149) аппроксимируют левые краевые условия; следующая цепочка уравнений аппроксимирует дифференциальное уравнение в регулярных узлах; последние два уравнения аппроксимируют правые краевые условия. Все разностные выражения симметричны, что обеспечивает аппроксимацию $O(h^2)$.

Решение задачи (1.149) легко угадывается. Разностные собственные функции совпадают с точными собственными функциями $v_k(x_n)$, а сеточные собственные значения равны

$$\mu_k = \left(\frac{16}{h^4}\right) \sin^4\left(\frac{\pi k h}{2a}\right), \; h = \frac{a}{N}, \; k = 1, 2, 3, \ldots .$$

Это решение легко проверяется подстановкой в (1.149). Видно, что младшие собственные значения с $k \ll N$ близки к точным значениям (1.145); с ростом k точность значений μ_k быстро ухудшается. Поэтому целесообразно вычислять сеточным методом только младшие собственные значения.

В первые два и последние два уравнения (1.149) значения λ не входят, поэтому схему (1.149) нельзя рассматривать как задачу на собственные значения матрицы. Однако можно исключить из регулярных уравнений значения u_{-1} и u_0 с помощью первых двух уравнений (граничных условий), а значения u_N и u_{N+1} — с помощью последних двух уравнений. Тогда остается задача на собственные значения для пятидиагональной матрицы. Поскольку реально нужны лишь младшие собственные значения, целесообразно использовать метод обратных итераций с переменным сдвигом, как и для задачи струны. Поскольку матрица пятидиагональна, то обратная итерация для нахождения вектора $\mathbf{u}^{(s+1)}$ легко решается методом Гаусса для ленточной матрицы; подобно преобразовывать эту матрицу к трехдиагональной невыгодно.

1.5.3. Обратные итерации

Значительная часть практически важных задач ставится для линейных дифференциальных уравнений. Коэффициенты этих уравнений могут быть переменными и даже разрывными (задачи в слоистых средах). Разностные сетки зачастую приходится выбирать неравномерными. Область определения задачи может быть как конечной, так и неограниченной. В этих случаях следует строить бикомпактные разностные схемы, как это описано

в п. 1.4.2. Поскольку дифференциальное уравнение линейно относительно $u(x)$ и ее производных, разностная схема будет линейной относительно сеточного решения u_n. Поэтому сеточная задача примет общий вид задачи (1.147) на собственные значения матрицы. Сама матрица в простейшем случае оказывается трехдиагональной аналогично (1.146). В общем случае она может иметь достаточно сложный вид, но чаще всего она оказывается ленточной с неширокой лентой.

В кн. 1 рассмотрены алгоритмы для эффективного нахождения спектров таких матриц. Напомним, что если матрица достаточно плотно заполнена, то ее сначала следует привести к верхней почти треугольной форме с помощью подобных преобразований отражения; если матрица A была эрмитовой, то преобразованная матрица также будет эрмитовой и, тем самым, трехдиагональной. Для пятидиагональных или других достаточно легко обращаемых матриц такое преобразование необязательно. Далее будем полагать, что матрица подобно преобразована к легко обратимой форме. Если требуется найти все собственные значения матрицы, то обычно используют стандартные программы QR-алгоритма. Однако на практике это редко требуется: только небольшая часть младших собственных значений матрицы близка к собственным значениям дифференциального уравнения. Точность остальных собственных значений мала, так что находить их нецелесообразно.

Если матрица A эрмитова, то ее собственные значения будут вещественными, даже если ее элементы комплексны. Если же матрица неэрмитова, то даже при вещественных элементах ее собственные значения могут оказаться комплексными. Поэтому во всех программах следует предусмотреть комплексную арифметику.

Младшие собственные значения обычно находят поочередно, используя алгоритм обратных итераций с переменным сдвигом. В этом алгоритме выбирают некоторые начальные приближения $\lambda^{(1)}$, $\left\{u_n^{(1)}\right\}$, по возможности близкие к искомым собственной функции и собственному значению. Итерации $s = 1, 2, \ldots$ вычисляют следующим образом. Сначала находят новое приближение для собственного вектора, используя обратные итерации:

$$\left(A - \lambda^{(s)} E\right) \mathbf{u}^{(s+1)} = \mathbf{u}^{(s)}, \ \mathbf{u} = \{u_n\}. \qquad (1.150)$$

Затем уточняют собственное значение:

$$\lambda^{(s+1)} = \lambda^{(s)} + \left(\mathbf{u}^{(s+1)}, \mathbf{u}^{(s)}\right) / \left(\mathbf{u}^{(s+1)}, \mathbf{u}^{(s+1)}\right). \qquad (1.151)$$

Уравнение (1.150) есть линейная система для $\mathbf{u}^{(s+1)}$. Поскольку предполагается, что матрица A заранее преобразована к форме узкой ленты, то система (1.150) экономично решается методом Гаусса. Формула (1.151) требует лишь вычисления скалярных произведений, поэтому алгоритм оказывается не трудоемким.

Вывод. Поясним вывод формулы обратной итерации (1.150). Обозначим k-е собственные векторы и собственные значения матрицы A через $\mathbf{v}_k, \mu_k$ (последние несколько отличаются от собственных значений дифференциального уравнения λ_k). Разложим итерируемые векторы по собственным векторам матрицы (предполагая, что последние образуют базис):

$$\mathbf{u}^{(s)} = \sum_k \alpha_k^{(s)} \mathbf{v}_k, \ \mathbf{u}^{(s+1)} = \sum_k \alpha_k^{(s+1)} \mathbf{v}_k. \qquad (1.152)$$

Подставляя эти выражения в (1.150) и учитывая, что $A\mathbf{v}_k = \mu_k \mathbf{v}_k$, получим

$$\alpha_k^{(s+1)} = \alpha_k^{(s)} / \left(\mu_k - \lambda^{(s)}\right).$$

Если $\lambda^{(s)}$ близко к μ_k, то амплитуда k-й компоненты разложения за одну итерацию сильно увеличивается. Если же $\lambda^{(s)}$ далеко от μ_k, то усиления соответствующей компоненты не происходит. Поэтому на обратной итерации (1.150) усиливается только одна компонента разложения, причем усиливается существенно. За небольшое число обратных итераций вектор $\mathbf{u}$ станет почти параллельным собственному вектору $\mathbf{v}_k$.

Формула сдвига (1.151) выводится из следующих соображений. Выберем $\lambda^{(s+1)}$ так, чтобы минимизировать невязку задачи на собственные значения: $\left(A - \lambda^{(s+1)}E\right)\mathbf{u}^{(s+1)} = \min$. Поскольку слева стоит вектор (причем в общем случае с комплексными компонентами), следует иметь в виду минимум длины вектора, т. е. скалярного произведения:

$$\left(\left(A - \lambda^{(s+1)}E\right)\mathbf{u}^{(s+1)}, \left(A - \lambda^{(s+1)}E\right)\mathbf{u}^{(s+1)}\right) = \min.$$

Для минимизации надо приравнять нулю производную от скалярного произведения по $\lambda^{(s+1)}$:

$$\left(\mathbf{u}^{(s+1)}, \left(A - \lambda^{(s+1)}E\right)\mathbf{u}^{(s+1)}\right) = 0. \qquad (1.153)$$

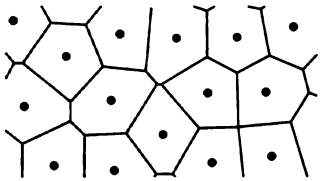


Рис. 1.7. Многоугольники Вороного

Подставляя сюда $\mathbf{u}^{(s+1)}$ из (1.150), после преобразований получим (1.151).

Сходимость. Приведем без доказательства следующие результаты. Как бы ни было выбрано начальное приближение, итерации (1.150) — (1.151) всегда сходятся. Собственное значение μ_k , к которому имеет место сходимость, определяется по следующему правилу. Рассмотрим всю совокупность μ_k на комплексной плоскости. Разобьем эту плоскость на такие области, что в каждой лежит только одно значение μ_k, а границы областей равноудалены от ближайших μ_k. Эти области называют многоугольниками Вороного (рис 1.7). Их границы являются ломаными, причем некоторые многоугольники являются неограниченными. Если $\lambda^{(1)}$ попало в некоторый многоугольник, то итерации сходятся к соответствующему μ_k независимо от выбора $\mathbf{u}^{(1)}$.

Если $\lambda^{(1)}$ лежит вблизи границы многогранника Вороного или $\mathbf{u}^{(1)}$ выбрано очень далеким от соответствующего вектора $\mathbf{v}_k$, то сначала итерации могут сходиться медленно. По мере приближения к решению сходимость ускоряется. Вблизи решения сходимость квадратичная, а если собственные векторы образуют ортогональный базис (например, для эрмитовых матриц) — даже кубическая. Поэтому на практике число итераций обычно оказывается небольшим (3 — 5).

Начальное приближение. Расположение многоугольников Вороного заранее неизвестно, поэтому, хотя процесс обязательно сойдется, он может сойтись не к тому собственному значению, которое нам нужно. По этой причине для выбора $\lambda^{(1)}$ надо использовать те оценки точного решения задачи, которые удается сделать.

Выбор $\mathbf{u}^{(1)}$ обычно слабо влияет на сходимость. Однако есть одно ограничение. В ряде прикладных задач собственный вектор $\mathbf{v} = \{v_n\}$ может иметь определенную симметрию по индексу n: четную или нечетную относительно центрального индекса $n = N/2$. Если нечаянно выбрать $\mathbf{u}^{(1)}$ противоположной четности, то итерационный процесс сначала долго не будет сходиться; затем, когда в результате итераций нарастут компьютерные ошибки округления, процесс сойдется к тому собственному

вектору, симметрия которого одинакова с заданным начальным приближением.

Чтобы избежать подобной ситуации, в качестве компонент вектора $\mathbf{u}^{(1)}$ рекомендуется выбирать псевдослучайные числа.

1.5.4. Дополненный вектор

Ранее были рассмотрены примеры задач на собственные значения для линейных дифференциальных уравнений. Однако существует много важных практических задач для нелинейных уравнений. Например, рассмотрим струну, коэффициент упругости которой не постоянен, а зависит от величины деформации $u(x)$: $\kappa(u) = \kappa_0 + \kappa_1 u^2$; в реальных задачах $0 < \kappa_1 \ll \kappa_0$. Это явление имеет место на самом деле: если струна оттянута вбок на величину $u(x)$, то ее длина увеличивается на величину $O(u^2)$; это приводит к дополнительному натяжению струны, что эквивалентно эффективному увеличению коэффициента упругости. Задача примет следующий вид:

$$\frac{d}{dx}\left[\left(\kappa_0 + \kappa_1 u^2\right)\frac{du}{dx}\right] + \lambda u = 0,\ 0 \le x \le a;\ u(0) = 0,\ u(a) = 0. \tag{1.154}$$

В отличие от линейного случая здесь необходимо явно написать и третье краевое условие; но его формулировка не очевидна.

Нетрудно написать разностную схему для задачи (1.154). Для аппроксимации дифференциального оператора следует воспользоваться бикомпактными разностными схемами аналогично п. 1.4.2. На равномерной сетке $h = a/N$ получим следующую схему аппроксимации $O(h^2)$:

$$\begin{aligned}&\left[\kappa_0 + \kappa_1\left(u_{n+1}^2 + u_n^2\right)/2\right](u_{n+1} - u_n) - \\ &- \left[\kappa_0 + \kappa_1\left(u_n^2 + u_{n-1}^2\right)/2\right](u_n - u_{n-1}) + \\ &+\lambda h^2 u_n = 0,\ 1 \le n \le N-1;\ u_0 = 0,\ u_N = 0.\end{aligned} \tag{1.155}$$

Здесь не хватает уравнения для третьего краевого условия. Однако и без него видно, что алгебраическая система (1.155) с неизвестными u_n и λ не является задачей на собственные значения матрицы. Это существенно нелинейная алгебраическая система, и встает вопрос о работоспособном алгоритме нахождения сеточного решения.

Эффективным оказался метод дополненного вектора. Дадим общее описание этого алгоритма. Пусть требуется найти сеточные собственные значения λ и собственную функцию $\mathbf{u} = \{u_n,\ 0 \le n \le N\}$, которая является вектором; общее число неизвестных равно $N+2$. Они удовлетворяют разностной схеме, т. е. системе $N+2$ нелинейных алгебраических уравнений, которую в общем виде можно записать так:

$$\mathbf{f}(\mathbf{u},\lambda) = 0,\ \mathbf{f} = \{f_n,\ 0 \le n \le N+2\},\quad \mathbf{u} = \{u_n,\ 0 \le n \le N\}. \tag{1.156}$$

В такой постановке величины $\mathbf{u}$ и λ являются элементами разных метрических пространств.

Введем дополненный вектор $\mathbf{U} = \{u_0, u_1, \ldots, u_N, u_{N+1} \equiv \lambda\}$, т. е. дополним вектор $\mathbf{u}$ еще одной компонентой. Тогда систему (1.156) можно формально записать как $\mathbf{F}(\mathbf{U}) = 0$. Единственным работоспособным методом решения нелинейных алгебраических систем общего вида является метод Ньютона или его вариант с усечением шага (см. п. 1.4.2, а также кн. 1).

Метод Ньютона хорошо сходится лишь в малой окрестности решения. Следует использовать сгущение сеток и многосеточный метод, также описанный в п. 1.4.2. В качестве нулевого приближения для сгущенной сетки следует взять сошедшееся решение с предыдущей сетки. При этом число итераций на каждой сетке обычно составляет 2—4; лишь на первой сетке число итераций оказывается заметно больше.

Контроль. Однако в многосеточном методе возможна одна неприятность. Задача обычно имеет не единственное собственное значение, а спектр $\lambda_k, k = 1, 2, \ldots$. Встречаются трудные задачи, в которых метод Ньютона на одной сетке сходится к сеточному решению с номером k_1, а на сгущенной сетке — к сеточному решению с совершенно другим номером k_2. В этом случае невозможно применять метод Ричардсона для оценки точности полученных сеточных решений. Для диагностики таких сбоев алгоритма целесообразно использовать визуальный контроль, выводя на дисплей графики численных решений на разных сетках. Если эти решения при сгущении сетки сходятся к предельной функции, то сбоя нет.

Визуальный контроль необходимо дополнять количественным контролем, обычным в методе Ричардсона. При последовательных сгущениях сетки вдвое нормы разности решений на сосед-

них сетках должны убывать в $\approx 2^p$ раз, где p — порядок аппроксимации схемы.

Замечание. Даже для линейных дифференциальных уравнений алгебраическая система всегда будет нелинейной хотя бы из-за члена λu. Поэтому метод дополненного вектора можно применять и к линейным дифференциальным уравнениям.

1.5.5. Другие методы

Хотя сеточный метод наиболее прост и универсален, для решения задач на собственные значения используют и другие методы.

Метод Галеркина применяют к линейным задачам. Рассмотрим его на примере струны (1.142). Выберем некоторый базис функций $\phi_m(x)$, $m = 1, 2, \ldots$. При этом базисные функции должны удовлетворять граничным условиям (1.142) $\phi_m(0) = 0$, $\phi_m(a) = 0$. Будем искать приближенное решение в виде обобщенного многочлена

$$\Phi(x) \approx \sum_{m=1}^{M} c_m \phi_m(x);$$

это выражение также удовлетворяет нулевым граничным условиям.

Подставим этот обобщенный многочлен в дифференциальное уравнение (1.142); полученное выражение будет не точно равняться нулю, а лишь приближенно. Потребуем, чтобы оно было ортогонально всем использованным $\phi_n(x)$:

$$\int_0^a \phi_n(x) \left\{ \frac{d}{dx}\left[q(x)\frac{d\Phi}{dx}\right] - r(x)\Phi(x) + \lambda\Phi(x) \right\} dx = 0,\ 1 \le n \le M.$$

Выполняя преобразования, получаем следующие выражения:

$$\sum_{m=1}^{M} (a_{nm} + \lambda b_{nm}) c_m = 0,\ 1 \le n \le M;$$

$$a_{nm} = \int_0^a \phi_n(x) \left\{ \frac{d}{dx}\left[q(x)\frac{d\phi_m}{dx}\right] - r(x)\phi_m(x) \right\} dx, \qquad (1.157)$$

$$b_{nm} = \int_0^a \phi_n(x)\phi_m(x)dx.$$

Первая строка системы (1.157) есть обобщенная задача на собственные значения для матрицы $A = (a_{nm})$ и $B = (b_{nm})$. Остальные строки определяют матричные элементы. Искомым решением являются собственное значение λ и соответствующий вектор коэффициентов $\mathbf{c} = (c_m)$.

Метод Галеркина может давать неплохие результаты, если удачно выбрана система базисных функций. Кроме того, матричные элементы должны достаточно хорошо вычисляться. Однако для такого выбора надо иметь хороший опыт расчетов. При формальном применении метода без учета этих соображений результаты обычно бывают неудовлетворительными.

В качестве базисных функций можно взять конечные элементы, например линейные B-сплайны. На этом принципе основаны многие пакеты прикладных программ.

Метод Ритца также включает выбор базисных функций и замену решения обобщенным многочленом. Однако, в отличие от метода Галеркина, необходимо найти функционал, который достигает минимума на точном решении. Последнее непросто сделать. Поэтому метод Ритца применять еще труднее, чем метод Галеркина.

Метод стрельбы основан на сведении к задаче Коши. Например, для струны (1.142) в силу однородности собственная функция определена с точностью до множителя. Поэтому величину производной $u_x(0)$ можно задать произвольно. Отбросим правое краевое условие и произвольно выберем некоторое значение λ. Тогда уравнение (1.142) с условиями $u(0) = 0$, $u_x(0) = \text{const}$ становится задачей Коши. Решим эту задачу до точки $x = a$. При этом мы наверняка получим $u(a) \neq 0$. Будем менять параметр λ и снова решать задачу Коши до тех пор, пока не добьемся выполнения условия $u(a) = 0$. Очевидно, окончательное λ будет некоторым собственным значением.

Ранее этот метод был широко распространен. Однако его трудно применять к уравнениям порядка выше второго, когда стрельба становится многопараметрической. Даже для уравнений второго порядка его почти невозможно применять, если задача Коши оказывается плохо обусловленной (именно к этому случаю относится уравнение Шредингера). Поэтому в настоящее время метод стрельбы применяют редко.

Фазовый метод. Пусть исходная задача имеет не одно собственное значение, а конечный или бесконечный спектр. В этом

случае нельзя заранее сказать, к какому именно собственному значению λ_k сойдутся все описанные выше численные методы. Для произвольной задачи это вообще вряд ли можно сделать. Однако для задачи струны (1.142) построен так называемый фазовый метод, позволяющий получить решение с заранее заданным k.

Метод основан на теоретическом знании качественного поведения точного решения. Точный спектр состоит из положительных λ_k, монотонно возрастающих с увеличением k от 1 до $+\infty$. Соответствующая λ_k собственная функция знакопеременна и имеет k полуволн. Это позволяет перейти к новым функциям — фазе $\phi(x)$ и амплитуде $\rho(x) > 0$. Они связаны с исходной функцией следующими соотношениями:

$$u(x) = \rho(x)\sin(\phi(x)),\ q(x)\frac{du}{dx} = \rho(x)\cos(\phi(x)). \tag{1.158}$$

Обращение $u(x)$ в нуль обусловлено только фазой. Условие наличия k полуволн означает, что граничные условия следует приписать только фазе; они имеют следующий вид:

$$\phi(0) = 0,\ \phi(a) = \pi k. \tag{1.159}$$

Эти условия выделяют единственное решение, соответствующее заранее заданному номеру k.

Далее для простоты ограничимся случаем $q(x) = \text{const} = 1$. Дифференцируя первое выражение (1.158) и сравнивая с вторым, получим следующее соотношение:

$$\rho'/\rho = \left(1 - \phi'\right)\operatorname{ctg}(\phi); \tag{1.160}$$

штрих означает дифференцирование по x. Подставляя (1.160) и (1.158) в уравнение струны (1.142), исключим из него ρ и получим уравнение только для фазы:

$$\phi' = \cos^2\phi + [\lambda - r(x)]\sin^2\phi. \tag{1.161}$$

Это уравнение с краевыми условиями (1.159) является задачей на собственные значения. Для заданного k эта задача имеет единственное решение. После решения этой задачи надо подставить найденное $\phi(x)$ в уравнение (1.160) и найти амплитуду $\rho(x)$ через квадратуры.

Замечания. 1. Поскольку (1.161) есть уравнение первого порядка, для него можно составить двухточечную разностную схему, интегрируя (1.161) на одном интервале по формуле трапеций. Это обеспечивает точность $O(h^2)$. На равномерной сетке с шагом $h = a/N$ схема имеет следующий вид:

$$2\left(\phi_n - \phi_{n-1}\right)/h = \cos^2\phi_n + \cos^2\phi_{n-1} + \\ + \left(\lambda - r_{n-1/2}\right)\left(\sin^2\phi_n + \sin^2\phi_{n-1}\right), \\ 1 \le n \le N; \phi_0 = 0,\ \phi_N = \pi k.$$

Задача для фазы нелинейная, так что ее целесообразно решать методом дополненного вектора.

2. Неограниченная область $a = \infty$ требует других граничных условий: $\phi(0) = 0$, $\rho(\infty) = 0$. Для этого случая приходится строить другой алгоритм.

3. Если $q(x)$ и $r(x)$ — кусочно гладкие функции, необходимо писать разностные уравнения на основе консервативных бикомпактных схем.

Глава 2

ТЕОРИЯ РАЗНОСТНЫХ СХЕМ

2.1. УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

2.1.1. Постановки задач

Движение систем малого числа частиц математически описывают обыкновенными дифференциальными уравнениями. Если число частиц очень велико, то следить за движением отдельных частиц практически невозможно. Удобнее рассматривать систему частиц как сплошную среду, характеризуя ее состояние средними величинами: плотностью, температурой в данной точке и т. д. Соотношения между ними описываются уравнениями в частных производных, интегральными или интегро–дифференциальными уравнениями.

К таким уравнениям приводят задачи газодинамики, теплопроводности, переноса излучения, распространения нейтронов, теории упругости, электромагнитных полей, процессов переноса в газах, квантовой механики и многие другие.

Независимыми переменными в физических задачах обычно являются время t и координаты $\mathbf{r}$; бывают и другие переменные, например скорости частиц $\mathbf{v}$ в задачах переноса.

Решение стационарной задачи ищется в пространственной области $G(\mathbf{r})$ с границей Γ; область может быть неограниченной. Нестационарные задачи обычно решаются в области $G(\mathbf{r}) \times$ $\times [0 \le t \le T]$. Для них возможны и другие границы области; например, для задач диффузии — сорбции область ограничена характеристиками уравнения диффузии.

Полная математическая постановка задачи включает в себя дифференциальное уравнение и дополнительные условия, позволяющие выделить единственное решение. В стационарных задачах дополнительные условия ставятся на границе Γ; такие задачи называют ***краевыми***. Для нестационарных задач в неограниченной области возможны задачи с ***начальными*** условиями

при $t = 0$. Их называют ***задачей Коши***. Нестационарные задачи в конечной области требуют начальных и краевых условий; их называют ***смешанными краевыми*** или ***начально-краевыми*** задачами.

Бóльшая часть этой книги посвящена простейшим уравнениям в частных производных. В этой главе изложены общие подходы к численному решению, применимые к любым типам уравнений. В последующих главах будут приведены конкретные реализации этих подходов для простейших типов уравнений.

Напомним классификацию простейших типов уравнений в частных производных. Для случая двух переменных они имеют следующий вид:

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + H = 0. \qquad (2.1)$$

Если коэффициенты зависят только от x и y, то уравнение (2.1) называется ***линейным***. Если коэффициенты зависят также от u, то уравнение называют ***квазилинейным***. При $A = B = C \equiv 0$ уравнение имеет первый порядок и называется ***уравнением переноса***. Уравнения второго порядка классифицируются по знаку дискриминанта $B^2 - AC$: у гиперболических уравнений дискриминант положителен, у параболических равен нулю, у эллиптических отрицателен. Уравнение с переменными коэффициентами может иметь разный тип в разных частях области G; в таких случаях довольно сложным является вопрос о корректности постановки задачи.

Физические процессы, которые описываются перечисленными здесь типами уравнений, существенно отличаются друг от друга. Соответственно полные постановки задач для этих типов уравнений имеют свои особенности, подробно рассмотренные далее.

2.1.2. Методы решения

Найти решение поставленной задачи в явном конечном виде удается лишь в исключительных случаях. Однако есть методы, позволяющие свести нахождение решения к задачам, изученным в кн. 1 или в гл. 1.

Точные методы. Пусть область прямоугольна, а коэффициенты, входящие в уравнения и граничные условия, постоянны. Тогда решение можно искать методом разделения переменных

в форме $u(x,y) = X(x)Y(y)$ или в виде суперпозиции таких решений. Для новых функций $X(x)$, $Y(y)$ при этом получаются задачи Коши или краевые задачи для обыкновенных дифференциальных уравнений. Решение последних задач нетрудно получить разложением в бесконечный ряд по гармоникам Фурье или экспонентам.

Метод разделения переменных можно успешно применять к любым типам линейных уравнений с постоянными коэффициентами. Однако число членов ряда, требуемое для получения высокой точности, может оказаться большим.

К таким задачам применяют также метод построения функции точечного источника. Эти методы применяют либо к уравнениям с постоянными коэффициентами, либо к квазилинейным уравнениям, коэффициенты которых зависят от функции u по степенному или иному несложному закону. В этих случаях нередко удается установить, что $u(x,y)$ по существу зависит не от двух переменных, а лишь от одной их комбинации. Такими комбинациями часто являются $\xi = x - cy$ или $\xi = x/y^{\alpha}$. Такие решения $u(\xi)$ называют ***автомодельными***. При этом уравнение в частных производных для $u(x,y)$ трансформируется в обыкновенное дифференциальное уравнение для $u(\xi)$. Последнее можно решить численными методами, приведенными в гл. 1.

Нередко удается найти такое преобразование ***подобия*** координат $(x,y) \to (\xi,\eta)$, при котором исходное уравнение в частных производных не меняется.

Пусть в этом случае известно некоторое частное решение уравнения $u(x,y)$. Применяя к его переменным это преобразование координат, получим $u(\xi,\eta)$, которое будет также частным решением исходного уравнения.

Разностные методы. Задачи для нелинейных уравнений с коэффициентами общего вида или даже линейные задачи, но в областях сложной формы, не удается решить классическими методами. Их решают численными методами. Чаще всего применяют *разностные методы* благодаря их универсальности.

Для этого в области $G(x,y)$ вводят некоторую сетку. Все производные, входящие в уравнение и краевые условия, заменяют разностями (или другими алгебраическими комбинациями) значений функции $u(\mathbf{r},\mathbf{t})$ в узлах сетки. Получающиеся алгебраические уравнения называют ***разностной схемой***. Решая эту ал-

гебраическую систему, находят приближенное (разностное) решение в узлах сетки.

Как и в гл. 1, возникают вопросы: существует ли решение алгебраической системы и единственно ли оно; как это решение фактически вычислить (за возможно меньшее число действий); при каких условиях это разностное решение стремится к точному и какова скорость сходимости? Есть еще два вопроса, которые для обыкновенных дифференциальных уравнений были несложными: как выбрать сетку и как составить разностную схему на этой сетке?

В данной главе рассмотрены общие способы составления и исследования разностных схем, применимые для разных типов задач. В следующих главах излагаются разностные схемы, которые дают хорошие результаты при решении наиболее распространенных типов уравнений математической физики.

Есть численные методы, близкие к разностным. В ***методе прямых*** сетка вводится только для пространственных переменных, а время t остается непрерывным. Производные по дискретным переменным заменяются разностями. При этом уравнение в частных производных аппроксимируется ***дифференциально-разностными уравнениями***, которые представляют собой систему большого числа обыкновенных дифференциальных уравнений. Метод прямых позволяет воспользоваться схемами интегрирования по времени, изложенными в гл. 1.

Метод конечных элементов использует разложение по пространственным базисным функциям на конечных носителях (обычно это линейные B-сплайны). Коэффициенты разложения находят методами Ритца или Галеркина.

Однако и метод конечных элементов, и метод прямых фактически также приводят к некоторым разностным схемам.

2.2. АППРОКСИМАЦИЯ

2.2.1. Сетка и шаблон

Одномерные задачи. Простейшими являются одномерные нестационарные задачи, в которых есть одна пространственная переменная x и временна́я t. Пространственная область при этом является отрезком $G(x) = [0 \le x \le a]$. Пространственно-временна́я область есть прямоугольник $[0 \le x \le a,\ 0 \le t \le T]$.

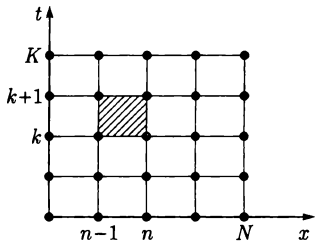


Рис. 2.1. Пример сетки и ячейки (нестационарная задача)

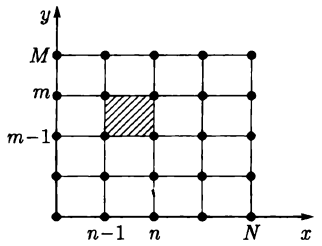


Рис. 2.2. Пример сетки и ячейки (стационарная задача)

Сетка в такой области строится естественно. Вводится пространственная сетка $\{x_n,\ 0 \le n \le N;\ x_0 = 0,\ x_N = a\}$ и временна́я сетка $\{t_k,\ 0 \le k \le K;\ t_0 = 0,\ t_K = T\}$; каждая сетка может быть неравномерной (рис. 2.1). Через указанные точки проводят прямые, параллельные осям x и t. Пересечения этих линий являются узлами пространственно-временно́й сетки $(x_n,\ t_k)$. Совокупность всех узлов сетки, лежащих на линии $t = t_k$, называют ***слоем***. Линию $t = t_0$ называют ***начальным*** слоем. Пространственные точки $x_0 = 0$ и $x_N = a$ называют ***граничными***, а точки x_n, $1 \le n \le N-1$, — ***внутренними***.

Пространственные шаги такой сетки обозначают через $h_n = x_n - x_{n-1}$, а временны́е — через $\tau_k = t_{k+1} - t_k$. Ячейкой такой сетки называют прямоугольник $[x_{n-1}, x_n; t_k, t_{k+1}]$. Он заштрихован на рис. 2.1. На равномерных сетках индекс шага опускают. Разностная схема может содержать величины из нескольких соседних ячеек или только из одной ячейки. Схемы, содержащие величины только из одной ячейки, называют ***бикомпактными***. У бикомпактных схем также нередко опускают индекс шага.

Стационарные задачи. Простейшей задачей, не содержащей времени, является двумерная стационарная краевая задача в области $G(x, y)$ с границей $\Gamma(G)$. Наиболее легким является случай, когда эта область есть прямоугольник $G(x, y) = [0 \le x \le a;\ 0 \le y \le b]$. В этом случае вводят одномерные сетки по каждой из переменных: $\{x_n,\ 0 \le n \le N\}$, $\{y_m,\ 0 \le m \le M\}$. Проводят линии $x = x_n$ и $y = y_m$; их называют ***направлениями***. Эти линии образуют прямоугольную сетку. Пересечение этих линий (x_n, y_m) являются узлами двумерной сетки (рис. 2.2). Шаги сетки по разным переменным обозначают h_x и h_y; на неравно-

мерных сетках при них надо ставить индекс интервала. Ячейка такой сетки заштрихована на рис. 2.2.

Во многих важных прикладных задачах граница $\Gamma(G)$ похожа на деформированный прямоугольник. К таким относится стационарное течение жидкости или газа в трубопроводах или соплах реактивных двигателей (рис. 2.3). В этом случае строят два семейства линий, похожих на соответствующие границы области. Точки пересечения этих линий берут узлами сетки. Такие сетки называют ***регулярными***. Каждая ячейка такой сетки похожа на деформированный прямоугольник. Обычно стараются построить такие семейства линий, чтобы они были взаимно ортогональными; при этом легче построить схемы высокой точности.

Если область G имеет криволинейную гладкую или кусочно-гладкую границу достаточно произвольной формы, то в ней трудно построить сетку с ячейками, похожими на прямоугольники. В этом случае размещают какое-то количество узлов на границы $\Gamma(G)$, причем в каждую точку излома границы обязательно ставят узел. Затем внутри области строят некоторую сетку с ячейками треугольной формы (рис. 2.4). Такие сетки далеко не всегда можно представить, как пересечение некоторых семейств линий. В разных узлах сетки может сходиться неодинаковое число ячеек. Подобные сетки называют ***нерегулярными***. Такие сетки получили широкое распространение в задачах гидро- и аэродинамики.

Многомерные задачи. Такие задачи могут быть стационарными и нестационарными. В нестационарных задачах пространственно-временна́я область является цилиндром $G(x, y) \times$ $\times [0 \le t \le T]$. В этом случае в области $G(x, y)$ выбирается одна

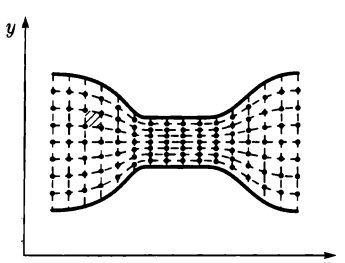


Рис. 2.3. Сетка в области с криволинейной границей

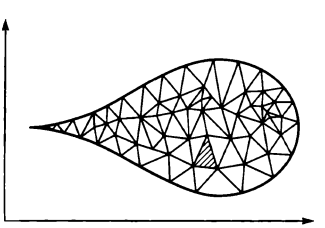


Рис. 2.4. Треугольная сетка

из двумерных сеток, описанных выше. По времени строят сетку $\{t_k\}$ как в простейших нестационарных задачах. Параллельно плоскости x, y проводят плоскости $t = t_k$. Лежащие на них узлы пространственной сетки называют, как и ранее, слоями. Пример такой сетки дан на рис. 2.5.

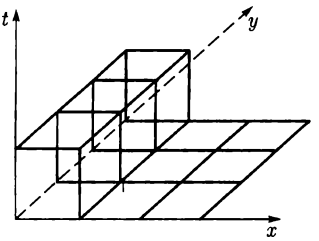


Рис. 2.5. Пример сетки для многомерной задачи

В трехмерных стационарных задачах поступают так же, как и в двумерных. Если область $G(x, y, z)$ является прямоугольным параллелепипедом, то в ней вводят ортогональную трехмерную сетку с плоскостями, параллельными граням параллелепипеда. Если область похожа на деформированный параллелепипед, стараются построить сетку из трех семейств взаимно ортогональных поверхностей. Если область ограничена произвольной криволинейной поверхностью, то строят на поверхности треугольную сетку, а область заполняют симплексами — фигурами, подобными тетраэдру.

Шаблон. Разностную схему получают, заменяя производные разностями сеточных значений функции, а также другими способами, описанными далее. Поэтому разностная схема есть алгебраическое уравнение, связывающее между собой значения сеточных функций в нескольких соседних узлах сетки. Простейшие примеры разностных схем были приведены в подразделе 1.4.

Внутри области надо заменять разностной схемой только само уравнение в частных производных. Для этого используют одну и ту же конфигурацию узлов, называемую ***шаблоном***. Эту конфигурацию можно отнести к некоторому узлу шаблона (обычно к центральному). Такой узел называют ***регулярным***.

В граничных узлах заменить уравнение в частных производных разностной схемой нельзя: в этих узлах нужно дополнительно учитывать граничное условие. Поэтому такие узлы зовут ***нерегулярными***, а стандартный шаблон в них приходится видоизменять. При некоторых краевых условиях не только граничные, но и близкие к граничным узлы могут оказаться нерегулярными.

Если шаблон разностной схемы содержит только узлы, являющиеся вершинами одной ячейки, то схема является бикомпактной. У такой схемы нет центрального узла, и следует говорить о

регулярности или нерегулярности ячейки. В этом случае разделение зависит от характера краевых условий. Если на границе заданы значения неизвестной функции, то обычно все ячейки являются регулярными. Если краевое условие содержит производные неизвестной функции, то нерегулярными оказываются ячейки, две вершины которых лежат на границе; если только одна вершина лежит на границе, то ячейка может оставаться регулярной.

2.2.2. Явные и неявные схемы

Обсудим вопрос о фактическом вычислении разностного решения. Бóльшая часть практических задач содержит время в качестве одной из переменных. Такие задачи содержат производную по времени. Следовательно, разностная схема должна включать значения с нескольких слоев. Если уравнение включает только первую производную du/dt, то разностная схема обычно является двуслойной, т. е. содержит только два слоя: исходный t_k и новый $t_k + \tau$. Если уравнение включает вторую производную d^2u/dt^2, то требуется по меньшей мере три слоя: к новому и исходному слоям добавляется предыдущий слой $t_k - \tau$. Для таких схем обычно используют следующие обозначения:

$$u(x_n, t_k) = u_n,\ u(x_n, t_k + \tau) = \hat{u}_n,\ u(x_n, t_k - \tau) = \check{u}_n. \qquad (2.2)$$

Схемы с бóльшим числом слоев не употребляют.

Для таких задач применяют послойный алгоритм решения: по известным значениям функции на исходном слое (или нескольких слоях) вычисляют значения функции на новом слое.

Разностные схемы при этом делят на явные и неявные. ***Явной*** называют схему, шаблон которой содержит только одну точку нового слоя. Это значение вычисляется по значениям с исходного слоя (или с предыдущих) за конечное число действий. Такие схемы очень просты в компьютерной реализации. Особенно полезны они на многопроцессорных компьютерах, так как расчет легко распараллеливается.

Схему называют ***неявной***, если ее шаблон содержит несколько точек нового слоя. Тогда для нахождения решения на новом слое возникает алгебраическая система большого числа уравнений. В одномерных нестационарных задачах структура такой системы проста и система легко решается (примеры будут рассмотрены далее). Но в многомерных задачах, решение таких

алгебраических систем нетривиально и требует разработки специальных алгоритмов. Неявные схемы гораздо хуже поддаются распараллеливанию, чем явные. Нетрудно видеть, что для стационарных задач все разностные схемы являются по существу неявными.

2.2.3. Составление схем

Существует несколько методов составления разностных схем. Проиллюстрируем их на простейшем примере — уравнении переноса:

$$\frac{\partial u}{\partial x} + c\frac{\partial u}{\partial t} = f(x,t),\ c = \mathtt{const} > 0. \tag{2.3}$$

Полную задачу пока не приводим, т. е. ограничиваемся только регулярными узлами (ячейками). Сетку выбираем согласно рис. 2.1. Будем строить двуслойную схему с использованием заштрихованной ячейки рисунка.

Метод разностной аппроксимации. Из вершин выбранной ячейки составим шаблон, позволяющий передать производные уравнения (2.3). Например, выберем шаблон (рис. 2.6). Вертикальный отрезок позволяет аппроксимировать $\partial u/\partial t$, а горизонтальный — $\partial u/\partial x$. Заменяя производные простейшими разностями и используя обозначения (2.2), получим следующую схему:

$$(\hat{u}_n - u_n)/\tau + c(\hat{u}_n - \hat{u}_{n-1})/h = \hat{f}_n; \tag{2.4}$$

правая часть для определенности отнесена к правому верхнему узлу шаблона. Индексы шагов опущены, поскольку схема бикомпактная. Полученная схема является неявной. Используя другой шаблон, нетрудно получить явную схему. Например, возьмем на рис. 2.6 две нижние точки и правую верхнюю. Заменяя пространственную производную разностью по нижнему отрезку, получим следующую схему:

$$(\hat{u}_n - u_n)/\tau + c(u_n - u_{n-1})/h = f_n. \tag{2.5}$$

Она содержит только одно значение на новом слое, т. е. является явной.

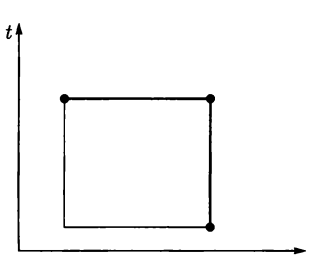


Рис. 2.6. Шаблон чисто неявной схемы

Описанный метод прост и в дополнительных пояснениях не нуждается.

Он легко позволяет составить схемы первого и второго порядков точности на прямоугольной сетке с непрерывными и достаточно гладкими коэффициентами. Однако этот метод трудно или даже невозможно применять в более сложных случаях: для уравнений с разрывными коэффициентами, на непрямоугольных сетках, для уравнений высокого порядка на неравномерных сетках и т. д.

Интегро-интерполяционный метод. Его называют также методом баланса. Этот метод применим к уравнениям с негладкими и даже разрывными коэффициентами. Особенно надежен он в том случае, когда с его помощью строят бикомпактную схему. Опишем этот случай.

Если уравнение в частных производных содержит производные выше первых, то заменим его эквивалентной системой уравнений с первыми производными (как это было сделано в п. 1.4.2). Сетку выберем так, чтобы все точки разрыва коэффициентов являлись узлами сетки (напомним, что такие сетки называют ***специальными***). Уравнение (2.3) уже имеет нужную форму. Проинтегрируем его по ячейке и точно возьмем интегралы от производных. Получим соотношение, являющееся интегральным законом сохранения:

$$\int\limits_{x_{n-1}}^{x_n} (\hat{u} - u)\, dx + c \int\limits_{t}^{\hat{t}} (u_n - u_{n-1})\, dt = \int\limits_{x_{n-1}}^{x_n} \int\limits_{t}^{\hat{t}} f dt dx.$$

Внутри ячейки все величины непрерывны и достаточно гладки, поэтому интегралы можно брать по любой квадратурной формуле. Снова используем шаблон рис. 2.6 с тремя черными точками. Беря все интегралы по формуле правых прямоугольников, опять получим разностную схему (2.4). Однако теперь мы знаем, что схема пригодна даже для разрывных коэффициентов.

В обоих методах были использованы несимметричные способы дифференцирования или интегрирования. Это позволяет надеяться лишь на первый порядок точности. Используя четырехточечный шаблон (все вершины ячейки) и симметричные формулы, можно в обоих методах получить второй порядок точности. Более сложными приемами можно добиться и более высокого порядка точности.

Краевые условия. На рассмотренном примере показано, как строить схему в регулярных узлах (ячейках). Вблизи гра-

ниц надо строить схему с учетом краевых условий. Если на границе задано значение искомой функции, вопрос является тривиальным. Если же граничные условия включают производную $\partial u/\partial x$, то для написания разностной схемы используют метод фиктивных точек (аналогично п. 1.4.2) или другие приемы.

2.2.4. Невязка

Невязка показывает согласованность разностной схемы с дифференциальным уравнением. Сначала введем некоторые обозначения. Для простоты обозначим совокупность временно́й и всех пространственных переменных одной буквой x. Пусть в области G задано уравнение

$$A\left(u\left(x\right)\right) = f\left(x\right),\ x \in G\left(x\right). \tag{2.6}$$

Здесь A — дифференциальный (интегральный, интегро-дифференциальный) оператор, в общем случае нелинейный; $u(x)$ — неизвестная функция; $f(x)$ — правая часть. Краевые и начальные условия, если они есть, предполагаются формально включенными в оператор A.

Введем в области $G(x)$ сетку $\omega_N = \{x_n,\ 0 \le n \le N\}$, имеющую шаг h. Построим на этой сетке разностную схему, которую обозначим следующим образом:

$$B(v(x)) = \phi(x),\ x \in \omega_N. \tag{2.7}$$

Здесь B — разностный оператор, который также может быть нелинейным. Сеточное (численное) решение $v(x_n) \equiv v_n$ и видоизмененная правая часть $\phi(x_n) \equiv \phi_n$ определены только на сетке ω_N.

Как сравнить уравнение (2.6) со схемой (2.7)? Нельзя применять к сеточному решению оператор A, определенный для всех $x \in G$. Зато точное решение $u(x)$ определено для всех значений x, в том числе и на сетке ω_N, поэтому к нему можно применить разностный оператор B. Выберем в качестве рассогласованности уравнения (2.6) со схемой (2.7) величину, называемую ***невязкой***:

$$\begin{aligned}\psi\left(x\right) = \left[B\left(u(x)\right) - \phi(x)\right] - \left[A\left(u(x)\right) - f(x)\right] \equiv \\ \equiv B\left(u(x)\right) - \phi(x),\ x \in \omega_N;\end{aligned} \tag{2.8}$$

вторая квадратная скобка равна нулю в силу уравнения (2.6), так что в правой части можно оставить только первую квадратную скобку.

Нахождение невязки. Если все коэффициенты уравнения и само точное решение являются достаточно гладкими функциями, то невязку можно найти разложением в ряд Тейлора. Например, найдем невязку явной схемы (2.4) для уравнения переноса (2.3). Выберем в качестве центра разложения вершину $(x_n, \hat{t})$ шаблона рис. 2.6 и запишем разложение с учетом вторых производных:

$$u_{n-1} \approx u_n - hu_x + \frac{h^2}{2}u_{xx},\ \hat{u}_n \approx u_n + \tau u_t + \frac{\tau^2}{2}u_{tt};$$

подразумевается, что все производные взяты в центре разложения, как и правая часть схемы (2.4). Подставим эти разложения в схему (2.4). Первые производные и правая часть сократятся в силу уравнения (2.3), и мы получим невязку

$$\psi(x,t) \approx \frac{\tau}{2}u_{tt} - \frac{h}{2}cu_{xx} = O(\tau + h) \tag{2.9}$$

с точностью до членов второго порядка. Полученная невязка имеет первый порядок малости и стремится к нулю при τ, h, стремящихся к нулю.

Если в рассмотренном примере точное решение не имеет вторых непрерывных производных, то определить невязку указанным способом нельзя. Вообще определение невязки в случае недостаточно гладких решений является сложной задачей. В этом случае рекомендуется исследовать интегральную форму уравнения, для которой легче получить оценки при негладких решениях.

Замечания. ***1.*** По определению, невязка оценивается на решении точной задачи $u(x,t)$. На практике это решение нам неизвестно. Поэтому можно сделать оценку лишь на некотором классе функций U, к которому должно принадлежать точное решение (например, на классе дважды или более непрерывно дифференцируемых функций). Такая оценка является мажорантной и, как правило, сильно завышенной.

2. Попробуем в нашем примере учесть, что $u(x,t)$ является решением точной задачи (2.3). Дифференцируя (2.3) по t или x, получим: $u_{tt} = -cu_{xt} + f_t$ и $u_{xt} = -cu_{xx} + f_x$. Подставляя эти выражения в (2.9), уточним невязку:

$$\psi(x,t) \approx \frac{c}{2}(c\tau - h)u_{xx} + \frac{\tau}{2}(f_t - cf_x) \tag{2.10}$$

Если мы выберем соотношение шагов $c\tau = h$, то первое слагаемое невязки исчезает. Второе слагаемое не исчезает, однако оно содержит только известную функцию f, поэтому его можно использовать как поправочный член. Вычитая его из правой части схемы (2.4), получим следующую схему:

$$(\hat{u}_n - u_n)/\tau + c(u_n - u_{n-1})/h = f_n - \frac{\tau}{2}(f_t - cf_x). \qquad (2.11)$$

Если выбрать в расчете шаги $c\tau = h$, то в невязке схемы (2.11) исчезнет член $O(\tau+h)$ и останутся лишь члены второго порядка малости. Схема будет иметь повышенную точность.

2.2.5. Аппроксимация

Пусть в области $G(x)$ с границей $\Gamma(G)$ задано дифференциальное уравнение $A(u(x)) = f(x)$. Полная постановка задачи включает задание граничных и начальных условий; но будем считать их формально включенными в оператор A. Введем в области сетку ω_N и запишем на этой сетке разностную схему $B(v) = \phi(x)$. Эта схема в регулярных узлах (ячейках) приближает исходное уравнение, а в нерегулярных — краевые условия. Близость разностной схемы к исходному уравнению будем определять по величине невязки (2.8). Введем следующие определения.

Определение 2.1. *Разностная схема аппроксимирует дифференциальное уравнение*, если

$$\|\psi\| \to 0 \text{ при } h \to 0. \ \bullet \qquad (2.12)$$

Определение 2.2. *Аппроксимация имеет порядок* p, если

$$\|\psi\| = O(h^p) \text{ при } h \to 0. \ \bullet \qquad (2.13)$$

Невязка есть сеточная функция. Поэтому нормы в данных определениях должны быть сеточными. Однако существует много различных норм. Обычно мы будем использовать одну из следующих двух норм:

$$\|\psi\|_c = \max_{0\le n\le N} |\psi_n|, \ \|\psi\|_{l_2} = \left(\sum_n h_n\psi_n^2 / \sum_n h_n\right)^{1/2}. \qquad (2.14)$$

Первая норма является сеточным аналогом чебышёвской нормы C в пространстве непрерывных функций; аппроксимацию в

ней называют ***локальной***. Вторая норма — сеточный аналог гильбертовой нормы L_2; аппроксимацию в ней называют ***среднеквадратичной***.

Нетрудно видеть, что $\|\psi\|_c \geq \|\psi\|_{l_2}$, т. е. сеточная чебышёвская норма является более сильной, чем сеточная гильбертова. Из аппроксимации в первой из этих норм следует аппроксимация во второй норме; но из аппроксимации во второй норме, вообще говоря, аппроксимация в первой норме не следует. Поэтому стараются доказать аппроксимацию в возможно более сильной норме.

Замечание. Невязка определяется на решении исходной задачи $u(x)$. Однако это решение неизвестно. Поэтому использовать его в нахождении невязки невозможно. В этом случае берут достаточно широкий класс U функций, которому $u(x)$ заведомо принадлежит (обычно это класс функций, непрерывных вместе с достаточным числом своих производных). Если на всех функциях класса U имеется аппроксимация порядка p, то аппроксимация на точном решении $u(x)$ имеет порядок не ниже p.

В подобных случаях аппроксимация на точных решениях может быть выше p. Например, рассмотрим уравнение переноса (2.3) и уточненную разностную схему (2.11). При $h = c\tau$ схема (2.11) аппроксимирует уравнение (2.3) на произвольных дважды непрерывно дифференцируемых функциях с порядком $p = 1$. Однако на точном решении уравнения (2.3) аппроксимация имеет более высокий порядок $p = 2$.

Многомерность. Многомерные сетки имеют шаги по разным переменным (например, h и τ). В определении аппроксимации надо требовать одновременного стремления шагов к нулю: $h \to 0$, $\tau \to 0$. Порядок аппроксимации может быть неодинаковым по разным переменным. Например, соотношение $\|\psi\| = O(\tau^p + h^q)$ означает аппроксимацию порядка p по переменной t и порядка q по переменной x. Например, для схемы (2.3) невязка (2.9) есть $O(\tau + h)$ и схема имеет первый порядок аппроксимации по каждой переменной. При этом $\|\psi\| \to 0$ при любых законах стремления τ и h к нулю; такую аппроксимацию называют ***безусловной***.

Иногда встречаются схемы, для которых норма невязки есть $O(\tau^p + h^q + \tau^r/h^s)$. В этом случае для $\|\psi\| \to 0$ недостаточно требовать просто $\tau \to 0$, $h \to 0$; необходимо выполнение дополнительного условия $\tau^r/h^s \to 0$. Такую аппроксимацию назы-

вают ***условной***. Контролировать выполнение дополнительного условия на практике обычно затруднительно. Поэтому схемы с условной аппроксимацией стараются не использовать без особой необходимости.

2.3. УСТОЙЧИВОСТЬ

2.3.1. Неустойчивость

Для некоторых разностных схем малые ошибки, допущенные на каком-либо этапе вычисления решения, при дальнейших выкладках сильно возрастают. При этом не только катастрофически снижается точность, но расчет может «развалиться» — возникают большие числа, не представимые на компьютере.

В кн. 1 рассматривались задачи численного дифференцирования и суммирования рядов Фурье, где малые ошибки входных данных приводили к большим ошибкам решения. Теперь рассмотрим пример уравнения переносы (2.3) без правой части: $f(x,t) \equiv 0$. Напишем явную разностную схему:

$$(\hat{u}_n - u_n)/\tau + c\,(u_n - u_{n-1})/h = 0. \qquad (2.15)$$

Разрешим его относительно значения на новом слое:

$$\hat{u}_n = (1-\kappa)\,u_n + \kappa u_{n-1},\ \kappa = c\tau/h.$$

Внесем в значение u_{n-1} малую ошибку δu_{n-1}. Тогда значение $\hat{u}_n$ также будет найдено с ошибкой: $\delta\hat{u}_n = \kappa\delta u_{n-1}$. Таким образом, при переходе на новый слой ошибка изменилась в κ раз.

Пусть расчет проводится до момента T. Тогда он содержит $K = T/\tau$ слоев. При переходе с нулевого слоя на последний ошибка изменится в κ^K раз. При $\kappa < 1$ это будет убывание ошибки, а при $\kappa > 1$ — возрастание.

Выберем такое отношение шагов τ и h, чтобы получить $\kappa > 1$. Будем одновременно сгущать сетки по τ и h, сохраняя указанное отношение. Тогда $K \to \infty$ и ошибка на последнем слое $\kappa^K \to \infty$.

Таким образом, если $\kappa > 1$, то при сгущении сеток расчет не стремится к точному решению. Вместо этого происходит неограниченное нарастание малых начальных ошибок. Это явление называют ***неустойчивостью***. При наличии неустойчивости численный расчет «разваливается»: ошибки быстро нарастают и становятся много больше самого решения.

2.3.2. Основные понятия

Рассмотрим на сетке ω_N произвольную (нелинейную) разностную схему (2.7) $B(v(x)) = \phi(x)$. Какими свойствами должна обладать схема, чтобы в ней не возникала неустойчивость, подобно описанной в п. 2.3.1? Для этого внесем в правую часть схемы ошибку $\delta\phi(x)$ и найдем сеточное решение с соответственно измененной правой частью:

$$B\left(v(x) + \delta v(x)\right) = \phi(x) + \delta\phi(x). \tag{2.16}$$

Оценим зависимость погрешности решения $\delta v(x)$ от погрешности правой части $\delta\phi(x)$.

Определение 2.3. Схема (2.7) называется *устойчивой*, если для сколь угодно малого $\epsilon > 0$ найдется такое δ, не зависящее от шага h, что если

$$\|\delta\phi\| < \delta, \text{ то } \|\delta v\| < \epsilon. \bullet \tag{2.17}$$

Это означает, что $\delta v(x)$ непрерывно зависит от $\delta\phi(x)$, причем эта зависимость равномерно ограничена по шагу h. Данное определение применимо для любой нелинейной схемы. Заметим также, что в этом определении не упоминается никакое дифференциальное уравнение, т. е. устойчивость является внутренним свойством только самой разностной схемы.

В определении могут использоваться различные нормы (чебышёвская, гильбертова и др.). Кроме того, для $v(x)$ и $\phi(x)$ могут использоваться неодинаковые нормы. Дальше мы встретимся с примерами разностных схем, устойчивых при одном выборе норм и неустойчивых — при другом.

Линейность. Пусть разностная схема линейна: $Bv(x) = \phi(x)$. Тогда $v(x)$ линейно зависит от $\phi(x)$ и то же относится к их вариациям. Поэтому для линейных схем определение устойчивости примет следующий вид:

$$\|\delta v\| \leq M \|\delta\phi\|, \tag{2.18}$$

где M — константа, не зависящая от шага h. Заметим, что эффективные методы исследования устойчивости пока разработаны только для линейных схем. Поэтому в практике обычно используется именно последнее определение.

Многомерность. Если независимых переменных несколько, то вводят понятия условной и безусловной устойчивости.

Устойчивость называется ***безусловной***, если (2.17) или (2.18) выполняется при произвольном соотношении шагов по различным переменным, лишь бы они были достаточно малы. Если для выполнения (2.17) или (2.18) шаги по разным переменным должны удовлетворять дополнительным соотношениям, то устойчивость называется ***условной***. Например, для уравнения переноса (2.3) ранее написаны неявная разностная схема (2.4) и явная разностная схема (2.15). В явной схеме ошибка неограниченно нарастала при $\kappa = c\tau/h > 1$ и не нарастала при $\kappa < 1$; тем самым эта схема оказалась условно устойчивой. Аналогичными рассуждениями можно показать, что для неявной схемы ошибка не нарастает при любом значении τ/h, т. е. неявная схема (2.4) безусловно устойчива.

2.3.3. Признаки устойчивости

В оператор B в (2.16) включались различные разностные уравнения, аппроксимирующие как дифференциальные уравнения, так и граничные или начальные условия. Зачастую бывает удобно разделять эти типы разностных уравнений и отдельно исследовать устойчивость для регулярного разностного оператора (ее называют устойчивостью по правой части), устойчивость по краевым условиям и по начальным данным. Это позволяет сформулировать удобные критерии устойчивости.

Устойчивость по начальным данным. Все простейшие типы уравнений, кроме эллиптического, в качестве одной из переменных содержат время. Для таких уравнений ставится эволюционная задача — смешанная задача Коши. Даже эллиптические уравнения нередко численно решают посредством счета на установление, т. е. при помощи постановки вспомогательной задачи Коши. Поэтому исследованию устойчивости эволюционных задач уделяют особое внимание.

Уравнение любого порядка можно свести к системе уравнений первого порядка. Для такой системы удобно строить схемы, содержащие только два слоя по времени — исходный и новый. Такие схемы называют ***двуслойными***. Для двуслойных схем решение смешанной задачи Коши на некотором слое t^* можно рассматривать как начальные данные для всех последующих слоев.

Для простоты ограничимся далее случаем линейных схем. Пусть $v_1(t)$ и $v_2(t)$ суть два решения схемы (2.7), соответству-

ющие разным начальным данным и одной и той же правой части. Схему называют ***равномерно устойчивой*** по начальным данным, если для $t^* < t < T$ выполняется

$$\|v_1(t) - v_2(t)\| \le K \|v_1(t^*) - v_2(t^*)\|, \tag{2.19}$$

где $K = \mathtt{const}$ не зависит от t^* и шага τ. Очевидно, из равномерной устойчивости по начальным данным следует обычная устойчивость по начальным данным (но не наоборот). Для двухслойной схемы справедлив следующий признак устойчивости.

Теорема 2.1. Пусть значения на исходном t и новом $\hat{t}$ слоях подчиняются соотношению

$$\|\hat{v}_1 - \hat{v}_2\| \le (1 + c\tau) \|v_1 - v_2\|, \; c \ge 0; \tag{2.20}$$

здесь $c = \mathtt{const}$ не зависит от t и τ. Тогда схема равномерно устойчива по начальным данным. •

Доказательство. За один шаг τ разность (2.20) возрастает не более чем в $1+c\tau$ раз. Промежуток $t-t^* < T$ и содержит не более T/τ шагов. Поэтому в данном промежутке решение возрастает не сильнее чем в

$$K = (1 + c\tau)^{T/\tau} \le \exp(cT) \tag{2.21}$$

раз. Поэтому условие (2.19) выполнено.

Из (2.20) видно, что если константа c велика, то, хотя схема устойчива, фактическая ошибка может существенно возрасти в ходе расчета, т. е. схема ***плохо обусловлена***. Чем больше промежуток T, тем меньшее значение c обеспечивает хорошую обусловленность. При $T \to \infty$ только $c = 0$ обеспечивает устойчивость.

Если разностное решение $v(t)$ сильно возрастает или убывает по времени, целесообразно классифицировать устойчивость не по абсолютной ошибке, а по относительной: $\|\delta v(t)\| \,/\, \|v(t)\|$. Пусть эта относительная погрешность возрастает за шаг не более чем в $1 + c\tau$ раз. Тогда схему называют плохо обусловленной при $cT \gg 1$, хорошо обусловленной при небольших cT и ***асимптотически*** (т. е. при $T \to \infty$) устойчивой при $c = 0$.

Устойчивость по правой части. Для нестационарных задач и двухслойных разностных схем она опирается на признак равномерной устойчивости по начальным данным (2.20). Рассмотрим схему $Bv = \phi$ с разными правыми частями ϕ_1 и ϕ_2; им

соответствуют решения v_1 и v_2. Справедлива следующая теорема.

Теорема 2.2. Пусть для схемы выполнен признак (2.20). Пусть также для любых двух решений, совпадающих на исходном слое $v_1(t) = v_2(t)$, на новом слое выполняется соотношение

$$\left\|v_1(\hat{t}) - v_2(\hat{t})\right\| \le \alpha\tau \left\|\phi_1 - \phi_2\right\|, \ \alpha \ge 0; \tag{2.22}$$

здесь $\alpha = \texttt{const}$ не зависит от t и τ. Тогда схема устойчива по правой части. •

Доказательство (нестрогое, которое нетрудно превратить в строгое). Согласно (2.22), погрешность правой части $\delta\phi$ за один шаг приводит к погрешности решения $\|\delta v\| \le \alpha\tau \|\delta\phi\|$. Но для всех последующих шагов погрешность $\delta v(\hat{t})$ может рассматриваться как погрешность начальных данных в момент $\hat{t}$. Согласно (2.20), на каждом шаге она умножается на $1 + c\tau$, а суммарный множитель к моменту T не превысит $\exp[c(T-t)]$. Поэтому вклад $\delta\phi(t)$ в конечную погрешность не превышает величины $\alpha\tau \exp[c(T-t)] \|\delta\phi\|$.

Просуммируем эти вклады по всем моментам $0 \le t \le T$. Учитывая, что число членов суммы равно T/τ, и приближенно заменяя сумму интегралом, получим окончательную оценку:

$$\|\delta v\| \le M \|\delta\phi\|, \ M = \alpha \frac{\exp(cT) - 1}{c}, \ \alpha > 0, \ c \ge 0; \tag{2.23}$$

при $c \to 0$ константа $M \to \alpha T$. Полученная оценка соответствует (2.18), т. е. означает устойчивость по правой части. ■

Формулировка теоремы включает равномерную устойчивость по начальным данным. Поэтому из теоремы следует одновременно устойчивость и по начальным данным, и по правой части. Устойчивость по краевым условиям можно проверять аналогично устойчивости по правой части; однако на практике для краевых условий заметно труднее доказать выполнение неравенства типа (2.22).

2.3.4. Метод гармоник

Ранее рассматривались общие признаки устойчивости; при этом не конкретизировались используемые нормы и не указывалось, какой математической техникой можно устанавливать тре-

буемые неравенства. Далее рассмотрим основные методы исследования устойчивости. Метод гармоник наиболее прост и применим к очень широкому кругу задач. Принцип максимума также прост, но применим далеко не ко всем схемам. Метод операторных неравенств дает наиболее сильные результаты, но требует высокой математической квалификации исследователя.

Метод гармоник применяется для строгого обоснования многих линейных схем и нестрогого, но плодотворного исследования большинства нелинейных задач. В нем используют норму $\|.\|_{l2}$ как для решения, так и для правых частей.

Простейший прием. Рассмотрим применение метода к линейным двуслойным схемам с постоянными коэффициентами на равномерной сетке. Запишем их в канонической форме:

$$B\frac{\hat{v}-v}{\tau}+Av=\phi, \tag{2.24}$$

где B, A — некоторые разностные операторы, действующие на v как на функцию пространственной переменной. Исследуя устойчивость по начальным данным, фиксируют правую часть ϕ. Тогда погрешность δv удовлетворяет однородному уравнению

$$B\left(\delta\hat{v}-\delta v\right)+\tau A\delta v=0. \tag{2.25}$$

Разложим $\delta v(x_n)$ в ряд Фурье по пространственным гармоникам $\exp\left(iqx_n\right)$. При переходе на новый слой q-я гармоника умножается на множитель роста ρ_q. Тогда в схему (2.25) надо подставить $\delta v(x_n)=\exp\left(iqx_n\right)$ и $\delta\hat{v}(x_n)=\rho_q\delta v(x_n)$. Получим уравнение для определения ρ_q:

$$\left(\rho_q-1\right)B\exp\left(iqx\right)+\tau A\exp\left(iqx\right)=0. \tag{2.26}$$

Когда многоточечные операторы A и B действуют на гармонику, получаются выражения, содержащие центральную точку шаблона x_n и соседние точки ($x_{n-1}=x_n-h, x_{n+1}=x_n+h$ и др.). Тогда в полученных выражениях можно сократить множитель $\exp(iqx_n)$. Окончательную формулу можно получить, делая в (2.26) следующую формальную замену:

$$\exp(iqx_n)\to 1,\ \exp(iqx_{n\pm1})\to\exp(\pm iqh)\text{ и т. д.} \tag{2.27}$$

Отсюда нетрудно выразить ρ_q через шаги τ и h и коэффициенты схемы.

Теорема 2.3. Для равномерной устойчивости схемы (2.24) по начальным данным необходимо и достаточно, чтобы для любых q выполнялось условие

$$|\rho_q| \leq 1 + c\tau, \; c \geq 0; \tag{2.28}$$

здесь константа c не зависит от q и шагов h и τ. •

Доказательство. ***Необходимость.*** Пусть хотя бы для одной q-й гармоники условие (2.28) не выполняется. Это означает, что сколь бы большим ни было c, для этой гармоники неравенство (2.28) будет иметь противоположное направление. Внесем на произвольном слое t возмущение в виде этой гармоники. Тогда к моменту T амплитуда этого возмущения возрастет более чем в

$$(1 + c\tau)^{(T-t)/\tau} \approx \exp\left[c\,(T - t)\right] \tag{2.29}$$

раз, где c сколь угодно велико. Это означает неустойчивость по начальным данным.

Достаточность. Разложим возмущение на некотором слое в ряд по гармоникам: $\delta v(t) = \sum\limits_q \alpha_q(t) \exp(iqx)$. На момент T возмущение станет равно $\delta v(T) = \sum\limits_q \alpha_q(T) \exp(iqx)$. Поскольку схема линейна, увеличение амплитуды каждой гармоники не зависит от других гармоник и мажорируется множителем (2.29), который одинаков для всех гармоник. Гармоники с различными q ортогональны. Норму L_2 за счет множителя перед интегралом можно определить так, чтобы норма $\|\exp(iqx)\|_{L_2}$ равнялась 1. Тогда для роста нормы возмущения получаем неравенство:

$$\begin{aligned} \|\delta v(T)\|_{L_2}^2 = \sum_q |\alpha_q(T)|^2 \leq e^{2c(T-t)} \sum_q |\alpha_q(t)|^2 = \\ = e^{2c(T-t)} \|\delta v(t)\|_{L_2}^2 . \end{aligned} \tag{2.30}$$

Это означает равномерную устойчивость по начальным данным в норме L_2. ■

Приведенное доказательство является строгим лишь при довольно жестких дополнительных ограничениях: дифференциальное уравнение линейное с постоянными коэффициентами, а задача поставлена на прямой $-\infty < x < \infty$, т. е. краевые усло-

вия отсутствуют. При этом используется не ряд Фурье, а интеграл Фурье. Однако справедлива аналогичная строгая теорема для задачи на ограниченном отрезке с краевыми условиями. В этом случае вместо гармоник нужно брать собственные функции разностного оператора, удовлетворяющие разностным краевым условиям. Их число конечно, а сами они определены только в узлах. Поэтому сходимость устанавливается в сеточной норме l_2. Однако найти такую систему разностных функций нелегко, поэтому строгий метод употребляется лишь для простейших задач.

Замораживание. Для уравнений с переменными коэффициентами доказанная теорема об устойчивости неприменима. Однако широко используют следующий прием. «Замораживают» коэффициенты уравнения, т. е. считают их постоянными, равными их значениям в какой-то выбранной точке x. Проверяют выполнимость условия теоремы 2.3. Если оно выполняется для значения коэффициентов в любой точке x, то схему считают устойчивой. Практика показывает, что в большинстве случаев расчет по этой схеме действительно оказывается устойчивым, т. е. при стремлении шагов к нулю не разваливается, а сходится к некоторой предельной функции.

Нелинейность. Только простейшие задачи описываются линейными уравнениями и линейными схемами. Рассмотрим нелинейную задачу, приведенную к форме, содержащую первую производную по времени. Напишем для нее неявную нелинейную двуслойную разностную схему:

$$(\hat{v} - v)/\tau = B(\hat{v}, v) + \phi(x, t),\ v = v(x, t). \tag{2.31}$$

Рассмотрим малые вариации решения δv, $\delta\hat{v}$. Разложим (2.31) в ряд Тейлора по вариациям, пренебрегая квадратичными членами. Получим для вариаций следующее уравнение:

$$\frac{\delta\hat{v} - \delta v}{\tau} = \frac{\partial B(\hat{v}, v)}{\partial\hat{v}}\delta\hat{v} + \frac{\partial B(\hat{v}, v)}{\partial v}\delta v. \tag{2.32}$$

Здесь производные от B по $\hat{v}$ и v суть матрицы Якоби. Уравнение (2.32) можно рассматривать как линейное уравнение с переменными коэффициентами для возмущения. Однако теперь коэффициенты уравнения зависят не только от x, но и от самого сеточного решения.

К уравнению (2.32) применяют метод замороженных коэффициентов и метод гармоник. Если для всех гармоник при разумных значениях неизвестного сеточного решения v выполняется условие устойчивости (2.28), то за один шаг возмущение нарастает слабо. Если при этом значение cT невелико, то и на всем протяжении расчета малое возмущение не сильно возрастает, т. е. остается малым. Тогда квадратом возмущения можно пренебрегать и считать линеаризованное уравнение (2.32) достаточно точным. В этом случае естественно ожидать, что нелинейная схема (2.31) устойчива. Эти рассуждения нестроги, но правдоподобны. На практике они позволяют уверенно опознать неустойчивую или условно устойчивую схему и решить, стоит ли эту схему программировать и тестировать. Если же условие устойчивости выполняется, причем с небольшим значением cT, то предсказанная таким образом устойчивость обычно подтверждается на практике (т. е. схему целесообразно программировать). Случай большого cT означает плохую обусловленность схемы; расчеты по такой схеме могут давать «срывы».

Метод замороженных коэффициентов и метод линеаризации широко используют на практике, поскольку математические выкладки при этом несложны. Однако построены примеры, в которых эти методы предсказывают устойчивость, а запрограммированная разностная схема оказывается неустойчивой. Поэтому эти методы применяют в первую очередь для отбраковки неудачных схем. Перспективную же схему программируют и тестируют.

2.3.5. Принцип максимума

Принцип максимума является строгим также лишь для линейных задач. Этим методом удается доказывать сходимость схем лишь первого порядка точности по времени. Зато при этом сходимость устанавливается в норме C, которая сильнее L_2. Рассмотрим его применение к двуслойной неявной разностной схеме. Запишем эту схему в следующей канонической форме:

$$\sum_k \alpha_k \hat{v}_{n+k} = \sum_m \beta_m v_{n+m} + \phi_n. \qquad (2.33)$$

Здесь n — центральный узел шаблона, а суммирование по k и m проводится по всем узлам шаблона. Центральным считаем тот

узел шаблона, в котором коэффициент на верхнем слое максимален по модулю:

$$|\alpha_0| \geq |\alpha_k| \tag{2.34}$$

при любых k. Тогда справедливы следующие теоремы.

Теорема 2.4. Схема (2.33) равномерно устойчива по начальным данным, если выполнено условие

$$(1 + c\tau)\,|\alpha_0| \geq \sum_{k\neq 0} |\alpha_k| + \sum_m |\beta_m|, \tag{2.35}$$

где $c = \mathtt{const} \geq 0$ не зависит от h и τ. •

Теорема 2.5. Если выполнены условия (2.35) и

$$|\alpha_0| - \sum_{k\neq 0} |\alpha_k| \geq \frac{\gamma}{\tau},\ \gamma = \mathtt{const} > 0, \tag{2.36}$$

то схема (2.33) устойчива по правой части. •

Доказательство. 1. Фиксируем правую часть (2.33) и внесем ошибку δv на исходном слое. Тогда ошибка $\delta\hat{v}$ удовлетворяет уравнению

$$\sum_k \alpha_k \delta\hat{v}_{n+k} = \sum_m \beta_m \delta v_{n+m}.$$

Отсюда для любого узла n следует обобщенное неравенство треугольника:

$$|\alpha_0|\,|\delta\hat{v}_n| \leq \sum_{k\neq 0} |\alpha_k|\,|\delta\hat{v}_{n+k}| + \sum_m |\beta_m|\,|\delta v_{n+m}|.$$

Применим это неравенство к тому узлу n, в котором $|\delta\hat{v}_n|$ достигает максимума; при этом в правой части заменим $|\delta\hat{v}_{n+k}|$ и $|\delta v_{n+m}|$ их максимальными значениями (т. е. нормами c), что только усилит неравенство. Тогда получим

$$|\alpha_0|\,\|\delta\hat{v}\|_c \leq \|\delta\hat{v}\|_c \sum_{k\neq 0} |\alpha_k| + \|\delta v\|_c \sum_m |\beta_m|.$$

Перепишем это неравенство в следующей форме:

$$\left(|\alpha_0| - \sum_{k\neq 0} |\alpha_k|\right) \|\delta\hat{v}\|_c \leq \left(\sum_m |\beta_m|\right) \|\delta v\|_c.$$

Но в силу неравенства (2.35)

$$\sum_m |\beta_m| \le (1+c\tau)\,|\alpha_0| - \sum_{k\neq 0} |\alpha_k| \le$$

$$\le (1+c\tau)\left(|\alpha_0| - \sum_{k\neq 0} |\alpha_k|\right).$$

Поэтому $\|\delta\hat{v}\|_c \le (1+c\tau)\,\|\delta v\|_c$, т. е. выполнен признак (2.20) равномерной устойчивости по начальным данным, причем в норме c.

2. Зафиксируем в (2.33) решение на исходном слое и внесем в правую часть погрешность $\delta\phi$. Тогда погрешность решения на новом слое удовлетворяет уравнению

$$\sum_k \alpha_k \delta\hat{v}_{n+k} = \delta\phi_n.$$

Отсюда следует обобщенное неравенство треугольника:

$$|\alpha_0|\,|\delta\hat{y}_n| \le \sum_{k\neq 0} |\alpha_k|\,|\delta\hat{y}_{n+k}| + |\delta\phi_n|.$$

Опять выберем узел n, в котором $|\delta\hat{v}_n|$ достигает максимума, и заменим справа все локальные погрешности их максимумами (нормами c). Тогда получим, что

$$\|\delta\hat{v}\|_c \left(|\alpha_0| - \sum_{k\neq 0} |\alpha_k|\right) \le \|\delta\phi\|_c.$$

Отсюда с учетом (2.36) следует, что

$$\|\delta\hat{v}\|_c \le \frac{\tau}{\gamma}\,\|\delta\phi\|_c.$$

Это означает выполнение признака устойчивости (2.22) по правой части, причем в норме c. ■

Замечания. ***1.*** Доказательство является строгим для уравнений с переменными коэффициентами, зависящими от x и t.

2. Краевые условия двуслойных линейных схем также имеют каноническую форму (2.33), поэтому данные теоремы позволяют устанавливать устойчивость по краевым условиям.

3. Принцип максимума дает достаточное условие устойчивости; невыполнение критериев (2.35) и (2.36) еще не означает неустойчивость схемы.

Принцип максимума в приведенной здесь форме применяют к уравнению переноса и параболическим уравнениям. Для эллиптических уравнений используется другая форма принципа максимума; ее не будем рассматривать. Для нелинейных схем можно применять принцип максимума, линеаризуя схему и приводя ее к виду (2.32). Для погрешностей получается уравнение с переменными коэффициентами, а к нему, в силу замечания 1, можно применять принцип максимума. Этот подход является нестрогим, но плодотворным.

2.3.6. Операторные неравенства

Метод основан на свойствах операторов и порождаемых ими энергетических нормах. Сами доказательства связаны с трудными и громоздкими преобразованиями и требуют высокой математической квалификации. Однако этот метод позволяет строго доказывать сходимость линейных схем с переменными коэффициентами на неравномерных сетках, что не удается другими методами. Кроме того, доказывается сходимость разностного решения в сильных нормах (зачастую более сильных, чем норма C) при использовании слабых норм для правой части. Возможны обобщения этого метода на нелинейные схемы. Этот метод развит в работах А. А. Самарского и его учеников.

Энергетические нормы. Рассмотрим основную идею на примере стационарной разностной схемы $Bv = \phi$. Если рассмотреть сеточные функции v и ϕ как векторы, то оператор B есть матрица. Пусть эта матрица симметричная и положительная. Для нее существует обратная матрица B^{-1}, которая также симметрична и положительна. Последнее означает, что скалярное произведение $(v, Bv) > 0$ для любого v, и можно ввести энергетическую и негативную нормы:

$$\|v\|_B^2 = (v, Bv) > 0, \ \|\phi\|_{B^{-1}}^2 = (\phi, B^{-1}\phi) > 0. \tag{2.37}$$

Проделаем цепочку преобразований:

$$\|v\|_B^2 = (v, Bv) = (B^{-1}\phi, \phi) = \|\phi\|_{B^{-1}}^2 .$$

Схема линейна, поэтому такое же соотношение будет выполняться для погрешностей

$$\|\delta v\|_B = \|\delta\phi\|_{B^{-1}}.$$

Последнее означает выполнение устойчивости (2.22) по правой части.

Рассмотрим конкретный пример. Пусть задана краевая задача для уравнения $u_{xx} = f$ (см. рис. 1.4). Тогда оператор B есть вторая разность, т. е. первая разность от первой разности. Это можно записать в форме $B = B^{1/2}B^{1/2}$, где $B^{1/2}$ есть первая разность. Тогда энергетическую норму (2.37) можно преобразовать так:

$$\|v\|_B^2 = \left(v,\ B^{1/2}B^{1/2}v\right) = \left(B^{1/2}v,\ B^{1/2}v\right) = \left\|B^{1/2}v\right\|_c^2. \quad (2.38)$$

Величина $B^{1/2}v$ есть разностный аналог первой производной, так что правая часть (2.38) является разностным аналогом нормы C для производной u_x. Это норма сильнее, чем норма C для $u(x)$.

Аналогичными рассуждениями можно показать, что оператор B^{-1} есть двукратная сумма, а негативная норма (2.37) является разностным аналогом однократного интеграла от $f(x)$. Эта норма много слабее, чем нормы C или L_2 для $f(x)$. Таким образом, метод позволяет доказывать устойчивость сеточного решения в сильной норме при слабой норме для правой части. В подобных случаях нередко удается доказать сходимость схем с более высоким порядком точности, чем при использовании других методов.

Операторные неравенства. Будем предполагать, что читатель знаком с операторами в гильбертовых пространствах. Рассмотрим применение этого метода к нестационарной задаче и двуслойной разностной схеме

$$B\frac{\hat{v} - v}{\tau} + Av = \phi. \quad (2.39)$$

Справедлива теорема.

Теорема 2.6. Если операторы B и A самосопряженные, не зависят от номера слоя и выполнено операторное неравенство

$$B \geq \frac{\tau}{2}A > 0, \quad (2.40)$$

то схема (2.39) устойчива по начальным данным. •

Доказательство. Для исследования устойчивости по начальным данным надо положить $\phi = 0$ в схеме (2.39). Затем умножим схему слева на оператор $A^{1/2}B^{-1}$ и введем обозначение $A^{1/2}v = w$. Тогда схема преобразуется в явную:

$$\frac{\hat{w} - w}{\tau} + Sw = 0,\ S = A^{1/2}B^{-1}A^{1/2}$$

или

$$\hat{w} = (E - \tau S)w, \tag{2.41}$$

где E — единичный оператор.

Перепишем неравенство (2.40) в следующем виде: $0 < B^{-1} \leq \leq (2/\tau)A^{-1}$. Умножая его слева и справа на положительный оператор $A^{1/2}$, получим: $0 < A^{1/2}B^{-1}A^{1/2} \equiv S \leq (2/\tau)E$, где E — единичный оператор. Умножая это неравенство на τ и вычитая его из E, получим: $E > E - \tau S \geq -E$. Таким образом, оператор в правой части (2.41) заключен между $-E$ и E. Следовательно,

$$\|\hat{w}\|_E \leq \|w\|_E, \tag{2.42}$$

что означает устойчивость по начальным данным.

Поскольку $\|w\|_E = (w, w) = \left(A^{1/2}v, A^{1/2}v\right) = (v, Av) = \|v\|_A^2$, то устойчивость доказана в энергетической норме оператора A. ■

Заметим, что при доказательстве не требовалось постоянства коэффициентов схемы или равномерности сетки. Тем самым, теорема 2.6 справедлива для схем с переменными коэффициентами на неравномерных сетках.

2.4. СХОДИМОСТЬ

2.4.1. Установление сходимости

Напомним, в чем суть численных расчетов. Пусть в области $G(x)$ с границей $\Gamma(G)$ задано дифференциальное (интегральное, интегро-дифференциальное и т. п.) уравнение с краевыми и начальными условиями. Символически его можно записать в форме

$$A(u(x)) = f(x),\ x \in G. \tag{2.43}$$

Оператор $A(u(x))$ в общем случае нелинейный и включает в себя как дифференциальное уравнение, так и краевые и начальные

условия. Вводится сетка $\omega = \{x_n\}$ с шагом h. На этой сетке задача (2.43) аппроксимируется разностной схемой

$$B(v(x)) = \phi(x),\ x \in \omega. \tag{2.44}$$

Оператор $B(v(x))$ также в общем случае нелинейный и аппроксимирует как дифференциальное уравнение, так и граничные и начальные условия. Сеточное решение устраивает нас, если оно становится сколь угодно близким к точному при $h \to 0$. Поскольку сеточное решение определено только на сетке, его сравнение с точным решением можно проводить только в сеточной норме.

Определение 2.4. *Разностное решение сходится к точному*, если

$$\|v - u\| \to 0 \text{ при } h \to 0. \quad \bullet \tag{2.45}$$

Определение 2.5. *Сходимость имеет порядок* p, если

$$\|v - u\| = O\left(h^p\right) \text{ при } h \to 0. \quad \bullet \tag{2.46}$$

На начальном слое в сеточное решение вносится погрешность начальных данных. На каждом последующем слое дополнительно вносятся погрешности аппроксимации дифференциального уравнения и краевых условий. Все эти погрешности передаются на последующие слои и могут усиливаться в ходе расчета. Для получения хорошей точности нужно, чтобы все эти погрешности были малы и в ходе расчета не сильно возрастали.

В любых расчетах также присутствуют ошибки округления. Поэтому, строго говоря, при $h \to 0$ надо одновременно увеличивать разрядность чисел. На практике это неудобно, и рекомендуется придерживаться следующего правила. Не следует проводить вычисления с 32-разрядными числами: это всего лишь 7 десятичных знаков, а даже при хорошо обусловленных алгоритмах теряется два — четыре знака. В большинстве случаев достаточно 64-разрядных чисел, имеющихся на современных персональных компьютерах. Если же алгоритм плохо обусловлен, то может потребоваться 128 разрядов и более.

Определение 2.6. *Разностная схема* (2.44) *корректна*, если ее решение существует и единственно при любых допустимых $\phi(x)$, и схема устойчива в смысле определения 2.3.

Это определение относится к любым нелинейным схемам. Для нелинейных схем решение может, вообще говоря, существо-

вать не при любых $\phi(x)$. Поэтому определение 2.6 можно уточнить, требуя существования, единственности и непрерывной зависимости от $\phi(x)$ только в некоторой окрестности решения. Ситуация напоминает два корня квадратного уравнения: каждый корень непрерывно зависит от коэффициентов уравнения, но оба корня сильно отличаются.

Приведем основную теорему о сходимости, справедливую для произвольных нелинейных задач. Ее кратко формулируют так: *из аппроксимации и устойчивости следует сходимость.*

Теорема 2.7. Пусть решение точной задачи (2.43) существует, а разностная схема (2.44) корректна и аппроксимирует точную задачу (2.43) в смысле определения 2.1. Тогда разностное решение сходится к точному. •

Доказательство. Вспомним определение невязки (2.8): $\psi(x) = B(u(x)) - \phi(x)$. Исключая отсюда $\phi(x)$ с помощью (2.44), получим

$$B(u(x)) - B(v(x)) = \psi(x). \tag{2.47}$$

Возьмем сколь угодно малое $\epsilon > 0$. В силу устойчивости (2.17) найдется такое $\delta > 0$, не зависящее от h, что при $\|\psi\| < \delta$ будет выполнено $\|v - u\| < \epsilon$. В силу аппроксимации (2.12) для данного ϵ найдется такой шаг h_0, что при всех шагах $h < h_0$ будет выполнено $\|\psi\| < \delta$. Таким образом, при всех шагах $h < h_0$ будет выполнено $\|v - u\| < \epsilon$. ■

Замечания. *1.* Некоторые начальные или граничные условия аппроксимируются точно. Примером являются граничные условия первого рода $u(x_0, t) = \mu(t)$, если узел x_0 расположен точно на границе. Устойчивости по таким условиям можно не требовать, ибо никакой ошибки в расчет они не вносят (кроме ошибки округления).

2. Устойчивости по правой части надо требовать всегда, ибо всегда присутствует погрешность аппроксимации, а она эквивалентна некоторой погрешности правой части.

3. Для доказательства теоремы 2.7 необходима аппроксимация на точном решении задачи (2.43). На практике обычно проверяют аппроксимацию на некотором широком классе функций U, которому принадлежит $u(x)$: например, на классе достаточно гладких функций. Такая аппроксимация достаточна для доказательства.

4. При исследовании аппроксимации и устойчивости конкретных разностных схем нередко используют разные нормы для одной и той же функции. Например, при установлении локальной аппроксимации для $\phi(x)$ берется $\|\phi\|_c$, а при спектральном исследовании устойчивости — $\|\phi\|_{l_2}$. Доказательство сходимости в этом случае справедливо,

только если аппроксимация установлена в нормах $\|\phi\|$, более сильных (или тех же самых), чем нормы, использованные для правых частей в определении устойчивости.

5. Если аппроксимация или устойчивость условные, то сходимость имеет место при выполнении условий устойчивости и аппроксимации (т. е. при определенных соотношениях между шагами по разным переменным).

6. Устойчивость является, как нетрудно убедиться, необходимым условием сходимости. Если схема неустойчива, то найдутся такие сколь угодно малые ошибки входных данных, которым соответствует значительная погрешность решения. Сходимости при этом не может быть.

2.4.2. Оценки точности

Для линейных задач оценки погрешности можно получить на основании приведенных ниже теорем. Одна теорема устанавливает априорные оценки погрешности; ее краткая формулировка: *для линейных задач порядок сходимости не ниже порядка аппроксимации.*

Теорема 2.8. Если условия теоремы 2.7 выполнены, операторы A и B линейны, а порядок аппроксимации равен p, то сходимость имеет порядок не ниже p. •

Доказательство. Для линейных операторов равенство (2.47) переписывается как $B(u-v)=\psi$. Для линейных схем условие устойчивости (2.18) имеет следующий вид: $\|\delta v\| \le M\|\delta\phi\|$, где константа M не зависит от h. Подставляя сюда ψ вместо $\delta\phi$ и $u-v$ вместо δv, получаем оценку: $\|u-v\| \le M\|\psi\|$. Но в силу p-го порядка аппроксимации, $\|\psi\| = O(h^p)$. Отсюда следует: $\|u-v\| \le O(h^p)$; это означает сходимость с порядком не ниже p. ■

Замечания. 1. Константа M не зависит от шага h, но она может зависеть от точного решения $u(x)$ и его производных.

2. Возможен разный порядок аппроксимации для самого дифференциального уравнения, краевых условий и начальных данных. Тогда теорема 2.8 справедлива для наименьшего из этих порядков аппроксимации.

3. Для доказательства требовалась линейность только разностных операторов, но фактически линейными разностными операторами можно аппроксимировать только линейные дифференциальные или интегральные операторы.

4. Если условия теоремы 2.8 выполнены, то порядок точности может быть выше порядка аппроксимации. В таких случаях более полное

исследование задачи нередко показывает, что для сходимости в выбранной норме решения $\|v\|$ достаточно устойчивости не в выбранной норме правой части $\|\phi\|$, а в некоторой более слабой норме, в которой порядок аппроксимации выше.

5. Для случая многих переменных порядок аппроксимации по разным переменным может быть не одинаковым. Очевидно, порядок точности по разным переменным также может быть различным.

Теорему 2.8 можно обобщить на нелинейные задачи, где условие устойчивости (2.17) имеет вид $\|\delta v\| \le \epsilon$, если $\|\delta\phi\| \le \delta(\epsilon)$. Например, справедлива следующая теорема, которую приведем без доказательства.

Теорема 2.9. Пусть нелинейный оператор B аппроксимирует оператор A с p-м порядком и выполняется условие $\delta(\epsilon) \le \delta_0 \epsilon^m$. Тогда имеет место сходимость с порядком $q \le p/m$. •

В гл. 1 для апостериорной асимптотически точной оценки погрешности использовалось сгущение сетки с применением метода Ричардсона. Обоснование этого подхода для линейных задач дает следующая теорема.

Теорема 2.10. Пусть операторы A и B линейны, а разностная схема (2.44) корректна и аппроксимирует задачу (2.43) с порядком p. Пусть существует функция непрерывного аргумента

$$\bar{\psi}(x) = \lim_{h\to 0} [\psi(x_n)/h^p],\ x_n \in \omega,\ x \in G. \tag{2.48}$$

Пусть также существует решение задачи

$$A\bar{z}(x) = \bar{\psi}(x),\ x \in G, \tag{2.49}$$

и на этом решении разностный оператор B аппроксимирует дифференциальный оператор A. Тогда погрешность численного решения имеет следующую асимптотику:

$$u(x) - v(x) \equiv z(x) = h^p \bar{z}(x) + o\left(h^p\right),\ h \to 0. \bullet \tag{2.50}$$

Доказательство. По определению невязки $\psi = Bu - \phi$. Вычтем из правой части $Bv - \phi \equiv 0$ и поделим равенство на h^p. Вычтем из полученного равенства соотношение (2.49). Тогда

$$\psi/h^p - \bar{\psi} = B\left[(u-v)/h^p\right] - A\bar{z} \equiv B\left[(u-v)/h^p - \bar{z}\right] + B\bar{z} - A\bar{z}. \tag{2.51}$$

При $h \to 0$ левая часть равенства стремится к нулю в силу (2.48). В правой части $B\bar{z} - A\bar{z} \to 0$ в силу условия аппроксимации.

Тогда последняя квадратная скобка также стремится к нулю, что доказывает теорему. ■

Замечания. *1.* Теорема 2.10 позволяет один раз сгустить сетку и применить метод Ричардсона. Но если функция $\bar{z}$ такова, что к ней самой применима эта теорема, то можно повторно сгустить сетку и применить рекуррентный метод Ричардсона.

2. Теорему можно обобщить на случай многих переменных, даже если порядок аппроксимации по разным переменным неодинаковый.

Изложенная в этой главе теория разностных схем применима к разностным схемам, аппроксимирующим корректно поставленные задачи для обыкновенных дифференциальных уравнений, уравнений в частных производных и интегральных уравнений. Теория переносится на решение уравнений в частных производных методом прямых. Хотя в большинстве формулировок фигурировало только одно уравнение и одна переменная, теория очевидным образом обобщается на системы уравнений или случай многих переменных.

Теория разностных схем применяется также для доказательства существования решения точной задачи (2.43) и установления его свойств. Например, приведем без доказательства следующий результат.

Теорема 2.11. Если для задачи (2.43) существует хотя бы одна корректная разностная схема (2.44), аппроксимирующая задачу на функциях $u(x) \in U$, то решение $u(x)$ задачи (2.43) в классе U существует и единственно. Если правая часть $\phi(x)$ непрерывна равномерно по h, то $u(x)$ непрерывно зависит от $f(x)$. •

2.4.3. Экспериментальная математика

Для любой задачи можно составить много разностных схем. Возникает вопрос: какую из схем использовать при решении реальной задачи? Как правило, теоретических оценок сходимости недостаточно. Это связано с несовершенством теоретических методов исследования схем. Укажем типичные трудности.

1. Для большинства нелинейных задач (например, газодинамики) нет строгого доказательства сходимости или хотя бы устойчивости разностных схем. Соображения об их устойчивости и сходимости основаны на анализе линеаризованных задач.

2. Оценки точности схем являются асимптотическими при стремлении шага к нулю. Однако для современных многомерных задач даже супер компьютеры не всегда могут обеспечить малость шага. Например, при трехмерном газодинамическом расчете обтекания самолета невозможно передать все малые вихри, образующиеся у кромок конструкции.

3. Теоретические априорные оценки обычно мажорантны и во много раз превышают асимптотически точные. Поэтому они создают пессимистическое впечатление о величине погрешности.

4. Даже наличие доказательства сходимости разностной схемы не гарантирует хорошего качества полученного по схеме решения. В расчетах нередко возникают пилообразные сеточные решения, качественно непохожие на ожидаемое точное решение.

На основе практики расчетов были сформулированы некоторые дополнительные, не всегда математически строгие требования к разностным схемам, способствующие их улучшению. Наиболее известны требования консервативности схем и монотонности. Полезны также требования бикомпактности, диссипативности и наличия аппроксимационной вязкости. Они будут рассмотрены в других главах.

Тестирование схем. Неполноту теоретического исследования компенсируют, испытывая схему на некоторой системе тестов. Тестами могут служить задачи, решение которых известно в явном виде или может быть вычислено с высокой точностью. Например, для уравнения в частных производных в качестве тестов нередко берут автомодельные решения. По схеме проводят серию расчетов данной задачи на последовательности сгущающихся сеток. Зная решение точной задачи, непосредственно вычисляют погрешность решения $z = u - v$. Для полученной погрешности вычисляют нормы (обычно C и L_2). Исследуют зависимость нормы $\|z\|$ от шага сетки (числа узлов N). Наиболее наглядным является построение графика с абсциссой $\log N$ и ординатой $\log \|z\|$.

Обычно сходимость степенная: $\|z\| = O(N^{-p})$. Тогда при увеличении N точки графика асимптотически выходят на прямую линию (рис. 2.7). Тангенс угла наклона этой прямой равен фактическому порядку точности p. При малых N точки могут заметно отличаться от асимптотической прямой. Это означает, что шаг слишком велик и пользоваться асимптотическими оценками еще нельзя. На очень подробных сетках возможен выход на

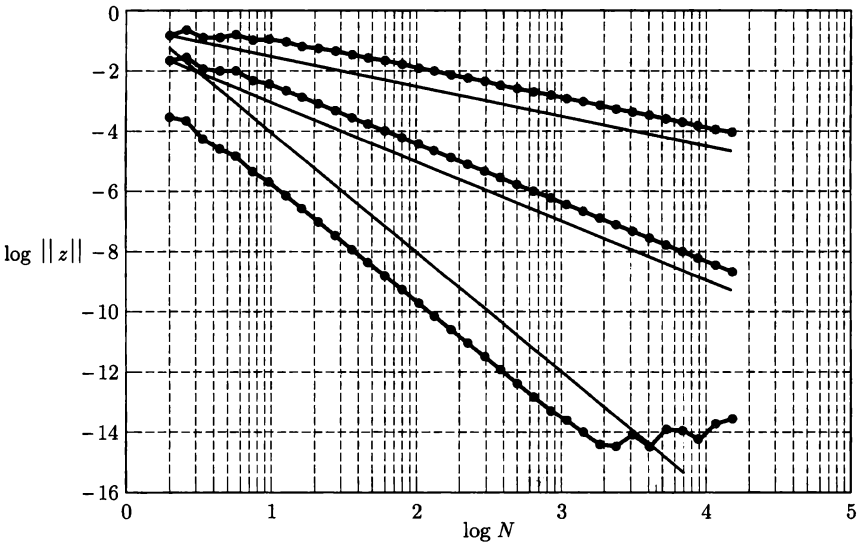


Рис. 2.7. Зависимость ошибки решения от числа узлов сетки (расчеты — точки, соединенные жирными линиями; теоретические наклоны — тонкие линии)

ошибки округления компьютера. Погрешность при этом достигает определенного уровня и далее перестает убывать при сгущении сеток.

Получив фактический порядок точности, его следует сравнить с теоретически ожидаемым. Их совпадение означает, что схема и программа написаны правильно. Однако может оказаться, что они не совпадают. Что это означает? Разберем типичные случаи.

1. Фактический порядок p оказался ниже теоретического. Возможны две причины. Первая — ошибка в программе. Надо перепроверить программу, особенное внимание обращая на написание тех слагаемых, порядок малости которых равен фактическому порядку точности. Вторая причина — тестовое решение недостаточно гладкое: не имеет тех производных, которые использовались в теоретическом анализе аппроксимации. Тогда следует взять тест с более гладким точным решением.

2. Фактический порядок точности оказался выше теоретического. Это означает, что теоретическое исследование было недостаточно полным. Например, аппроксимация исследовалась на слишком широком классе функций, где она хуже, чем на точ-

ном решении. Возможно также, что в доказательствах использовались более сильные нормы для правых частей, чем это в действительности необходимо.

Установление сходимости. Ранее отмечалось, что для нелинейных задач редко удается доказать устойчивость схемы. Тогда сходимость остается недоказанной. Кроме того, часто приходится решать прикладные задачи, где не доказано даже существование точного решения. Однако численные расчеты при этом зачастую выполняются и дают разумные результаты. Можно ли им доверять?

Изложенная ранее теория сходимости позволяет выработать следующую стратегию. Составим разностную схему и установим ее аппроксимацию на некотором классе функций, предположительно включающем ожидаемое решение. Если решение ожидается достаточно гладким, обосновать аппроксимацию несложно (например, находя невязку разложением в ряды Тейлора по степеням шага). Если гладкость ожидаемого решения невелика, следует переходить к интегральной форме вывода разностных схем. Тем самым исследование аппроксимации будет достаточно полным и строгим.

Затем составим программу и проверим ее на тестовых задачах, проводя расчеты на последовательности сгущающихся сеток. При этом возможны следующие ситуации.

1. Последовательность численных решений не будет стремиться ни к какой предельной функции при $h \to 0$. Обычно при этом решение «разбалтывается», причем при $h \to 0$ амплитуда этой «разболтки» возрастает и возможен выход за пределы представимых на компьютере чисел. Это показывает, что схема неустойчива и непригодна для расчетов.

2. Если сеточные решения сходятся к предельной функции при $h \to 0$, то данный расчет оказался устойчивым. Согласно теореме 2.7, из аппроксимации и устойчивости следует сходимость. Тем самым наблюдаемый предел является решением точной задачи, а наши сеточные расчеты сходятся к этому точному решению.

Аппроксимацию мы обосновали теоретически, а сходимость установили экспериментально. Поэтому такой способ обоснования сходимости является сочетанием теоретического анализа и математического эксперимента. Он не является полностью строгим: быть может, при дальнейшем сгущении сеток расчеты пе-

рестанут сходиться. Однако если расчеты проведены на достаточно подробных сетках, то результат является убедительным.

Согласно теореме 2.11, эту методологию можно применять в прикладных расчетах, где нам заранее неизвестно точное решение (и даже неизвестно, существует ли оно). При этом мы, хоть не совсем строго, доказываем даже существование неизвестного точного решения и получаем хорошее приближение к нему. Поэтому данным способом широко пользуются на практике.

Такая методология практически безошибочно действует на задачах с гладкими решениями. Однако при недостаточно гладких (например, разрывных) решениях этот способ может давать неправильные результаты. В гл. 3 будет приведен пример, когда имеется сходимость численного решения к предельной функции, но эта предельная функция не является точным решением исходной задачи. Ошибка была вызвана тем, что искомое решение было разрывным, а исследование аппроксимации проводилось на классе достаточно гладких решений, к которому точное решение не принадлежит.

Контроль точности. Пусть мы решаем задачу с неизвестным ответом, сгущаем сетки и визуально наблюдаем сходимость сеточных решений к предельной функции. Метод Ричардсона позволяет автоматизировать эту процедуру, т. е. заменить визуальный контроль алгоритмическим и при этом получить оценку погрешности. Напомним соответствующую процедуру (см. кн. 1).

Пусть построены расчеты на достаточно большом числе равномерных или квазиравномерных сеток, сгущающихся ровно в два раза. Тогда узел каждой сетки будет четным узлом следующей сетки. Выберем три соседние сетки с числами интервалов N, $2N$, $4N$. Всем узлам первой сетки x_n, $0 \le n \le N$, будут соответствовать узлы x_{2n} второй сетки и x_{4n} третьей сетки.

Обозначим сеточные решения в узле x_n на трех выбранных сетках соответственно $v_1(x_n)$, $v_2(x_n)$, $v_3(x_n)$. Пусть схема имеет порядок аппроксимации p. Вычислим ричардсоновские поправки по соседним парам сеток: $\delta_2(x_n) = (v_2(x_n) - v_1(x_n)) / (2^p - 1)$, $\delta_3(x_n) = (v_3(x_n) - v_2(x_n)) / (2^p - 1)$. Их индексы показывают, что в каждой паре они приписаны более подробной сетке. Поведение этих поправок целесообразно контролировать, строя график вида рис. 2.7, где вместо $\log \|z\|$ берется $\log \|\delta\|$. Если при увеличении N точки выходят на прямую с правильным накло-

ном, метод Ричардсона можно считать надежно применимым. В этом случае $\|\delta\|$ на каждой сетке можно считать асимптотически точной погрешностью сеточного решения.

Если в каждой точке с удовлетворительной точностью выполняется соотношение $\delta_2(x_n)/\delta_3(x_n) \approx 2^p$, то оценку погрешности можно производить поточечно, а также вводить экстраполяционное уточнение решения: $u(x) \approx v_3(x_n)+\delta_3(x_n)$. При этом уточняется решение на самой подробной сетке, но лишь в тех узлах, которые совпадают с узлами самой грубой из трех.

Для одномерных задач ресурсы компьютеров позволяют сгущать сетку очень много раз. При этом экспериментальное установление сходимости и процедура контроля точности работают надежно. Для многомерных задач трудоемкость расчетов при каждом сгущении сетки стремительно растет, поэтому для них приходится ограничиваться небольшим числом сгущений. Однако и в этом случае удается получать неплохие результаты.

Глава 3

УРАВНЕНИЕ ПЕРЕНОСА

3.1. ЛИНЕЙНОЕ УРАВНЕНИЕ ПЕРЕНОСА

3.1.1. Задачи и решения

Существует много задач о распространении частиц в веществе: определение нейтронных потоков в реакторе, теплопроводности в газах, обусловленной диффузией атомов и электронов, и т. д. Такие задачи приводят к уравнению переноса, которое может быть интегро-дифференциальным. Решение подобных задач достаточно сложно, и здесь ограничимся простейшим линейным уравнением переноса:

$$\frac{\partial u}{\partial t} + \mathbf{c}(\mathbf{x},t)\mathtt{grad}u = f(\mathbf{x},t),\ \mathbf{x} = \{x_1, x_2, \ldots, x_p\}, \qquad (3.1)$$

где $\mathbf{c}$ — вектор скорости переноса. Как будет видно в дальнейшем, для этого уравнения многомерность не вносит принципиальных осложнений. Все основные идеи можно пояснить на одномерном уравнении

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = f(x,t), \qquad (3.2)$$

где скорость c будем считать постоянной, если специально не оговорено противное.

Пусть уравнение (3.2) однородно, т. е. правая часть $f \equiv 0$. Тогда его решение ***автомодельно*** (зависит от одной комбинации аргументов) и имеет вид бегущей волны:

$$u(x,t) = \phi(\xi),\ \xi = x - ct, \qquad (3.3)$$

где $\phi(\xi)$ — произвольная дифференцируемая функция. При $c>0$ (что обычно будем предполагать) волна бежит направо, а при $c < 0$ — налево. Решение остается постоянным вдоль линий $\xi = \mathtt{const}$, называемых ***характеристиками***; они показаны

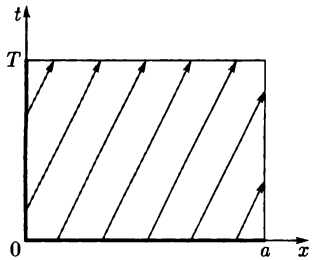


Рис. 3.1. Представление решения уравнения в виде характеристик

стрелками на рис. 3.1. Вид решения (3.3) подсказывает, как можно корректно поставить полную задачу для уравнения (3.2).

Смешанная задача Коши. Уравнение (3.2) содержит только первые производные, поэтому требуется задавать по одному дополнительному условию на каждый аргумент. Зададим начальное и граничное условия на отрезках, показанных на рис. 3.1 жирными линиями:

$$u(x,0) = \mu_1(x),\ 0 \le x \le a;\ u(0,t) = \mu_2(t),\ 0 \le t \le T. \qquad (3.4)$$

При $c = \mathtt{const} > 0$ характеристики являются параллельными прямыми. Начальные и граничные условия переносятся по характеристикам. Поэтому решение задачи (3.2), (3.4) однозначно определено в области $G = [0 \le x \le a] \times [0 \le t \le T]$.

Пусть начальные и граничные данные являются p-гладкими, т. е. непрерывны вместе со своими p-ми производными. Тогда решение $u(x,t)$ в области G будет также p-гладким всюду, за исключением характеристики, исходящей из стыка границ области ($x = 0$, $t = 0$). Для обеспечения гладкости на этой характеристике необходимо обеспечить условие согласования; для однородного уравнения они имеют следующий вид:

$$\frac{d^q \mu_2(0)}{dt^q} = (-c)^q \frac{d^q \mu_1(0)}{dx^q},\ 0 \le q \le p. \qquad (3.5)$$

Для неоднородного уравнения надо требовать $(p-1)$-гладкости от $f(x,t)$ и соответственно изменить условие согласования (3.5).

Пусть начальные данные заданы на полубесконечной прямой $-\infty < x \le a$. Левая граница отсутствует, так что это «чистая» задача Коши. Тогда решение однозначно определено в области $G = (-\infty < x \le a] \times [0 \le t < +\infty)$. Гладкость решения соответствует гладкости начальных данных $\mu(x)$ и правой части $f(x,t)$.

Интегрирование по характеристикам. Решение неоднородного уравнения (3.2) меняется вдоль характеристик. Это изменение легко найти, если перейти к новым координатам, связанным с характеристиками:

$$\xi = x - ct,\ \eta = x + ct. \tag{3.6}$$

При их помощи уравнение (3.2) преобразуется к виду

$$2c\frac{\partial u}{\partial \eta} = \phi(\xi, \eta),\ \phi(\xi, \eta) = f\left(\frac{\xi + \eta}{2}, \frac{\eta - \xi}{2c}\right). \tag{3.7}$$

Следовательно, вдоль характеристики $\xi = \texttt{const}$ решение u можно найти, интегрируя по η обыкновенное дифференциальное уравнение (3.7), в котором ξ играет роль параметра. Так можно определить решение в любой точке области G, поскольку при $c = \texttt{const}$ характеристики покрывают всю область.

Этот способ построения точного решения легко обобщается на уравнение с переменным коэффициентом $c(x, t)$. Пусть $c(x, t) > 0$ и непрерывно в области G. Тогда через каждую точку области G проходит одна и только одна характеристика, причем эта характеристика исходит из какой-либо граничной точки (жирные линии на рис. 3.1). Видно, что этого достаточно для корректности смешанной задачи Коши.

Метод интегрирования по характеристикам нередко используется при численном решении задач переноса.

Сохранение монотонности является важным свойством однородного уравнения переноса. Если для него поставлена задача Коши с монотонными начальными данными $u(x, 0) = \mu(x)$, $-\infty < x \leq a$, то в любой момент времени t профиль $u(x, t)$, т. е. зависимость решения от аргумента x, тоже будет монотонным. Монотонность сохраняется и в смешанной задаче Коши, если граничное условие $u(0, t)$ также монотонно зависит от t и направление монотонности противоположно направлению монотонности начальных данных (т. е. при возрастающих начальных данных граничные условия должны быть убывающими).

В уравнении переноса монотонность является тривиальным следствием из вида общего решения (3.3). Однако во многих уравнениях начальная монотонность решения сохраняется, хотя общее решение не имеет вида одной бегущей волны. При определенных условиях это имеет место, например, в задачах теплопроводности и газодинамики. Поэтому монотонность — достаточно общее и важное свойство многих уравнений. Нарушение монотонности в прикладных расчетах может приводить к абсурдным результатам. Например, пусть $u(x, t)$ есть концентрация частиц. На немонотонном профиле могут оказаться значения $u(x, t) < 0$, т. е. получатся отрицательные концентрации.

3.1.2. Схемы бегущего счета

Схемы бегущего счета предназначены для решения смешанной задачи Коши (3.2), (3.4). Они легко обобщаются на случай любого числа измерений. Эти схемы являются наиболее простыми и позволяют численно решать даже очень сложные задачи переноса с неплохой точностью при умеренном объеме вычислений.

Рассмотрим задачу (3.2), (3.4) и построим в области $G = [0 \le x \le a] \times [0 \le t \le T]$ прямоугольную сетку. Сетка может быть неравномерной по каждой из переменных. Однако схемы будут использовать узлы только одной ячейки и содержать шаги h_n и τ_k, относящиеся только к этой ячейке. Поэтому в дальнейшем индексы у шагов будем опускать. Выберем шаблоны, показанные на рис. 3.2. Для составления схемы используем метод простейшей замены производных разностями. Составим на трехточечных шаблонах простейшие схемы с использованием односторонних разностей:

$$(\hat{u}_n - u_n)/\tau + c(u_n - u_{n-1})/h = \phi_n, \tag{3.8}$$

$$(\hat{u}_{n-1} - u_{n-1})/\tau + c(\hat{u}_n - \hat{u}_{n-1})/h = \phi_n, \tag{3.9}$$

$$(\hat{u}_n - u_n)/\tau + c(\hat{u}_n - \hat{u}_{n-1})/h = \phi_n; \tag{3.10}$$

здесь $\phi_n = f(x_n - h/2, t_k + \tau/2)$. На четырехточечном шаблоне (см. рис. 3.2) напишем схему с симметризованными разностями:

$$\begin{gathered}(\hat{u}_n + \hat{u}_{n-1} - u_n - u_{n-1})/\tau + \\ + c(\hat{u}_n + u_n - \hat{u}_{n-1} - u_{n-1})/h = 2\phi_n.\end{gathered} \tag{3.11}$$

Правую часть выбираем в центре ячейки. Для схемы (3.11) это необходимо по соображениям точности, а для схем (3.8) — (3.10) допустимо вычислять ее в любой точке ячейки.

Порядок расчета. Расчет по этим схемам очень прост. Хотя формально схема (3.8) является явной, а остальные три — неяв-

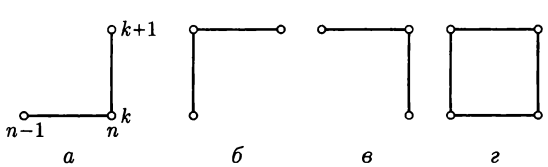


Рис. 3.2. Шаблоны ($а$ — $г$) схем бегущего счета

ными, фактически при расчете смешанной задачи Коши они ведут себя как явные.

В самом деле, во всех четырех схемах значение $\hat{u}_n$ явно выражается через значения $\hat{u}_{n-1}$, u_n, u_{n-1} (или любые два из них). Значение решения на нулевом слое $u_n^0 = \mu_1(x_n)$ известно из начального условия. На следующем (первом) слое значение $\hat{u}_0 = \mu_2(t_1)$ в силу граничного условия, и можно вычислить $\hat{u}_1$; зная $\hat{u}_1$, можно вычислить $\hat{u}_2$, затем $\hat{u}_3$. Так последовательно вычисляются слева направо все $\hat{u}_n$ первого слоя. Зная решение на первом слое, точно так же вычисляем его на втором слое, на третьем и т. д.

Заметим, что явная схема (3.8) пригодна для решения задачи Коши на полубесконечной (или бесконечной) прямой; неявные схемы бегущего счета к такой задаче неприменимы.

Из описанного алгоритма видно, что для каждой из схем (3.8) — (3.11) разностное решение существует и единственно при любых правых частях и начальных и граничных данных. Поэтому для доказательства сходимости остается исследовать аппроксимацию и устойчивость схем. Заметим, что краевое условие $u(0,t) = \mu_2(t)$ для всех схем аппроксимируется точно, поэтому устойчивости по нему не требуется.

Явная схема (3.8). Исследуем ее погрешность аппроксимации. Пусть начальные и граничные данные дважды непрерывно дифференцируемы и удовлетворяют условиям согласования типа (3.5) с $p = 2$, а правая часть $f(x,t)$ имеет непрерывные первые производные. Тогда решение $u(x,t)$ дважды непрерывно дифференцируемо. Разложим его по формуле Тейлора в узле (x_n, t_k):

$$\hat{u}_n = u_n + \tau u_t + \frac{1}{2}\tau^2 u_{tt},$$
$$u_{n-1} = u_n - h u_x + \frac{1}{2}h^2 u_{xx},$$
$$\phi_n = f_n + \frac{1}{2}\tau f_t - \frac{1}{2}h f_x.$$

Подставляя эти формулы в определение невязки (2.8), получим

$$\psi_n = \frac{1}{\tau}(\hat{u}_n - u_n) + \frac{c}{h}(u_n - u_{n-1}) - \phi_n = \\ = \frac{\tau}{2}(u_{tt} - f_t) - \frac{h}{2}(c u_{xx} - f_x) + o(\tau + h) = O(\tau + h). \tag{3.12}$$

При сделанных предположениях схема (3.8) имеет аппроксимацию в норме c с первым порядком.

Сделаем интересное замечание. Невязка вычисляется на точном решении уравнения (3.2). Дифференцируя это уравнение по t, получим: $u_{tt} + cu_{tx} = f_t$. Дифференцируя по x, получим: $u_{tx} + cu_{xx} = f_x$. Подставляя найденные значения f_t и f_x в (3.12), преобразуем ее:

$$\psi = \frac{1}{2}(h - c\tau)\, u_{tx} \tag{3.13}$$

с точностью до малых высшего порядка. При $c\tau = h$ главный член невязки обращается в нуль и погрешность аппроксимации имеет более высокий порядок малости. Нетрудно показать, что в этом случае на трижды непрерывно дифференцируемых решениях $\psi = O\left(\tau^2 + h^2\right)$.

Устойчивость исследуем методом гармоник (см. п. 2.3.4), поскольку он дает необходимое и достаточное условие. Полагая $\phi_n = 0$ и заменяя $u_n \to 1$, $\hat{u}_n \to \rho$, $u_{n-1} \to \exp(-iqh)$, получим множитель роста q-й гармоники:

$$\rho_q = 1 - \frac{c\tau}{h}\left(1 - e^{-iqh}\right). \tag{3.14}$$

Возможные значения величины ρ в комплексной плоскости лежат на окружности радиуса $c\tau/h$, центр которой имеет вещественную координату $1 - c\tau/h$. Если $c\tau < h$, эта окружность целиком лежит внутри единичной окружности (рис. 3.3), так что $|\rho_q| < 1$. Если $c\tau > h$, то окружность ρ_q лежит вне единичной окружности и $|\rho_q| > 1$. Таким образом, для устойчивости необходимо и достаточно, чтобы так называемое ***число Куранта***

$$\kappa \equiv c\tau/h \le 1. \tag{3.15}$$

Заметим, что ошибка при этом не нарастает, так что схема хорошо обусловлена.

Метод гармоник предполагает, что $h = \texttt{const}$. Применение его к неравномерным сеткам основано на приближении «замороженных» шагов. Число Куранта $\kappa_{k,n} = c\tau_k/h_n$ будет свое для каждого интервала и каждого слоя. В этом случае для устойчивости требуют, чтобы на всех слоях и интервалах выполнялось $\kappa_{k,n} \le 1$. Правда, нарушение этого условия в отдельных интервалах в отдельные моменты времени может и не привести к «разболтке» расчетов, но злоупотреблять этим не следует.

Метод гармоник дает устойчивость в норме l_2. Однако можно применить принцип максимума (см. п. 2.3.5). Подставим в условие равномерной устойчивости по начальным данным (2.35) коэффициенты схемы (3.8). Снова получим условие Куранта (3.15), что доказывает устойчивость в норме c. Напомним, что доказательство принципа максимума является строгим для произвольных неравномерных сеток.

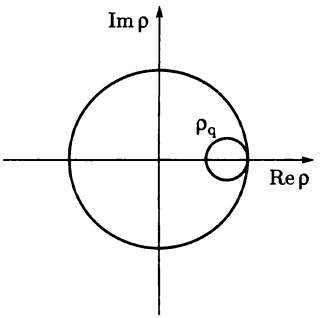


Рис. 3.3. Множитель роста (3.14)

Непосредственно видно, что дополнительное условие по правой части (2.22) выполняется, причем $\alpha = 1$. Поэтому схема устойчива по правой части в норме c при выполнении условия Куранта (3.15).

Тогда из теорем о сходимости следует, что если решение $u(x,t)$ непрерывно вместе со своими вторыми производными, то схема (3.8) при выполнении условия Куранта (3.15) сходится в норме c со скоростью $O(\tau + h)$, т. е. с первым порядком точности.

Неявная схема (3.9) исследуется аналогично; при исследовании аппроксимации разложение по формуле Тейлора удобнее вести около узла $(x_{n-1}, \hat{t})$. На дважды непрерывно дифференцируемых решениях эта схема при выполнении условия устойчивости $\kappa \equiv c\tau/h \geq 1$ обеспечивает сходимость со скоростью $O(\tau + h)$.

Чисто неявная схема (3.10) исследуется аналогично. Она оказывается безусловно устойчивой и сходится с точностью $O(\tau + h)$. Однако коэффициенты перед производными в выражениях невязки для нее оказываются больше, чем в схемах (3.8), (3.9). Поэтому по точности она уступает этим схемам, хотя порядок точности также первый.

Безусловная устойчивость ценна тем, что она обеспечивает надежность расчета. Это свойство особенно важно на быстродействующих компьютерах, где невозможна оперативная корректировка расчетов. Поэтому пользователи предпочитают чисто неявную схему условно устойчивым схемам, несмотря на ее несколько меньшую точность.

Составная схема. Условия устойчивости схем (3.8) и (3.9) противоположны. Составим из них единую схему с альтерна-

тивными формулами расчета. Если $\kappa < 1$, то вычисления будем проводить по явной схеме (3.8), а при $\kappa \geq 1$ — по неявной схеме (3.9). Тогда ошибка не будет нарастать. Тем самым составная схема является устойчивой при любом отношении τ к h, т. е. она безусловно устойчива, и ее сходимость также является безусловной, а точность есть $O(\tau + h)$.

Величина остаточного члена в составной схеме такова же, как в исходных схемах, т. е. меньше, чем в чисто неявной схеме (3.10). Поэтому составная схема не уступает чисто неявной схеме по устойчивости и превосходит ее по точности. Эта схема часто используется на практике.

Симметричная схема (3.11). При исследовании ее аппроксимации целесообразно разлагать $u(x,t)$ по формуле Тейлора около центра ячейки $(x_n - h/2,\ t_k + \tau/2)$, предполагая наличие третьих непрерывных производных решения и вторых производных правой части. Громоздкие выкладки дают

$$\psi = \tau^2 \left(\frac{1}{24} u_{ttt} + \frac{c}{8} u_{ttx} \right) + h^2 \left(\frac{1}{8} u_{txx} + \frac{c}{24} u_{xxx} \right) = O\left(\tau^2 + h^2\right) \tag{3.16}$$

Схема имеет второй порядок аппроксимации. Это лучше, чем у предыдущих схем.

Устойчивость исследуем методом гармоник. Замена, аналогичная явной схеме, дает следующий множитель роста q-й гармоники:

$$\rho_q = e^{-iqh} \frac{(h + c\tau) + (h - c\tau)e^{iqh}}{(h + c\tau) + (h - c\tau)e^{-iqh}}. \tag{3.17}$$

Видно, что $|\rho_q| = 1$ для любой гармоники при любых соотношениях шагов. Следовательно, схема равномерно устойчива по начальным данным в $\|\cdot\|_{l_2}$, причем устойчивость безусловная.

Дополнительный критерий устойчивости по правой части (2.22) после умножения на τ принимает для схемы (3.11) следующий вид:

$$1 + \frac{c\tau}{h} - \left| 1 - \frac{c\tau}{h} \right| \geq \alpha,\ \alpha = \text{const} > 0.$$

Убедимся, что для $\alpha = 2$ это неравенство выполняется при любых τ и h. В самом деле, если $c\tau \leq h$, то левая часть неравенства равна 2. Если же $c\tau > h$, то левая часть неравенства равна $(2c\tau/h) > 2$. Поскольку критерий выполнен, то схема безусловно устойчива по правой части.

Из изложенного ранее следует, что на трижды непрерывно дифференцируемых решениях $u(x,t)$ схема (3.11) безусловно сходится в норме $\|\cdot\|_{l_2}$ со скоростью $O(\tau^2+h^2)$. Более сложными методами можно доказать ее сходимость в $\|\cdot\|_c$.

Замечания. *1.* Схемы бегущего счета сходятся на решениях меньшей гладкости и даже на разрывных решениях (разумеется, не равномерно, а в среднем). Например, теоретический анализ и примеры численных расчетов показали, что схема (3.10) сходится на кусочно-непрерывных решениях в $\|\cdot\|_{l_p}$ с погрешностью $O\left((\tau+h)^{\frac{1}{(2p)}}\right)$. Любопытно, что порядок точности оказался не целым!

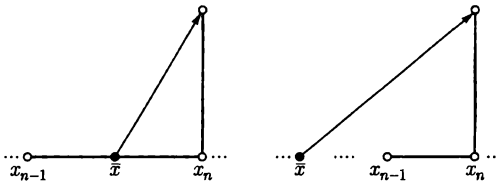
2. Оценим условия оптимального применения различных схем бегущего счета. Симметричная схема (3.11) имеет второй порядок точности и безусловно устойчива. На достаточно гладких решениях при не слишком больших шагах τ и h она дает лучшие результаты. Однако для разрывных решений или для решений с большими производными при недостаточно подробных сетках она оказывается плохой (далее будет рассмотрена причина этого — немонотонность схемы). В этом случае удовлетворительные результаты дает чисто неявная схема (3.10) или составная схема; последняя предпочтительнее из-за лучшей точности.

3. Напомним, что метод гармоник является нестрогим вариантом метода разделения переменных — разложения погрешности по собственным функциям (векторам) разностного оператора. Неявные схемы (3.9) — (3.11) содержат два значения на новом слое: $\hat{u}_{n-1}$ и $\hat{u}_n$. Поэтому формально они определяются из линейной алгебраической системы, матрица которой содержит главную диагональ и поддиагональ, т. е. является жордановой. Жорданова матрица имеет только один собственный вектор (см. кн. 1), т. е. ее собственные векторы не образуют базиса. Тем самым метод гармоник для них является необоснованным. Однако фактические результаты по устойчивости, полученные методом гармоник, отлично подтверждаются в практике расчетов.

3.1.3. Геометрическая интерпретация устойчивости

Ограничимся устойчивостью по начальным данным. Рассмотрим однородное уравнение (3.2). Его общее решение имеет вид бегущей волны (3.3) и распространяется по характеристикам $x-ct=\mathtt{const}$ без изменения.

Рассмотрим схему (3.8) с шаблоном, изображенным на рис. 3.4. Построим характеристику, проходящую через искомый узел $(x_n,\hat{t})$; она обозначена стрелкой на рис. 3.4. Эта характеристика пересекает исходный слой t в точке $\bar{x}=x_n-c\tau$. Схему (3.8)



а *б*

Рис. 3.4. Шаблоны (*а*, *б*) схемы (3.8)

с $\phi_n = 0$ можно интерпретировать следующим образом. Линейно интерполируя разностное решение между узлами исходного слоя, найдем

$$\bar{u} = \frac{x_n - \bar{x}}{h} u_{n-1} + \frac{\bar{x} - x_{n-1}}{h} u_n = \frac{c\tau}{h} u_{n-1} + \left(1 - \frac{c\tau}{h}\right) u_n. \quad (3.18)$$

Затем найденное значение перенесем без изменения по характеристике в искомый узел, т. е. положим $\hat{u}_n = \bar{u}$.

Если выполнено условие устойчивости схемы $c\tau \le h$, то характеристика проходит между точками x_{n-1} и x_n, а значение $\bar{u}$ вычисляется интерполяцией в узком смысле слова. Если $c\tau > h$, то характеристика пересекает слой t вне исходного отрезка, а $\bar{u}$ вычисляется экстраполяцией. В кн. 1 уже пояснялась ненадежность экстраполяции. Здесь видно, что экстраполяция приводит к неустойчивости.

Схемы (3.9) и (3.10) тоже можно интерпретировать как линейную интерполяцию по двум уже вычисленным значениям, с последующим переносом по характеристике. В частности, безусловная устойчивость схемы (3.10) связана с тем, что приходящая в искомый узел характеристика (стрелка на рис. 3.5) при любых τ и h пересекает отрезок, соединяющий исходные узлы (штриховая линия на рисунке).

Схема (3.11) интерпретируется тоже как интерполяция, но не двухточечная линейная, а трехточечная квадратичная (что, естественно, приводит к более высокому порядку точности). Какую бы сторону ячейки на рис. 3.2, *г* не пересекала приходящая в узел $(x_n, \hat{t})$ характеристика — горизонтальную или вертикальную, эта сторона связывает узлы с ранее вычисленными значениями u; поэтому экстраполяции здесь нет, что приводит к безусловной устойчивости схемы (3.11).

В явно-неявной схеме проверка условия Куранта по существу выясняет, какую из сторон четырехугольной ячейки пересекает

искомая характеристика. Если пересекается горизонтальная сторона, выбирается шаблон на рис. 3.2, *а*, а если вертикальная — шаблон на рис. 3.2, *б*. Это обеспечивает безусловную устойчивость.

Таким образом, прослеживая положение характеристик, нетрудно так выбрать шаблон и составить на нем разностную схему, чтобы соблюдалось условие интерполяционности. Если оно нарушено, не стоит тратить время на исследование такого шаблона: схема окажется неустойчивой. Перспективными являются шаблоны, удовлетворяющие условию интерполяционности. Приведем некоторые примеры.

Случай $c < 0$. В этом случае наклон характеристик на плоскости (x, t) отрицателен; характеристики зеркально отражены относительно вертикали по сравнению со случаем $c > 0$. Соответственно меняется постановка задачи: для отрезка $0 \le x \le a$ граничное условие теперь должно задаваться справа при $x = a$.

Очевидно, шаблоны для устойчивых схем можно получить зеркальным отражением соответствующих шаблонов на рис. 3.2. Например, вместо шаблона на рис. 3.2, *а* берут шаблон рис. 3.6, получая устойчивую при $|c|\,\tau \le h$ схему. Направление бегущего счета также меняется: расчет на каждом слое ведут справа налево.

Отметим, что шаблоны рис. 3.2, *б* и *в* зеркальны друг другу; это означает, что при $c < 0$ схема (3.9) становится безусловно устойчивой, а схема (3.10) — условно устойчивой. Симметричная схема (3.11) не меняется при отражении, так что она устойчива при любом знаке скорости; но направление счета, разумеется, зависит от знака скорости.

Замечание. Геометрическая интерпретация не является строгим математическим методом. Она лишь позволяет не тратить время на исследование бесперспективных шаблонов и выбрать перспективные.

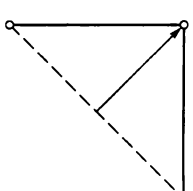


Рис. 3.5. Шаблон схемы (3.10)

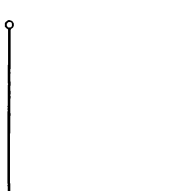


Рис. 3.6. Шаблон (замена рис. 3.2, *а*)

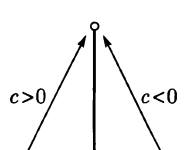


Рис. 3.7. Шаблон схемы (3.19)

Однако интерполяционность еще не гарантирует устойчивости. Перспективные схемы надо проверять обычными методами. Покажем это на следующем примере.

Пример. Рассмотрим следующую явную схему на шаблоне рис. 3.7:

$$(\hat{u}_n - u_n)/\tau + c\,(u_{n+1} - u_{n-1})/(2h) = \phi_n. \tag{3.19}$$

При $|c|\,\tau \le h$ она соответствует интерполяции на исходном слое. Поэтому можно ожидать ее условной устойчивости, причем даже в случае знакопеременного c. Однако методом гармоник легко доказать, что она неустойчива при любом соотношении шагов, т. е. абсолютно неустойчива. В самом деле, выполним стандартную упрощенную подстановку: $u_n \to 1$, $u_{n\pm1} \to \exp(\pm \mathrm{i} q h)$, $\hat{u}_n \to \rho_q$. Получим множитель роста q-й гармоники:

$$\rho_q = 1 - \mathrm{i}\frac{c\tau}{h}\sin(qh)\,,\ |\rho_q|^2 = 1 + (c\tau/h)^2 \sin^2(qh). \tag{3.20}$$

Таким образом $|\rho_q| > 1$ для всех гармоник при любом числе Куранта, т. е. схема действительно абсолютно неустойчива.

3.1.4. Монотонность схем

В п. 3.1.1 отмечалось, что решение однородного уравнения переноса (3.2), соответствующее монотонным начальным данным, в любой момент времени имеет монотонный профиль. Сохраняется ли это свойство у разностного решения? Иными словами, пусть профиль u_n монотонен; будет ли монотонным профиль $\hat{u}_n$?

Определение 3.1. Однородные разностные схемы, сохраняющие монотонность профиля разностного решения, называются *монотонными.* •

Теорема 3.1. Для монотонности явной двуслойной линейной однородной схемы

$$\hat{u}_n = \sum_m \beta_m u_{n+m} \tag{3.21}$$

необходимо и достаточно, чтобы для всех m выполнялось условие $\beta_m \ge 0$. •

Доказательство. Из (3.21) следует равенство

$$\hat{u}_{n-1} - \hat{u}_n = \sum_m \beta_m\,(u_{n-1+m} - u_{n+m}). \tag{3.22}$$

Если профиль u_n монотонен (для определенности — невозрастающий), то все скобки в правой части (3.22) неотрицательны. Тогда если все $\beta_m \geq 0$, то $\hat{u}_{n-1} - \hat{u}_n \geq 0$, и профиль $\hat{u}_n$ также невозрастающий. Достаточность условия $\beta_m \geq 0$ доказана.

Предположим, что хотя бы один коэффициент $\beta_k < 0$. Выберем такой невозрастающий профиль:

$$u_{n+m} = 1 \text{ при } m \leq k-1,$$
$$u_{n+m} = 0 \text{ при } m \geq k.$$

Подставляя его в (3.22), получим

$$\hat{u}_{n-1} - \hat{u}_n = \beta_k \left(u_{n-1+k} - u_{n+k}\right) = \beta_k < 0.$$

Монотонность нарушена: имеется локальное возрастание профиля $\hat{u}_n$. Это доказывает необходимость. ■

Замечания. ***1.*** Признак монотонности относится к разностным схемам, аппроксимирующим как уравнение переноса, так и любые другие типы уравнений.

2. Если двуслойная линейная однородная схема неявна, то ее можно преобразовать к явной форме (3.21), где пределы суммы по m бесконечны, и затем применить теорему 3.1.

Теорема 3.2. Двуслойная линейная монотонная схема для уравнения переноса $u_t + cu_x = 0$ не может иметь порядок точности $p \geq 2$. •

Доказательство. Предположим, что имеется линейная монотонная схема порядка точности $p \geq 2$. Запишем ее в форме (3.21), где все $\beta_m \geq 0$. Построим равномерную сетку $x_n = nh$. Выберем в качестве начальных данных задачи Коши квадратичную функцию

$$u(x,0) = \left(\frac{x}{h} - \frac{1}{2}\right)^2 - \frac{1}{4},\ u_n = \left(n - \frac{1}{2}\right)^2 - \frac{1}{4} \geq 0. \qquad (3.23)$$

В этом случае решение есть также квадратичная функция, и его третьи производные равны нулю. Невязка схем второго порядка точности выражается через третьи производные. Поэтому при квадратичных начальных данных (3.23) разностное решение для нашей схемы должно совпадать с точным решением.

На первом слое точное и разностное решения равны соответственно

$$\hat{u}(x,\tau) = \left(\frac{x - c\tau}{h} - \frac{1}{2}\right)^2 - \frac{1}{4},\ \hat{u}_n = \left(n - \frac{c\tau}{h} - \frac{1}{2}\right)^2 - \frac{1}{4}. \qquad (3.24)$$

Подставляя разностные решения на исходном (3.23) и новом (3.24) слоях в разностную схему (3.21), получим равенство

$$\left(n - \frac{c\tau}{h} - \frac{1}{2}\right)^2 - \frac{1}{4} = \sum_m \beta_m \left[\left(n + m - \frac{1}{2}\right)^2 - \frac{1}{4}\right].$$

В правой части этого равенства стоит неотрицательная величина. Но левая часть при нецелом $c\tau/h$ в одной из точек x_n отрицательна. Полученное противоречие доказывает теорему. ■

Примеры. *1.* Явная схема (3.8) легко записывается в форме (3.21). Нетрудно проверить, что при выполнении условия устойчивости Куранта $c\tau \le h$ ее коэффициенты неотрицательны. Следовательно, она монотонна.

2. Безусловно устойчивая схема (3.10) неявная. Запишем ее в следующем виде:

$$\hat{u}_n = \frac{1}{h + c\tau}\left(c\tau \hat{u}_{n-1} + h u_n\right). \tag{3.25}$$

Уменьшая индексы на единицу, получим выражение $\hat{u}_{n-1}$ через $\hat{u}_{n-2}$. Подставим его в правую часть (3.25). Продолжая процедуру уменьшения индекса, приведем схему к явной форме:

$$\hat{u}_n = \frac{h}{h + c\tau} \sum_{m=0}^{\infty} \left(\frac{c\tau}{h + c\tau}\right)^m u_{n-m}. \tag{3.26}$$

Все коэффициенты здесь положительны; следовательно, схема (3.10) монотонна при любых τ и h.

3. Схема (3.11) линейна, и на достаточно гладких решениях имеет второй порядок точности. Следовательно, она немонотонна.

Теорема 3.2 ставит перед вычислителем альтернативу: вести расчет по схеме высокой точности ($p \ge 2$), с риском нарушить монотонность профиля решения, или перейти на монотонные схемы малой точности ($p = 1$). Сделать выбор помогает пример, показанный на рис. 3.8. Различие монотонных и немонотонных схем особенно четко проявляется при расчетах задач с разрывными точными решениями (жирная линия).

Расчет по монотонной схеме (3.10) дает сглаженное разностное решение (кружки), а расчет по немонотонной схеме (3.11) — характерную «разболтку» — пилообразный профиль. Несмотря

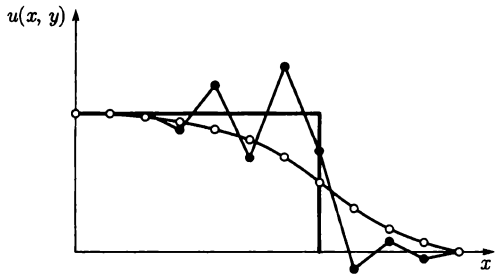


Рис. 3.8. Расчет разрывного решения по монотонной (кружки) и немонотонной (точки) схемам

на «разболтку», устойчивость не теряется: амплитуда «пилы» не увеличивается со временем.

Сходную «разболтку» могут давать немонотонные схемы даже на гладких решениях, если имеются участки с большой кривизной или большим градиентом профиля. Такие задачи целесообразно решать с помощью монотонных схем, несмотря на их невысокую точность $O(\tau+h)$. Обычно немонотонность не возникает, если $|hu_{xx}/u_x| \ll 1$. Таким образом, немонотонность фактически проявляется при расчетах на довольно грубых сетках. Поэтому для получения монотонного разностного решения надо либо считать по немонотонным схемам с достаточно малым шагом, либо переходить на расчет по монотонным схемам.

Теорема 3.2 строго доказана только для линейных схем. Были попытки построения нелинейных схем, обеспечивающих монотонность и претендующих на второй порядок точности. Однако они оказались неудачными. Расчеты на сгущающихся сетках показали, что эффективный порядок точности этих схем на грубых сетках может быть близким к второму, но при $h \to 0$ стремится к первому.

Более перспективно направление, связанное с использованием схем третьего порядка точности. Их фактическая немонотонность на разрывных решениях оказалась слабее, чем у схем второго порядка точности: амплитуда «разболтки» меньше и «разболтка» захватывает меньшее число интервалов вокруг разрыва.

3.1.5. Диссипативность схем

Немонотонность схем проявляется в появлении пилообразных профилей (см. рис. 3.8). Эта пила представляет собой высокочастотное колебание. Напомним, что при исследовании устойчиво-

сти методом гармоник мы разлагаем решение в сумму Фурье. При аккуратном рассмотрении надо учитывать, что на ограниченной сетке узлов $0 \le n \le N$ существует лишь конечное число гармоник. На равномерной сетке они имеют вид

$$\exp\left(\mathrm{i}\pi q n/N\right),\ 0 \le q \le N-1; \tag{3.27}$$

$q = 0$ соответствует профилю $u_n = \mathtt{const}$. Зубцы пилообразного профиля соответствуют значениям q, близким к N.

У точного решения однородного уравнения переноса любая гармоника не нарастает и не убывает. А как воздействуют на гармоники различные схемы? Это описывается множителями роста ρ_q. Сравним различные схемы.

У явной схемы (3.8) при выполнении условия устойчивости $c\tau \le h$ нулевая гармоника имеет $\rho_0 = 1$, как и в точном решении. Для всех остальных допустимых гармоник значения ρ_q лежат на верхней внутренней полуокружности на рис. 3.3. Если выполняется строгое неравенство $c\tau < h$, то амплитуды всех остальных гармоник строго убывают: $\rho_q < 1$; затухают все гармоники, кроме нулевой. В этом случае выполняется более сильное утверждение: $1 = |\rho_0| > |\rho_1| > |\rho_2| > \ldots > |\rho_{N-1}|$. Это означает, что чем выше гармоника, тем быстрее затухает ее амплитуда.

Аналогичными свойствами, только при других значениях числа Куранта, обладает неявная схема (3.9). А у чисто неявной схемы (3.10) при любых соотношениях шагов амплитуды гармоник затухают тем быстрее, чем больше q.

Схемы с затуханием гармоник называют ***диссипативными***; говорят также, что они обладают ***аппроксимационной вязкостью***. Подавление высоких гармоник избавляет расчетные профили от пилообразности и делает схемы монотонными. Однако это улучшение качественного вида решения одновременно связано с ухудшением количественной точности, поскольку в точном решении гармоники не затухают.

У симметричной схемы (3.11) для всех гармоник $|\rho_q| = 1$, т. е. ни одна из гармоник не затухает. Высокие гармоники не подавляются. Однако множители ρ_q комплексны и отличны друг от друга. Это означает, что сеточные гармоники сдвинуты по фазе относительно точных, причем величина сдвига фазы зависит от номера гармоники q. Эти неодинаковые фазовые сдвиги гармоник в отсутствие затухания и отвечают за появление немонотонности.

3.1.6. Перенос с поглощением

Для неоднородного уравнения переноса (3.2) от способа аппроксимации правой части $f(x,t)$ зависит только порядок аппроксимации. Для получения второго порядка надо в качестве ϕ_n брать значения $f(x,t)$ в центре ячейки, а для первого порядка — в любой точке ячейки. На устойчивость это не влияет.

Положение меняется, если правая часть f зависит от u. Рассмотрим это на примере простейшей линейной зависимости $f = -bu$, что в задачах физики соответствует поглощению частиц. Уравнение переноса с поглощением принимает вид:

$$u_t + cu_x = -bu. \tag{3.28}$$

Будем искать решение этого уравнения в виде $u(x,t) = v(x,t)\exp(-bt)$. Подставляя его в (3.28), получим для $v(x,t)$ однородное уравнение переноса $v_t + cv_x = 0$, общее решение которого является бегущей волной $v(x,t) = v(x-ct,0)$. Следовательно, общее решение задачи Коши для уравнения (3.28) имеет вид

$$u(x,t) = \exp(-bt)u(x-ct,0). \tag{3.29}$$

Оно описывает перенос частиц по характеристикам при наличии поглощения (если $b > 0$) или размножения (если $b < 0$).

Дальше ограничимся случаем $b > 0$, когда точное решение экспоненциально убывает со временем. Рассмотрим два варианта явной схемы (3.8):

$$\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{c}{h}(u_n - u_{n-1}) = -b\hat{u}_n, \tag{3.30}$$

$$\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{c}{h}(u_n - u_{n-1}) = -bu_n. \tag{3.31}$$

Они отличаются только аппроксимацией члена с поглощением. Обе схемы имеют первый порядок аппроксимации. Исследуем их устойчивость методом гармоник.

Делая стандартную подстановку гармоники $\exp(\mathrm{i}qx)$, получим для схемы (3.30) множитель роста

$$\rho_q = \left[1 - \frac{c\tau}{h}\left(1 - e^{-\mathrm{i}qh}\right)\right]/(1+b\tau).$$

Если выполнено условие Куранта $c\tau \le h$, то для любых гармоник справедливо неравенство $|\rho_q| \le (1+b\tau)^{-1} < 1$, так что схема

(3.30) не только устойчива, но и хорошо обусловлена: ошибки не нарастают, а при $t \to \infty$ неограниченно убывают.

Для схемы (3.31) множитель роста есть

$$\rho_q = 1 - b\tau - \frac{c\tau}{h}\left(1 - e^{-\mathrm{i}qh}\right).$$

Геометрическое положение чисел ρ_q в комплексной плоскости можно понять, если на рис. 3.3 сдвинуть внутреннюю окружность влево на величину $b\tau$. Видно, что самой опасной будет гармоника с $\exp(-\mathrm{i}qh) = -1$. Для нее $\rho_q = 1 - 2c\tau/h - b\tau$. На грани устойчивости $c\tau = h$, так что $|\rho_q| = 1 + b\tau$. Схема оказывается формально устойчивой, но плохо обусловленной. Таким образом, характер устойчивости схем (3.30) и (3.31) является не вполне одинаковым.

Это различие проявляется сильнее, если рассмотреть асимптотическую устойчивость схем (соответствующую поведению относительной погрешности $\|\delta u\| / \|u\|$ при $t \to \infty$). Точное решение убывает как $e^{-b\tau}$, так что его гармоники за один шаг затухают как $e^{-b\tau} \approx (1 + b\tau)^{-1}$. Гармоники схемы (3.30) затухают с той же скоростью, так что схема (3.30) асимптотически устойчива при $c\tau \le h$. Наоборот, у схемы (3.31) при $c\tau = h$ нет асимптотической устойчивости: гармоника с $\exp(\mathrm{i}qh) = -1$ не только не убывает, а даже возрастает.

Этот пример показывает, что на устойчивость может влиять способ аппроксимации не только высших производных данного уравнения, но и всех остальных членов. Из схем (3.30) — (3.31) и (3.8) — (3.11) угадывается также следующая тенденция: чем неявнее схема (т. е. чем больше величин взято на новом слое), тем лучшей устойчивости можно ожидать.

Замечание. В физических задачах о поглощении начальные данные положительны и общее решение (3.29) также положительно. Нетрудно показать, что схема (3.30) сохраняет это свойство общего решения, и вдобавок монотонна. Если же схему (3.31) переписать в форме

$$\hat{u}_n = \left(1 - \frac{c\tau}{h} - b\tau\right) u_n + \frac{c\tau}{h} u_{n-1},$$

то видно, что при достаточно большом коэффициенте $b > 0$ и не слишком малом шаге τ монотонность может нарушиться, а численное решение — стать отрицательным. Фактически это приводит к дополнительному ограничению на шаг τ схемы (3.31). В задачах с сильным поглощением это ограничение, формально не связанное с устойчивостью, может оказаться достаточно жестким.

3.1.7. Многомерность

Схемы бегущего счета естественно обобщаются на многомерное уравнение переноса. Рассмотрим, для определенности, задачу с двумя пространственными переменными в области $G = [0 \le x \le a] \times [0 \le y \le b] \times [0 \le t \le T]$:

$$\frac{\partial u}{\partial t} + c_x \frac{\partial u}{\partial x} + c_y \frac{\partial u}{\partial y} = f(x, y, t); \tag{3.32}$$

$$u(x, y, 0) = \mu_1(x, y), \; u(0, y, t) = \mu_2(y, t), \; u(x, 0, t) = \mu_3(x, t). \tag{3.33}$$

Скорости переноса по осям c_x, c_y считаем положительными и, для простоты, постоянными.

Построим, например, многомерный аналог абсолютно устойчивой схемы (3.10). Введем по переменной x сетку $\{x_n, 0 \le n \le N\}$, а по переменной y — сетку $\{y_m, \; 0 \le m \le M\}$. Значения решения в узлах этой сетки обозначим следующим образом:

$$u(x_n, y_m, t) = u_{nm}, \; u(x_n, y_m, t + \tau) = \hat{u}_{nm}. \tag{3.34}$$

Возьмем шаблон, изображенный жирными линиями на рис. 3.9, и составим на нем схему

$$\frac{1}{\tau}(\hat{u}_{nm} - u_{nm}) + \frac{c_x}{h_x}(\hat{u}_{nm} - \hat{u}_{n-1,m}) + \frac{c_y}{h_y}(\hat{u}_{nm} - \hat{u}_{n,m-1}) = \phi_{nm}, \tag{3.35}$$

где h_x, h_y — шаги по соответствующим направлениям. В качестве ϕ_{nm} целесообразно брать $f(x, y, t)$ в центре ячейки.

Исследовать схему (3.35) несложно. Из принципа максимума сразу следует безусловная устойчивость этой схемы. Ее невязка определяется разложением по формуле Тейлора и равна $O(\tau + h_x + h_y)$. Следовательно, схема (3.35) сходится в $\|\cdot\|_c$ с первым порядком точности.

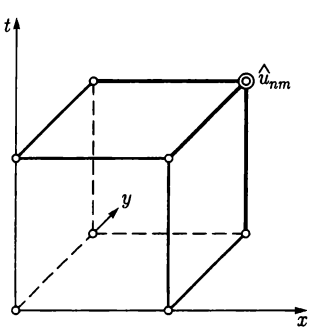


Рис. 3.9. Состав схемы по шаблону

Вычисления проводятся послойно. Значение $\hat{u}_{nm}$ в узле, отмеченном на

рис. 3.9 двойным кружком, выражается по формуле (3.35) через значения в нескольких других вершинах ячейки. Когда решение на слое t вычислено, то его значения на слое $\hat{t}$ можно вычислять по этой формуле вдоль направлений x (рис. 3.10, *а*), где последовательность вычислений указана стрелками.

Заметим, что последовательность вычислений может быть иной. Например, можно вести расчет на слое вдоль направления y (рис. 3.10, *б*). В принципе не обязательна даже послойная организация расчета, достаточно, чтобы последовательность расчета соответствовала какому-то порядку заполнения первого координатного угла в пространстве (x, y, t) ячейками, при котором новая ячейка прикладывается тремя гранями к ранее уложенным ячейкам или координатным плоскостям.

Двумерный аналог симметричной схемы (3.11), имеющий второй порядок точности, нетрудно написать интегро-интерполяционным методом. Для этого проинтегрируем уравнение (3.32) по ячейке, преобразуем трехкратные интегралы в двукратные и вычислим последние по формуле трапеций. Детали настолько очевидны, что мы на них не будем останавливаться. Алгоритм расчета здесь точно такой же, как для чисто неявной схемы, аппроксимация имеет второй порядок по всем переменным (для решений с непрерывными третьими производными), а устойчивость доказывается методом гармоник.

Таким образом, в линейном уравнении переноса многомерность не приводит к принципиальным усложнениям. Вычислительный алгоритм остается простым и экономичным. В декартовых координатах даже формулы расчета имеют обычно простой вид, хотя в криволинейных координатах они могут быть громоздкими.

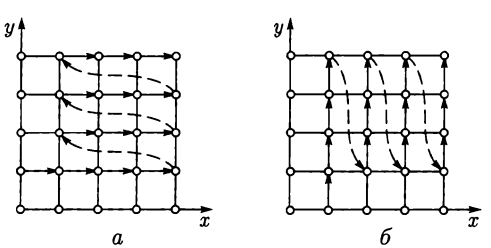


Рис. 3.10. Различная последовательность вычислений:

а — на слое вдоль направления x; *б* — на слое вдоль направления y

3.2. КВАЗИЛИНЕЙНОЕ УРАВНЕНИЕ ПЕРЕНОСА

3.2.1. Сильные и слабые разрывы

Решение линейного уравнения переноса может иметь разрывы только в том случае, если они содержатся в начальных или граничных данных. В квазилинейном уравнении при многократно дифференцируемых начальных данных могут возникать разрывы решения. Характер этих разрывов удобно исследовать на простейшем квазилинейном уравнении переноса (которое подробно изучил ван Хопф):

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0,\ x, t, u > 0. \tag{3.36}$$

Оно напоминает линейное уравнение переноса, в котором роль скорости переноса играет величина самого решения $u(x,t)$.

Полная постановка задачи при знакопеременной скорости сложна; мы рассмотрим только наиболее важный случай $u > 0$. Для простоты мы также ограничимся задачей Коши для полуплоскости $t \geq 0$, когда заданы только начальные данные $u(x,0)$ на прямой $-\infty < x < \infty$. Все последующие рассуждения нетрудно перенести на смешанную задачу Коши в первом квадранте.

Начальные значения переносятся по характеристикам $x - ut = \text{const}$. Их тангенс угла наклона (крутизна) в каждой точке плоскости равен $1/u(x,t)$. Рассмотрим характер решения при четырех основных типах начальных данных.

Гладкие решения. Начальное значение $u(x,0)$ — непрерывная монотонно неубывающая функция. В данном случае наклон характеристик монотонно и непрерывно убывает слева направо. Поэтому верхняя полуплоскость всюду плотно покрыта характеристиками (рис. 3.11), причем через каждую ее точку проходит одна и только одна характеристика. Эта характеристика переносит в данную точку начальное значение. Решение однозначно определено и непрерывно во всей верхней полуплоскости. Если начальные значения гладки, то решение также будет гладким.

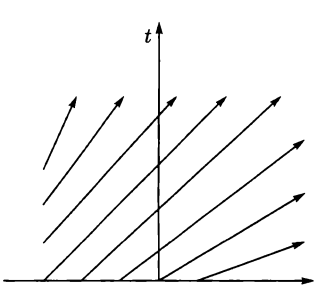


Рис. 3.11. Гладкое решение

Слабые разрывы. Начальные данные монотонно невозрастающие, но имеют разрывы. Для простоты положим $u(x,0)=a$ при $x<0$ и $u(x,0)=b$ при $x>0$; здесь $a<b$, так что разрыв не нарушает условия неубывания.

Левее разрыва характеристики на плоскости (x,t) имеют наклон $1/a$, а правее разрыва — меньший наклон $1/b$ (рис. 3.12, *а*).

Проведем обе характеристики из точки разрыва начальных данных; на рисунке они показаны жирными стрелками. Левее левой и правее правой из них через каждую точку плоскости проходит одна и только одна характеристика, т. е. решение определено и единственно. Но между ними нет ни одной характеристики, и решение не определено.

Потребуем корректности задачи, т. е. устойчивости решения относительно бесконечно малых возмущений начальных данных. Это позволит нам доопределить решение. Сгладим разрыв начальных данных, заменив его непрерывным монотонным переходом на бесконечно узком интервале. Тогда в пустом угле появится «веер» характеристик и наклон каждой характеристики определит значение решения на ней (рис. 3.12, *б*).

Легко видеть, что доопределенное решение будет иметь следующий вид:

$$\begin{gathered} u(x,t)=a \text{ при } x\le at,\\ u(x,t)=x/t \text{ при } at\le x\le bt,\\ u(x,t)=b \text{ при } bt\le x. \end{gathered} \tag{3.37}$$

Поэтому оно непрерывно на всей верхней полуплоскости, кроме точки $x=0$, $t=0$. Значит, такой разрыв начальных данных

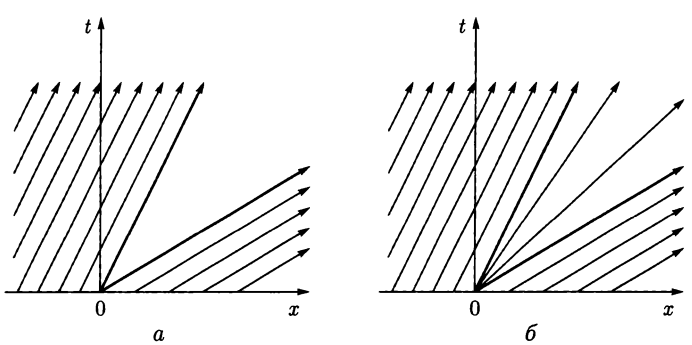


Рис. 3.12. Слабые разрывы:

а — различный угол; *б* — появление «веера» характеристик в зоне разрыва

сглаживается со временем. Но след начального разрыва остается: на жирных характеристиках производные решения будут разрывны. Во всех остальных точках решение будет гладким, если начальные данные были гладкими.

Разрыв самого решения называют ***сильным разрывом***, а разрыв его производных — ***слабым разрывом*** решения. Слабые разрывы квазилинейного уравнения переноса распространяются по характеристикам, как и в линейном уравнении переноса.

Решение (3.37) нередко описывают так: сильный разрыв начальных данных распался на два слабых разрыва решения.

Сильный разрыв. Пусть начальные данные являются убывающими. Опять возьмем $u(x,0) = a$ при $x < 0$ и $u(x,0) = b$ при $x > 0$, но теперь $a > b$. Тогда характеристики будут иметь вид, изображенный на рис. 3.13. В угле, образованном жирными характеристиками, через каждую точку проходят две характеристики, приносящие в нее разные значения функции. Вне этого угла решение однозначно определено, а внутри угла оно неоднозначно.

В этом случае непрерывное решение построить не удается. Сглаживание разрыва начальных данных не помогает: ход характеристик на некотором расстоянии от точки $x = 0$, $t = 0$ все равно не меняется, так что неоднозначность остается. Значит, однозначное решение должно быть разрывным, т. е. оно будет обобщенным решением дифференциального уравнения.

Обобщенное решение удовлетворяет некоторому интегральному уравнению, которое получается из определенной дивергентной формы записи данного дифференциального уравнения. Разные дивергентные формы записи одного и того же уравнения приводят к разным разрывным решениям, хотя гладкие решения для всех дивергентных форм одинаковы. Дивергентная форма, соответствующая физическому закону сохранения, определяет правильное решение (его называют также допустимым).

Уравнение (3.36) не имеет физического смысла, и естественного закона сохранения для него нет. Постулируем такую дивергентную форму:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0. \tag{3.38}$$

Будем искать решение, имеющее единственный разрыв. Пусть наклон линии разрыва соответствует скорости D, т. е. разрыв

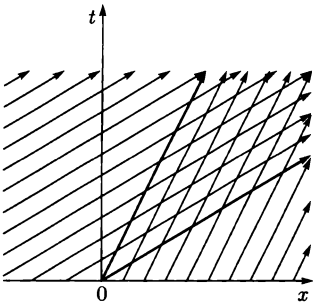


Рис. 3.13. Сильный разрыв

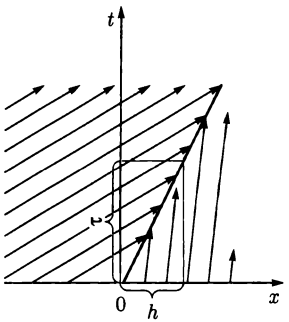


Рис. 3.14. Решение уравнения (3.38)

бежит, как волна. По поведению характеристик видно (рис. 3.14), что искомое решение имеет вид

$$\begin{aligned} u(x,t) &= a \text{ при } x < Dt, \\ u(x,t) &= b \text{ при } x > Dt. \end{aligned} \tag{3.39}$$

Проинтегрировав (3.38) по площади прямоугольника со сторонами τ и $h = D\tau$, диагональю которого является линия разрыва, получим

$$\int (\hat{u} - u)\,dx + \frac{1}{2}\int \left(u^2_{\text{прав}} - u^2_{\text{лев}}\right) dt = (a-b)h + \frac{1}{2}\left(b^2 - a^2\right)\tau = 0.$$

Отсюда скорость распространения разрыва равна

$$D = \frac{1}{2}(a+b). \tag{3.40}$$

Таким образом, построено обобщенное решение с сильным разрывом (в газодинамике сильный разрыв называют ***ударной волной***). В теории квазилинейных уравнений доказывают, что только такое решение является устойчивым относительно возмущения начальных данных. Сильный разрыв распространяется не по характеристике.

При другом выборе дивергентной формы (закона сохранения) мы также получим сильный разрыв, но движущийся с другой скоростью.

Общий случай. Пусть $u(x,0)$ сколь угодно гладкя, но убывающая. Тогда характеристики будут пересекаться, а это приводит к образованию сильных разрывов. Местная скорость дви-

жения разрыва будет определяться по формуле типа (3.40) приносимыми в данную точку значениями решения и уже не будет постоянной. Существенно, что здесь при непрерывных и гладких начальных данных с течением времени возникают сильные разрывы решения. Число разрывов со временем также может изменяться.

3.2.2. Однородные схемы

На недостаточно гладких решениях разностные схемы могут иметь меньший порядок аппроксимации и точности. Особенно сильно ухудшается точность расчета, если решение содержит сильные или слабые разрывы; далее увидим, что некоторые схемы при этом приводят даже к грубо ошибочным результатам. Для расчета подобных задач строят два типа схем: схемы с явным выделением особенностей и так называемые однородные схемы.

Однородные схемы, называемые также схемами сквозного счета, строят следующим образом. Шаблон и разностные аналоги производных выбирают так, чтобы нужная аппроксимация была обеспечена всюду, в том числе и на особенностях решения. Поэтому весь расчет ведется по однотипным разностным уравнениям без явного выделения особенностей. Например, для линейного уравнения переноса схемы (3.8) — (3.10) позволяют неплохо рассчитывать решение с разрывами начальных данных (см. рис 3.8) без явного выделения этого разрыва.

Однородные схемы не громоздки, требуют умеренного объема вычислений, и каждая хорошая схема пригодна для широкого класса задач. Компьютерные программы, составленные на их основе, также позволяют без заметных переделок рассчитывать широкий круг задач. По точности они уступают схемам с явным выделением особенностей, но благодаря своей простоте широко используются в практике расчетов. По однородным схемам успешно проводят расчеты даже таких сложных задач, как задачи многомерной магнитной газодинамики, в которых возникает большое число ударных, тепловых и других волн, являющихся разрывами.

Выделение особенностей. В задачах, где требуется особо высокая точность, используют схемы с явным выделением особенностей. В них каждую особенность решения выделяют,

и по отдельным формулам рассчитывают ее движения. В промежутках между особенностями решение непрерывно и достаточно гладко; в этих промежутках дифференциальное уравнение аппроксимируют разностной схемой. Уравнения, описывающие особенности, служат своеобразными внутренними краевыми условиями, связывающими между собой разностные уравнения в соседних промежутках.

Особенности решения могут быть связаны с разрывами или нарушением гладкости начальных данных и коэффициентов уравнения, с возникновением ударных волн, с образованием слабых разрывов при столкновении ударной волны с какой-либо особенностью решения. Число особенностей с течением времени может меняться. К каждому типу особенностей нужен индивидуальный подход. Очевидно, явно учесть все особенности можно только в наиболее простых задачах, поэтому подобные схемы достаточно сложны и каждая схема рассчитана лишь на узкий круг задач.

3.2.3. Ложная сходимость

Для нелинейных задач отсутствуют строгие доказательства устойчивости и тем самым сходимости. Для таких случаев в п. 2.4.3 рекомендовалось доказать аппроксимацию, а устойчивость проверить, проводя расчеты на последовательности сгущающихся сеток. Стремление численного решения к предельной функции при $\tau \to 0, h \to 0$ свидетельствует об устойчивости схемы. Тогда из теоремы 2.7 делается вывод о сходимости.

Эти рассуждения справедливы для достаточно гладких решений, где разложением в ряды Тейлора можно надежно доказать аппроксимацию. Если же решение имеет сильные или слабые разрывы, то локальной аппроксимации в точках разрыва нет и предыдущие рассуждения могут привести к неверному результату.

Пример. Построим схему, которая дает сходимость к пределу, отличающемуся от точного решения. Для квазилинейного уравнения переноса построим явную схему на шаблоне рис. 3.2, *б*, аналогичную схеме (3.8). В качестве значения скорости c возьмем значение $u(x_n, t)$. Получим следующую схему:

$$\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{1}{h}u_n(u_n - u_{n-1}) = 0,\ x_n = nh,\ -\infty < n < \infty. \tag{3.41}$$

Это однородная схема. В ней нет явного выделения особенностей решения, т. е. предусматривается сквозной расчет.

Проведем по схеме (3.41) расчет сильного разрыва (3.39). Начальные данные на сетке для определенности зададим следующим образом:

$$\begin{aligned} u_n(0) &= a \text{ при } n \le 0, \\ u_n(0) &= b \text{ при } n \ge 1. \end{aligned} \tag{3.42}$$

Выберем соотношение шагов $h/\tau = b$. Подставляя (3.42) в (3.41), нетрудно убедиться, что на первом слое разностное решение будет равно

$$\begin{aligned} \hat{u}_n &= a \text{ при } n \le 1, \\ \hat{u}_n &= b \text{ при } n \ge 2. \end{aligned} \tag{3.43}$$

Таким образом, за один шаг по времени разрыв продвинулся ровно на один шаг по пространству и сохранил свою форму. Очевидно, то же будет и на всех других шагах по времени. Следовательно, сеточный разрыв (3.42) будет сохранять свою форму и двигаться со скоростью b.

Будем сгущать сетку (т. е. $\tau \to 0, h \to 0$), сохраняя указанное отношение $h/\tau = b$. Тогда сеточный разрыв стремится к разрывной функции, но этот разрыв движется со скоростью b, а не $D = (a+b)/2$. Имеет место сходимость к предельной функции, не являющейся искомым точным решением!

Это явление называется ложной сходимостью. Причина в данном случае очевидна. На фронте разрыва аппроксимацию нельзя исследовать разложением в ряды Тейлора. Такая ситуация возникает, когда ищутся обобщенные решения дифференциального уравнения. В этих случаях следует помнить, что:

1) для недостаточно гладких и тем более разрывных решений необходимо особое внимание уделять обоснованию аппроксимации;

2) если аппроксимация обоснована недостаточно аккуратно, то полагаться на визуальную сходимость численных расчетов к предельной функции при сгущении сеток опасно.

3.2.4. Консервативные схемы

Как построить однородную схему, чтобы избежать ложной сходимости на обобщенных решениях? Это делают с помощью ***консервативных*** разностных схем. Эти схемы составляют интегро-интерполяционным методом, исходя из физических зако-

нов сохранения и соблюдая дополнительное правило, описанное ниже.

Сначала разберем законы сохранения на примере уравнения (3.36). Запишем ту дивергентную форму этого уравнения (3.38), которая в п. 3.1.1 была условно принята за правильную:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0. \tag{3.44}$$

Интегрируя по одной ячейке сетки и точно беря интегралы от производных, получим точное интегральное соотношение:

$$\int\limits_{x_{n-1}}^{x_n} (\hat{u} - u)\, dx + \frac{1}{2}\int\limits_{t}^{\hat{t}} \left(u_n^2 - u_{n-1}^2\right) dt = 0. \tag{3.45}$$

Уравнение (3.44) можно проинтегрировать не по отдельной ячейке, а по всей области. Для смешанной задачи Коши это будет область

$$G = [x_0 \le x \le x_N] \times [t_0 \le t \le t_K].$$

Получим аналогичное интегральное соотношение:

$$\int\limits_{x_0}^{x_N} \left(u^K - u^0\right) dx + \frac{1}{2}\int\limits_{t_0}^{t_K} \left(u_N^2 - u_0^2\right) dt = 0. \tag{3.46}$$

Это соотношение напоминает физические законы сохранения: первый интеграл дает изменение $\int u dx$ за истекшее время, а второй есть разность потоков $\frac{1}{2}\int u^2 dt$ через правую и левую границы. Соотношение (3.46) является законом сохранения для данной задачи.

Очевидно, соотношение (3.45) является законом сохранения для каждой отдельной ячейки; оно содержит потоки и другие величины на границах этой ячейки. Просуммируем (3.45) по всем ячейкам области G:

$$\sum_{n=1}^{N}\sum_{k=0}^{K-1}\left[\int\limits_{x_{n-1}}^{x_n}\left(u^{k+1} - u^k\right) dx + \frac{1}{2}\int\limits_{t_k}^{t_{k+1}} \left(u_n^2 - u_{n-1}^2\right) dt\right] = 0. \tag{3.47}$$

Легко видеть, что интегралы по тем границам ячеек, которые лежат внутри G, попарно уничтожаются; остаются только интегралы по наружной границе, и (3.47) совпадает с (3.46). Иными словами, закон сохранения во всей области есть ***точное*** следствие закона сохранения в отдельных ячейках.

Схемы, обладающие таким свойством, называют ***консервативными***.

Неконсервативность. Не всякая схема является консервативной. Например, возьмем схему с ложной сходимостью (3.41). Перепишем ее в следующем виде:

$$\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{1}{2h}\left(u_n^2 - u_{n-1}^2\right) + \frac{1}{2h}(u_n - u_{n-1})^2 = 0. \qquad (3.48)$$

Просуммируем (3.48) по всем ячейкам области G аналогично (3.47); при этом надо умножить на τ и h, что эквивалентно интегрированию. При суммировании первого слагаемого по слоям все значения на внутренних слоях сокращаются, оставляя только разность значений u на последнем и нулевом слоях. При суммировании второго слагаемого аналогично сокращаются члены во внутренних точках и остается только разность потоков в точках x_N и x_0. Если не было бы третьего слагаемого, это означало бы консервативность. Однако третье слагаемое в (3.48) является положительным во всех узлах на всех слоях. Оно ни с чем не сокращается, а его суммирование по области G, с учетом умножения на τ и h, дает остаток

$$\Delta = \frac{\tau}{2}\sum_{n=1}^{N}\sum_{k=0}^{K-1}\left(u_n^k - u_{n-1}^k\right)^2 > 0, \qquad (3.49)$$

называемый ***дисбалансом***. Тем самым суммирование по ячейкам не дает закона сохранения во всей области.

Если бы схема обеспечивала сходимость, то при $h \to 0$, $\tau \to 0$ должно выполняться $\Delta \to 0$. Если решение гладкое или даже со слабыми разрывами, то $u_n - u_{n-1} \approx h u_x$; заменяя суммирование интегрированием, получаем

$$\Delta \approx \frac{h}{2}\iint u_x^2 dx dt = O(h),$$

что не препятствует сходимости. Если же решение имеет сильный разрыв, то на каждом слое найдется интервал, в котором

$u_n - u_{n-1} = O(1)$; тогда суммирование дает $\Delta = O(1)$ и сходимость невозможна.

Построим примеры консервативных схем, у которых дисбаланс равен нулю.

Явная схема. Построим аналог линейной схемы (3.8). В интегральном соотношении (3.45) аппроксимируем интегралы по формуле прямоугольников на шаблоне рис. 3.2, *а*. Получим следующую схему:

$$\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{1}{2h}\left(u_n^2 - u_{n-1}^2\right) = 0. \tag{3.50}$$

Левая часть этой схемы совпадает с первыми двумя слагаемыми неконсервативной схемы (3.48). Третье слагаемое, приводившее к дисбалансу Δ, отсутствует. Поэтому при суммировании по области G мы получаем интегральный закон сохранения, т. е. схема консервативна.

Численные расчеты показывают, что эта схема условно устойчива; для устойчивости во всех точках должно выполняться соотношение $\tau u_n < h$. В этом случае при $\tau \to 0$, $h \to 0$ даже на решениях с сильным разрывом имеет место сходимость к правильному решению.

Чисто неявная схема. Построим аналог линейной схемы (3.10). Аппроксимируем интегралы в (3.45) по формуле прямоугольников на шаблоне рис. 3.2, *б*. Получим следующую схему:

$$\frac{1}{\tau}(\hat{u}_n - u_n) + \frac{1}{2h}\left(\hat{u}_n^2 - \hat{u}_{n-1}^2\right) = 0. \tag{3.51}$$

Это неявная схема, содержащая два значения решения на новом слое. Тем самым она является схемой бегущего счета. Схема нелинейная, так что для нахождения $\hat{u}_n$ требуется решить квадратное уравнение (3.51). Один из корней оказывается отрицательным и отбрасывается; положительным решением является корень

$$\hat{u}_n = \sqrt{\frac{h^2}{\tau^2} + \frac{2h}{\tau}u_n + \hat{u}_{n-1}^2} - \frac{h}{\tau}. \tag{3.52}$$

Нетрудно показать, что при положительных начальных данных всегда будет $\hat{u}_n > 0$.

Легко видеть, что суммирование соотношения (3.51) также приводит к интегральному закону сохранения, т. е. схема кон-

сервативна. Численные расчеты показывают, что схема (3.51) — (3.52) безусловно устойчива, а при $\tau \to 0, h \to 0$ численное решение сходится к точному даже при наличии сильных или слабых разрывов. Эта схема также является монотонной.

Замечания. *1.* Гладкие решения нелинейных уравнений можно считать и по неконсервативным схемам. Однако консервативные схемы выражают закон сохранения на сетке, т. е. они качественно передают свойства исходного интегрального уравнения. Неконсервативные схемы этим свойством не обладают. Поэтому даже на таких решениях, которые можно рассчитывать по неконсервативным схемам, обычно сходные консервативные схемы дают более высокую точность.

2. Для систем нелинейных уравнений могут существовать одновременно несколько различных законов сохранения. Например, для уравнений газодинамики одновременно выполняются законы сохранения массы, импульса и энергии. Для таких систем можно построить схемы, одновременно удовлетворяющие всем этим законам сохранения. Такие схемы называют полностью консервативными.

3.2.5. Псевдовязкость

Основную трудность для вычислений по разностным схемам представляют сильные разрывы решения. Эффективный прием расчета задач с разрывными решениями заключается в следующем. Подберем такую «малую» добавку к исходному уравнению, чтобы его разрывные решения превратились в непрерывные и достаточно гладкие. Тогда составить разностную схему для численного расчета этих гладких решений уже несложно.

Гладкие решения присущи уравнениям с диссипативными членами типа вязкого трения. Поэтому добавляемый в исходное уравнение член должен играть роль вязкости. Его называют ***псевдовязкостью***, а также ***искусственной*** или ***математической вязкостью***. Обычно этот член содержит более высокую производную, чем входящие в исходное уравнение, но с малым коэффициентом. Возможны разные формы вязкого члена. Рассмотрим указанный способ на примере квазилинейного уравнения переноса (3.36).

Линейная вязкость. Добавим в правую часть уравнения (3.36) линейный член со второй производной (его ввел Бюрерс):

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \epsilon\frac{\partial^2 u}{\partial x^2},\ 0 < \epsilon \ll 1. \tag{3.53}$$

Уравнение (3.53) напоминает уравнение теплопроводности, все решения которого неограниченно дифференцируемы. Очевидно, на дважды непрерывно дифференцируемых решениях исходного уравнения (3.36) член ϵu_{xx} мал в силу малости ϵ. Поэтому все достаточно гладкие решения уравнений (3.36) и (3.53) близки.

Рассмотрим, имеет ли уравнение (3.53) автомодельное решение вида бегущей волны

$$u(x,t) = f(\xi),\ \xi = x - Dt,\ D = (a+b)/2, \tag{3.54}$$

а профиль функции f напоминает сильный разрыв (3.39), бегущий со скоростью D; разумеется, этот разрыв должен быть сглаженным. Для этого решение должно удовлетворять уравнению (3.54) и граничным условиям $f(-\infty) = a$, $f(+\infty) = b$; потребуем также, чтобы оно было монотонно убывающим.

Подставляя (3.54) в (3.53), получим для $f(\xi)$ обыкновенное дифференциальное уравнение с граничными условиями на бесконечности:

$$\epsilon f'' + Df' - ff' = 0,\ f(-\infty) = a,\ f(+\infty) = b. \tag{3.55}$$

Первый интеграл от этого уравнения тривиален:

$$\epsilon f' + Df - f^2/2 = \mathtt{const}. \tag{3.56}$$

Поскольку $f(\xi)$ монотонна и асимптотически выходит на константы, то $f'(-\infty) = 0$ и $f'(+\infty) = 0$. Подставляя эти производные в (3.56) и сравнивая с краевыми условиями (3.55), получим $\mathtt{const} = \dfrac{ab}{2}$. Тогда с учетом значения D (3.54) соотношение (3.56) принимает следующий вид:

$$2\epsilon f' - (f-a)(f-b) = 0,\ f(-\infty) = a,\ f(+\infty) = b. \tag{3.57}$$

Легко проверить, что решение последней задачи имеет вид

$$f(\xi) = \frac{a+b}{2} - \frac{a-b}{2}\,\mathrm{th}\left(\frac{a-b}{4\epsilon}\xi\right),\ \xi = x - Dt. \tag{3.58}$$

Полученное решение монотонно, качественно напоминает сильный разрыв (3.39) и при $\epsilon \to 0$ имеет пределом этот сильный разрыв.

Таким образом, уравнение (3.53) действительно имеет гладкие решения, которые при $\epsilon \to 0$ переходят в сильный разрыв

(3.39), движущийся с правильной скоростью. Поэтому для уравнения (3.53) можно составлять разностные схемы, рассчитанные на гладкие решения, и производить по ним расчеты решений с сильными разрывами. Разумеется, в расчете эти разрывы будут сглажены: из (3.58) видно, что эффективная ширина фронта скачка составляет $\delta x \approx 8\epsilon/(a-b)$. Чтобы ширина скачка стремилась к нулю при сгущении сетки, обычно полагают $\epsilon = \mathtt{const} \cdot h$.

Достоинством линейной вязкости является то, что решения уравнения (3.53) многократно дифференцируемы. Поэтому при разумно составленных схемах качественное поведение численного решения будет хорошим (не будет пилообразности). Недостатком является то, что разрывы разной амплитуды сглаживаются неодинаково: разрывы большой амплитуды размываются на малое число пространственных интервалов, а разрывы малой амплитуды — на большое.

Квадратичная вязкость. Добавим в правую часть квазилинейного уравнения переноса не линейный член, а квадратичный:

$$u_t + uu_x = -\frac{\epsilon^2}{2}\frac{\partial}{\partial x}\left(u_x^2\right) \equiv -\epsilon^2 u_x u_{xx},\ \epsilon^2 \ll 1. \tag{3.59}$$

Это уравнение также имеет гладкие решения, похожие на бегущий сильный разрыв (3.39). Подставляя (3.39) в (3.59), получим следующее обыкновенное дифференциальное уравнение:

$$\left(\epsilon^2 f'' + f - D\right) f' = 0,\ f(-\infty) = a,\ f(+\infty) = b. \tag{3.60}$$

Приравнивая каждый из сомножителей нулю, получим два типа решений:

$$f_1(\xi) = \mathtt{const}, f_2(\xi) = D + \mathtt{const}\sin\left(\frac{\xi}{\epsilon}\right) + \mathtt{const}. \tag{3.61}$$

Из них можно сконструировать решение, похожее на размытую волну шириной $\sim \epsilon$:

$$u(x,t) = \begin{cases} a & \text{при } x - Dt < -\dfrac{\pi\epsilon}{2}, \\ \dfrac{a+b}{2} - \dfrac{a-b}{2}\sin\left(\dfrac{x-Dt}{\epsilon}\right) & \text{при } -\dfrac{\pi\epsilon}{2} < x - Dt < \dfrac{\pi\epsilon}{2}, \\ b & \text{при } x - Dt > \dfrac{\pi\epsilon}{2}. \end{cases} \tag{3.62}$$

Справедливость решения (3.62) проверяется непосредственной подстановкой в (3.59). Фронт скачка сглажен на величину $\delta x \approx$ $\approx 2\epsilon$. Поэтому в расчетах полагают $\epsilon = \texttt{const} \cdot h$. Тогда фронт будет сглажен на одно и то же число интервалов сетки, так что при сгущении сетки ($h \to 0$) численное решение будет стремиться к правильному разрывному.

Достоинством квадратичной вязкости является то, что ширина сглаживания фронта не зависит от амплитуды скачка. Недостатком же является то, что решение (3.62) имеет разрывы второй производной в точках $x - Dt = \pm\pi\epsilon/2$. Это может приводить к пилообразности численного решения.

Есть много важных прикладных задач, решения которых имеют сильные разрывы. Например, это система уравнений газодинамики и магнитной газодинамики. В таких случаях часто вводят в уравнения искусственную вязкость. Разумеется, вид этой искусственной вязкости подбирают отдельно для каждого класса уравнений. Выбор этого вида обосновывают аналогичными способами.

Введение искусственной вязкости понижает требование к однородным разностным схемам. Нередко даже неконсервативные схемы позволяют получать удовлетворительные результаты, хотя консервативные схемы обычно надежнее и точнее.

Глава 4

ПАРАБОЛИЧЕСКИЕ УРАВНЕНИЯ

4.1. ОДНОМЕРНЫЕ УРАВНЕНИЯ

4.1.1. Постановки задач

К параболическим уравнениям приводят задачи теплопроводности, диффузии и ряд других. Простейшим случаем является линейное уравнение теплопроводности для однородной среды, заданное в ограниченной области:

$$\begin{gathered} u_t(x,t) = \kappa u_{xx}(x,t) + f(x,t), \\ \kappa = \mathtt{const} > 0,\ 0 < x < a,\ 0 < t \le T. \end{gathered} \tag{4.1}$$

Рассмотрим полную постановку задачи. Наличие первой производной по времени требует постановки одного начального условия, а наличие второй производной по пространству — двух граничных условий. Простейшие условия имеют следующий вид:

$$u(x,0) = \mu(x),\ 0 \le x \le a; \tag{4.2}$$

$$u(0,t) = \mu_1(t),\ u(a,t) = \mu_2(t),\ 0 \le t \le T. \tag{4.3}$$

Формулы (4.1) — (4.3) определяют простейшую смешанную задачу Коши.

Элементы теории. В курсах математической физики хорошо изучены линейные задачи; в них уравнение и краевые условия линейны. Для таких задач рассматривают три типа краевых условий. Условия первого рода (4.3) применительно к уравнению теплопроводности означают, что на границах задана зависимость температуры u от времени. Условия второго рода

$$u_x(0,t) = \mu_1(t),\ u_x(a,t) = \mu_2(t) \tag{4.4}$$

соответствуют заданию тепловых потоков через границы. Условия третьего рода

$$u(0,t) + \alpha_1 u_x(0,t) = \mu_1(t),\ u(a,t) + \alpha_2 u_x(a,t) = \mu_2(t) \tag{4.5}$$

возникают, если на границах имеется линейный (ньютоновский) теплообмен с окружающей средой. Для задачи (4.1) — (4.2) с краевыми условиями (4.3), (4.4) или (4.5) корректность постановки доказана.

Часто встречаются и нелинейные задачи. Например, задача остывания черного тела приводит к нелинейному краевому условию: в (4.5) вместо ньютоновского теплообмена с окружающей средой $u(0,t)$ появляется лучистый теплообмен $u^4(0,t)$. Само уравнение также может стать нелинейным: например, в задачах физики высокотемпературной плазмы коэффициент теплопроводности зависит от температуры как $\kappa(u) \sim u^{5/2}$. В последнем случае само уравнение вместо (4.1) приобретает более сложную форму.

Параболическое уравнения (4.1) имеет важную особенность: разрывы начальных и граничных данных быстро сглаживаются. Если $\mu(x)$, $\mu_1(t)$ или $\mu_2(t)$ разрывны, но $f(x,t)$ неограниченно дифференцируема, то при сколь угодно малом отступлении от границ внутрь области G решение $u(x,t)$ также будет неограниченно дифференцируемым. Это позволяет использовать разложение в ряды Тейлора для исследования аппроксимации.

Упомянем другое интересное свойство уравнения (4.1) на прямой $-\infty < x < +\infty$. Возьмем дельта-функцию в качестве начальных данных: $\mu(x) = \delta(x - x_0)$. В начальный момент тепло имеется только в точке x_0, а во всех остальных точках оно отсутствует. Но уже через сколь угодно малый промежуток времени в любой точке пространства $u(x,t)$ будет отлично от нуля. Поэтому формально при $\kappa = \texttt{const}$ скорость распространения тепла бесконечна.

Другой важный результат следует из однородной задачи (4.1) — (4.3), в которой положено $f(x,t) = 0$, $\mu_1(t) = 0$, $\mu_2(t) = 0$. Методом разделения переменных можно построить точное решение такой задачи:

$$u(x,t) = \sum_{m=1}^{\infty} c_m e^{\lambda_m t} \sin(\pi m x/a),\ \lambda_m = -\kappa\pi^2 m^2/a^2, \qquad (4.6)$$

здесь c_m суть коэффициенты Фурье начальных данных $\mu(x)$. Решение разложено в ряд Фурье по пространственным гармоникам. Гармоники затухают со временем, причем чем больше номер гармоники, тем быстрее она затухает. Затухание высоких гармоник ($m \to \infty$) становится неограниченно быстрым.

Это свидетельствует о жесткости данной задачи. Последнее надо учитывать при выборе разностных схем.

Существует обратная задача теплопроводности: по профилю температуры восстановить температуру в прошлом. Это соответствует расчету в обратном направлении по оси t. Однако при подстановки в (4.6) значений $t < 0$ амплитуды всех гармоник оказываются возрастающими. При $m \to \infty$ этот рост неограниченно велик; высшие гармоники, порожденные рябью начального профиля, неограниченно усиливаются. Это означает, что обратная задача теплопроводности некорректна. Ее можно решать только специальными методами регуляризации.

4.1.2. Простейшие схемы

Метод прямых. Рассмотрим задачу (4.1) — (4.3) и покажем, как можно построить простейшие схемы с помощью метода прямых. Введем в области G равномерную сетку $\{x_n = nh, 0 \leq \leq n \leq N, h = a/N; t_k = k\tau, k = 0, 1, \ldots\}$. Рассмотрим пространственный шаблон, содержащий три точки: x_{n-1}, x_n, x_{n+1}. Заменим в уравнении (4.1) пространственную производную второй разностью, а производную по времени сохраним. Тогда (4.1) заменится системой обыкновенных дифференциальных уравнений для узловых значений решения $u_n(t) = u(x_n, t)$:

$$\begin{gathered} \frac{du_n}{dt} = \frac{\kappa}{h^2}(u_{n-1} - 2u_n + u_{n+1}) + f_n(t), \\ f_n(t) = f(x_n, t),\ 1 \leq n \leq N-1. \end{gathered} \tag{4.7}$$

Систему (4.7) можно писать только во внутренних узлах сетки. Она дополняется граничными условиями, следующими из (4.2) — (4.3):

$$u_0(t) = \mu_1(t),\ u_N(t) = \mu_2(t);\ u_n(0) = \mu(x_n),\ 1 \leq n \leq N-1. \tag{4.8}$$

Формулы (4.7)–(4.8) составляют задачу Коши для обыкновенных дифференциальных уравнений. Такое сведение уравнений в частных производных к системе ОДУ называют *методом прямых.*

Система ОДУ (4.7) аппроксимирует исходное уравнение (4.1). Аппроксимация заключалась в замене второй пространственной производной на вторую разность. Поскольку решение внутри области G неограниченно дифференцируемо, погрешность этой

аппроксимации можно исследовать разложением в ряды Тейлора аналогично п. 1.4.2. Нетрудно проверить, что невязка есть $\psi(x) \approx \dfrac{\kappa h^2}{12} u_{xxxx} = O(h^2)$.

Легко убедиться, что система (4.7) жесткая. Для этого рассмотрим однородную задачу, полагая $f(x,t) = 0$, $\mu_1(t) = 0$, $\mu_2(t) = 0$. Тогда точное решение задачи Коши можно получить методом разделения переменных, аналогично (4.6). Для этого ищем частное решение в виде $du_n/dt = \exp(\nu t)\sin(\pi m x_n/a)$; пространственные гармоники те же, что и в (4.6), но они берутся от дискретного аргумента x_n. Число этих гармоник $(1 \le m \le N-1)$ равно числу внутренних узлов x_n. Подстановка такого решения в (4.7) позволяет получить спектральное число ν_m. Полное решение получается суперпозицией частных решений; оно имеет следующий вид:

$$u_n(t) = \sum_{m=1}^{N-1} \gamma_m e^{\nu_m t} \sin(\pi m x_n/a),$$
$$\nu_m = -\frac{4\kappa}{h^2}\sin^2\left(\frac{\pi m}{2N}\right),\ 1 \le n \le N-1. \tag{4.9}$$

Спектральные значения ν_m при $m \ll N$ близки к соответствующим значениям λ_m; с увеличением m отличие увеличивается и при максимально возможном $m = N-1$ составляет $\sim 1,6$ раза. Отношение наибольшего и наименьшего спектральных чисел

$$\frac{\nu_{N-1}}{\nu_1} = \operatorname{ctg}^2\left(\frac{\pi}{2N}\right) \approx \left(\frac{2N}{\pi}\right)^2 \tag{4.10}$$

на подробных сетках становится большим. Сами числа ν_m отрицательны. Поэтому система ОДУ (4.7) содержит компоненты с сильно отличающимися скоростями затухания, т. е. является жесткой.

Двуслойные схемы. Поскольку система (4.7) жесткая, для ее решения следует применять A-устойчивые методы, описанные в подразделе 1.2. Воспользуемся простым и надежным семейством одностадийных схем Розенброка, описанных в пп. 1.2.3, 1.2.4. Для этого введем следующие обозначения:

$$\mathbf{u}(t) = (u_m(t),\ 1 \le m \le N-1)^T,$$
$$(\Lambda \mathbf{u})_n = \frac{\kappa}{h^2}(u_{n-1} - 2u_n + u_{n+1}). \tag{4.11}$$

Здесь $\mathbf{u}$ есть решение системы ОДУ (4.7), записанное в виде вектора-столбца, а Λ является трехдиагональной матрицей порядка $N-1$, действующей на $\mathbf{u}$; ненулевые элементы этой матрицы соответственно равны $a_{nn} = -2\kappa/h^2$, $a_{n,n\pm1} = \kappa/h^2$. В этих обозначениях систему (4.7) можно записать в канонической форме:

$$d\mathbf{u}/dt = \mathbf{F}(u,x,t) \equiv \Lambda\mathbf{u} + \mathbf{f}. \tag{4.12}$$

В формулах (4.11) — (4.12) нужно полагать $u_0(t) = \mu_1(t)$ и $u_N(t) = \mu_2(t)$ всюду, где встречаются эти граничные величины.

Учтем, что матрица Якоби $\partial\mathbf{F}/\partial\mathbf{u} = \Lambda$. Тогда одностадийная схема Розенброка для системы (4.12) примет следующий вид:

$$\hat{\mathbf{u}} = \mathbf{u} + \tau\mathtt{Re}\mathbf{w};\ (E - \sigma\tau\Lambda)\,\mathbf{w} = \Lambda\mathbf{u} + \mathbf{F}. \tag{4.13}$$

Здесь E есть единичная матрица. Для A-устойчивости необходимо $\mathtt{Re}\sigma \geq 1/2$. В подразделе 1.2 рекомендовались следующие значения σ:

$$\sigma = \frac{1+\mathtt{i}}{2},\ \sigma = 1,\ \sigma = 1/2. \tag{4.14}$$

Алгоритм вычисления сеточного решения одинаков при всех σ. Схема (4.13) при $\sigma \neq 0$ неявная: значения $\mathbf{w}$ определяются из линейной системы, матрица которой $E - \sigma\tau\Lambda$ трехдиагональна. Эта линейная система экономично решается прогонкой или методом Гаусса для ленточной матрицы (см. кн. 1). С учетом приведенных выше коэффициентов матрицы Λ нетрудно проверить, что в матрице системы имеется преобладание диагонального коэффициента. При этом решение линейной системы существует и единственно, а методы прогонки и Гаусса устойчиво вычисляют это решение.

Заметим, что при $\sigma = 0$ схема явная. Но это значение параметра непригодно для расчетов, так как дает схему, не обладающую A-устойчивостью.

Уравнения (4.13) были получены из метода прямых. Однако сами они являются некоторой разностной схемой. Эта схема содержит три точки исходного слоя, а также три точки нового слоя. Тем самым она соответствует шеститочечному шаблону, показанному на рис. 4.1. При вещественном σ

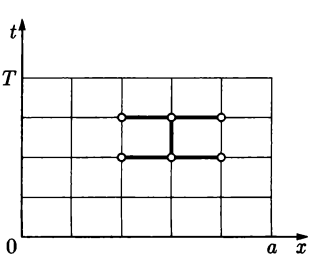


Рис. 4.1. Шаблон схемы с весами

они переходят в хорошо известные схемы с весами: производная u_t заменяется разностью по вертикальному отрезку, а пространственная разность $\Lambda\mathbf{u}$ суммируется с весами σ на новом слое и $1-\sigma$ на исходном слое.

Рассмотрим схемы, соответствующие рекомендованным значениям параметра σ (4.14).

Комплексная схема. В п. 1.2.4 показано, что комплексная схема Розенброка (CROS) имеет аппроксимацию $O(\tau^2)$, а ее функция устойчивости есть

$$\rho(z) = 1/\left(1 - z + z^2/2\right),\ z = \mu\tau. \tag{4.15}$$

Здесь вещественные $\mu < 0$ — спектральные множители в линейном тесте Далквиста. Отсюда видно, что $\rho(z)$ вещественно, причем $0 < \rho < 1$ при любом τ. В терминах разностных схем это означает, что схема CROS безусловно устойчива по начальным данным. Поскольку множители ρ не зависят от номера временного слоя, эта устойчивость равномерна.

Можно также проверить, что выполняется условие теоремы 2.2, т. е. схема устойчива по правой части. Краевые условия первого рода аппроксимируются точно, поэтому устойчивости по ним не требуется. Таким образом, полностью доказана безусловная устойчивость схемы CROS.

С учетом пространственной невязки $O(h^2)$ полная погрешность аппроксимации будет $O(\tau^2 + h^2)$. В силу теоремы 2.8 схема CROS безусловно сходится с точностью $O(\tau^2 + h^2)$.

Сравним это заключение с классическим выводом устойчивости методом гармоник. Сделаем традиционную подстановку $u_n \to 1$, $u_{n\pm1} \to \exp(\pm \mathrm{i} qh)$, $\hat{u} = \rho u$; здесь q — волновое число гармоники. Подставим эти выражения в (4.13), вводя еще одно обозначение: $w_n = \gamma u_n$. Сначала вычислим действие оператора Λ:

$$\begin{aligned} \Lambda u \to \frac{\kappa}{h^2}\left(e^{-\mathrm{i}qh} - 2 + e^{\mathrm{i}qh}\right) = \\ = -\frac{2\kappa}{h^2}\left(1 - \cos(qh)\right) = -\frac{4\kappa}{h^2}\sin^2\left(\frac{qh}{2}\right). \end{aligned} \tag{4.16}$$

Подставляя это выражение в линейную систему (4.13) и учитывая выражение для w, получим

$$\left(1 + \frac{1+\mathrm{i}}{2}\frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right)\right)\gamma = -\frac{4\kappa}{h^2}\sin^2\left(\frac{qh}{2}\right).$$

Подставляя полученное выражение в первую из формул (4.13), получим

$$\rho = 1 + \tau \mathrm{Re}\gamma = \\ = 1 - \mathrm{Re}\left[\frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right) / \left(1 + \frac{1+\mathrm{i}}{2}\frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right)\right)\right] = \\ = 1 / \left[1 + \frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right) + \frac{1}{2}\left(\frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right)\right)^2\right]. \tag{4.17}$$

Если учесть, что допустим только дискретный спектр гармоник с $q = \pi m/a$, то полученное выражение совпадает с функцией устойчивости (4.15).

Напомним, что схема CROS является $L2$–устойчивой, а (4.15) есть двучленная Паде-аппроксимация e^z. Это одновременно обеспечивает хорошую точность и быстрое затухание высоких гармоник решения. Это затухание является качественно правильным. В самом деле у точного решения скорость затухания тем больше, чем выше номер гармоники. При этом амплитуды гармоник не меняют знак ($\rho > 0$). Все эти свойства выполняются и для гармоник (4.15), (4.17). Это обеспечивает хорошее качественное поведение численного решения. Последнее позволяет хорошо рассчитывать решения даже с разрывами начальных данных: разрыв сглаживается со временем, причем численное решение остается практически монотонным в районе разрыва, как и должно быть в точном решении.

Особо скрупулезный анализ показывает, что для линейного уравнения теплопроводности в схеме CROS при расчете разрывных решений возможно ничтожное нарушение пространственной монотонности, не превышающее 0,01 % от амплитуды разрыва. Однако это практически незаметно в обычных расчетах. Другой недостаток схемы проявляется при расчете квазилинейного уравнения теплопроводности; об этом будет сказано далее.

Таким образом, схема CROS является безусловно устойчивой и диссипативной, имеет второй порядок точности, а расчеты по ней отличаются высокой надежностью. Эту схему можно в первую очередь рекомендовать для расчетов.

Чисто неявная схема соответствует $\sigma = 1$. Формулы (4.13) легко преобразуются к традиционной форме записи этой схемы:

$$\frac{\hat{u}_n - u_n}{\tau} = (\Lambda\hat{u})_n + \hat{f}_n \equiv \frac{\kappa}{h^2}(\hat{u}_{n-1} - 2\hat{u}_n + \hat{u}_{n+1}) + f(x_n, \hat{t}). \quad (4.18)$$

Из п. 1.2.3 следует, что схема (4.18) имеет аппроксимацию $O(\tau)$; с учетом пространственной невязки это дает полную аппроксимацию $O(\tau + h^2)$. Последнее легко проверяется непосредственным разложением в ряды Тейлора в точке $(x_n, \hat{t})$. Поэтому по точности чисто неявная схема существенно уступает схеме CROS.

Устойчивость схемы по начальным данным исследуем методом гармоник. Стандартная замена с учетом соотношения (4.16) после подстановки в (4.13) дает

$$\rho = 1/\left(1 + \frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right)\right). \quad (4.19)$$

Это вещественная величина, причем $0 < \rho < 1$ для всех гармоник при любом шаге τ. Поэтому схема безусловно устойчива по начальным данным и диссипативна. Равномерная устойчивость по начальным данным и устойчивости по правой части устанавливаются так же, как и для схемы CROS. В силу теоремы 2.8 схема безусловно сходится с точностью $O(\tau + h^2)$. Выражение (4.19) есть Паде-аппроксимация e^z, но лишь одночленная. Высокие гармоники затухают качественно правильно, но медленнее, чем для схемы CROS.

Схема «с полусуммой», называемая также схемой Кранка—Никольсон, соответствует $\sigma = 1/2$. Для нее подстановка (4.16) также легко приводит к традиционной форме

$$\frac{\hat{u}_n - u_n}{\tau} = \frac{\kappa}{2h^2}[(\hat{u}_{n-1} - 2\hat{u}_n + \hat{u}_{n+1}) + (u_{n-1} - 2u_n + u_{n+1})] + f(x_n, t + \tau/2). \quad (4.20)$$

Из п. 1.2.3 следует, что аппроксимация по времени есть $O(\tau^2)$; полная аппроксимация с учетом пространственной невязки есть $O(\tau^2 + h^2)$. Это нетрудно проверить, разлагая (4.20) в ряды Тейлора в точке $(x_n, t + \tau/2)$: в силу симметрии схемы все члены первого порядка сокращаются. Таким образом, порядок точности схемы «с полусуммой» таков же, как у схемы CROS. Однако коэффициент в остаточном члене перед τ^2 меньше, чем для

схемы CROS. Поэтому на решениях высокой гладкости ее фактическая точность может оказаться несколько лучше.

Исследование устойчивости традиционным методом гармоник дает следующий множитель роста (функцию устойчивости):

$$\rho = \left(1 - \frac{2\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right)\right) / \left(1 + \frac{2\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right)\right). \qquad (4.21)$$

Видно, что для всех гармоник ρ вещественно и выполняются строгие неравенства $-1 < \rho < 1$. Следовательно, схема диссипативна и безусловно устойчива по начальным данным. Равномерная устойчивость по начальным данным и устойчивость по правой части устанавливаются так же, как и для предыдущих схем. В силу теоремы 2.8 она безусловно сходится с точностью $O\left(\tau^2 + h^2\right)$.

Схема имеет серьезный качественный недостаток. Гармоники точного решения затухают тем быстрее, чем больше номер гармоники. Младшие гармоники (4.21) схемы «с полусуммой» также затухают указанным образом. Однако при разумных величинах шага $\tau \sim h$ множитель $\dfrac{2\kappa\tau}{h^2} \sim h^{-1}$ является большой величиной. Поэтому, начиная с некоторого номера гармоники, множитель ρ становится отрицательным, а модуль ρ начинает увеличиваться с возрастанием номера гармоники. Эта качественная неправильность мало сказывается при расчете решений с небольшими пространственными производными. Однако если пространственные производные велики, т. е. решение имеет крутой фронт, возможно появление пилообразной немонотонности в численном решении (подробнее это рассмотрено далее в п. 4.1.4).

Поэтому надежность схемы «с полусуммой» не всегда оказывается достаточной.

Явная схема приводится для сравнения. Она соответствует $\sigma = 0$. При этом (4.13) легко преобразуется к традиционной форме:

$$\frac{\hat{u}_n - u_n}{\tau} = \frac{\kappa}{h^2}\left(u_{n-1} - 2u_n + u_{n+1}\right) + f\left(x_n, t\right). \qquad (4.22)$$

Аппроксимация $O\left(\tau + h^2\right)$ легко устанавливается разложением в ряды Тейлора. Видно, что точность схемы невелика. Однако

гораздо хуже ее устойчивость. Методом гармоник нетрудно найти

$$\rho = 1 - \frac{4\kappa\tau}{h^2}\sin^2\left(\frac{qh}{2}\right). \tag{4.23}$$

Величина ρ вещественная, причем $\rho < 1$. Однако для высоких гармоник со значениями $\sin\left(\frac{qh}{2}\right) \approx 1$ величина $\rho \approx 1-(4\kappa\tau)/h^2$ может стать отрицательной и очень большой. Чтобы соблюдалось $-1 \le \rho$, должно выполняться условие

$$2\kappa\tau \le h^2. \tag{4.24}$$

Таким образом, явная схема лишь условно устойчива. При этом для устойчивости требуется очень малый шаг по времени: для хорошей точности надо брать небольшой пространственный шаг h. В этом случае из (4.24) следует, что схема устойчива лишь при крайне малых временных шагах $\tau \sim h^2$. Расчет с таким малым шагом τ неприемлемо трудоемок.

4.1.3. Асимптотическая устойчивость

Некоторые задачи (например, прогнозирование таяния вечной мерзлоты) требуют расчета на очень большие времена. Исследуем, при каких условиях схемы п. 4.1.2 позволяют рассчитывать задачи с нулевыми краевыми значениями для очень больших промежутков времени, т. е. каковы условия асимптотической устойчивости схемы.

Выход решения параболического уравнения (4.1) на асимптотику при $t \to \infty$ определяется скоростью затухания начальных данных. Приведенное в п. 4.1.1 разложение решения в ряд Фурье (4.6) показывает, что медленнее всего затухает первая гармоника. Ей соответствует множитель роста

$$\tilde{\rho}_1 = \exp\left(-\kappa\tau\pi^2/a^2\right) = 1 - \kappa\pi^2\tau/a^2 + O\left(\tau^2\right). \tag{4.25}$$

Чтобы схемы были асимптотически устойчивы, все их гармоники должны затухать не медленнее, чем $\tilde{\rho}_1$, т. е. должно выполняться условие

$$|\rho| \le \tilde{\rho}_1. \tag{4.26}$$

Разумеется, соблюдение этих неравенств достаточно с точностью до членов $O\left(\tau^2\right)$, потому что наличие таких членов

приведет к умножению амплитуд гармоник на величину $[1 + O(\tau^2)]^{t/\tau} = 1 + tO(\tau)$, чем при $\tau \to 0$ можно пренебречь, даже если t велико. При этом надо учитывать, что волновое число q принимает лишь дискретный ряд значений $q = \pi m/a$, $1 \le m \le N-1$. Соответственно значения синуса монотонно возрастают от $\dfrac{\pi}{2N} \ll 1$ до приблизительно $1 - O(N^{-2})$.

Рассмотрим, при каких условиях соотношение (4.26) выполняется для схем из п. 4.1.2.

Схема CROS. Ее множители роста (4.15), (4.17) положительны и монотонно убывают при возрастании номера m. Наибольшим является множитель ρ_1. Нетрудно видеть, что $\rho_1 = \tilde{\rho}_1 + O(\tau^2)$. Тем самым первая гармоника затухает со скоростью первой гармоники точного решения, а высшие гармоники затухают еще быстрее. Следовательно, схема CROS асимптотически безусловно устойчива.

Чисто неявная схема (4.18). Ее множители роста (4.19) ведут себя аналогично: они положительны, первая гармоника затухает со скоростью гармоники точного решения, а остальные гармоники затухают еще быстрее. Следовательно, чисто неявная схема также асимптотически безусловно устойчива.

Схема «с полусуммой» (4.20). Ее множители роста (4.21) убывают монотонно, причем опять $\rho_1 = \tilde{\rho_1} + O(\tau^2)$. Поэтому низшие гармоники, у которых $\rho_m > 0$, не нарушают асимптотической устойчивости.

Однако у старших гармоник $\rho_m < 0$. Сами значения ρ_m убывают с ростом m, но их модули возрастают. Наиболее опасной для устойчивости является последняя гармоника. Учтем, что при разумном выборе шагов расчета $h \ll 1$ и $\tau \sim h$, так что $h^2 \ll \kappa\tau$. Тогда для этой гармоники получаем

$$\rho_{N-1} \approx \frac{1 - 2\kappa\tau/h^2}{1 + 2\kappa\tau/h^2} \approx -\left(1 - \frac{h^2}{\kappa\tau}\right). \tag{4.27}$$

Требуя выполнения условия (4.26) с учетом (4.25) и (4.27), получаем условие устойчивости

$$\pi\kappa\tau \le ah. \tag{4.28}$$

Тем самым схема «с полусуммой» асимптотически условно устойчива (хотя ее обычная устойчивость безусловна).

Явная схема. Поскольку ее условие устойчивости (4.24) требует очень малого шага τ, эта схема совершенно непригодна для расчета задач на большие времена.

4.1.4. Монотонность

Точное решение $u(x,t)$ однородного уравнения $u_t = \kappa u_{xx}$ при определенных условиях сохраняет пространственную монотонность. Например, если профиль начальных данных $u(x,0)$ монотонен, а граничные условия $u(0,t)$ и $u(a,t)$ постоянны, то профиль температуры $u(x,t)$ в любой момент t будет монотонен. То же будет при аналогичной постановке задачи Коши на бесконечной прямой. Ограничимся случаем бесконечной прямой и выясним, сохраняют ли монотонность схемы, приведенные в п. 4.1.2. При этом будем опираться на теорему 3.1: для монотонности двуслойной разностной схемы, записанной в форме (4.22)

$$\hat{u}_n = \sum_m \beta_m u_{n+m}, \tag{4.29}$$

необходимо и достаточно выполнения условия $\beta_m \geq 0$ для всех m.

Явная схема (4.22) уже записана в требуемой форме (4.29). Сумма содержит только три индекса $m = -1, 0, 1$. Соответствующие коэффициенты равны

$$\beta_{-1} = \beta_1 = \kappa\tau/h^2 > 0,\ \beta_0 = 1 - 2\kappa\tau/h^2. \tag{4.30}$$

Два коэффициента всегда положительны. Чтобы получить $\beta_0 \geq \geq 0$, необходимо выполнение условия

$$2\kappa\tau \leq h^2. \tag{4.31}$$

Таким образом, для явной схемы условие монотонности совпадает с условием устойчивости.

Чисто неявная схема (4.18) не имеет требуемого вида (4.29). Однако ее можно привести к форме (4.29) с бесконечной суммой по m следующим образом. В формуле (4.18) перенесем все члены с $\hat{u}_n$ в левую часть, а все остальные — в правую часть. Получим выражение $\hat{u}_n$ в центральной точке шаблона через значения в соседних:

$$\hat{u}_n = \frac{\hat{u}_{n-1} + \hat{u}_{n+1} + \dfrac{h^2}{\kappa\tau} u_n}{2 + \dfrac{h^2}{\kappa\tau}}. \tag{4.32}$$

С учетом делителя коэффициенты перед всеми значениями u в правой части положительны, причем их сумма с учетом делителя равна 1. Перепишем (4.32), сдвигая индексы на $+1$ и -1; это даст выражения $\hat{u}_{n-1}$ и $\hat{u}_{n+1}$ через значения в соседних с ними точках. Продолжим эту процедуру неограниченно и подставим все получившиеся выражения в (4.32). Поскольку на всех стадиях коэффициенты положительны, в итоге получим бесконечную сумму (4.29), все коэффициенты которой также будут положительны.

Если проделать все описанные выкладки (что достаточно громоздко), то получим следующие значения коэффициентов:

$$\beta_0 = 1 - \frac{4\kappa\tau}{\gamma(h+\gamma)},\ \beta_{\pm m} = \frac{4h\kappa\tau}{\gamma(h+\gamma)^2}\left[\frac{4\sigma\kappa\tau}{(h+\gamma)^2}\right]^{m-1} \text{ для } m \geq 1;$$
$$\gamma = \sqrt{h^2 + 4\sigma\kappa\tau},\ \sigma = 1. \tag{4.33}$$

Положительность коэффициентов с $m \neq 0$ очевидна, а положительность β_0 легко проверяется с учетом данного значения σ. Отсюда следует монотонность схемы (4.18) при любых значениях τ и h, т. е. безусловная монотонность.

Схема «с полусуммой» (4.20). Ее также можно преобразовать к форме, где $\hat{u}_n$ выражается через значения в остальных точках шаблона, и провести аналогичную процедуру приведения к бесконечной сумме. При этом снова получаются выражения (4.33), но со значением $\sigma = 1/2$. Опять очевидно, что $\beta_m > 0$ при $m \neq 0$. Однако коэффициент β_0 теперь может стать отрицательным. Нетрудно проверить, что условие его неотрицательности имеет следующий вид:

$$2\kappa\tau \leq 3h^2. \tag{4.34}$$

Таким образом, схема «с полусуммой» лишь условно монотонна, причем условие ее монотонности (4.34) является очень жестким: $\tau = O\left(h^2\right)$. В реальных расчетах оно обычно нарушается. Поэтому на практике эта схема легко может нарушить монотонность решения, особенно если оно имеет крутые фронты или разрывы начальных данных.

Схема CROS. Для нее теоретическое исследование гораздо более громоздко. Удается показать, что пространственная монотонность может нарушаться, хотя монотонность по времени (t-монотонность) сохраняется. Однако, во-первых, это нарушение монотонности очень мало (менее 0,1 % от амплитуды крутого фронта), так что на графике оно незаметно. Во-вторых, это нарушение происходит не вблизи фронта, а на расстоянии многих интервалов сетки от фронта. В-третьих, благодаря сильной диссипативности схемы вид профиля не пилообразный, а плавный.

Пример. На рис. 4.2 показан первый шаг по времени при расчете уравнения с разрывными начальными данными. Видно, что схема «с полусуммой» дает пилообразный профиль в районе разрыва. Амплитуда пилы на первом временном шаге может доходить до 100 % величины разрыва начальных данных. На последующих шагах амплитуда пилы уменьшается, но достаточно долго остается заметной. Чисто неявная схема дает строго монотонный профиль, однако точность ее невысока, и этот профиль заметно отличается от точного решения. Схема CROS лежит гораздо ближе к точному решению, а ее немонотонность на графике незаметна.

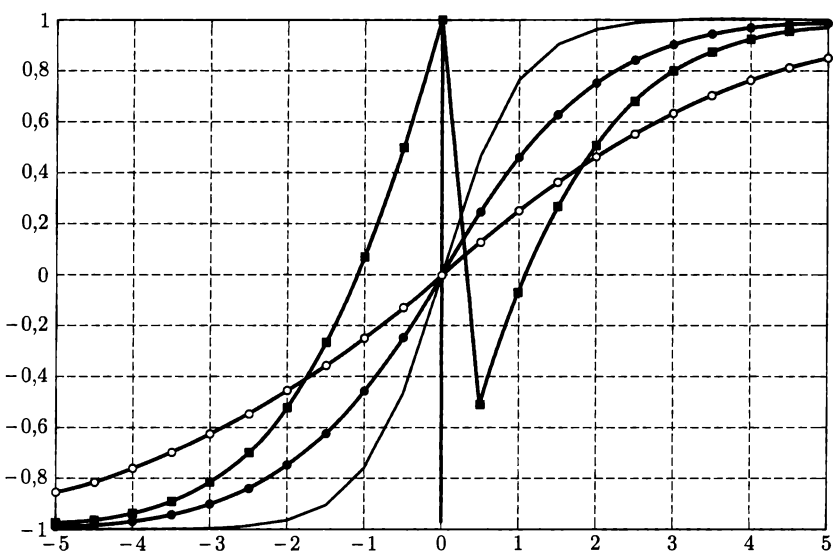


Рис. 4.2. Первый шаг: тонкая кривая — точное решение, • — схема CROS, ○ — чисто неявная схема, ■ — схема «с полусуммой»; жирные линии — разрывные начальные данные

Для уравнения переноса была доказана теорема 3.2 о том, что монотонная схема не может иметь второй порядок точности. Для параболического уравнения аналогичной теоремы не доказано. Однако не удалось построить ни одной схемы точностью $O(\tau^2 + h^2)$, которая была бы строго монотонной (хотя пример схемы CROS показывает, что есть схемы с очень малой немонотонностью). Более того, ни для какого типа эволюционных уравнений в частных производных до сих пор не построено ни одной строго монотонной схемы точности $O(\tau^2 + h^2)$ или выше.

4.1.5. Бикомпактные схемы

Обобщенное решение. Простейшую задачу теплопроводности можно последовательно усложнять. Во-первых, сетку можно взять неравномерной. Во-вторых, коэффициент теплопроводности и источник тепла могут быть переменными. В-третьих, существуют слоистые среды; при этом на границах слоев коэффициент теплопроводности и источник тепла имеют разрывы. Рассмотрим общий случай, содержащий все эти усложнения.

Пусть среда состоит из слоев с неподвижными границами. В каждом слое коэффициент теплопроводности $\kappa(x,t)$ и источник тепла $f(x,t)$ будем считать ограниченными непрерывными многократно дифференцируемыми функциями. Однако на границах слоев x^* они разрывны. Внутри каждого слоя уравнение теплопроводности с переменными коэффициентами имеет следующий вид:

$$c(x,t)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\kappa(x,t)\frac{\partial u}{\partial x}\right) + f(x,t). \tag{4.35}$$

Здесь дополнительно введена теплоемкость среды $c(x,t)$. В точках x^*, являющихся границами слоев, дифференциальное уравнение неприменимо: дифференцировать разрывное $\kappa(x,t)$ нельзя. Поэтому решение, которое мы ищем, является обобщенным решением дифференциального уравнения. Для его построения необходимо выяснить, каким внутренним граничным условиям оно должно удовлетворять в точках x^*.

Заменим уравнение второго порядка по пространству (4.35) эквивалентной системой двух уравнений первого порядка по пространству:

$$c(x,t)\frac{\partial u}{\partial t} = \frac{\partial w}{\partial x} + f(x,t); \tag{4.36}$$

$$w = \kappa(x,t)\frac{\partial u}{\partial x}. \tag{4.37}$$

Если u является температурой, то величина w имеет физический смысл теплового потока. Форма записи (4.36) — (4.37) позволяет сформулировать внутренние краевые условия.

Предположим, что $u(x,t)$ разрывно в некоторой точке. Тогда в этой точке $u_x = \infty$, а из (4.37) следует, что и $w = +\infty$. Однако тепловой поток есть энергия, переносимая движущимися частицами среды, и он не может быть бесконечным. Поэтому $u(x,t)$ должно быть непрерывно по пространству во всех точках среды, включая границы слоев. Это есть первое внутреннее краевое условие.

Необходимо, чтобы $u(x,t)$ было также непрерывно по времени. В самом деле, температура u пропорциональна объемной энергии среды. Энергия не может меняться скачком, поэтому u_t должно оставаться конечным. Но тогда из уравнения (4.36) следует, что w_x всюду конечно. Тем самым $w(x,t)$ должно быть непрерывным по пространству всюду, включая границы слоев. Это второе внутреннее граничное условие.

Полученные два внутренних граничных условия выделяют из всех формально возможных обобщенных решений системы (4.36) — (4.37) физически правильное обобщенное решение.

Простейшая схема. Для нахождения обобщенного решения построим бикомпактную схему аналогично п. 1.4.2. Уравнение второго порядка по пространству (4.35) уже заменено системой двух уравнений первого порядка (4.36) — (4.37). Введем специальную сетку $\omega = \{x_n, 0 \le n \le N; h_n = x_n - x_{n-1}\}$. Напомним, что специальными называют сетки, в которых в каждую особенность решения поставлен некоторый узел x_n. Очевидно, такая сетка будет неравномерной.

Температуру и тепловой поток будем определять в узлах сетки: u_n, w_n. Коэффициенты $c(x,t)$,$\kappa(x,t)$ и $f(x,t)$ на специальной сетке непрерывны и многократно дифференцируемы внутри каждого интервала. Однако в тех узлах сетки, которые являются границами слоев, эти коэффициенты разрывны.

Построим бикомпактную схему с помощью интегро-интерполяционного метода. Для вывода схемы применим метод прямых. Рассмотрим один пространственный интервал $[x_{n-1}, x_n]$. Проинтегрируем уравнение (4.36) по этому интервалу, беря интеграл от w_x точно. Получим следующее точное соотношение:

$$\int\limits_{x_{n-1}}^{x_n} c(x,t)\frac{\partial u}{\partial t}dx = w_n - w_{n-1} + \int\limits_{x_{n-1}}^{x_n} f(x,t)dx. \qquad (4.38)$$

Поскольку $f(x,t)$ многократно дифференцируема внутри интервала, интеграл в правой части (4.38) возьмем по формуле средних, имеющей точность $O(h^2)$. Первый интеграл нельзя брать по формуле средних, так как мы определяем u_n только на границах интервала. Но $c(x,t)$ брать в узлах нежелательно: на границах областей оно разрывно, а использование односторонних пределов слишком громоздко. Поэтому для интеграла в левой части (4.38) используем гибридную формулу, в которой u_n берутся аналогично формуле трапеций, а $c(x,t)$ — в середине интервала; это также обеспечивает аппроксимацию $O(h^2)$. Получаем следующие дифференциально-разностные уравнения:

$$\frac{h_n}{2}c_{n-1/2}\left(\frac{du_{n-1}}{dt} + \frac{du_n}{dt}\right) = w_n - w_{n-1} + h_n f_{n-1/2}. \qquad (4.39)$$

Здесь вместо частной производной по времени стоит обыкновенная, так как она относится к узловому значению температуры. Уравнение (4.39) аппроксимирует дифференциальное уравнение (4.36) с невязкой $O(h^2)$.

Уравнение (4.37) перепишем в виде $u_x = w/\kappa(x,t)$. Проинтегрируем это уравнение по интервалу сетки, точно беря интеграл от u_x:

$$u_n - u_{n-1} = \int\limits_{x_{n-1}}^{x_n} \frac{w}{\kappa(x,t)}dx.$$

Интеграл в правой части, аналогично сказанному выше, вычисляем по гибриду формулы средних и трапеций. Получаем следующее разностное соотношение, имеющее аппроксимацию $O(h^2)$:

$$u_n - u_{n-1} = \frac{h_n}{2\kappa_{n-1/2}}\left(w_{n-1} + w_n\right). \qquad (4.40)$$

Разностные соотношения (4.39) — (4.40) являются бикомпактной формой метода прямых. Эти выражения справедливы для всех интервалов сетки ($1 \le n \le N$), в том числе и примыкающих к границам слоев. Внутренние краевые условия автоматически учтены при написании этих выражений.

Однако форма (4.39) — (4.40) неудобна для расчетов. Она содержит неизвестные величины u_n, w_n с $0 \le n \le N$, т. е. полное число неизвестных равно $2N+2$. Число разностных уравнений равно $2N-2$. Еще два уравнения следуют из граничных условий для $u(0,t)$ и $u(a,t)$. Для определения неизвестных не хватает двух уравнений. Можно получить два дополнительных условия для w на внешних границах, но все равно разностная схема остается алгоритмически сложной.

Есть более простой способ. Приведем (4.39) — (4.40) к трехточечной форме, исключая w_n. Перепишем эти уравнения в следующем виде:

$$w_n - w_{n-1} = A_n,\ A_n \equiv \frac{h_n}{2} c_{n-1/2} \left(\frac{du_{n-1}}{dt} + \frac{du_n}{dt} \right) - h_n f_{n-1/2}; \tag{4.41}$$

$$w_{n-1} + w_n = B_n,\ B_n \equiv \frac{2\varkappa_{n-1/2}}{h_n} (u_n - u_{n-1}). \tag{4.42}$$

Напишем уравнение (4.41) с индексом, увеличенным на 1: $w_{n+1} - w_n = A_{n+1}$. Складывая его с уравнением (4.41), получим соотношение $w_{n+1} - w_{n-1} = A_n + A_{n+1}$.

Увеличивая на единицу индекс в (4.42), получим: $w_n + w_{n+1} = B_{n+1}$. Вычитая из этого выражения (4.42), получим $w_{n+1} - w_{n-1} = B_{n+1} - B_n$. Сравнивая это выражение с последней формулой предыдущего абзаца, исключим w и получим соотношение $A_n + A_{n+1} = B_{n+1} - B_n$.

Подставляя в последнее соотношение выражения для A и B из (4.41) — (4.42), получим разностную схему метода прямых, не содержащую значений w. Запишем эту схему в матричной форме:

$$M\frac{du}{dt} = \Lambda u + \phi;$$

$$\left(M\frac{du}{dt} \right)_n = \frac{h_n}{4} c_{n-1/2} \frac{du_{n-1}}{dt} + \left(\frac{h_n}{4} c_{n-1/2} + \frac{h_{n+1}}{4} c_{n+1/2} \right) \frac{du_n}{dt} + \\ + \frac{h_{n+1}}{4} c_{n+1/2} \frac{du_{n+1}}{dt};$$

$$(\Lambda u)_n = \frac{\varkappa_{n+1/2}}{h_{n+1}} (u_{n+1} - u_n) - \frac{\varkappa_{n-1/2}}{h_n} (u_n - u_{n-1}),$$

$$\phi_n = \frac{1}{2} \left(h_n f_{n-1/2} + h_{n+1} f_{n+1/2} \right),\ 1 \le n \le N-1. \tag{4.43}$$

Видно, что матрицы M и Λ являются трехдиагональными. Схема написана во внутренних узлах. Она дополняется заданием $u_0(t)$ и $u_N(t)$ из условий на внешних границах. Тем самым задаются и производные в этих точках. Неизвестными остаются $u_n(t)$ во внутренних узлах, а их число равно числу уравнений.

Замечание. В литературе часто приводятся схемы, в которых значения температуры u_n берутся в узлах, а значения тепловых потоков $w_{n-1/2}$ — в серединах интервалов (или наоборот). Этот сдвиг на полшага означает, что при вычислении производных вместо шага h фактически используется шаг $h/2$. Поэтому коэффициент в члене $O\left(h^2\right)$ уменьшается в 4 раза, что приводит к соответствующему улучшению точности схемы. Однако такие схемы не являются бикомпактными. Их также можно сделать пригодными для расчета слоистых сред, однако они менее надежны.

Интегрирование по времени. Аналогично п. 4.1.2, воспользуемся семейством одностадийных схем Розенброка с параметром α. Тогда получим для (4.43) следующее семейство разностных схем:

$$\hat{u} = u + \tau \mathrm{Re} v, \ (M - \alpha\tau\Lambda)\, v = \Lambda u + \phi. \tag{4.44}$$

Поскольку матрицы M и Λ трехдиагональные, приращение решения v находится из линейной системы уравнений методом прогонки (или методом Гаусса для ленточной матрицы). Поэтому алгоритм нахождения решения столько же прост, как и для уравнения теплопроводности с постоянными коэффициентами. Нетрудно проверить, что для матрицы $M - \alpha\tau\Lambda$ выполнено условие преобладания диагонального элемента, поэтому решение линейной системы существует и единственно, а метод прогонки устойчив.

Напомним, что рекомендуется одно из трех следующих значений параметра. В первую очередь это $\alpha = (1 + \mathrm{i})/2$ (схема CROS); оно дает L2-устойчивую схему с аппроксимацией $O(\tau^2 + h^2)$. Эта схема почти монотонна. Эта схема сочетает хорошую надежность и высокую точность.

Если требуется особенно высокая надежность, следует взять $\alpha = 1$. Получается L1–устойчивая, строго монотонная чисто неявная схема. Однако она имеет существенно худшую точность $O(\tau + h^2)$.

Употребляют также схему с $\alpha = 1/2$. Это A-устойчивая схема «с полусуммой», имеющая хорошую точность $O(\tau^2 + h^2)$; коэффициент в невязке у нее даже несколько меньше, чем в ком-

плексной схеме. Однако эта схема сильно немонотонна; это может сказаться при решении задач с крутыми фронтами, приводя к появлению пилообразных профилей.

Устойчивость по начальным данным схемы (4.44) исследуем только для наиболее рекомендуемого варианта — схемы CROS. Применим метод гармоник, замораживая коэффициенты, т. е. беря постоянными коэффициенты и шаги сетки. Положим в уравнении (4.44) $\alpha = (1+\mathrm{i})/2$, умножим уравнение для v слева на оператор $M - \alpha^*\tau\Lambda$ и учтем, что при постоянных коэффициентах и шагах операторы M и Λ симметричны и, тем самым, коммутативны. Тогда получим следующее соотношение:

$$\left(M^2 - \tau M\Lambda + \frac{1}{2}\tau^2\Lambda^2\right) v = \left(M - \frac{1-\mathrm{i}}{2}\tau\Lambda\right)\Lambda u.$$

Возьмем от этого уравнения вещественную часть, для этого справа отбросим i, а слева вместо v поставим $\mathrm{Re}v = (\hat{u} - u)/\tau$. Получим вещественную запись схемы:

$$\left(M^2 - \tau M\Lambda + \frac{1}{2}\tau^2\Lambda^2\right)\frac{\hat{u}-u}{\tau} = \left(M - \frac{1}{2}\tau\Lambda\right)\Lambda u,$$

или

$$\left(M^2 - \tau M\Lambda + \frac{1}{2}\tau^2\Lambda^2\right)\hat{u} = M^2 u. \tag{4.45}$$

Сделаем стандартную подстановку $u_n \to 1$, $u_{n\pm1} \to \exp(\pm \mathrm{i} qh)$, $\hat{u} \to \rho u$. Учитывая выражения трехточечных операторов M и Λ, нетрудно получить результат их действия: $\Lambda u \to \dfrac{-4\kappa}{h}\sin^2\left(\dfrac{qh}{2}\right)u$, $Mu \to ch\cos^2\left(\dfrac{qh}{2}\right)u$. Подстановка этих выражений в (4.45) дает множитель роста q-й гармоники:

$$\rho_q = 1/\left(1 - z + z^2/2\right),\ z = \frac{-4\kappa\tau}{ch^2}\,\mathrm{tg}^2\left(\frac{qh}{2}\right). \tag{4.46}$$

Поскольку для всех гармоник $z < 0$, то $0 < \rho_q < 1$, и схема является безусловно устойчивой.

Множитель роста для бикомпактной схемы похож на множитель роста (4.15) для классической схемы CROS, только в выражении для z вместо синуса стоит тангенс. При увеличении

q тангенс возрастает быстрее синуса; соответственно ρ_q убывает сильнее. Поэтому в бикомпактной схеме высшие гармоники затухают быстрее, чем в классической схеме, и качественное поведение решения улучшается.

Сетки. Бикомпактные схемы обеспечивают погрешность $O(h^2)$ на произвольной неравномерной сетке. Это означает, что погрешность мажорируется $\mathtt{const}\max(h_n^2)$. Однако на практике мажорантные оценки обычно сильно завышены. Для фактического определения погрешности нужны асимптотически точные оценки. Только они позволяют применять метод Ричардсона во всех его формах и получать надежные неулучшаемые апостериорные оценки погрешности при сгущении сеток.

Их нельзя получить на произвольной сетке. Для этого нужны равномерные или квазиравномерные сетки. Покажем, как строить требуемые сетки в слоистых средах. Отрезок $[0, a]$ может делиться границей слоев x^* на несоизмеримые части. Внутренняя граница должна быть узлом специальной сетки. При этом невозможно построить равномерную сетку на всем отрезке. Однако можно построить равномерную сетку на отрезке $[0, x^*]$ и равномерную сетку, но с другим шагом, на отрезке $[x^*, a]$. Суммарная сетка будет кусочно-равномерной. Будем сгущать сетки на каждой части отрезка в одно и то же число раз. В этом случае метод Ричардсона можно применять, подставляя вместо шага h величину $1/N$, где N — суммарное число узлов сетки.

Аналогично можно строить кусочно-квазиравномерные сетки. В каждом слое выбираем свое порождающее преобразование $x(\xi)$ и строим свою квазиравномерную сетку. Затем сгущаем сетку по вспомогательной переменной ξ в одно и то же число раз в каждом слое.

Неограниченная область. Пусть одна или обе границы являются бесконечно удаленными точками. Бикомпактные схемы пригодны и для таких задач, если использовать квазиравномерные сетки. Например, пусть уравнение (4.35) задано на полупрямой $0 \le x < +\infty$ при граничных условиях первого рода: $u(0,t) = \mu_1(t)$, $u(+\infty, t) = \mu_2(t)$. Построим квазиравномерную сетку с помощью следующего порождающего преобразования:

$$x(\xi) = \beta\xi / \left(1 - \xi^2\right)^{\gamma}, \ 0 \le \xi \le 1, \ \beta > 0, \ \gamma > 0. \qquad (4.47)$$

По вспомогательной переменной введем равномерную сетку $\xi_n = n/N$, $0 \le n \le N$. Она порождает квазиравномерную сетку

$x_n = x(\xi_n)$. Нулевой узел этой сетки есть левая граница полупрямой, а последний является бесконечно удаленной точкой. Бикомпактная схема (4.44) позволяет вести расчеты на этой сетке, непосредственно подставляя в расчет точные граничные условия $u_0(t) = \mu_1(t)$, $u_N(t) = \mu_2(t)$. Однако в схеме (4.44) надо иначе определить шаг сетки. Нельзя полагать $h_n = x_n - x_{n-1}$: при таком определении получается последний шаг $h_N = \infty$. Необходимо ввести иное определение шага, причем во всех интервалах, а не только в последнем:

$$h_n = \left(\frac{dx}{d\xi}\right)_{n-1/2} \neq h_n - h_{n-1}. \tag{4.48}$$

Такое переопределение шага сохраняет порядок точности схемы, хотя несколько увеличивает коэффициент в остаточном члене. Заметим также, что середины интервалов, в которых вычисляются коэффициенты уравнения, нужно определять с помощью того же преобразования: $x_{n-1/2} = x(\xi_{n-1/2}) \neq (x_n + x_{n-1})/2$. При таком определении середина последнего бесконечного интервала $[x_{N-1}, x_N]$ также остается конечной точкой.

В практических расчетах обычно выбирают параметр $\gamma = 1/2$; при этом особенно простой вид имеет производная $dx/d\xi$. Параметр β стараются выбрать так, чтобы большая часть точек была сосредоточена в практически значимой области, а меньшая — на удаленной части полупрямой.

4.1.6. Квазилинейное уравнение

При высоких температурах коэффициент теплопроводности сам обычно является функцией температуры $\kappa(u)$. Например, для высокотемпературной плазмы коэффициент электронной теплопроводности $\kappa(u) \sim u^{5/2}$, а коэффициент лучистой теплопроводности еще сильнее зависит от температуры. В этом случае уравнение теплопроводности приобретает следующий вид:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\kappa(u,x,t)\frac{\partial u}{\partial x}\right] + f(u,x,t). \tag{4.49}$$

Расчет таких задач весьма труден, причем основная трудность связана с зависимостью коэффициентов от u, а не от x и t.

Поясним трудности на примере задачи для уравнения (4.49) на полупрямой $0 \leq x < +\infty$. Пусть коэффициент теплопровод-

ности имеет вид $\kappa = \kappa_0 u^m$. Начальное условие и правое краевое условие взяты нулевыми: $u(x,0) = 0$, $u(+\infty, t) = 0$. Для этой задачи известно точное решение, имеющее вид бегущей волны:

$$u(x,t) = \begin{cases} \left[\dfrac{Dm}{\kappa_0}(Dt - x)\right]^{1/m} & \text{при } x \le Dt, \\ 0, & \text{при } x > Dt. \end{cases} \tag{4.50}$$

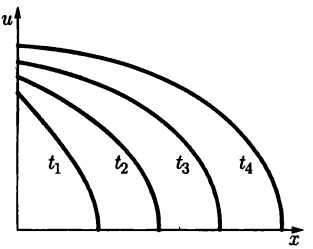


Рис. 4.3. Профиль тепловой волны (4.50) в разные моменты времени

Это решение реализуется, если левое граничное условие имеет вид

$$u(0,t) = \left(D^2 m t/\kappa_0\right)^{1/m}. \tag{4.51}$$

Видно, что скорость бегущей волны D тем больше, чем быстрее нарастает температура на левой границе. На рис. 4.3 показан профиль бегущей волны при $m > 1$. Решение является непрерывным, но лишь кусочно-гладким. Оно имеет резко выраженный фронт — точку $x^* = Dt$. В ней разрывна производная u_x. При этом левый предел $(u_x)_{x^*-0}$ равен бесконечности. Наличие бесконечной производной напоминает разрыв решения, и при расчетах таких задач по немонотонным схемам возможна сильная пилообразность расчетного профиля. Поэтому для решения таких задач нужны схемы особо высокой надежности.

Включение точки. Для линейных задач наиболее надежной считается чисто неявная схема (4.18), являющаяся одностадийной схемой Розенброка с $\alpha = 1$. Запишем ее для однородного уравнения ($f \equiv 0$) на равномерной сетке с учетом того, что $\kappa \neq \mathtt{const}$:

$$(\hat{u}_n - u_n)/\tau = \left[\kappa_{n+1/2}(\hat{u}_{n+1} - \hat{u}_n) - \kappa_{n-1/2}(\hat{u}_n - \hat{u}_{n-1})\right]/h^2. \tag{4.52}$$

Здесь переменные коэффициенты теплопроводности берутся в серединах интервалов и считаются зависящими от температуры на исходном (не новом!) слое. Такая схема не требует итераций и решается прогонкой.

Применим эту схему к расчету бегущей тепловой волны (4.50). Очевидно следующее: если в двух соседних узлах температуры нулевые ($u_{n-1} = 0$, $u_n = 0$), то $\kappa_{n-1/2} = 0$. Если же одна из этих

температур не нулевая, то при разумном определении среднего $\kappa_{n-1/2} \neq 0$.

Волна (4.50) бежит по фону нулевой температуры. Возьмем на исходном слое три соседние точки, лежащие на нулевом фоне: $u_{n-1} = u_n = u_{n+1} = 0$. Тогда $\kappa_{n-1/2} = \kappa_{n+1/2} = 0$. Подставим эти значения κ в схему (4.52). При этом правая часть обращается в нуль, откуда следует $\hat{u}_n = u_n = 0$. Тепло в точку x_n не проникло.

Расчетный фронт бегущей волны есть такая точка, в которой $u = 0$, но левее которой $u \neq 0$. Значит, если на исходном слое фронт лежал в точке x_{n-1}, то он за один шаг τ может переместиться в точку x_n, но не может переместиться в точку x_{n+1}: левее точки x_{n+1} лежит значение $\hat{u}_n = 0$. Отсюда следует теорема.

Теорема 4.1. При расчетах по схеме (4.52) фронт бегущей волны (4.50) за один шаг τ может пройти не более одного интервала h. •

Аналогичные теоремы справедливы для всего семейства одностадийных безытерационных схем Розенброка (4.13), в которых $\kappa(u)$ берется на предыдущем слое, в том числе для схем «с полусуммой» и CROS.

Теорема 4.2. Каково бы ни было отношение $c \equiv h/\tau$, существует такое точное решение уравнения (4.49), к которому расчет по схеме (4.13) не сойдется при $\tau \to 0$, $h = c\tau \to 0$. •

Доказательство. В качестве точного решения возьмем бегущую волну (4.50). Нарастание (4.51) на левой границе возьмем настолько большим, чтобы получить $D > c$. Фронт точного решения движется со скоростью D. Согласно теореме 4.1, скорость движения разностного фронта не может быть больше c. Поэтому сходимость разностного решения к точному при указанном сгущении сеток невозможна. ■

Отсюда видно, что безытерационные схемы фактически непригодны для решения квазилинейного уравнения теплопроводности. Они являются лишь условно устойчивыми. При этом условие устойчивости оказывается нетривиальным. Для уравнения переноса условие устойчивости было типа Куранта, т. е. отношение τ/h зависело от констант, входящих в само уравнение. Это условие можно было заранее проверить, а его нарушение легко определялось по неограниченному нарастанию «разболтки». Здесь же условие устойчивости зависит от конкретно-

го решения. При нарушении этого условия визуально может наблюдаться сходимость к пределу при сгущении сеток, но этот предел отличен от точного решения. Тем самым опознать нарушение устойчивости по результатам расчета далеко не всегда возможно.

Итерационная схема. Преодолеть указанные трудности можно, беря коэффициент теплопроводности не с исходного слоя, а с нового. Например, запишем чисто неявную схему аппроксимации $O\left(\tau + h^2\right)$ следующего вида:

$$\left(\hat{u}_n - u_n\right)/\tau = \left[\hat{\kappa}_{n+1/2}\left(\hat{u}_{n+1} - \hat{u}_n\right) - \hat{\kappa}_{n-1/2}\left(\hat{u}_n - \hat{u}_{n-1}\right)\right]/h^2,$$
$$\hat{\kappa}_{n-1/2} = \left[\kappa\left(\hat{u}_{n-1}\right) + \kappa\left(\hat{u}_n\right)\right]/2. \tag{4.53}$$

Для определения $\hat{\kappa}$ возможны и другие формулы, однако приведенная формула дает, по-видимому, более высокую точность. Схема (4.53) является системой нелинейных уравнений для определения $\hat{u}$. Ее необходимо решать каким-либо итерационным методом. Практика расчетов показывает, что наилучшие результаты дает итерационный метод Ньютона; его формулы довольно сложны, так как включают производные $\partial\hat{\kappa}/\partial\hat{u}$. Метод простых итераций (последовательных приближений) выглядит проще: надо в формулах (4.53) на s-й итерации вычислить $\kappa^{(s)}$ по значениям $u^{(s-1)}$ с предыдущей итерации. Однако метод простых итераций менее надежен: он далеко не всегда обеспечивает сходимость итераций. Проводить итерации надо до получения высокой точности. Недопустимо ограничивать расчет по заданному числу итераций: при этом получаются схемы, не обеспечивающие сходимость. Аналогично теореме 4.1 легко доказывается следующая теорема.

Теорема 4.3. При расчете простых итераций по схеме (4.53) фронт волны (4.50), бегущей по нулевому фону, за одну итерацию может пройти не более одного интервала h. •

Отсюда следует, что при фиксированном числе простых итераций s скорость движения разностного фронта не может превышать $c = h/\tau$. Тогда справедлив аналог теоремы 4.2 об отсутствии сходимости при сгущении сеток для произвольных решений.

Аналогично (4.53) можно записать схему «с полусуммой» точности $O\left(\tau^2 + h^2\right)$. Ее можно неплохо применять на решениях

с некрутыми фронтами. Однако на решениях с крутыми фронтами схема «с полусуммой» может давать пилообразные профили, и более надежной оказывается схема (4.53).

4.2. МНОГОМЕРНЫЕ УРАВНЕНИЯ

4.2.1. Схема с весами

Для уравнения переноса хорошие одномерные схемы — схемы бегущего счета — естественно обобщались на случай многих измерений. Однако попытка обобщить хорошие одномерные схемы для уравнения теплопроводности — семейство одностадийных схем Розенброка — наталкивается на принципиальные трудности. Рассмотрим их на примере двумерного уравнения теплопроводности с постоянным коэффициентом, для которого задана первая краевая задача в прямоугольной области:

$$\begin{gathered} u_t = \varkappa\left(u_{xx} + u_{yy}\right) + f(x,y,t),\ \varkappa = \mathtt{const}, \\ 0 < x < a,\ 0 < y < b,\ 0 < t \le T; \end{gathered} \tag{4.54}$$

$$\begin{gathered} u(0,y,t) = \mu_1(y,t),\ u(a,y,t) = \mu_2(y,t), \\ u(x,0,t) = \mu_3(x,t),\ u(x,b,t) = \mu_4(x,t), \\ u(x,y,0) = \mu(x,y). \end{gathered} \tag{4.55}$$

Введем прямоугольную сетку $\{x_n, y_m, 0 \le n \le N, 0 \le m \le M\}$, причем для простоты возьмем шаги по каждой переменной h_x, h_y постоянными. Значения функции в узлах обозначим $u_{nm} = u\left(x_n, y_m, t\right)$.

Для построения схемы используем метод прямых. Возьмем пятиточечный пространственный шаблон, имеющий форму креста с центром в точке (x_n, y_m). На этом шаблоне заменим пространственные производные вторыми разностями:

$$\begin{aligned} \left(\Lambda_x u\right)_{nm} &= \varkappa\left(u_{n-1,m} - 2u_{nm} + u_{n+1,m}\right)/h_x^2, \\ \left(\Lambda_y u\right)_{nm} &= \varkappa\left(u_{n,m-1} - 2u_{nm} + u_{n,m+1}\right)/h_y^2. \end{aligned} \tag{4.56}$$

Подставляя аппроксимации (4.56) в уравнение (4.54), получим систему обыкновенных дифференциальных уравнений для внутренних узлов сетки:

$$\frac{du_{nm}}{dt} = \left[\left(\Lambda_x + \Lambda_y\right)u\right]_{nm} + f_{nm}. \tag{4.57}$$

Входящие сюда значения u_{0m}, u_{Nm}, u_{n0}, u_{nM} берутся из граничных условий (4.55). Начальные условия для (4.57) есть $u_{nm}(0) = \mu(x_n, y_m)$. Система (4.57) аппроксимирует задачу (4.54) — (4.55) с погрешностью $O\left(h_x^2 + h_y^2\right)$.

Для системы (4.57) нетрудно написать одностадийное семейство схем Розенброка. Для простоты ограничимся случаем вещественного параметра α и в (4.13) сделаем замену $w = (\hat{u} - u)/\tau$. Схема примет следующий вид:

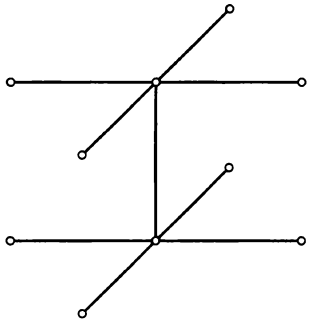


Рис. 4.4. Шаблон схемы (4.58)

$$[E - \alpha\tau(\Lambda_x + \Lambda_y)]\frac{\hat{u} - u}{\tau} = (\Lambda_x + \Lambda_y)u + f. \qquad (4.58)$$

Этой схеме соответствует шаблон, показанный на рис. 4.4. Легко проверить, что погрешность аппроксимации этой схемы на решениях с непрерывными четвертыми производными равна $O(\tau^\nu + h_x^2 + h_y^2)$, где $\nu = 2$ при $\alpha = 1/2$ и $\nu = 1$ при $\alpha \neq 1/2$. Напомним, что схема является A-устойчивой при $\alpha \geq 1/2$, $L1$-устойчивой при $\alpha = 1$ и t-монотонной при $\alpha \geq 1$. Схему (4.58) можно переписать в следующем виде:

$$\frac{\hat{u} - u}{\tau} = (\Lambda_x + \Lambda_y)[\alpha\hat{u} + (1 - \alpha)u] + f. \qquad (4.59)$$

Эту форму называют схемой с весами, а коэффициент α имеет смысл веса для нового слоя в пространственных производных. Чтобы вес каждого слоя был неотрицательным, надо брать $0 \leq \alpha \leq 1$.

Исследовать устойчивость семейства (4.59) можно методом гармоник, подставляя двумерную гармонику $\exp(\mathrm{i}qx + \mathrm{i}ry)$. Такое исследование будет проведено далее для факторизованных схем. Однако напомним, что A-устойчивость означает безусловную устойчивость схемы. А отсутствие A-устойчивости означает условную устойчивость. Поэтому схемы «с полусуммой» ($\alpha = 1/2$) и чисто неявная ($\alpha = 1$) будут безусловно устойчивыми.

Трудоемкость расчета принято измерять количеством операций, приходящихся на один узел сетки при переходе со слоя на слой при шаге $\tau \sim h$. В одномерном случае решение сеточных уравнений при любом α проводилось методом прогонки, и

на один узел сетки приходилось фиксированное число операций, т. е. $O(1)$. Такие схемы называют ***экономичными***.

Покажем, что многомерные схемы Розенброка не являются экономичными.

Сначала рассмотрим явную схему ($\alpha = 0$). Значение $\hat{u}$ вычисляется по явной формуле (4.59), т. е. за $O(1)$ действий. Однако даже одномерная схема (4.22) была лишь условно устойчива при шаге $2\kappa\tau_0 \le h^2$. Поэтому для выполнения условного шага $\tau \sim h$ надо сделать $N \sim 1/h$ шагов τ_0. В итоге число операций на один узел сетки будет $O(N) \gg O(1)$, т. е. схема не экономична.

Рассмотрим безусловно устойчивые схемы (4.59) с $\alpha \ge 1/2$. Для них можно брать шаг $\tau \sim h$. Схема представляет собой систему линейных уравнений для определения $\hat{u}$. Однако матрица этой системы не является трехдиагональной. Чтобы привести систему к классическому виду линейной алгебры $Au = \phi$, надо двумерный массив u_{nm} превратить в вектор. Это можно сделать, выписывая строки матрицы u_{nm} одну за другой. Тогда матрица A принимает вид, изображенный на рис. 4.5. Она состоит из $M-1$ квадратных клеток порядка $N-1$. Каждая диагональная клетка есть трехдиагональная матрица, кодиагональные клетки являются диагональными матрицами, а остальные клетки — нулевые. Матрица A является сильно разреженной: ее порядок есть $(N-1)(M-1)$, а каждая строка содержит лишь пять ненулевых элементов. Матрица имеет ленточную структуру с шириной ленты $2N-1$. Бóльшая часть этой ленты заполнена нулями. Однако метод Гаусса для ленточной матрицы и другие аналогичные методы не умеют учитывать эти нули. Поэтому полное число операций на один узел сетки пропорционально квадрату

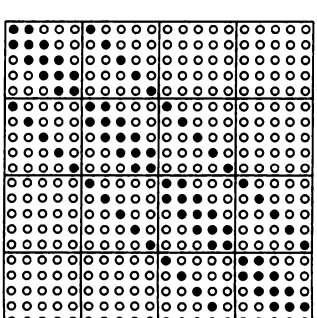


Рис. 4.5. Структура матрицы

полуширины ленты и составляет $O(N^2)$. Это еще больше, чем для явной схемы! Тем самым схема неэкономична и практически непригодна для расчетов.

Трехмерный случай рассматривается аналогично. Схему с весами написать нетрудно, но ширина ленты матрицы A есть $O(N^2)$, и число действий на один узел сетки есть $O(N^4)$.

Рассуждения для комплексной схемы Розенброка (CROS) более сложны, но дают те же самые результаты. Таким образом, семейство схем Розенброка (схем с весами) непригодно для двумерных и трехмерных расчетов.

4.2.2. Эволюционная факторизация

Экономичными являются схемы, в которых алгоритм сводится к выполнению одной или нескольких одномерных прогонок. Прогонка применима только к одномерному трехточечному оператору. Поэтому многомерный оператор надо представить в виде произведения одномерных; такое представление называется ***факторизацией***, или расщеплением.

Факторизацию нельзя выполнить точно. Можно лишь приближенно заменить нужный оператор другим, который поддается факторизации.

Наиболее удобной является факторизация, называемая эволюционной. Она единообразно строится для двумерных и трехмерных задач.

Построим эволюционную факторизацию, опираясь на трехмерную схему Розенброка с $\alpha = 1/2$, которую запишем аналогично (4.58):

$$\left(E - \frac{\tau}{2}\Lambda\right)\frac{\hat{u}-u}{\tau} = \Lambda u + f(x,y,z,t+\tau/2),\ \Lambda = \Lambda_x+\Lambda_y+\Lambda_z. \quad (4.60)$$

Эта схема имеет аппроксимацию $O\left(\tau^2 + h_x^2 + h_y^2 + h_z^2\right)$ и безусловно устойчива. Однако эта схема неэкономична. Поэтому наряду с ней рассмотрим схему, где в левой части вместо многомерного оператора стоит произведение одномерных:

$$\left(E - \frac{\tau}{2}\Lambda_z\right)\left(E - \frac{\tau}{2}\Lambda_y\right)\left(E - \frac{\tau}{2}\Lambda_x\right)\frac{\hat{u}-u}{\tau} = \Lambda u + f. \quad (4.61)$$

Факторизации подвергся тот оператор схемы (4.60), который стоит перед разностной производной по времени. Поэтому схе-

ма (4.61) называется эволюционно факторизованной. Она отличается от схемы «с полусуммой». Двумерная схема получается из (4.61) отбрасыванием членов, содержащих оператор Λ_z. При этом в левой части (4.61) остается только два одномерных сомножителя.

Исследуем новую схему.

Аппроксимация. Перемножим одномерные операторы в (4.61):

$$\left(E-\frac{\tau}{2}\Lambda_z\right)\left(E-\frac{\tau}{2}\Lambda_y\right)\left(E-\frac{\tau}{2}\Lambda_x\right)=$$

$$=E-\frac{\tau}{2}\left(\Lambda_z+\Lambda_y+\Lambda_x\right)+\frac{\tau^2}{4}\left(\Lambda_z\Lambda_y+\Lambda_z\Lambda_x+\Lambda_y\Lambda_x\right)-\frac{\tau^3}{8}\Lambda_z\Lambda_y\Lambda_x=$$

$$=E-\frac{\tau}{2}\Lambda+O\left(\tau^2\right). \tag{4.62}$$

Факторизованный оператор отличается от оператора в левой части схемы «с полусуммой» (4.60) членом $O\left(\tau^2\right)$. Этот член имеет тот же порядок малости по τ, что и в аппроксимации схемы «с полусуммой». Поэтому аппроксимация факторизованной схемы есть $O\left(\tau^2+h^2\right)$. Правда, теперь в оценку невязки входят произведения операторов Λ, парные в двумерном случае и тройные в трехмерном. Это требует существования более высоких производных решения: u_{txxyy} в двумерном случае и $u_{txxyyzz}$ в трехмерном случае.

Однако решение уравнения теплопроводности имеет неограниченное количество производных внутри области, так что требуемые производные существуют.

Устойчивость исследуем методом гармоник, причем в строгой форме. Для одномерного оператора Λ_x при постоянном коэффициенте и равномерной сетке k-й собственной функцией будет синусоида $\sin\left(\pi kx/a\right)$. Тогда в трехмерном прямоугольном параллелепипеде собственной функцией будет произведение одномерных собственных функций:

$$u_{klm}\left(x,y,z\right)=\sin\left(\pi kx/a\right)\sin\left(\pi ly/b\right)\sin\left(\pi mz/c\right),$$
$$1\le k\le N_x-1,\ 1\le l\le N_y-1,\ 1\le m\le N_z-1. \tag{4.63}$$

Каждый одномерный оператор воздействует только на свою переменную, поэтому

$$
\begin{gathered}
\Lambda_x u_{klm}(x,y,z) = \lambda_{xk} u_{klm}(x,y,z), \\
\lambda_{xk} = -\frac{4\kappa}{h_x^2} \sin^2\left(\frac{\pi k}{2N_x}\right),\ 1 \le k \le N_x - 1, \\
\Lambda_y u_{klm}(x,y,z) = \lambda_{yl} u_{klm}(x,y,z), \\
\lambda_{yl} = -\frac{4\kappa}{h_y^2} \sin^2\left(\frac{\pi l}{2N_y}\right),\ 1 \le l \le N_y - 1, \\
\Lambda_z u_{klm}(x,y,z) = \lambda_{zm} u_{klm}(x,y,z), \\
\lambda_{zm} = -\frac{4\kappa}{h_z^2} \sin^2\left(\frac{\pi m}{2N_z}\right),\ 1 \le m \le N_z - 1.
\end{gathered}
\tag{4.64}
$$

Подставим в эволюционно-факторизованную схему (4.61) собственную функцию (4.63) и положим $\hat{u} = \rho u$. Учитывая соотношения (4.64), получим выражение для множителя роста гармоники:

$$
\rho_{klm} = 1 + \frac{\tau(\lambda_{xk} + \lambda_{yl} + \lambda_{zm})}{(1 - \tau\lambda_{xk}/2)(1 - \tau\lambda_{yl}/2)(1 - \tau\lambda_{zm}/2)}. \tag{4.65}
$$

Перемножим все скобки в знаменателе (4.65) и сгруппируем все члены по степеням τ. Напомним, что все $\lambda < 0$. Тогда каждое слагаемое в знаменателе будет положительным. При этом слагаемое $O(\tau)$ будет по модулю вдвое меньше числителя. Следовательно, вся дробь в (4.65) будет заключена в пределах $(-2, 0)$. Отсюда вытекает $-1 < \rho_{klm} < 1$. Это строгое неравенство справедливо при любом шаге τ, т. е. схема безусловно устойчива.

Доказательство остается справедливым для двумерного случая. При этом надо из всех выражений вычеркнуть члены, содержащие Λ_z и λ_z. Аналогично проводится доказательство для одномерного случая.

Замечание. Для эволюционно-факторизованной схемы устойчивость в обычном смысле является безусловной; однако ее асимптотическая устойчивость лишь условна. Это хорошо видно на одномерном случае, когда схема превращается в одномерную схему «с полусуммой», у которой множитель роста и условие асимптотической устойчивости имеют следующий вид (см. пункт 4.1.3)

$$
\rho_k = (1 + \tau\lambda_{xk}/2) / (1 - \tau\lambda_{xk}/2),\ \tau \le a^2 / (\pi\kappa_x N_x) = O(h). \tag{4.66}
$$

В двумерном случае множитель роста (4.65) после приведения к общему знаменателю представляется как произведение одномерных множителей типа (4.66):

$$\rho_{kl} = \frac{1+\tau\lambda_{xk}/2}{1-\tau\lambda_{xk}/2} \cdot \frac{1+\tau\lambda_{yl}/2}{1-\tau\lambda_{yl}/2}. \tag{4.67}$$

Критичными для устойчивости являются первая и последняя гармоники по каждому направлению. Учтем, что для первых гармоник $|\tau\lambda| \ll 1$, а для последних — $|\tau\lambda| \gg 1$. Тогда из (4.67) следует:

$$\rho_{11} \approx 1 + \tau\left(\lambda_{x1} + \lambda_{y1}\right),$$
$$\rho_{N_x-1,N_y-1} \approx 1 + (4/\tau)\left(1/\lambda_{x,N_x-1} + 1/\lambda_{y,N_y-1}\right).$$

Оба множителя роста меньше 1, поскольку все λ отрицательны. Для асимптотической устойчивости последние гармоники должны затухать быстрее первых, т. е. надо требовать $\rho_{N_x-1,N_y-1} \leq \leq \rho_{11}$. Отсюда следует условие асимптотической устойчивости в двумерном случае:

$$\tau^2 \leq \tau_0^2 \equiv 4\left(1/\lambda_{x,N_x-1} + 1/\lambda_{y,N_y-1}\right) / \left(\lambda_{x1} + \lambda_{y1}\right). \tag{4.68}$$

Подставляя сюда спектральные значения (4.64) и принимая, что все числа интервалов N много больше единицы, получим

$$\tau \leq \tau_0 \approx \frac{1}{\pi}\left(\frac{h_x^2/\kappa_x + h_y^2/\kappa_y}{\kappa_x/a^2 + \kappa_y/b^2}\right)^{1/2} = O(h). \tag{4.69}$$

Условие асимптотической устойчивости оказалось практически таким же, как в одномерном случае.

Трехмерный случай более сложен. Множитель роста уже не удается представить в виде произведения одномерных сомножителей. Громоздкими выкладками можно показать, что в этом случае предельно допустимый шаг $\tau_0 = O\left(h^{4/3}\right)$. Это более жесткое ограничение, чем в одномерном и двумерном случаях.

Алгоритм нахождения разностного решения сводит задачу к последовательности одномерных прогонок. Вводя вспомогательные неизвестные сеточные функции v и w, перепишем факторизованную схему (4.61) в виде трех уравнений:

$$\left(E - \frac{\tau}{2}\Lambda_z\right) w = \left(\Lambda_x + \Lambda_y + \Lambda_z\right) u + f\left(x, y, z, t + \frac{\tau}{2}\right), \tag{4.70}$$

$$\left(E - \frac{\tau}{2}\Lambda_y\right) v = w, \tag{4.71}$$

$$\left(E - \frac{\tau}{2}\Lambda_x\right)\frac{\hat{u}-u}{\tau} = v. \tag{4.72}$$

Уравнение (4.70) написано для неизвестной трехмерной функции w. Его правая часть вычисляется на исходном слое, т. е. известна. Левая часть содержит лишь одномерный оператор по направлению z. Разобьем трехмерный массив w на одномерные массивы по направлению z. По этому направлению оператор $(E - \tau\Lambda_x/2)$ является трехдиагональной матрицей. Поэтому вдоль каждого направления схема (4.70) решается одномерной прогонкой. Выполняя такие прогонки по всем направлениям z, получим трехмерный массив w. При этом на один узел сетки придется девять арифметических операций.

Теперь в уравнении (4.71) найдена правая часть w. На неизвестную функцию v действует одномерный оператор по направлению y. Он также является трехдиагональной матрицей. Поэтому вычисление трехмерного массива v опять сводится к одномерным прогонкам по каждому из направлений y.

В заключение подставляем найденное v в правую часть (4.72). Теперь выражение $(\hat{u} - u)/\tau$ находится одномерными прогонками по направлениям x. Таким образом, для нахождения $\hat{u}$ приходится выполнить последовательность трех одномерных прогонок по разным направлениям. В результате полное число действий на один узел трехмерной сетки будет равно 27. Это число является фиксированным (не зависит от размеров сетки), т. е. факторизованная схема является экономичной.

В двумерном случае не надо вводить функцию w и уравнение (4.70). В уравнении (4.71) левая часть не меняется, а в правую часть вместо w подставляется $(\Lambda_x + \Lambda_y)u + f$. Уравнение (4.72) сохраняет свой вид. Таким образом, остаются два уравнения с одномерными трехдиагональными операторами.

4.2.3. Дополнения

Граничные условия первого рода для факторизованной схемы (4.70) — (4.72) являются нетривиальными. Для не факторизованной схемы (4.60) эти граничные условия можно поставить точно: достаточно $\hat{u}$ в граничных точках взять из граничных условий на гранях параллелепипеда в момент $t+\tau$. Однако в факторизованной схеме только для последнего уравнения (4.72), т. е. для прогонки по направлению x, можно брать $\hat{u}$ из точных

граничных значений в момент $t+\tau$ на гранях y, x параллелепипеда.

Для выполнения прогонки по направлению y в уравнении (4.71) нужно поставить граничное условие для v на гранях (x, z) параллелепипеда. Для получения этих условий воспользуемся (4.72). Подставим в его левую часть значения $\hat{u}, u$ из граничного условия на этих гранях:

$$v_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\Lambda_x\right)\frac{\hat{u}-u}{\tau}\right]_{\text{гран}}. \tag{4.73}$$

Опускать в правой части оператор Λ_x нельзя: это ухудшает погрешность аппроксимации до $O(\tau)$. Однако вторая разность на границе не всегда удобна. С точностью до h^2 ее можно заменить второй производной, что дает следующий вариант граничного условия:

$$v_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\varkappa\frac{\partial^2}{\partial x^2}\right)\frac{\hat{u}-u}{\tau}\right]_{\text{гран}}. \tag{4.74}$$

Дифференцирование граничного значения при этом надо выполнять точно. Условие (4.74) вносит дополнительную погрешность $O\left(\tau h^2\right)$; она имеет третий порядок малости, чем можно пренебречь.

Для двумерной задачи этого достаточно. В трехмерном случае надо построить граничные условия для w на гранях (x, y). Для этого подставляем v из (4.72) в левую часть (4.71) и получаем

$$w_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\Lambda_y\right)\left(E - \frac{\tau}{2}\Lambda_x\right)\frac{\hat{u}-u}{\tau}\right]_{\text{гран}}.$$

Перемножим операторы в круглых скобках. Член $\tau^2\Lambda_y\Lambda_x/4$ имеет порядок малости $O\left(\tau^2\right)$. Его можно отбросить, не ухудшая порядка аппроксимации всей схемы. Это дает следующее граничное условие для w:

$$w_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\Lambda_y - \frac{\tau}{2}\Lambda_x\right)\frac{\hat{u}-u}{\tau}\right]_{\text{гран}}. \tag{4.75}$$

Здесь также можно заменить разностное дифференцирование точным: $w_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\varkappa\frac{\partial^2}{\partial y^2} - \frac{\tau}{2}\varkappa\frac{\partial^2}{\partial x^2}\right)\frac{\hat{u}-u}{\tau}\right]_{\text{гран}}$. Эволюционно-факторизованная схема с описанными граничными условиями обеспечивает аппроксимацию $O\left(\tau^2 + h^2\right)$.

Замечания. ***1.*** Первой экономичной неодномерной схемой была двумерная продольно-поперечная схема. В ней вводился промежуточный шаг $t + \tau/2$. Значение $\hat{u}$ на промежуточном шаге вычислялось по схеме

$$\frac{\bar{u} - u}{\tau/2} = \Lambda_x u + \Lambda_y \bar{u} + \frac{\bar{f}}{2}. \tag{4.76}$$

Переход с промежуточного на новый слой производился по схеме

$$\frac{\hat{u} - \bar{u}}{\tau/2} = \Lambda_x \hat{u} + \Lambda_y \bar{u} + \frac{\bar{f}}{2}. \tag{4.77}$$

Схема (4.76) является явной по направлению x и неявной по направлению y; поэтому значения $\bar{u}$ находятся одномерными прогонками по направлениям y. Аналогично, схема (4.77) является явной по направлению y и неявной по направлению x; поэтому значения $\hat{u}$ находятся одномерными прогонками по направлениям x. Тем самым эта схема экономична. Можно доказать, что она безусловно устойчива, а ее точность есть $O\left(\tau^2 + h_x^2 + h_y^2\right)$. Однако перенести эту идею на трехмерный случай не удавалось. Кроме того, довольно трудно было написать граничные условия для $\bar{u}$, обеспечивающие точность $O\left(\tau^2\right)$.

Можно исключить из (4.76) — (4.77) значение $\bar{u}$. Тогда оставшаяся схема точно совпадет с двумерной эволюционно-факторизованной схемой (4.61). Поэтому эволюционно-факторизованную схему можно трактовать как обобщение продольно-поперечной схемы на трехмерный случай (и даже на случай произвольного числа измерений). Одновременно при этом выясняется вид граничных условий, обеспечивающих точность $O\left(\tau^2\right)$.

2. Известны также локально-одномерные схемы, пригодные для любого числа измерений. Однако их простейший вид, не симметризованный по времени, дает точность лишь $O(\tau)$. Для получения точности $O\left(\tau^2\right)$ нужно симметризовать эти схемы по времени, а также писать довольно сложные граничные условия. Это приводит к довольно громоздким выражениям.

3. Эволюционно-факторизованная схема столь же сильно немонотонна, как и схема «с полусуммой». Поэтому при расчетах задач с разрывами начальных данных, а также с разрывами коэффициентов могут возникать пилообразные профили решения. Со временем амплитуда этой пилообразности уменьшается благодаря диссипативным свойствам уравнения теплопроводности. Однако пилообразность может оставаться заметной спустя значительное время.

Можно построить не факторизованную схему Розенброка с $\alpha = (1 + i)/2$. Немонотонность этой схемы пренебрежимо мала. Однако эта схема не экономична, а ее факторизация приводит к сильно немонотонным схемам.

Неоднородная среда. Эволюционно-факторизованные схемы можно обобщить на уравнения с переменными коэффициентами при неравномерных сетках. При гладких коэффициентах для этого достаточно взять для одномерных разностных операторов выражения из бикомпактной схемы (4.43). Для этого вокруг центрального узла шаблона возьмем по каждой переменной по одному интервалу справа и слева от центрального узла шаблона, например $[x_{n-1}, x_{n+1}]$. Тогда в точке x_n

$$(\Lambda_x u)_n = \frac{2}{h_{xn} + h_{x,n+1}} \left(\varkappa_{n+1/2} \frac{u_{n+1} - u_n}{h_{x,n+1}} - \varkappa_{n-1/2} \frac{u_n - u_{n-1}}{h_{xn}} \right); \tag{4.78}$$

значения y и z в коэффициенте $\varkappa$ относятся к центральному узлу пространственного шаблона. Аналогично записываются вторые разности по другим направлениям. Однако значения разностной производной по времени $(\hat{u} - u)/\tau$ следует брать полностью в центральном узле пространственного шаблона, в то время как в бикомпактной схеме берется линейная комбинация в разных точках шаблона. Причина разницы в том, что для полностью бикомпактной схемы пока не найдено хорошего способа факторизации. Тогда окончательная дифференциально-разностная система примет вид

$$\frac{du}{dt} = (\Lambda_x + \Lambda_y + \Lambda_z)\, u + \bar{f}. \tag{4.79}$$

Схема «с полусуммой» и эволюционно-факторизованная схема для нее имеют вид (4.60) и (4.61) соответственно, где пространственные операторы пишутся аналогично (4.78).

Глава 5

ЭЛЛИПТИЧЕСКИЕ УРАВНЕНИЯ

5.1. ЭВОЛЮЦИОННОЕ РЕШЕНИЕ СТАЦИОНАРНЫХ ЗАДАЧ

5.1.1. Счет на установление

Постановка задачи. К эллиптическим уравнениям приводит ряд физических задач: определение прогиба нагруженной мембраны, давления газа в неоднородном силовом поле, стационарного (не зависящего от времени) распределения тепла в теле, электростатического поля между электродами и т. д. Все эти задачи имеют общее свойство: предполагается, что внешние воздействия не зависят от времени, а начальные условия были заданы достаточно давно, так что физическая система успела выйти на стационарное решение $u(\mathbf{r})$, не зависящее от времени.

Примером полной математической постановки является задача с краевыми условиями первого рода, называемая задачей Дирихле. Пусть задана область $G(\mathbf{r})$ с границей Γ; здесь $\mathbf{r}$ есть радиус-вектор точки. Требуется найти решение задачи

$$\Delta u(\mathbf{r}) = -f(\mathbf{r}),\ \mathbf{r} \in G;\ u(\mathbf{r}) = \mu(\mathbf{r}), \mathbf{r} \in \Gamma. \tag{5.1}$$

В отличие от эволюционных задач, разобранных в предыдущих главах, постановка (5.1) не содержит начальных условий. Обобщением задачи (5.1) на случай неоднородной среды является следующая задача:

$$\begin{gathered} \mathtt{div}\left[\kappa(\mathbf{r})\,\mathtt{grad} u(\mathbf{r})\right] = -f(\mathbf{r}),\ \kappa(\mathbf{r}) > 0,\ \mathbf{r} \in G; \\ u(\mathbf{r}) = \mu(\mathbf{r}),\ \mathbf{r} \in \Gamma. \end{gathered} \tag{5.2}$$

В кристаллической среде $\kappa(\mathbf{r})$ может быть не скаляром, а тензором. Задачи с другими краевыми условиями мы не будем рассматривать. Задачи (5.1), (5.2) будем называть стационарными.

Эволюционная задача. Наряду с (5.2) рассмотрим эволюционную задачу для параболического уравнения с теми же гра-

ничными условиями и произвольно выбранными начальными данными:

$$\frac{\partial U(\mathbf{r},t)}{\partial t} = \mathtt{div}\left[\kappa(\mathbf{r})\mathtt{grad}U(\mathbf{r},t)\right] + f(\mathbf{r}),\ \mathbf{r}\in G,\ 0\le t<+\infty,$$
$$U(\mathbf{r},t)=\mu(\mathbf{r}),\ \mathbf{r}\in\Gamma;\ U(\mathbf{r},0)=U_0(\mathbf{r}). \tag{5.3}$$

Сравним решения эволюционной задачи (5.3) и стационарной задачи (5.2). Вычитая (5.2) из (5.3), получим

$$\frac{\partial}{\partial t}\left[U(\mathbf{r},t)-u(\mathbf{r})\right] = \mathtt{div}\left\{\kappa(\mathbf{r})\mathtt{grad}\left[U(\mathbf{r},t)-u(\mathbf{r})\right]\right\},\mathbf{r}\in G;$$
$$U(\mathbf{r},t)-u(\mathbf{r})=0,\ \mathbf{r}\in\Gamma. \tag{5.4}$$

Таким образом, $U(\mathbf{r},t)-u(\mathbf{r})$ удовлетворяет однородному уравнению с нулевыми граничными условиями. Поскольку начальные данные в (5.3) были выбраны произвольно, то без ограничения общности можно считать, что начальные данные задачи (5.4) также выбраны произвольно.

Параболическое уравнение является диссипативным. Для одномерного уравнения ранее было приведено разложение в ряд Фурье (4.6). Все гармоники затухали, причем медленнее всего затухала первая. В курсах математической физики показано, что для многомерной задачи (5.3) решение разлагается аналогичным образом, только вместо одномерных гармоник берутся многомерные собственные функции. Поэтому при $t\to\infty$ все гармоники неограниченно затухают, и $U(\mathbf{r},t)-u(\mathbf{r})\to 0$. Решение нестационарной задачи (5.3) сходится к решению стационарной задачи (5.2) при $t\to\infty$.

Поэтому идея счета на установление такова: вместо стационарной задачи (5.2) решаем нестационарную задачу (5.3) с произвольными начальными данными. Решение проводим до тех пор, пока $\|\partial U/\partial t\|$ не станет достаточно малой. Тогда $U(\mathbf{r},t)$ будет близким к искомому решению $u(\mathbf{r})$. Для решения нестационарной задачи можно воспользоваться разностными схемами, разработанными в подразделе 4.2.

Установление стационарного решения происходит довольно быстро благодаря экспоненциальному характеру затухания начальных данных. Наиболее медленно затухает первая гармоника, так что требуемое время расчета определяется первым собственным значением. Однако сейчас не будем делать оценку требуемого времени, потому что на самом деле мы будем решать не

дифференциальное уравнение, а разностную схему, аппроксимирующую это уравнение.

Начальное условие формально можно брать произвольно. Однако если $U(\mathbf{r},t) \neq u(\mathbf{r})$ при $\mathbf{r} \in \Gamma$, то начальные и граничные условия для задачи (5.3) не согласуются между собой и решение $U(\mathbf{r},t)$ будет разрывно на границе при $t = 0$. Конечно, благодаря диссипативности уравнения этот разрыв затухнет со временем. Но этот начальный разрыв приведет к тому, что амплитуды высоких гармоник в разложении $U(\mathbf{r},t)$ будут большими. Это фактически замедлит выход решения на стационар. Поэтому начальные данные рекомендуется выбирать так, чтобы $U(\mathbf{r},0)$ удовлетворяло граничному условию.

5.1.2. Разностная схема

Далее решение нестационарной задачи снова будем обозначать буквой u, а не U. Для расчета нестационарного уравнения (5.3) ограничимся распространенным, но достаточно простым случаем. Пусть область есть прямоугольный параллелепипед $G = [0 \leq x \leq a, 0 \leq y \leq b, 0 \leq z \leq c]$, а коэффициенты уравнения $\varkappa(\mathbf{r})$ и $f(\mathbf{r})$ дважды непрерывно дифференцируемы. Поскольку $\varkappa(\mathbf{r})$ мы считаем скаляром, а не тензором, то уравнение (5.3) в прямоугольных координатах принимает следующий вид:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\varkappa\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\varkappa\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(\varkappa\frac{\partial u}{\partial z}\right) + f(x,y,z). \quad (5.5)$$

Пространственный оператор не содержит смешанных пространственных производных. Таким образом, в случае изотропной среды пространственный оператор не содержит пространственных производных.

Для кристаллической среды, когда $\varkappa(\mathbf{r})$ есть тензор, вместо (5.5) получится уравнение со смешанными производными. Этот случай гораздо более сложен и его рассматривать не будем.

Для численного расчета задачи составим разностную схему. Введем в области G прямоугольную неравномерную сетку $\omega = \{x_n, 0 \leq n \leq N; y_m, 0 \leq m \leq M; z_k, 0 \leq k \leq K\}$. Шаги этой сетки обозначим как $h_{xn} = x_n - x_{n-1}, h_{ym} = y_m - y_{m-1}, h_{zk} = z_k - z_{k-1}$. Вторые пространственные разности вычислим аналогично бикомпактной схеме (4.78):

$$(\Lambda_x u)_{nmk} = \frac{2}{h_{xn} + h_{x,n+1}} \left[\varkappa \left(x_{n+1/2}, y_m, z_k \right) \frac{u_{n+1,mk} - u_{nmk}}{h_{x,n+1}} - \right.$$
$$\left. - \varkappa \left(x_{n-1/2}, y_m, z_k \right) \frac{u_{nmk} - u_{n-1,mk}}{h_{xn}} \right]. \tag{5.6}$$

Сходный вид имеют разности по остальным направлениям. Эволюционно-факторизованная схема имеет традиционный вид:

$$(E - \frac{\tau}{2}\Lambda_z)(E - \frac{\tau}{2}\Lambda_y)(E - \frac{\tau}{2}\Lambda_x)\frac{\hat{u} - u}{\tau} = (\Lambda_x + \Lambda_y + \Lambda_z)\, u + f. \tag{5.7}$$

Алгоритм ее решения сводится к трем одномерным прогонкам для системы, переписанной в виде трех уравнений (4.70) — (4.72); надо только помнить, что написанная в них функция u имеет смысл решения нестационарного уравнения.

Граничные условия для u являются стационарными: $u(\mathbf{r}, t) = \mu(\mathbf{r})$, $\mathbf{r} \in \Gamma$. Поэтому на границе $\hat{u} = u$. В обозначениях системы (4.70) — (4.72) это можно записать как следующее граничное условие для прогонки по x:

$$[\hat{u} - u]_{\text{гран}} \equiv 0. \tag{5.8}$$

Тогда действие произвольного оператора на последнюю разность также дает нуль. Отсюда следуют граничные условия для прогонки по направлению y (4.71):

$$v_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\Lambda_x \right) \frac{\hat{u} - u}{\tau} \right]_{\text{гран}} = 0. \tag{5.9}$$

Аналогично, для прогонки по направлению z

$$w_{\text{гран}} = \left[\left(E - \frac{\tau}{2}\Lambda_y \right) \left(E - \frac{\tau}{2}\Lambda_x \right) \frac{\hat{u} - u}{\tau} \right]_{\text{гран}} = 0. \tag{5.10}$$

Граничные условия для всех трех прогонок оказались нулевыми. Таким образом, граничные условия в счете на установление записываются существенно проще, чем для произвольной нестационарной параболической задачи.

5.1.3. Оптимальный шаг

Эволюционно-факторизованная схема безусловно устойчива, причем при любом шаге τ все гармоники строго затухают (см.

п. 4.2.2). Поэтому при расчете любым шагом τ влияние начальных данных будет строго затухать, и через достаточно большое число временных шагов расчет выйдет на стационарный режим. В качестве практического критерия окончания итераций целесообразно выбрать условие малости невязки:

$$\|\Lambda u - f\| < \epsilon, \tag{5.11}$$

где ϵ — требуемая точность. Невязку не нужно специально вычислять, так как она является правой частью разностной схемы (5.7).

Однако от величины шага τ сильно зависит общий объем вычислений, т. е. число шагов, которое надо сделать для выполнения критерия (5.11). Найдем величину оптимального шага τ_0, при которой требуемое число шагов будет минимальным. Для этого необходимо рассмотреть, как зависит скорость затухания гармоники от ее номера и величины τ. Напомним, что мы ограничиваемся случаем прямоугольной сетки при отсутствии смешанных производных в уравнении.

Одномерный случай. Сам по себе он неинтересен, поскольку одномерное уравнение решается одномерной прогонкой, но методически полезен. Множитель роста k-й гармоники имеет вид (4.66):

$$\rho_k = \frac{1 + \tau\lambda_{xk}/2}{1 - \tau\lambda_{xk}/2}, 1 \le k \le N_x - 1,\ \lambda_{xk} < 0. \tag{5.12}$$

Множитель роста ρ_k является монотонно убывающей функцией τ и k (рис. 5.1). Чем больше τ, тем быстрее убывает первая гармоника. Однако при слишком больших τ множитель ρ_{N-1} становится слишком близким к минус единице и последняя гармоника будет медленно убывать.

Очевидно, что оптимален шаг τ_0, при котором первая и последняя гармоники убывают с одинаковой скоростью. С учетом знаков множителей для этого должно выполняться $\rho_1 = -\rho_{N-1}$ (рис. 5.2). Подставляя в это соотношение (5.12), получим

$$\tau_0 = 2/\sqrt{\lambda_{x1}\lambda_{x,N-1}} \to \frac{h^2}{\kappa \sin(\pi/N)} \approx \frac{ah}{\pi\kappa} = O(h). \tag{5.13}$$

Стрелкой показано значение шага τ_0 при равномерной сетке и постоянном коэффициенте κ. Нетрудно заметить, что значение

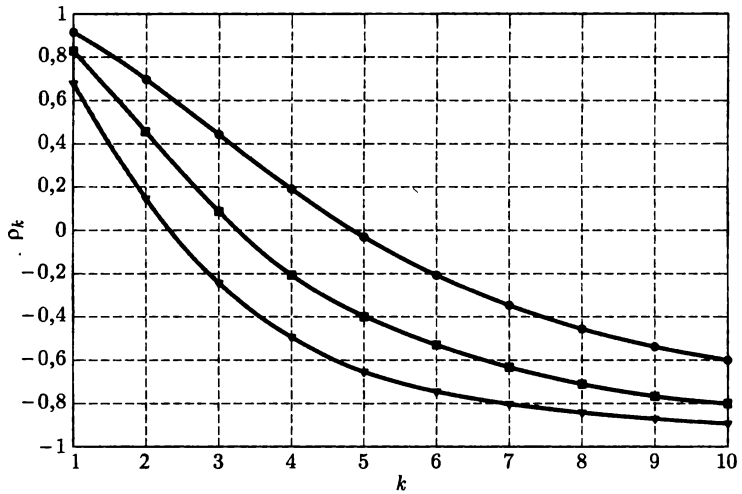


Рис. 5.1. График зависимости множителя роста от k и τ (верхняя кривая: $\tau < \tau_0$, средняя: $\tau = \tau_0$, нижняя: $\tau > \tau_0$)

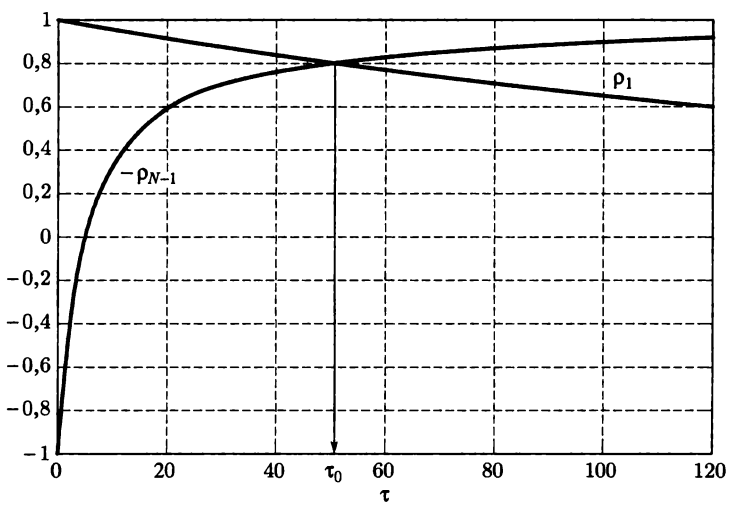


Рис. 5.2. Нахождение оптимального шага

τ_0 совпадает с границей асимптотической устойчивости нестационарной схемы (4.66).

Оценим полное число шагов S, нужное для получения точности ϵ. При оптимальном шаге все гармоники затухают не медленнее, чем первая. Поэтому надо положить $\rho_1^S(\tau_0) = \epsilon$. Учтем, что $\ln(\rho_1(\tau_0)) = \tau_0\lambda_1[1 + O(\tau_0^2\lambda_1^2)]$. Окончательно получаем

$$S = \ln(\epsilon)/\ln(\rho_1(\tau_0)) = \frac{\ln(1/\epsilon)}{2}\sqrt{\frac{\lambda_{N-1}}{\lambda_1}} \to$$
$$\to \frac{\ln(1/\epsilon)}{2}\operatorname{ctg}\left(\frac{\pi}{2N}\right) \approx \frac{\ln(1/\epsilon)}{\pi}N. \tag{5.14}$$

Число шагов оказалось пропорциональным числу интервалов сетки. Такие методы называют *экономичными*. Их трудоемкость является приемлемой. При разумных значениях, $N \sim 100-1000$, число шагов оказывается не слишком большим.

При расчетах с 64-разрядными числами можно рекомендовать значение $\epsilon \sim 10^{-8} - 10^{-12}$. Худшую точность брать нецелесообразно: возможно, потребуется проводить сгущение сеток и применять метод Ричардсона, а это требует достаточного числа достоверных знаков. Меньшее значение ϵ также нецелесообразно: во-первых, процесс установления сходится к решению лишь сеточного уравнения, а не дифференциального. Во-вторых, ошибки округления становятся соизмеримыми с ϵ.

Двумерный случай. В п. 4.2.2 показано, что множитель роста в двумерном случае (4.67) является произведением одномерных: $\rho_{kl} = \rho_{xk}\rho_{yl}$, $1 \le k \le N_x - 1$, $1 \le l \le N_y - 1$. Зависимость этих множителей от τ и индексов соответствует одномерному случаю. Поэтому оптимальным оказывается шаг, при котором произведение множителей роста двух первых гармоник равно произведению двух последних: $\rho_{11} = \rho_{N_x-1,N_y-1}$. Это условие совпадает с условием границы асимптотической устойчивости (4.68) — (4.69):

$$\tau_0 = 2\sqrt{\left(1/\lambda_{x,N_x-1} + 1/\lambda_{y,N_y-1}\right)/\left(\lambda_{x1} + \lambda_{y1}\right)};$$
$$\tau \le \tau_0 \approx \frac{1}{\pi}\left(\frac{h_x^2/\kappa_x + h_y^2/\kappa_y}{\kappa_x/a^2 + \kappa_y/b^2}\right)^{1/2}. \tag{5.15}$$

Число шагов S для достижения точности ϵ определяется аналогичным соотношением $\rho_{11}^S(\tau_0) = \epsilon$. Учитывая, что $\ln(\rho_{11}) \approx \approx \tau(\lambda_{x1} + \lambda_{y1})$, получаем

$$S = \frac{\ln(1/\epsilon)}{2}/\sqrt{\left(1/\lambda_{x,N_x-1} + 1/\lambda_{y,N_y-1}\right)\left(\lambda_{x1} + \lambda_{y1}\right)} \to$$
$$\to \frac{\ln(1/\epsilon)}{\pi}/\sqrt{\left(\frac{\kappa_x}{a^2} + \frac{\kappa_y}{b^2}\right)\left(\frac{h_x^2}{\kappa_x} + \frac{h_y^2}{\kappa_y}\right)}. \tag{5.16}$$

Нетрудно видеть, что $\tau_0 = O(N_x + N_y)$, аналогично одномерному случаю, т. е. алгоритм экономичен.

Более того, при $\kappa_x = \kappa_y$, $h_x = h_y$, $a = b$ число итераций (5.16) в два раза меньше, чем в одномерном случае (5.14)! Этот эффект объясняется тем, что в двумерном случае каждая гармоника гасится произведением сразу двух одномерных (по каждому из направлений) множителей роста. Поэтому в двумерном случае счет на установление более эффективен, чем в одномерном.

Трехмерный случай. Множитель роста (4.65) для произведения трех одномерных гармоник не разлагается на произведение трех одномерных множителей роста. Однако и в этом случае для достижения оптимальной скорости затухания надо сопоставить множители роста для произведения трех младших и трех старших гармоник. Учтем, что при оптимальном шаге величина $\tau|\lambda_1| \ll 1$, а $\tau|\lambda_{N-1}| \gg 1$. Тогда указанные множители роста равны

$$\begin{gathered}\rho_{111} \approx 1 + \tau(\lambda_{x1} + \lambda_{y1} + \lambda_{z1}), \\ \rho_{N_x-1,N_y-1,N_z-1} \approx 1 - 8\frac{\lambda_{x,N_x-1} + \lambda_{y,N_y-1} + \lambda_{z,N_z-1}}{\tau^2\lambda_{x,N_x-1}\lambda_{y,N_y-1}\lambda_{z,N_z-1}}.\end{gathered} \tag{5.17}$$

Оба они положительны и несколько меньше единицы. Приравнивая их, получаем величину оптимального шага:

$$\tau_0 \approx 2\left[\frac{\lambda_{x,N_x-1} + \lambda_{y,N_y-1} + \lambda_{z,N_z-1}}{-\lambda_{x,N_x-1}\lambda_{y,N_y-1}\lambda_{z,N_z-1}(\lambda_{x1} + \lambda_{y1} + \lambda_{z1})}\right]^{1/3} \to$$

$$\to \left(\frac{1}{2\pi^2}\right)^{1/3}\left[\frac{h_x^2h_y^2h_z^2\left(\kappa_x/h_x^2 + \kappa_y/h_y^2 + \kappa_z/h_z^2\right)}{\kappa_x\kappa_y\kappa_z\left(\kappa_x/a^2 + \kappa_y/b^2 + \kappa_z/c^2\right)}\right]^{1/3} = O\left(h^{4/3}\right). \tag{5.18}$$

Заметим, что этот шаг является границей асимптотической устойчивости трехмерной эволюционно-факторизованной схемы. Видно, что он имеет более высокий порядок малости, чем в одномерном (5.13) и двумерном (5.15) случаях. Однако ухудшение не слишком велико. Это намного лучше, чем условие устойчивости явной схемы $\tau = O\left(h^2\right)$.

Число шагов по-прежнему определяем из соотношения $\rho_{111}^S(\tau_0) = \epsilon$. Подставляя сюда (5.17) и (5.18), получим

$$S = \frac{\ln(1/\epsilon)}{2}\left[\frac{\lambda_{x,N_x-1}\lambda_{y,N_y-1}\lambda_{z,N_z-1}}{\left(\lambda_{x,N_x-1}+\lambda_{y,N_y-1}+\lambda_{z,N_z-1}\right)\left(\lambda_{x1}+\lambda_{y1}+\lambda_{z1}\right)^2}\right]^{1/3} \to$$

$$\to \ln\left(\frac{1}{\epsilon}\right)\left(\frac{2}{\pi^4}\right)^{1/3} \times$$

$$\times\left[\frac{\kappa_x\kappa_y\kappa_z}{h_x^2h_y^2h_z^2\left(\kappa_x/h_x^2+\kappa_y/h_y^2+\kappa_z/h_z^2\right)\left(\kappa_x/a^2+\kappa_y/b^2+\kappa_z/c^2\right)^2}\right]^{1/3} =$$

$$= O\left(N^{4/3}\right). \tag{5.19}$$

Видно, что алгоритм нельзя назвать экономичным.

Сравним для наглядности числа шагов при разных размерностях, полагая равными коэффициенты, пространственные шаги и числа интервалов по всем направлениям. Для одномерного, двумерного и трехмерного случаев получим соответственно

$$S_1 = \frac{N}{\pi}\ln(1/\epsilon),\ S_2 = \frac{N}{2\pi}\ln(1/\epsilon),\ S_3 = \frac{\ln(1/\epsilon)}{3}\left(\frac{2N^4}{\pi^4}\right)^{1/3}. \tag{5.20}$$

Двумерный случай требует вдвое меньшего числа шагов, чем одномерный. Однако трехмерный случай при типичных $N \sim 100-1000$ требует в три — шесть раз больше шагов, чем двумерный. Это существенное увеличение трудоемкости. Таким образом, в трехмерном случае трудоемкость расчета с оптимальным шагом нельзя считать хорошей.

Границы спектра. В формулы для определения τ_0 и S входят границы спектра. Они легко вычисляются лишь при постоянных коэффициентах и пространственных шагах (эти случаи показаны стрелкой в формулах). Если либо коэффициенты, либо шаги сетки непостоянны, надо использовать приближенные оценки границ спектра. Эти оценки можно делать методом «замораживания» коэффициентов и шагов. Например, можно полагать

$$\lambda_{x1} \approx -(\pi/a)^2 \min(\kappa_x),\ \lambda_{x,N_x-1} \approx -4\max\left(\kappa_x/h_x^2\right); \tag{5.21}$$

такая оценка для модуля λ_{x1} будет заниженной, а для λ_{x,N_x-1} — завышенной. По другим координатам берутся аналогичные выражения. Такие оценки являются достаточно грубыми, но лучшие оценки получить трудно.

К сожалению, зависимость τ_0 от границ спектра достаточно сильна. Например, возьмем в одномерном случае вместо точного значения λ_{x1} приближенную оценку снизу $\tilde{\lambda}_{x1}$. Из (5.13) и (5.14) видно, что для соответствующего оценочного значения $\tilde{\tau}_0$ выполняется соотношение $\tilde{\tau}_0/\tau_0 = \tilde{S}/S = \left(\lambda_{x1}/\tilde{\lambda}_{x1}\right)^{1/2}$. Если оценка $\tilde{\lambda}_{x1}$ занижена в 10 раз по сравнению с истинным значением, то расчет с найденным по ней $\tilde{\tau}_0$ потребует почти в три раза больше шагов. Это существенно увеличивает трудоемкость.

Такая достаточно сильная зависимость τ_0 от границ спектра ясна из рис. 5.2. Значение τ_0 определяется из пересечения двух гипербол. Изменение каждой границы спектра сдвигает соответствующую гиперболу. При этом точка пересечения сдвигается столь же сильно.

Из изложенного выше следует, что расчет с постоянным оптимальным шагом фактически оказывается довольно трудоемким, хотя в современной практике расчетов такая трудоемкость является хорошей.

5.1.4. Логарифмический набор шагов

Счет на установление может сходиться гораздо быстрее, если вместо постоянного оптимального шага τ_0 использовать специальный набор шагов τ_s, $1 \le s \le S$. Рассмотрим способ построения так называемого ***логарифмического набора***, являющегося наиболее эффективным в данное время. Поясним идею на примере одномерного случая.

Если установление выполнено с помощью набора шагов τ_s, то полный множитель затухания каждой гармоники будет равен произведению одношаговых множителей роста:

$$R_k = \prod_{s=1}^{S} \rho_k(\tau_s) = \prod_{s=1}^{S} \frac{1+\tau_s\lambda_k/2}{1-\tau_s\lambda_k/2},\ 1 \le k \le N-1\ (\lambda_k < 0). \quad (5.22)$$

Напомним, что каждый из сомножителей лежит в пределах $(-1,+1)$, т. е. каждая гармоника на каждом шаге строго затухает. Возьмем $\tau_1 = -2/\lambda_1$; тогда на первом шаге первая гармоника будет полностью погашена, и на последующих итерациях она не возникнет. На втором шаге возьмем $\tau_2 = -2/\lambda_2$ и погасим вторую гармонику. Продолжив этот процесс, мы погасим все гармоники за $N-1$ шагов, и метод установления сойдется

точно. Однако такой процесс не осуществим на практике, ибо он требует знания всех спектральных значений; мы же можем рассчитывать лишь на оценку границ спектра.

Видоизменим процесс следующим образом. Возьмем первый и последний шаги следующим образом:

$$\tau_1 \approx -2/\lambda_1,\ \tau_S \approx -2/\lambda_{N-1} \tag{5.23}$$

(значения S и N пока не связаны друг с другом). Очевидно, $\tau_1 > \tau_S$, поскольку $|\lambda_1| > |\lambda_{N-1}|$. Остальные шаги набора расположим между крайними: $\tau_1 > \tau_2 > \tau_3 > \ldots > \tau_S$. Качественно рассмотрим затухание гармоник при использовании такого набора шагов.

Очевидно, первый шаг почти гасит первую гармонику, а последний шаг — последнюю гармонику. Поскольку $\rho_k(\tau_s)$ монотонно зависит от номера гармоники k и шага τ, то на каждом промежуточном шаге τ_s будет довольно сильно затухать гармоника с некоторым промежуточным номером k. Одновременно при этом будут заметно, но более слабо затухать гармоники с ближайшими номерами k. Таким образом, каждый шаг τ_s заметно гасит некоторую группу гармоник. На всех остальных шагах эта группа гармоник также будет ослабевать, хотя не столь сильно. Возникает задача: как построить последовательность τ_s, $1 \le s \le S$, на отрезке $[\tau_S, \tau_1]$, чтобы после выполнения всех шагов затухание всех гармоник оказывалось наибольшим.

Положим $S = 1$. Если выполняется только один шаг, то наибольшее затухание всех гармоник, включая первую и последнюю, дает оптимальный шаг τ_0. Он определяется формулой (5.13). Сравнивая его с (5.23), получим $\tau_0 = \sqrt{\tau_1 \cdot \tau_S}$, или $\lg(\tau_0) = (\lg(\tau_S) + \lg(\tau_1))/2$. При переходе к логарифмическому масштабу оптимальный шаг оказался серединой отрезка, определяемого максимальным и минимальным шагами (5.23). Отсюда для произвольного S можно по аналогии построить набор шагов, являющийся равномерным в логарифмическом масштабе:

$$\lg(\tau_s) = \lg(\tau_{\min}) + \frac{s-1}{S-1}\lg\left(\frac{\tau_{\max}}{\tau_{\min}}\right), \tag{5.24}$$
$$1 \le s \le S;\ \tau_{\min} = 2/\lambda_{\max},\ \tau_{\max} = 2/\lambda_{\min}.$$

Этот набор шагов называют логарифмическим. Его можно применять для задачи произвольной размерности. Обсудим, как надо выбирать границы спектра разной размерности.

Одномерный случай. В одномерном случае множитель роста на каждом слое имеет вид (5.12), а полный множитель роста определяется формулой (5.22). Поэтому следует принимать

$$\lambda_{\min} = \lambda_{x1} \to (\pi/a_x)^2 \kappa_x,\ \lambda_{\max} = \lambda_{x,N_x-1} \to 4\kappa_x/h_x^2. \qquad (5.25)$$

Здесь стрелкой показаны предельные значения при $\kappa = \mathtt{const}$, $h = \mathtt{const}$; они приведены с точностью до $O\left(h^2\right)$, что достаточно для практических целей. В этом случае удается доказать оценку числа итераций S, необходимого для получения точности ϵ:

$$S = \frac{1}{4} \ln(1/\epsilon) \ln(\lambda_{\max}/\lambda_{\min}) \to \frac{1}{2} \ln(1/\epsilon) \ln(2N/\pi). \qquad (5.26)$$

При высокой точности $\epsilon = 10^{-10}$ и типичных числах интервалов $N \sim 100 - 1000$ получаем $S \approx 50 - 75$. Видно, что логарифмический набор требует гораздо меньшего числа шагов, чем расчет постоянным оптимальным шагом.

Двумерный случай. В этом случае множитель роста ρ_{kl} на одном шаге (4.67) равен произведению двух одномерных множителей. Соответственно полный множитель роста есть

$$R_{kl} = \prod_{s=1}^{S} \frac{1 + \tau_s \lambda_{xk}/2}{1 - \tau_s \lambda_{xk}/2} \times \frac{1 + \tau_s \lambda_{yl}/2}{1 - \tau_s \lambda_{yl}/2}. \qquad (5.27)$$

Его можно рассматривать как одно произведение, содержащее $2S$ сомножителей-дробей. При этом наибольший шаг определяется из условия обращения в нуль либо ρ_{x1}, либо ρ_{y1}; выбирается тот из двух сомножителей, который дает большее значение τ. Соответственно наименьший шаг получаем из условия обращения в нуль сомножителя ρ_{x,N_x-1} или ρ_{y,N_y-1}. Выбирается тот сомножитель, который дает меньшее значение τ. В итоге получаем

$$\lambda_{\min} = \min\left(|\lambda_{x1}|, |\lambda_{y1}|\right),\ \lambda_{\max} = \max\left(|\lambda_{x,N_x-1}|, |\lambda_{y,N_y-1}|\right). \qquad (5.28)$$

Полученные значения надо подставлять в логарифмический набор шагов (5.24). В этом случае остается справедливой оценка числа шагов (5.26), приведенная для одномерного случая.

Однако на практике установление обычно будет достигаться быстрее, чем в одномерных задачах. Это связано с тем, что обычно границы спектра по направлениям x и y не слишком сильно отличаются друг от друга. Тогда множители двух произведений в (5.27) перемежаются. В окрестности каждого τ_s оказывают-

ся малые множители ρ как из первого произведения, так и из второго; это приводит к более быстрому затуханию. Например, если рассматривается эллиптическое уравнение в квадрате при $\kappa_x = \kappa_y = \mathtt{const}$ и $N_x = N_y$, то первое и второе произведения в R_{kl} оказываются одинаковыми и в одномерной оценке (5.26) надо уменьшить число итераций в два раза. Например, при $N \sim 100 - -1000$ для сходимости с $\epsilon = 10^{-10}$ потребуется всего $25 - 30$ итераций, что является превосходным результатом. Получается парадоксальный результат: в двумерном случае логарифмический набор шагов требует меньшего числа шагов, чем в одномерном!

Трехмерный случай существенно сложнее для рассмотрения. В трехмерных эволюционно-факторизованных схемах множитель роста не разлагается на произведение одномерных. Однако на практике логарифмический набор шагов дает хорошие результаты, если аналогично двумерному случаю полагать $\lambda_{\min} = \min(|\lambda_{x1}|, |\lambda_{y1}|, |\lambda_{z1}|)$, $\lambda_{\max} = \max(|\lambda_{x,N_x-1}|, |\lambda_{y,N_y-1}|, |\lambda_{z,N_z-1}|)$. Практика расчетов показывает, что требуемое число шагов S при этом больше, чем в двумерном случае; оно примерно таково же, как в одномерном случае согласно оценке (5.26). Тем самым в трехмерном случае установление также происходит быстро.

Практические рекомендации. При счете на установление с логарифмическим набором шагов требуется заранее оценить границы спектра, вычислить $\tau_{\min}$ и $\tau_{\max}$ и задать желательную точность ϵ. Тогда по оценке (5.26) можно вычислить требуемое число шагов S, а затем построить весь набор τ_s (5.24). При этом нельзя использовать только часть набора шагов; необходимо вычислить все шаги до конца. Использование части набора может оставить непогашенными значительную часть гармоник, а тогда результат будет грубо ошибочным. Напомним, что на каждом шаге амплитуда каждой гармоники строго уменьшается. Поэтому в расчетах не возникает накопления погрешностей. Тем самым выполнять шаги τ_s можно в произвольном порядке. Необходимость заранее вычислять S является некоторым (небольшим) недостатком метода.

Более существенным недостатком является необходимость предварительного задания границ одномерных спектров. В задачах с переменными коэффициентами и неравномерными сетками надо пользоваться оценками границ одномерных спектров (5.21). Они же являются точными значениями при постоянных коэффициентах и равномерных сетках. При практических расчетах

рекомендуется перестраховываться: вместо $\lambda_{\min}$ брать несколько меньшую величину, а вместо $\lambda_{\max}$ — несколько бóльшую (приблизительно в три раза). Это обеспечивает более надежное затухание самых младших и самых старших гармоник.

Рекомендуемое расширение границ спектра приводит к некоторому увеличению S. Однако это увеличение невелико. В самом деле, $\lambda_{\max}/\lambda_{\min}$ приблизительно равно $(2N/\pi)^2$. При типичных $N \sim 100-1000$ величина $\ln(\lambda_{\max}/\lambda_{\min}) = 2\ln(N) \approx 9-14$. Уменьшение $\lambda_{\min}$ и увеличение $\lambda_{\max}$ втрое дает $\ln(\lambda_{\max}/\lambda_{\min}) \approx 11-16$. Соответственно величина S увеличивается на $\sim 20\,\%$. Такое несущественное увеличение трудоемкости вполне искупается повышением надежности расчетов.

Таким образом, объем расчетов при логарифмическом наборе шагов мало чувствителен к границам спектра. Этим он выгодно отличается от расчетов с постоянным оптимальным шагом, для которых трехкратное изменение обоих границ спектра увеличивало бы число шагов в три раза. Однако самым большим преимуществом логарифмического набора является уникально малый объем расчетов. В нем $S \sim \ln(\lambda_{\max}/\lambda_{\min})$, в то время как в лучших известных методах $S \sim \sqrt{\lambda_{\max}/\lambda_{\min}}$ (к лучшим методам относится и расчет с постоянным оптимальным шагом), а во многих методах даже $S \sim \lambda_{\max}/\lambda_{\min}$. Поэтому для решения эллиптических уравнений без смешанных производных в прямоугольном параллелепипеде на прямоугольных сетках рекомендуется пользоваться эволюционно-факторизованными схемами с логарифмическим набором шагов (напомним, что этот метод применим при переменных коэффициентах и неравномерных сетках).

Заметим также, что этот метод можно применять в односвязных областях ступенчатой формы, т. е. «склеенных» из прямоугольников: в этих областях также можно ввести неравномерную прямоугольную сетку и применять эволюционную факторизацию.

5.2. ИТЕРАЦИОННЫЕ МЕТОДЫ

5.2.1. Сложные задачи

Существует много актуальных прикладных стационарных задач, к которым не применим счет на установление с логарифмическим набором шагов. Во-первых, это задачи в областях с

криволинейными границами: в них практически невозможно построить хорошую прямоугольную сетку и приходится строить треугольную (тетраэдральную). Во-вторых, это задачи со смешанными производными; для них не разработано подходящего метода факторизации. В-третьих, это несамосопряженные задачи. Например, задачи диффузии-конвекции. В них в уравнение типа (5.2) добавляется конвективный член переноса. В-четвертых, это задачи для уравнений более высоких порядков. Разностные схемы для таких уравнений имеют общий вид

$$A\mathbf{u} = \mathbf{b},\ A = \{a_{nm}\},\ \mathbf{b} = \{b_n\}. \tag{5.29}$$

Здесь $\mathbf{u}, \mathbf{b}$ — векторы, имеющие размерность M, равную полному числу внутренних узлов многомерной сетки (очевидно, $M \sim N^{\nu}$, где ν есть размерность пространства, а N есть характерное число интервалов пространственной сетки по одному направлению). Квадратная матрица A имеет размер $M \times M$. В зависимости от исходной задачи она может быть эрмитовой или неэрмитовой; для эрмитовых матриц можно разделять случаи знакоопределенной и знаконеопределенной матрицы.

Однако во всех случаях матрица A является очень сильно разреженной: шаблон разностной схемы включает в себя лишь несколько соседних пространственных узлов, поэтому каждая строка матрицы содержит лишь несколько (обычно 5 — 20) ненулевых элементов при огромных порядках матрицы $M \sim 10^4 - -10^9$. В этом случае умножение матрицы A на вектор является недорогой процедурой, по трудоемкости мало отличающейся от сложения векторов и вычисления скалярных произведений.

Для решения задачи (5.29) разработано много итерационных методов, основанных на умножении матрицы на вектор. Из них наиболее быстро сходятся многошаговые итерационные методы сопряженных направлений. Идея их построения такова. Задается некоторая квадратичная форма $\Phi(\mathbf{u})$ в пространстве векторов $\mathbf{u} \in R_M$. Выбирается некоторое начальное приближение $\mathbf{u}_1$ и некоторая прямая, проходящая через эту точку. По этой прямой производят движение, уменьшающее функционал $\Phi(\mathbf{u})$. Точку $\mathbf{u}_2$, в которой функционал достигает минимума на этой прямой, принимают за следующее приближение. Из нее строят новую прямую, перпендикулярную первой. Эти две прямые определяют плоскость. Следующий спуск производится в этой плоскости до достижения минимума функционала. Полученная точка является новым приближением $\mathbf{u}_3$. Через нее проводят прямую,

перпендикулярную этой плоскости. Очередной спуск проводят в трехмерном пространстве и т. д. Таким образом, в пространстве R_M строится ортогональный базис.

Такие методы обладают так называемым ***исчерпыванием***: после M шагов строится полный ортогональный базис пространства, так что $(M+1)$-й шаг является спуском во всем пространстве и дает точное решение. Однако в неодномерных задачах M настолько велико, что расчеты не успевают дойти до исчерпывания. Поэтому на практике ограничиваются умеренным числом итераций, обеспечивающим нужную погрешность сходимости ϵ.

Существует несколько типов таких многошаговых методов, которые удается преобразовать к одношаговой форме. Это обеспечивает простоту алгоритма. Далее для справочных целей приведем без вывода расчетные формы трех наиболее рекомендуемых методов, рассчитанных на разные классы матриц A. Из различных форм записи, имеющихся в литературе, здесь выбраны так называемые рекуррентные формы, в которых число умножений матрицы на вектор минимально. При этом нормирующие множители в формулах выбраны так, чтобы ошибки компьютерного округления были наименьшими.

5.2.2. Сопряженные градиенты

Этот метод рассчитан на эрмитовы знакоопределенные матрицы: $A = A^H > 0$ (или < 0). Напомним, что все собственные значения эрмитовых матриц вещественны; если матрица знакоопределенна, все ее собственные значения имеют тот же знак.

Геометрически искомое решение есть точка экстремума квадратичной формы $(\mathbf{u}, A\mathbf{u} - 2\mathbf{b})$; это минимум при $A > 0$ и максимум при $A < 0$. Для знакопеременной матрицы A экстремум является седловой точкой.

В расчетах на s-й итерации используют следующие векторы: решение $\mathbf{u}_s$, невязка $\mathbf{r}_s$, направление очередного спуска $\mathbf{p}_s$ и вспомогательный вектор $\mathbf{q}_s$. Для начала расчета полагают $\mathbf{p}_0 = 0$ и произвольно выбирают $\mathbf{u}_1$. Итерации $s = 1, 2, 3, \ldots$ вычисляют по следующим формулам:

$$\begin{gathered}\mathbf{r}_s = \begin{cases} A\mathbf{u}_s - \mathbf{b} & \text{для } s = 1, \\ \mathbf{r}_{s-1} - \mathbf{q}_{s-1}/(\mathbf{q}_{s-1}, \mathbf{p}_{s-1}) & \text{для } s = 2, 3, \ldots; \end{cases} \\ \mathbf{p}_s = \mathbf{p}_{s-1} + \mathbf{r}_s/(\mathbf{r}_s, \mathbf{r}_s); \\ \mathbf{q}_s = A\mathbf{p}_s; \quad \mathbf{u}_{s+1} = \mathbf{u}_s - \mathbf{p}_s/(\mathbf{q}_s, \mathbf{p}_s). \end{gathered} \tag{5.30}$$

На стандартных итерациях $s \geq 2$ требуется всего одно умножение матрицы на вектор. Лишь на первой итерации возникает второе умножение. Поэтому формулы (5.30) наименее трудоемки. Итерации можно выполнять до тех пор, пока евклидова норма невязки не станет достаточно малой: $\|\mathbf{r}_s\|_2 < \epsilon$. Для погрешности решения в этом случае справедлива оценка $\|\mathbf{u}_s - \mathbf{u}_{\text{точн}}\|_2 \leq \epsilon\lambda_{\max}/\lambda_{\min}$. Поэтому для хорошо обусловленных матриц, у которых отношение $\lambda_{\max}/\lambda_{\min}$ не очень велико, остановка счета по малости невязки дает разумные результаты.

Сходимость. Для метода (5.30) известна строгая оценка скорости сходимости метода. Она точно совпадает с оценкой (5.14) для счета на установление по эволюционно-факторизованной схеме с постоянным оптимальным шагом. Однако метод сопряженных градиентов имеет важное преимущество: он автоматически обеспечивает данную сходимость, в то время как для нахождения оптимального шага τ_0 требовалось явно задавать значения $\lambda_{\max}$ и $\lambda_{\min}$. Напомним, что на практике эти значения обычно неизвестны, а использование для них приближенных оценок приводит к заметному увеличению числа шагов счета на установление. Практика численных расчетов показывает, что фактическая погрешность немонотонно убывает с возрастанием номера итерации s. Однако для матриц $A = A^H > 0$ (или < 0) эта немонотонность обычно настолько невелика, что почти незаметна на графиках.

В расчетах наблюдается любопытное явление. Для знаконеопределенных эрмитовых матриц $A = A^H$ сходимость метода не доказана. Однако в численных расчетах сходимость зачастую имеет место, хотя при этом немонотонное поведение погрешности очень хорошо заметно на графике.

Для неэрмитовых матриц метод сопряженных градиентов не сходится.

5.2.3. Сопряженные невязки

Этот метод предназначен для эрмитовых знаконеопределенных матриц $A = A^H$. Знаконеопределенность матрицы означает, что одна часть ее собственных значений положительна, а другая — отрицательна. В этом методе вводят еще один дополнительный вектор $\mathbf{g}_s$. Для начала расчета задают $\mathbf{p}_0 = 0$ и произвольно выбирают $\mathbf{u}_1$. Затем считают итерации $s = 1, 2, \ldots$:

$$\mathbf{r}_s = \begin{cases} A\mathbf{u}_s - \mathbf{b} & \text{для } s = 1, \\ \mathbf{r}_{s-1} - \mathbf{q}_{s-1}\dfrac{(r_{s-1}, q_{s-1})}{(q_{s-1}, q_{s-1})} & \text{для } s = 2, 3, \ldots; \end{cases}$$

$$\mathbf{g}_s = A\mathbf{r}_s;$$

$$\mathbf{p}_s = \mathbf{p}_{s-1} + \mathbf{r}_s / (\mathbf{r}_s, \mathbf{g}_s); \qquad (5.31)$$

$$\mathbf{q}_s = \begin{cases} A\mathbf{p}_s & \text{для } s = 1, \\ \mathbf{q}_{s-1} + \mathbf{g}_s / (\mathbf{r}_s, \mathbf{g}_s) & \text{для } s = 2, 3, \ldots; \end{cases}$$

$$\mathbf{u}_{s+1} = \mathbf{u}_s - \mathbf{p}_s (\mathbf{r}_s, \mathbf{q}_s) / (\mathbf{q}_s, \mathbf{q}_s).$$

Стандартная итерация $s \geq 2$ содержит только одно умножение матрицы на вектор. Первая итерация содержит три умножения. Тем самым трудоемкость стандартной итерации невелика.

Сходимость метода определяется той же формулой, что и для сопряженных градиентов; при этом подразумевается $\lambda_{\max} = \max |\lambda_j|$, $\lambda_{\min} = \min |\lambda_j|$. Эта скорость сходимости также обеспечивается автоматически, задавать значения границ спектра не требуется. Численные расчеты показывают, что норма погрешности убывает практически монотонно с возрастанием s; небольшая немонотонность есть, но на графике она обычно незаметна.

Метод построен так, что в отсутствие ошибок округления норма невязки в нем строго убывает; если в расчете наблюдается немонотонность нормы $\|\mathbf{r}_s\|_2$, это означает, что расчет вышел на ошибки округления и его надо останавливать. Однако на практике в рекуррентных формулах (5.31) немонотонность убывания невязки наблюдается крайне редко. Если у формул под фигурными скобками отбросить нижние строчки и при всех s брать верхние строчки формул, получится нерекуррентный вариант; в нем выход невязки на ошибки округления действительно наблюдается. Но нерекуррентный вариант более трудоемок, и применять его нецелесообразно.

Видно, что метод сопряженных невязок применим к более широкому классу матриц, чем метод сопряженных градиентов. При этом его скорость сходимости для любых матриц не хуже. Поэтому в практике расчетов можно не использовать метод сопряженных градиентов.

Для неэрмитовых матриц метод сопряженных невязок не обеспечивает сходимости. Это подтверждается расчетами.

5.2.4. Метод Крейга

Этот метод рассчитан на произвольные матрицы A, в том числе неэрмитовые и даже прямоугольные с числом строк больше числа столбцов; в последнем случае линейная система (5.29) является переопределенной. Переопределенные системы часто возникают в задачах экономики с противоречивыми критериями эффективности.

Для начала расчета полагают $\mathbf{p}_0 = 0$ и произвольно выбирают $\mathbf{u}_1$. Стандартная итерация $s = 1, 2, \ldots$ выполняется по следующим формулам:

$$\begin{gathered} \mathbf{r}_s = A\mathbf{u}_s - \mathbf{b}, \\ \mathbf{p}_s = \mathbf{p}_{s-1} + \mathbf{r}_s / (\mathbf{r}_s, \mathbf{r}_s), \\ \mathbf{q}_s = A^H \mathbf{p}_s, \\ \mathbf{u}_{s+1} = \mathbf{u}_s - \mathbf{q}_s / (\mathbf{q}_s, \mathbf{q}_s). \end{gathered} \tag{5.32}$$

Стандартная итерация содержит два умножения матрицы на вектор; меньше сделать невозможно. Форма (5.32) является нерекуррентной; переход к рекуррентной форме здесь не уменьшает трудоемкости и поэтому нецелесообразен.

Доказано, что итерации (5.32) сходятся для произвольной матрицы. При этом погрешность $\|\mathbf{u}_s - \mathbf{u}_{\text{точн}}\|_2$ строго убывает с возрастанием s.

Если система (5.29) переопределенная, то у нее не существует решения в обычном смысле; существует только ***квазирешение***, минимизирующее норму невязки $\|\mathbf{r}\|_2$; сама невязка при этом не может обратиться в нуль. Для переопределенных систем процесс Крейга сходится к квазирешению.

Оценок сходимости в литературе не удалось найти. Численные расчеты показывают, что для симметричных матриц метод Крейга сходится медленнее, чем оба предыдущих метода: судя по расчетам, погрешность убывает как геометрическая прогрессия со знаменателем $(1 - 2\lambda_{\min}/\lambda_{\max})$, что дает $S = O(\lambda_{\max}/\lambda_{\min})$. Поэтому его целесообразно применять лишь для несимметричных матриц.

5.2.5. Погрешности

В задачах математической физики можно считать, что компоненты матрицы A и вектора $\mathbf{b}$ заданы с точностью до ошибок компьютерного округления. Все три итерационных метода

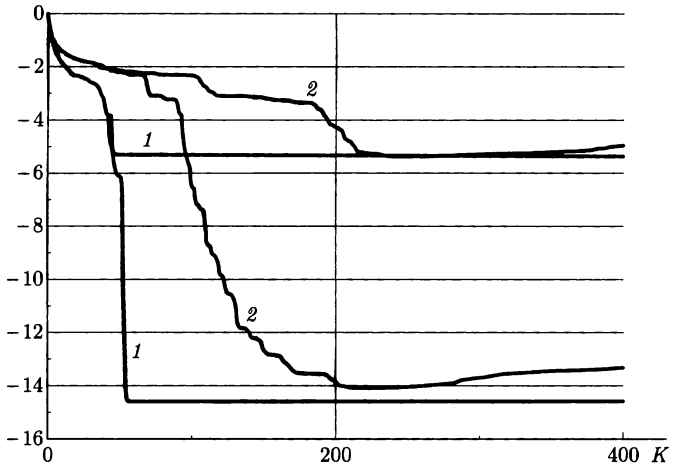


Рис. 5.3. Зависимость логарифма погрешности от номера итерации. Расчет на 32 (тонкие линии)- и 64 (толстые линии)-разрядном компьютерах:

1 — метод сопряженных градиентов или минимальных невязок; *2* — метод Крейга

обладают исчерпыванием, т. е. для матрицы порядка M в отсутствие ошибок округления $(M+1)$-я итерация дает точный ответ. При этом погрешность сравнительно медленно убывает вплоть до M-й итерации, а на следующей итерации скачком обращается в нуль.

Ошибки округления меняют картину. В методе сопряженных градиентов погрешность действительно медленно убывает до M-й итерации. Затем следует быстрый (за $1-5$ итераций) спад погрешности, но не до нуля, а до некоторого фонового значения. Это фоновое значение тем больше, чем хуже обусловленность матрицы A (рис. 5.3). Найти решение с точностью выше фоновой практически невозможно.

При расчете того же примера методом Крейга на первых M итерациях, т. е. до исчерпывания, погрешность убывает также. Далее вместо быстрого спада следует довольно медленный. Но в конце концов погрешность выходит на то же фоновое значение. Причина такого влияния ошибок округления на область быстрого спада не объяснена. При расчете с 32 разрядами этот выход происходит в пределах рис. 5.3, а при расчетах с 64 разрядами — за пределом графика.

После выхода на фон дальнейший расчет не имеет смысла. Для нахождения момента такого выхода существует хороший эвристический критерий. Пусть δ есть относительная ошибка однократного округления (для 32-разрядных чисел $\delta \sim 10^{-7}$, а для 64-разрядных чисел $\delta = 10^{-16}$). Считаем, что округления происходят случайно. Тогда влияния многократных умножений на округление нет, а при большом числе арифметических сложений (вычитаний) складываются квадраты абсолютных погрешностей. В результате абсолютная погрешность вычисления вектора невязки будет

$$\Delta^2 = \delta^2 \sum_{n=1}^{M} \left(b_N^2 + \sum_{m=1}^{M} a_{nm}^2 u_m^2 \right). \tag{5.33}$$

На начальных итерациях $\Delta / \|\mathbf{r}\|_2 \ll 1$. При возрастании s это отношение растет, хотя не строго монотонно. По мере приближения погрешности к фоновому значению это отношение начинает возрастать все быстрее и наконец может превысить единицу. Это означает приблизительный выход погрешности на фон.

Вычислять значения Δ на каждой итерации невыгодно: объем вычислений соответствует дополнительному умножению матрицы на вектор. Учтем, что когда решение стремится к точному, величина $\Delta \to \texttt{const}$. Поэтому достаточно вычислить Δ только один раз, когда норма $\|\mathbf{r}_s\|$ станет достаточно малой, но еще не фоновой. Далее расчет ведется до выполнения следующего критерия:

$$\Delta^2 \sum_{s} \frac{1}{|\mathbf{r}_s|^2} = 1. \tag{5.34}$$

Суммирование по s введено в (5.34) для того, чтобы убрать эффект немонотонного поведения невязки. Заметим, что этот критерий особенно хорошо работает для рекуррентных форм записи итерационных методов; для нерекуррентной записи он может вызвать несколько преждевременную остановку расчета.

Пусть известны границы спектра матрицы; допустимо также использовать нижнюю оценку $\lambda_{\min}$ и верхнюю оценку для $\lambda_{\max}$. Тогда по значению невязки можно оценить погрешность решения:

$$\|\mathbf{u}_s - \mathbf{u}_{\text{точн}}\|_2 \le (\lambda_{\max}/\lambda_{\min}) \|\mathbf{r}_s\|_2 . \tag{5.35}$$

Это мажорантная оценка; фактическая погрешность обычно оказывается заметно меньшей. Однако если расчет не останов-

лен по критерию (5.34), а существенно продвинулся в область фона, то оценкой (5.35), особенно при расчете по рекуррентным формулам, нельзя пользоваться: в этом случае невязка на фоне может существенно убывать и фактическая погрешность будет многократно превышать правую часть (5.35).

Замечание. Счет на установление в подразделе 5.1 тоже можно рассматривать как итерационный процесс. Расчет по эволюционно-факторизованным схемам с постоянным оптимальным шагом τ_0 имеет ту же скорость сходимости, что и методы сопряженных направлений; но он менее удобен, так как требует априорного знания или оценки границ спектра матрицы. Расчет же с логарифмическим набором шагов является несравненно более быстрым итерационным процессом. Он также требует знания границ спектра, но его скорость перевешивает этот недостаток.

5.3. ДРУГИЕ МЕТОДЫ

5.3.1. Метод Ритца

Вариационные методы применяются к эллиптическим уравнениям в частных производных независимо от числа измерений. Рассмотрим задачу:

$$\begin{gathered}\mathtt{div}\left[\varkappa(\mathbf{r})\,\mathtt{grad}u(\mathbf{r})\right]-\rho(\mathbf{r})\,u(\mathbf{r})=-f(\mathbf{r}),\ \varkappa(\mathbf{r})>0,\ \mathbf{r}\in G;\\ u(\mathbf{r})=\mu(\mathbf{r}),\ \mathbf{r}\in\Gamma.\end{gathered}\tag{5.36}$$

Дифференциальный оператор является самосопряженным. Поэтому задача (5.36) эквивалентна задаче на минимум функционала $\Phi[u]=(u,\ Au-2f)$. С помощью формул векторного анализа ее можно записать в следующем виде:

$$\int\limits_G\left[\varkappa(\mathbf{r})\,(\mathtt{grad}u)^2+\rho(\mathbf{r})\,u^2(\mathbf{r})-2f(\mathbf{r})\,u(\mathbf{r})\right]d\mathbf{r}=\min,\ u_\Gamma=\mu(\mathbf{r}).\tag{5.37}$$

В интегральной форме (5.37) задача сформулирована лучше, чем в исходной дифференциальной форме (5.36): она лежит ближе к исходным законам сохранения. Поэтому уравнение (5.36) описывает лишь те решения $u(\mathbf{r})$, которые имеют вторые производные, а задача (5.37) допускает решения $u(\mathbf{r})$, имеющие лишь первые производные (которые должны быть интегрируемыми), т. е. она имеет более широкий класс решений.

Возьмем некоторую функцию $\phi_0(\mathbf{r})$, удовлетворяющую граничному условию из (5.36), и полную систему функций $\phi_m(\mathbf{r})$, $m = 1, 2, \ldots$, обращающихся в нуль на границе. Будем искать приближенное решение задачи (5.37) в следующем виде:

$$u(\mathbf{r}) \approx u_M(\mathbf{r}) \equiv \phi_0(\mathbf{r}) + \sum_{m=1}^{M} c_m \phi_m(\mathbf{r}); \qquad \phi_0(\mathbf{r}) = \mu(\mathbf{r}),\ \phi_m(\mathbf{r}) = 0,\ \mathbf{r} \in \Gamma. \tag{5.38}$$

Подставляя (5.38) в (5.37), получим задачу на минимум квадратичной функции неизвестных коэффициентов c_m; для простоты ограничимся случаем $\phi_0(\mathbf{r}) \equiv 0$, соответствующим $u_\Gamma = 0$. Задача на минимум примет следующий вид:

$$\int_G \left[\sum_{k=1}^{M} \sum_{m=1}^{M} c_k c_m \left(\kappa \mathbf{grad}\phi_k \mathbf{grad}\phi_m + \rho \phi_k \phi_m \right) - 2f \sum_{k=1}^{M} c_k \phi_k \right] d\mathbf{r} = \min. \tag{5.39}$$

Приравнивая нулю производные по коэффициентам, получим для определения c_m систему линейных уравнений

$$\sum_{m=1}^{M} c_m \int_G \left(\kappa \mathbf{grad}\phi_k \mathbf{grad}\phi_m + \rho \phi_k \phi_m \right) d\mathbf{r} = \int_G f \phi_k d\mathbf{r},\ 1 \le k \le M. \tag{5.40}$$

Обоснование сходимости метода Ритца при $M \to \infty$ рассматривалось в кн. 1.

При практическом применении метода Ритца успех сильно зависит от выбора системы функций $\phi_m(\mathbf{r})$. Если область имеет несложную форму, то нередко выбирают систему с разделяющимися переменными; например, для прямоугольника полагают $\phi_{km}(\mathbf{r}) = \xi_k(x)\,\eta_m(y)$, а для круга: $\phi_{km}(\mathbf{r}) = \xi_k(r)\,\eta_m(\theta)$. Метод Ритца применяется даже в сложнейших задачах квантовой химии, где требуется решать уравнение Шредингера для системы всех валентных электронов молекулы. Число функций M может при этом превышать 10^4.

Замечания. *1.* Пусть в одномерной задаче для получения удовлетворительной точности требовалось брать M членов ряда. Тогда в аналогичной двумерной задаче базис есть произведение одномерных базисов, и полное число используемых одномерных функций будет

$\sim M^2$, а в трехмерном — M^3. Однако здесь возможно существенное уменьшение числа членов. Пусть базис выбран удачно и коэффициенты c_m достаточно быстро убывают с ростом индекса. Тогда в двумерном произведении $c_m\gamma_k$ малыми будут те произведения, где велик хотя бы один из индексов. Тем самым можно использовать не полную матрицу коэффициентов $c_m\gamma_k$, а лишь ее часть:

$$u(x,y) \approx \sum_{m=1}^{M} \sum_{k=1}^{K(m)} c_m\gamma_k\xi_m(x)\eta_k(y). \tag{5.41}$$

Оптимальный вид кривой $K(m)$ зависит от специфики задачи. При удачном подборе базисов зачастую оказывается, что оптимальная кривая является гиперболой. Например, $K(m) = M/m$. При этом число учтенных коэффициентов составляет $M\ln(M) - M + 1$, что много меньше M^2. Еще большая экономия числа коэффициентов возможна в трехмерном случае.

2. Если оператор в задаче (5.36) не самосопряженный, то вместо метода Ритца применяют метод Галеркина (см. п. 1.4.3.).

3. Метод Ритца применим к задаче Штурма — Лиувилля (задаче на собственные значения).

5.3.2. Быстрое преобразование Фурье

Прямые методы. В подразделе 5.2 дифференциальное уравнение аппроксимировалось разностной схемой — системой линейных алгебраических уравнений. Полученная алгебраическая система решалась итерационными методами. Однако для систем достаточно общего вида итерационные методы сходятся не слишком быстро. Лишь в достаточно простом случае существует логарифмический набор шагов (п. 5.1.4), сходящийся быстро. Поэтому представляет интерес нахождение таких систем, для которых возможно очень быстрое решение специализированными прямыми методами.

Самым быстрым из известных прямых методов является так называемое быстрое преобразование Фурье. Этот метод применим к задаче Дирихле в прямоугольнике (прямоугольном параллелепипеде), решаемой на равномерных сетках. Метод экономичен в том случае, если число узлов N_α по каждой стороне является произведением малых целых чисел; особенно он эффективен при $N_\alpha = 2^{r_\alpha}$.

Условия применимости метода кажутся довольно искусственными. Для задач математической физики это обычно серьезное

ограничение. Однако метод нашел широкое применение в задачах обработки изображений. Поэтому метод включен в книгу.

Одномерный случай лежит в основе многомерных построений, поэтому начинаем рассмотрение с него.

1. Возьмем одномерную краевую задачу с постоянными коэффициентами $u_{xx} - \mu u = -f(x)$ и краевыми условиями первого рода; без ограничения общности краевые условия можно взять периодическими. Составим разностную схему на равномерной сетке $\{x_n = nh,\ 0 \le n \le N\}$:

$$\frac{1}{h^2}\left(u_{n-1} - 2u_n + u_{n+1}\right) - \mu u_n = -\phi_n,\ 1 \le n \le N-1, \qquad u_0 = \phi_0,\ u_N = \phi_0. \tag{5.42}$$

Система (5.42) с учетом периодичности содержит N неизвестных. Будем искать разностное решение в виде суммы Фурье с таким же числом членов:

$$u_n = \sum_{q=0}^{N-1} a_q w^{nq}, \text{ где } w = \exp(2\pi i/N). \tag{5.43}$$

Подставим сумму (5.43) в соотношение (5.42), умножим на $w^{-np} = \exp(-2\pi i n p/N)$ и просуммируем по n от 0 до $N-1$.

При выполнении этих выкладок учтем одно соотношение. Легко взять следующую сумму как геометрическую прогрессию и проверить, что

$$\sum_{n=0}^{N-1} w^{nq} = 0 \text{ при } q \ne 0 \text{ и равна } N \text{ при } q = 0. \tag{5.44}$$

Аналогичные соотношения справедливы для произведения гармоник, просуммированных по пространственным узлам. Это суть условия ортогональности дискретных гармоник на равномерных сетках. Выполняя предписанные в предыдущем абзаце действия с учетом условий ортогональности, получим коэффициент Фурье решения a_p через коэффициент Фурье правой части b_p:

$$a_p = b_p \Big/ \left(\frac{4}{h^2}\sin^2\left(\frac{\pi p}{N}\right) + \mu\right), \tag{5.45}$$

где

$$b_p = \frac{1}{N}\sum_{n=0}^{N-1} \phi_n w^{-np},\ 0 \le p \le N-1. \tag{5.46}$$

Это классические формулы Бесселя. Однако они неэкономичны. Необходимо вычислить N коэффициентов b_p, причем нахождение каждого коэффициента по формуле (5.46) требует примерно $2N$ операций. Следовательно, задача (5.42) решается за $2N^2$ операций, т. е. намного медленнее, чем в методе прогонки.

2. Если число интервалов сетки составное, $N = KL$, то формулу (5.46) можно преобразовать так, что требуемое количество операций уменьшится. Представим индексы n и p в следующем виде:

$$\begin{aligned} &n = l_1 + Ll_2, \quad 0 \le l_1 \le L-1, \quad 0 \le l_2 \le K-1, \\ &p = p_1 + Kp_2, \quad 0 \le p_1 \le K-1, \quad 0 \le p_2 \le L-1. \end{aligned} \tag{5.47}$$

Запишем формулу (5.46) в виде двойной суммы:

$$b_p = \frac{1}{KL} \sum_{l_1=0}^{L-1} \sum_{l_2=0}^{K-1} \phi_{l_1+Ll_2} w^{-p_1 l_1 - Lp_1 l_2 - Kl_1 p_2 - LKl_2 p_2}. \tag{5.48}$$

Отбросим в показателе степени последнее слагаемое, ибо $w^{LK} = 1$. Получим следующее выражение коэффициентов Фурье:

$$b(p) \equiv b_p = \frac{1}{L} \sum_{l_1=0}^{L-1} b(l_1, p_1)\, w^{-pl_1}, \; 0 \le p \equiv p_1 + Kp_2 \le N-1, \tag{5.49}$$

где

$$b(l_1, p_1) = \frac{1}{K} \sum_{l_2=0}^{K-1} \phi_{l_1+Ll_2} w^{-Lp_1 l_2}, \; 0 \le l_1 \le L-1, \; 0 \le p_1 \le K-1. \tag{5.50}$$

Вычисление N коэффициентов $b(p)$ по формуле (5.49) требует $2NL$ операций. Вычисление $LK = N$ вспомогательных коэффициентов $b(l_1, p_1)$ по формуле (5.50) производится еще за $2NK$ операций. Следовательно, число операций, необходимое для нахождения коэффициентов Фурье по формулам (5.49) — (5.50), равно $2N(L+K)$. Оно существенно меньше, чем $2N^2$ (например, при $K = L = \sqrt{N}$ меньше в $\sqrt{N}/2$ раз).

3. Если K в свою очередь разбивается на множители, то формулу (5.50) следует преобразовать аналогичным образом. Это позволит дополнительно уменьшить объем вычислений.

Приведем без вывода рекуррентные формулы вычисления коэффициентов Фурье для случая $N = L^r$:

$$
\begin{gathered}
b(p) = \frac{1}{L} \sum_{l_1=0}^{L-1} b\,(l_1, p_1)\, w^{-pl_1}, \\
b(l_1, l_2, \ldots, l_k, p_k) = \\
= \frac{1}{L} \sum_{l_{k+1}=0}^{L-1} b(l_1, l_2, \ldots, l_{k+1}, p_{k+1}) w^{-L^k l_{k+1} p_k}, \\
1 \le k \le r-2, \\
b(l_1, l_2, \ldots, l_{r-1}, p_{r-1}) = \\
= \frac{1}{L} \sum_{l_r=0}^{L-1} \Phi_{l_1 + Ll_2 + L^2 l_3 + \cdots + L^{r-1} l_r} w^{-L^{r-1} l_r p_{r-1}}, \\
0 \le l_k \le L-1,\ 0 \le p_k \le L^{r-k} - 1.
\end{gathered}
\tag{5.51}
$$

Число вспомогательных коэффициентов k-го ранга $b(l_1, \ldots, l_k, p_k)$ равно N, поэтому для вычисления коэффициентов всех рангов по формулам (5.51) требуется около $2NLr$ операций.

Если учесть, что $L = N^{1/r}$, то нетрудно найти оптимальное число сомножителей $r_{\text{опт}} \approx \ln(N)$ и оптимальное значение $L_{\text{опт}} \approx e \approx 3$. Но для программирования считается более удобным, если $N = 2^r$ и $L = 2$; в последнем случае требуемое число операций равно $4N \log_2(N)$, что мало отличается от оптимального случая и почти не уступает по скорости методу прогонки (которая требует $9N$ операций).

Двумерный случай. Обобщение этого метода на случай многих измерений очевидно. Пусть, например, для уравнения с постоянными коэффициентами

$$u_{xx} + u_{yy} - \mu u = -f(x, y)$$

поставлена первая краевая задача в прямоугольной области. Введем равномерную сетку $x_n = nh_x$, $y_m = mh_y$, $0 \le n \le N$, $0 \le m \le M$, и составим разностную схему

$$
\begin{gathered}
\frac{1}{h_x^2} (u_{n-1,m} - 2u_{nm} + u_{n+1,m}) + \\
+ \frac{1}{h_y^2} (u_{n,m-1} - 2u_{nm} + u_{n,m+1}) - \mu u_{nm} = -\phi_{nm}.
\end{gathered}
\tag{5.52}
$$

Будем искать разностное решение в виде разложения Фурье

$$u_{nm} = \sum_{p=0}^{N-1} \sum_{q=0}^{M-1} a_{pq} w_x^{np} w_y^{mq}, \qquad (5.53)$$
$$w_x = \exp(2\pi i/N),\ w_y = \exp(2\pi i/M).$$

Аналогично одномерному случаю, получим следующие выражения для коэффициентов Фурье:

$$a_{pq} = b_{pq} / \left(\frac{4}{h_x^2} \sin^2 \left(\frac{\pi p}{N} \right) + \frac{4}{h_y^2} \sin^2 \left(\frac{\pi q}{M} \right) + \mu \right), \qquad (5.54)$$

где

$$b_{pq} = \frac{1}{NM} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \phi_{nm} w_x^{-np} w_y^{-mq}. \qquad (5.55)$$

Запишем последнюю формулу в следующем виде:

$$b_{pq} = \frac{1}{N} \sum_{n=0}^{N-1} \beta_{nq} w_x^{-np},$$
$$\beta_{nq} = \frac{1}{M} \sum_{m=0}^{M-1} \phi_{nm} w_y^{-mq}. \qquad (5.56)$$

Каждая сумма в формулах (5.56) имеет тот же вид, что и в формуле (5.48). Поэтому если N и M разлагаются на множители, то каждую сумму можно вычислить по рекуррентным формулам типа (5.51). Если при этом $N = L_x^{r_x}$ и $M = L_y^{r_y}$, то число операций на каждый узел сетки, аналогично одномерному случаю, есть $O(r_x L_x + r_y L_y) = O(\log(NM))$. Следовательно, быстрое преобразование Фурье даже в многомерном случае по экономичности мало уступает самому быстрому одномерному методу — прогонке.

Замечания. 1. При обработке изображений сигнал записывают с помощью ПЗС матрицы. Ячейки матрицы расположены равномерно, а числа ячеек по каждой стороне выбираются из требований техники. Обычно их выбирают из условия $N = 2^r$: это позволяет применить быстрое преобразование Фурье к обработке сигналов. Экономичность данного алгоритма важна в задачах космическсй связи.

2. Область применимости быстрого преобразования Фурье ýже, чем у счета на установление с логарифмическим набором шагов. Однако быстрое преобразование Фурье является прямым методом, и в нем не возникает вопроса о точности сходимости итераций.

3. Существуют прямые методы, по скорости и области применения аналогичные БПФ. Это метод декомпозиции, нечетно-четной редукции и др.

5.3.3. Чебышёвский набор шагов

Факторизованные схемы не приспособлены для расчетов на многопроцессорных компьютерах: они плохо распараллеливаются. Кроме того, факторизованные схемы строятся лишь на прямоугольных сетках, причем при отсутствии смешанных производных.

Для многопроцессорных компьютеров удобна явная схема: она естественно распараллеливается. Такую схему нетрудно написать для уравнения со смешанными производными и даже для случая не скалярного, а тензорного коэффициента $\kappa(\mathbf{r})$. Схему можно записать для произвольной области G, в том числе с криволинейной границей, а также для непрямоугольных сеток (например, треугольных, которые хорошо аппроксимируют криволинейные границы). Схема единообразно пишется при произвольном числе измерений. Поэтому явным схемам уделяют серьезное внимание.

Однако у явных параболических схем есть серьезный недостаток — условная сходимость с сильным ограничением на шаг τ. Например, напишем простейшую явную трехмерную схему на прямоугольной сетке:

$$(\hat{u} - u)/\tau = \Lambda u + f,\ \Lambda = \Lambda_x + \Lambda_y + \Lambda_z. \tag{5.57}$$

Ее условие устойчивости легко получается методом гармоник:

$$\tau \le \tau_0,\ 2\tau_0 \left(\kappa_x/h_x^2 + \kappa_y/h_y^2 + \kappa_z/h_z^2\right) = 1. \tag{5.58}$$

Если вести счет на установление с постоянным шагом τ, то оптимальным будет значение $\tau = \tau_0 = O\left(h^2\right)$. На подробных сетках $h = a/N \ll 1$, так что для расчета потребуется $S = O\left(N^2\right)$ шагов. Это в $\sim N$ раз превышает число шагов, необходимое для расчета по неявной эволюционно-факторизованной схеме с оптимальным шагом. Поскольку типичные $N \sim 100 - 1000$, такая трудоемкость неприемлема, т. е. явная схема с постоянным шагом непригодна для практических расчетов.

Однако для явных схем был построен так называемый чебышёвский набор шагов τ_s, обеспечивающий установление за

$S = O(N)$ шагов. Приведем этот набор шагов без обоснования, как справочный материал. Оператор Λ является самосопряженным и отрицательным. Пусть известны границы его спектра, модули которых обозначим через $\lambda_{\min}$ и $\lambda_{\max}$; это значит, что выполняются операторные неравенства

$$\lambda_{\min} E \le -\Lambda \le \lambda_{\max} E. \tag{5.59}$$

Тогда величины $1/\tau_s$ являются корнями чебышёвского многочлена первого рода степени S, построенного на отрезке $[\lambda_{\min}, \lambda_{\max}]$:

$$\tau_s = 2/\left[(\lambda_{\max} + \lambda_{\min}) + (\lambda_{\max} - \lambda_{min}) \cos\left(\frac{\pi(2s-1)}{2S}\right)\right], \tag{5.60}$$
$$1 \le s \le S.$$

Число шагов S, нужное для установления с точностью ϵ, является точно таким же, как для эволюционно-факторизованной схемы с постоянным оптимальным шагом:

$$S = \frac{\ln(1/\epsilon)}{2} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}. \tag{5.61}$$

Для области произвольной формы получить оценки $\lambda_{\min}$ и $\lambda_{\max}$ крайне трудно. Лишь для прямоугольного параллелепипеда и задач с разделяющимися переменными можно написать строгие выражения:

$$\lambda_{\min} = |\lambda_{x1}| + |\lambda_{y1}| + |\lambda_{z1}|, \ \lambda_{\max} = |\lambda_{x,N_x-1}| + |\lambda_{y,N_y-1}| + |\lambda_{z,N_z-1}|. \tag{5.62}$$

При постоянных коэффициентах и равномерных сетках можно выразить эти значения через параметры задачи и сетки, как это было сделано для эволюционно-факторизованных схем.

Для практического использования чебышёвского набора шагов (5.60) надо знать границы $\lambda_{\min}$ и $\lambda_{\max}$. Допустимо также использовать их оценки: для $\lambda_{\min}$ надо брать заниженную оценку, а для $\lambda_{\max}$ — завышенную. Чувствительность числа шагов S к точности этих оценок столь же велика, как для расчетов по эволюционно-факторизованным схемам с τ_0. Она значительно сильнее, чем для логарифмического набора шагов. Таким образом, явные схемы с чебышёвским набором шагов существенно уступают эволюционно-факторизованным схемам с логарифмическим набором шагов как по скорости установления, так и по

чувствительности к границам спектра. Зато явные схемы позволяют решать более сложные задачи и эффективно использовать ресурсы многопроцессорных компьютеров.

Для эволюционно-факторизованных схем на каждом шаге все гармоники строго затухали; поэтому порядок выполнения шагов был несуществен. Однако в явных схемах с чебышёвским набором шагов часть шагов τ_s имеет величину больше τ_0, и на этих шагах появляется неустойчивость: ошибки нарастают. На остальных шагах $\tau_s < \tau_0$ и ошибки затухают. В конечном итоге затухание должно преобладать, и при полном выполнении S шагов произойдет затухание ошибок до расчетного уровня погрешности ϵ. Однако на промежуточных шагах ошибка в принципе может оказаться большой. При расчетах на компьютерах с недостаточной разрядностью ошибка может выйти за пределы представимых чисел.

Преодолеть эту трудность можно двумя способами. Первый — использовать компьютер с достаточно большой разрядностью чисел и не думать о порядке выполнения шагов. Второй способ пригоден для компьютеров с невысокой разрядностью чисел. Надо выполнять шаги в определенной последовательности, чтобы в каждой паре шагов один давал нарастание ошибки, а второй — соответствующее погашение. Например, простейший такой способ — выбор последовательности

$$\tau_1, \tau_S;\ \tau_2, \tau_{S-1};\ \tau_3, \tau_{S-2}; \ldots . \tag{5.63}$$

Еще лучше устойчивость расчета, если в этих парах шагов провести дальнейшую перегруппировку: после первой пары поставить последнюю, после второй — предпоследнюю и т. д. В полученной последовательности четверок шагов можно произвести аналогичную группировку и продолжить этот процесс. Такие ***упорядоченные*** чебышёвские наборы шагов обеспечивают хорошую устойчивость даже на компьютерах с очень малой разрядностью.

Чебышёвский набор шагов можно рассматривать как итерационный метод того же класса, что и метод сопряженных градиентов. Он имеет ту же скорость сходимости. Но чебышёвский набор шагов требует: во-первых, априорного знания границ спектра матрицы; во-вторых, задания желательного числа итераций S еще до начала расчета; в-третьих, необходимо упорядочивать шаги. Поэтому в настоящее время этот метод редко применяется.

Глава 6

ГИПЕРБОЛИЧЕСКИЕ УРАВНЕНИЯ

6.1. ТРЕХСЛОЙНЫЕ СХЕМЫ

6.1.1. Постановка задачи

К гиперболическим уравнениям приводят задачи колебания струны, движения сжимаемого газа, распространения возмущений электромагнитных полей и многие другие.

Типичным примером одномерной задачи является задача малых колебаний натянутой струны с распределенной по длине нагрузкой $f(x,t)$:

$$u_{tt} = c^2 u_{xx} + f(x,t),\ 0 \le x \le a,\ 0 \le t \le T; \qquad (6.1)$$

величина c имеет размерность скорости. Это же уравнение описывает плоские акустические волны в газе при наличии внешнего силового поля f. Если уравнение (6.1) однородно ($f \equiv 0$), то нетрудно заметить, что его решение является автомодельным:

$$u(x,t) = \phi_1(x+ct) \text{ и } u(x,t) = \phi_2(x-ct), \qquad (6.2)$$

здесь $\phi_{1,2}$ — произвольные функции. Оно имеет вид волны, бегущей с постоянной скоростью. Знак «плюс» соответствует волне, бегущей влево, а знак «минус» дает волну, бегущую вправо. Общее решение есть сумма двух волн, бегущих в разных направлениях.

Аналогично уравнению переноса бегущая волна требует постановки граничного условия на той границе, с которой она бежит. Две бегущие волны требуют постановки условия на обоих границах отрезка. Простейшими являются граничные условия первого рода:

$$u(0,t) = \mu_1(t),\ u(a,t) = \mu_2(t),\ 0 \le t \le T; \qquad (6.3)$$

они соответствуют заданным законам движения концов струны. Возможны также граничные условия других типов.

В отличие от параболического уравнения уравнение (6.1) имеет второй порядок по t. Поэтому оно требует постановки двух начальных условий. Обычно ими являются начальное смещение и начальная скорость вещества:

$$u(x,0) = \mu_3(x),\ u_t(x,0) = \mu_4(x),\ 0 \le x \le a. \tag{6.4}$$

Уравнение (6.1) с краевыми условиями (6.3) и начальными условиями (6.4) составляют полную постановку задачи.

6.1.2. Схема «крест»

Составим несложную и эффективную разностную схему для задачи (6.1). Выберем по x, t прямоугольную сетку, для простоты равномерную, и возьмем изображенный на рис. 6.1 шаблон. Он содержит три слоя по времени: новый, исходный (средний) и предыдущий (нижний). Значения решения на предыдущем слое обозначим через $\check{u}$. Аппроксимируя вторые производные разностями, получим трехслойную схему:

$$\frac{1}{\tau^2}(\hat{u}_n - 2u_n + \check{u}_n) = \frac{c^2}{h^2}(u_{n+1} - 2u_n + u_{n-1}) + f_n,\ 1 \le n \le N-1, \tag{6.5}$$

с граничными условиями

$$\hat{u}_0 = \mu_1(\hat{t}),\ \hat{u}_N = \mu_2(\hat{t}). \tag{6.6}$$

По форме шаблона эту схему называют «крест». Исследуем ее.

Вычисление решения. На нулевом слое решение известно из начального условия:

$$u_n^0 = \mu_3(x_n),\ 0 \le n \le N. \tag{6.7}$$

На первом слое решение также можно вычислить по начальным данным. Простейший способ состоит в том, что полагают

$$\hat{u}_n \approx u_n + \tau u_t(x_n, 0) = \\ = \mu_3(x_n) + \tau\mu_4(x_n). \tag{6.8}$$

Более хорошие результаты дает использование следующего члена разложения:

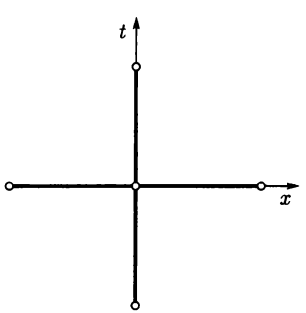


Рис. 6.1. Шаблон схемы «крест» (одномерный случай)

$$\hat{u}_n \approx u_n + \tau u_t(x_n, 0) + \frac{\tau^2}{2} u_{tt}(x_n, 0);$$

выражение для u_{tt} в это соотношение надо подставить из уравнения (6.1). Окончательно получим

$$\hat{u}_n = \mu_3(x_n) + \tau\mu_4(x_n) + \\ + \frac{\tau^2}{2}\left[c^2 \frac{d^2\mu_3(x_n)}{dx^2} + f(x_n, 0)\right], \qquad (6.9)$$
$$1 \le n \le N-1.$$

Здесь μ_{3xx} можно заменить второй пространственной разностью.

Схема «крест» (6.5) явная и позволяет выразить $\hat{u}_n$ через значения u с двух предыдущих слоев. Поэтому, начиная со второго слоя, разностное решение вычисляется по этой схеме.

Описанный алгоритм показывает, что после того как выбрана одна из начальных формул (6.8), (6.9), разностное решение существует и единственно.

Аппроксимация. Разложим точное решение по формуле Тейлора с центром в узле (x_n, t), предполагая наличие непрерывных четвертых производных:

$$u_{n\pm1} = u \pm h u_x + \frac{h^2}{2} u_{xx} \pm \frac{h^3}{6} u_{xxx} + \frac{h^4}{24} u_{xxxx}, \\ u_n(t \pm \tau) = u \pm \tau u_t + \frac{\tau^2}{2} u_{tt} \pm \frac{\tau^3}{6} u_{ttt} + \frac{\tau^4}{24} u_{tttt}. \qquad (6.10)$$

Подставим в определение невязки (2.8) разностный оператор (6.5) и разложения (6.10). Тогда получим следующую невязку схемы в регулярных узлах:

$$\psi_n = \frac{\tau^2}{12} u_{tttt} - \frac{c^2 h^2}{12} u_{xxxx} = O\left(\tau^2 + h^2\right), \ 1 \le n \le N-1. \qquad (6.11)$$

Невязка в узлах первого слоя для простейшей аппроксимации (6.8) равна

$$\psi_n = \frac{\tau}{2} u_{tt} = O(\tau), \ 1 \le n \le N-1; \qquad (6.12)$$

для уточненной аппроксимации (6.9) невязка составляет

$$\frac{1}{6}\tau^2 u_{ttt} = O(\tau^2), \ 1 \le n \le N-1. \qquad (6.13)$$

Значения u_n на нулевом слое и краевые условия u_0, u_N на всех слоях вычисляются точно.

Таким образом, схема «крест» (6.5) с улучшенной аппроксимацией первого слоя (6.9) имеет аппроксимацию $O\left(\tau^2 + h^2\right)$. Упрощенное вычисление первого слоя (6.8) ухудшает аппроксимацию до $O\left(\tau + h^2\right)$.

Устойчивость исследуем методом гармоник. Делаем стандартную замену $u_n \to 1$, $u_{n\pm1} \to \exp(\pm \mathrm{i} q h)$, $\hat{u} \to \rho u$; поскольку исходный слой является новым по отношению к предыдущему, надо также положить $u_n \to \rho \check{u}_n$ или $\check{u}_n \to u_n/\rho$. Для множителя роста q-й гармоники ρ получим квадратное уравнение:

$$\rho_q^2 - 2\gamma_q \rho_q + 1 = 0, \ \gamma_q = 1 - 2\left(\frac{c\tau}{h}\sin\frac{qh}{2}\right)^2. \tag{6.14}$$

По теореме Виета произведение его корней $\rho_q' \rho_q'' = 1$. Значит, условие устойчивости $|\rho_1| < 1$ может быть выполнено, только если $\left|\rho'\right| = \left|\rho''\right| = 1$. Для уравнения с действительными коэффициентами (6.14) это означает, что корни образуют комплексно-сопряженную пару; для этого дискриминант уравнения не должен быть положительным: $|\gamma_q| \le 1$. Поскольку γ_q вещественно, это можно переписать так:

$$-1 \le 1 - 2\left(\frac{c\tau}{h}\sin\frac{qh}{2}\right)^2 \le 1.$$

Правое неравенство всегда выполняется. Чтобы левое неравенство соблюдалось для любых гармоник (т. е. при любых значениях синуса), необходимо и достаточно выполнение ***условия Куранта***:

$$c\tau \le h. \tag{6.15}$$

Таким образом, схема «крест» условно устойчива.

Сходимость. Из сказанного выше следует, что схема (6.5) с улучшенной аппроксимацией первого слоя при выполнении условия Куранта (6.15) сходится со скоростью $O\left(\tau^2 + h^2\right)$.

Поскольку устойчивость мы доказали методом гармоник, т. е. в норме $\|\cdot\|_{l_2}$, то и сходимость доказана в $\|\cdot\|_{l_2}$. Методом операторных неравенств можно доказать, что сходимость равномерная.

Схема (6.5) обеспечивает хорошую точность расчета решений $u(x,t)$, имеющих непрерывные четвертые производные. Она позволяет рассчитывать менее гладкие и даже разрывные решения, хотя в последнем случае точность расчетов невелика, и обычно возникает легкая пилообразность, связанная с немонотонностью схемы. Условие устойчивости (6.15) естественное, поскольку для получения хорошей точности тоже надо полагать $c\tau \sim h$. Поэтому схему «крест» часто используют для практических расчетов.

Замечания. 1. Схема (6.5) написана для случая постоянных шагов τ и h. Если шаги переменные, то надо заменить производные по пространству и времени соответствующими выражениями (см. кн. 1), которые обеспечивают локальную аппроксимацию $O\left(\tau^2 + h^2\right)$ только в случае квазиравномерных сеток по x и t. Поэтому для трехслойных схем, в отличие от двуслойных, резкая неравномерность шагов по времени или пространству опасна: это может привести к ухудшению точности.

2. Для задач с краевыми условиями первого рода (6.3) удобно выбирать сетку так, чтобы узлы x_0 и x_N были концами отрезка $[0,a]$. Пусть на одном из концов задано краевое условие второго рода

$$u_x(0,t) = \mu_1(t) \tag{6.16}$$

(например, для колебаний столба воздуха в органной трубе это условие соответствует открытому концу трубы). Тогда целесообразно полагать $x_0 = -h/2$ и $x_1 = h/2$, чтобы граница была полуцелой точкой. Естественное разностное краевое условие примет следующий вид:

$$(u_1 - u_0)/h = \mu_1(t). \tag{6.17}$$

Фактически здесь использован метод фиктивных точек. Он обеспечивает аппроксимацию $O\left(h^2\right)$. Такой выбор сетки полезен и для других типов уравнений.

6.1.3. Неявная схема

Если схема условно устойчива, то случайное небольшое нарушение условия устойчивости может привести к быстрому нарастанию погрешностей, вплоть до переполнения при расчетах на компьютерах. Поэтому многие вычислители предпочитают использовать безусловно устойчивые схемы. По опыту предыдущих глав мы видели, что такие схемы должны быть неявными.

Построим хорошую неявную схему для задачи (6.1). Возьмем изображенный на рис. 6.2 шаблон и составим схему с весами при

пространственных производных на разных слоях. Интуитивно ясно, что симметрия схемы по времени должна дать аппроксимацию $O(\tau^2)$. Поэтому веса на новом и предыдущем слоях возьмем одинаковые. Получится следующая схема:

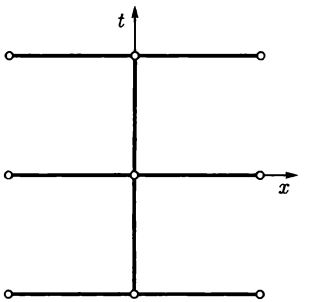


Рис. 6.2. Шаблон трехслойной неявной схемы

$$\frac{1}{\tau^2}(\hat{u} - 2u + \check{u}) = \\ = \Lambda[\sigma\hat{u} + (1-2\sigma)u + \sigma\check{u}] + f, \\ \Lambda u_n = \frac{c^2}{h^2}(u_{n+1} - 2u_n + u_{n-1}). \tag{6.18}$$

В граничных узлах решение определяется из краевых условий (6.3). Первый слой вычисляется так же, как для схемы «крест».

Вычисление решения. Регулярный расчет начинается со второго слоя. При этом значения решения на исходном и предыдущем слоях известны. В левой части уравнения (6.18) стоит $\hat{u}_n/\tau^2$, а в правой части имеется трехточечное выражение $\Lambda\hat{u}_n$. Для нахождения $\hat{u}_n$ получается система линейных уравнений с трехдиагональной матрицей. Эта система легко решается методом Гаусса для трехдиагональной матрицы или прогонкой (см. кн. 1). Этот алгоритм является экономичным (число операций на узел сетки постоянно, т. е. имеет тот же порядок трудоемкости, что и явная схема «крест»).

Аппроксимация. Разложением решения по формуле Тейлора нетрудно установить, что на решениях с непрерывными четвертыми производными схема (6.18) аппроксимирует уравнение (6.1) с погрешностью $O\left(\tau^2 + h^2\right)$ при любом σ.

Устойчивость исследуем методом гармоник с той же подстановкой, что для схемы «крест». Для множителя роста q-й гармоники получаем квадратное уравнение:

$$\rho_q^2 - 2\gamma_q\rho_q + 1 = 0, \\ \gamma_q = \frac{1 - 2(1-2\sigma)\beta_q^2}{1 + 4\sigma\beta_q^2},\ \beta_q = \frac{c\tau}{h}\sin\frac{qh}{2}. \tag{6.19}$$

На основании тех же рассуждений, что и в п. 6.1.2, можно сделать следующий вывод: устойчивость будет только при комплексно-сопряженных корнях, т. е. при $|\gamma_q| \leq 1$. Отсюда вытекает условие устойчивости схемы:

$$\left(\frac{c\tau}{h}\right)^2 (1 - 4\sigma) \leq 1. \tag{6.20}$$

Из неравенства (6.20) видно, что при $\sigma \geq 1/4$ схема (6.18) безусловно устойчива. Если $\sigma < 1/4$, то схема условно устойчива при $c\tau \leq h(1 - 4\sigma)^{-1/2}$.

Сходимость. Рассмотрим разумные пределы для веса σ. Поскольку неявность вводилась ради безусловной устойчивости, целесообразно выбирать $\sigma \geq 1/4$. Но если взять $\sigma > 1/2$, то вес центрального слоя схемы (6.18), равный $1 - 2\sigma$, будет отрицательным; это неразумно. Поэтому целесообразно выбирать вес в пределах $1/4 \leq \sigma \leq 1/2$.

При таком выборе веса неявная схема (6.18) безусловно сходится с точностью $O\left(\tau^2 + h^2\right)$.

Замечания. 1. Схема (6.18) при $\sigma = 0$ переходит в схему «крест», а условие устойчивости (6.20) — в условие Куранта (6.15).

2. Явную и неявную схемы можно обобщить на случай неоднородной среды и неравномерных пространственных сеток. Но для этого надо правильно написать закон сохранения, на чем здесь не будем останавливаться.

6.2. ДВУСЛОЙНЫЕ СХЕМЫ

6.2.1. Преобразование уравнения

Уравнение второго порядка (6.1) можно заменить эквивалентной ему парой уравнений первого порядка. Для этого введем потенциалы скоростей и правой части:

$$v(x,t) = \int_0^x u_t(\xi,t)d\xi, \; F(x,t) = \int_0^x f(\xi,t)d\xi. \tag{6.21}$$

Функции $u(x,t)$, $v(x,t)$ удовлетворяют следующей системе уравнений первого порядка по пространству и времени:

$$u_t = v_x, \; v_t = c^2 u_x + F(x,t). \tag{6.22}$$

Начальные условия (6.4) с учетом (6.21) примут следующий вид:

$$u(x,0)=\mu_3(x),\ v(x,0)=\int\limits_0^x \mu_4(\xi)d\xi, \tag{6.23}$$

а граничные условия (6.3) останутся без изменения:

$$u(0,t)=\mu_1(t),\ u(a,t)=\mu_2(t). \tag{6.24}$$

Задача акустики (6.22) — (6.24) нередко оказывается более удобной для численного решения, чем волновое уравнение (6.1); в частности, она позволяет использовать неравномерные сетки не только по x, но и по t.

6.2.2. Пространственная аппроксимация

Будем рассматривать в узлах неравномерной пространственной сетки величины $u_n \approx u(x_n,t)$, $0 \le n \le N$, а в серединах интервалов — величины $v_{n-1/2} \approx v\left(x_{n-1/2},t\right)$, $1 \le n \le N$. Сетку предполагаем неравномерной; будем использовать ее целые шаги $h_n = x_n - x_{n-1}$ $(1 \le n \le N)$ и полуцелые $h_{n+1/2} = x_{n+1/2} - x_{n-1/2}$ $(1 \le n \le N-1)$.

Для построения разностной схемы применим метод прямых. Сохраним в системе (6.22) производные по времени, а пространственные производные заменим простейшими разностными аппроксимациями. Получим две системы дифференциально-разностных уравнений:

$$\frac{du_n}{dt} = \left(v_{n+1/2} - v_{n-1/2}\right)/h_{n+1/2},\ 1 \le n \le N-1, \tag{6.25}$$

$$\frac{dv_{n-1/2}}{dt} = c^2\left(u_n - u_{n-1}\right)/h_n + F_{n-1/2}(t),\ 1 \le n \le N. \tag{6.26}$$

Обе системы в сумме содержат $2N-1$ неизвестную функцию. В уравнения (6.25) — (6.26) при $n=1$ входит $u_0(t)$, а при $n=N$ — $u_N(t)$; они берутся из граничных условий (6.24).

Очевидно, на равномерной сетке уравнения (6.25)–(6.26) имеют аппроксимацию $O\left(h^2\right)$; при этом $h_n = h_{n+1/2} = \texttt{const} \equiv h$. Нетрудно проверить, что на квазиравномерных сетках уравнения также имеют аппроксимацию $O\left(h^2\right)$. На произвольных неравномерных сетках аппроксимация ухудшается до $O(h)$.

Начальные значения для уравнений (6.25) очевидны: $u_n(0) = \mu_3(x_n)$, $0 \le n \le N$. Начальные значения для потенциала скоростей $v_{n-1/2}(0)$ находятся численным интегрированием $\mu_4(x)$ согласно (6.23). Можно воспользоваться следующим гибридом формул трапеций и средних:

$$\begin{aligned} v_{1/2}(0) &= \frac{1}{2}\left[\mu_4(x_0) + \mu_4(x_{1/2})\right](x_{1/2} - x_0); \\ v_{n+1/2}(0) &= v_{n-1/2}(0) + \mu_4(x_n)h_{n+1/2},\ 1 \le n \le N-1. \end{aligned} \tag{6.27}$$

Аналогично вычисляется правая часть $F(x_{n-1/2}, t)$:

$$\begin{aligned} F_{1/2}(t) &= \frac{1}{2}\left[f(x_0,t) + f(x_{1/2},t)\right](x_{1/2} - x_0); \\ F_{n+1/2}(t) &= F_{n-1/2}(t) + f(x_n,t)h_{n+1/2},\ 1 \le n \le N-1. \end{aligned} \tag{6.28}$$

Квадратуры (6.27) — (6.28) имеют аппроксимацию $O(h^2)$ на равномерных или квазиравномерных сетках, и $O(h)$ — на произвольных сетках. Это совпадает с аппроксимацией основной системы (6.25) — (6.26).

Единая система. Две системы (6.25) — (6.26) удобнее записать в виде единой системы дифференциально-разностных уравнений. Для этого все значения $u_n(t)$ и $v_n(t)$ обозначим единой буквой $w_k(t)$ со следующей индексацией:

$$w_{2n}(t) = u_n(t),\ 0 \le n \le N;\ w_{2n-1}(t) = v_{n-1/2}(t),\ 1 \le n \le N. \tag{6.29}$$

Таким образом, индекс k лежит в пределах $0 \le k \le 2N$. Для функции $w_k(t)$ система (6.25) — (6.26) примет следующий вид:

$$\frac{dw_k}{dt} = \Phi_k(w_{k-1}, w_{k+1}) \equiv$$

$$\equiv \begin{cases} \dfrac{w_{k+1} - w_{k-1}}{h_{(k+1)/2}} \text{ при четных } k = 2, 4, \ldots, 2N-2; \\ \dfrac{c^2\,(w_{k+1} - w_{k-1})}{h_{(k+1)/2}} + F_{k/2}(t) \text{ при нечетных } k = 1, 3, \ldots, 2N-1. \end{cases} \tag{6.30}$$

Это система $2N-1$ уравнений с таким же числом неизвестных функций $w_k(t)$, $1 \le k \le 2N-1$. Граничными условиями для этой системы являются

$$w_0(t) \equiv u_0(t) = \mu_1(t),\ w_{2N}(t) \equiv u_N(t) = \mu_2(t). \tag{6.31}$$

Начальные условия очевидны:

$$w_{2n}(0) = u_n(0) = \mu_3(x_n),\ w_{2n-1}(0) = v_{n-1/2}(0). \tag{6.32}$$

Последняя величина берется согласно (6.27).

6.2.3. Разностная схема

Хотя система дифференциальных уравнений (6.30) не является жесткой, используем для ее решения семейство одностадийных схем Розенброка (см. п. 1.2.3). Напомним его векторную запись:

$$\hat{\mathbf{w}} = \mathbf{w} + \tau \mathtt{Re}\mathbf{z},\ \left[E - \alpha\frac{\partial \mathbf{\Phi}}{\partial \mathbf{w}}\right]\mathbf{z} = \mathbf{\Phi}\left(\mathbf{w}, t + \tau/2\right). \tag{6.33}$$

Это двуслойная схема. Значение параметра $\alpha = 0$ дает явную схему, $\alpha = 1/2$ — схему «с полусуммой», $\alpha = 1$ — чисто неявную схему, $\alpha = (1 + \mathtt{i})/2$ — комплексную схему Розенброка (CROS). Последние три схемы являются неявными, так как в них для нахождения приращения решения $\mathbf{z}$ требуется решить систему линейных уравнений, в которую входит матрица Якоби правых частей $\partial\mathbf{\Phi}/\partial\mathbf{w}$.

При $\alpha \neq 0$ требуется решать систему линейных уравнений. Однако правые части зависят только от двух кодиагоналей: $\Phi_k(w_{k-1}, w_{k+1})$. Поэтому матрица Якоби $\partial\mathbf{\Phi}/\partial\mathbf{w}$ является трехдиагональной, так что линейная система (6.33) легко решается прогонкой или методом Гаусса для ленточной матрицы. Это экономичный алгоритм, по простоте и трудоемкости почти не уступающий явной схеме.

Поскольку схема двуслойная, в ней не требуется постоянства шагов по времени τ. В ходе расчета можно произвольно менять шаг по времени. Это является существенным преимуществом перед трехслойными схемами.

Есть и другое важное преимущество двуслойной схемы: не нужно рассчитывать первый слой по особым формулам. Начальные данные естественно ставятся на нулевом слое.

Явная схема не является A-устойчивой. Остальные три схемы являются A-устойчивыми; это обеспечивает их безусловную устойчивость.

Гармонический анализ. Устойчивость можно исследовать также методом гармоник. Для комплексного α это исследование

довольно громоздко. Ограничимся исследованием только для вещественных $\alpha (0 \le \alpha \le 1)$. Запись в виде единой системы (6.33) для этого неудобна. Лучше воспользоваться системой двух уравнений (6.25) — (6.26) для двух функций. Тогда одностадийная схема Розенброка с вещественным α принимает следующий вид:

$$\hat{u}_n - u_n = \frac{\tau}{h}\left[\alpha\left(\hat{v}_{n+1/2} - \hat{v}_{n-1/2}\right) + (1-\alpha)\left(v_{n+1/2} - v_{n-1/2}\right)\right]; \tag{6.34}$$

$$\hat{v}_{n-1/2} - v_{n-1/2} = \\ = \frac{\tau c^2}{h}\left[\alpha\left(\hat{u}_n - \hat{u}_{n-1}\right) + (1-\alpha)\left(u_n - u_{n-1}\right)\right] + \tau F_{n-1/2}\left(t + \tau/2\right). \tag{6.35}$$

Ранее рассматривались задачи только с одной функцией u, поэтому амплитуда возмущения ее q-й гармоники без ограничения общности принималась равной 1. Теперь имеется вторая функция v, поэтому надо ввести амплитуду возмущения ее гармоники β. Тогда стандартная замена для метода гармоник примет следующий вид:

$$u(x) = \exp(\mathtt{i}qx),\ v(x) = \beta\exp(\mathtt{i}qx),\ \hat{u} = \rho_q u,\ \hat{v} = \rho_q v; \tag{6.36}$$

множитель роста ρ_q для обеих функций одинаков. Подставим (6.36) в (6.34) — (6.35) и сократим все пространственные множители в уравнении (6.34) на $\exp\left(\mathtt{i}qx_n\right)$, а в уравнении (6.35) — на $\exp\left(\mathtt{i}qx_{n-1/2}\right)$. Получим систему уравнений для ρ_q и β:

$$\rho_q - 1 = 2\mathtt{i}\beta\left(\tau/h\right)\left(\alpha\rho_q + 1 - \alpha\right)\mathtt{sin}\left(qh/2\right),$$
$$\beta\left(\rho_q - 1\right) = 2\mathtt{i}\left(c^2\tau/h\right)\left(\alpha\rho_q + 1 - \alpha\right)\mathtt{sin}\left(qh/2\right).$$

Перемножим эти уравнения. Тогда амплитуда β сократится, и останется уравнение для множителя роста:

$$\left(\rho_q - 1\right)^2 = -\left(\alpha\rho_q + 1 - \alpha\right)^2 s_q,\ s_q = 4\left(c\tau/h\right)^2\mathtt{sin}^2\left(qh/2\right) > 0. \tag{6.37}$$

Преобразуем это квадратное уравнение к стандартной форме:

$$\left(1 + \alpha^2 s_q\right)\rho_q^2 - 2\left[1 - \alpha\left(1-\alpha\right)s_q\right]\rho_q + \left[1 + \left(1-\alpha\right)^2 s_q\right] = 0. \tag{6.38}$$

Дискриминант этого квадратного уравнения равен $-4s_q$; он отрицателен, так что корни образуют комплексно-сопряженную пару. Их модули равны, и из теоремы Виета следует:

$$|\rho_q| = \left\{ \left[1 + (1-\alpha)^2 s_q \right] / \left(1 + \alpha^2 s_q \right) \right\}^{1/2}. \qquad (6.39)$$

Отсюда следуют три случая: 1) если $1/2 < \alpha \leq 1$, то для всех гармоник $|\rho_q| < 1$; схема безусловно устойчива и диссипативна (т. е. ошибки строго затухают); 2) если $\alpha = 1/2$, то $|\rho_q = 1|$; схема безусловно устойчива, но не диссипативна; 3) если $0 \leq \alpha < 1/2$, то $|\rho_q| > 1$ для всех гармоник; схема безусловно неустойчива.

Заметим, что устойчивость схемы при $\alpha < 1/2$ оказалась хуже, чем для параболического уравнения. В параболическом уравнении одностадийное семейство схем Розенброка при $0 \leq$ $\leq \alpha < 1/2$ было условно устойчивым. По-видимому, это объясняется сильными диссипативными свойствами самого параболического уравнения.

Остановимся на свойствах отдельных схем из семейства Розенброка.

Схема CROS. Эта схема имеет аппроксимацию $O(\tau^2)$; при этом сетка по t может быть произвольной неравномерной. С учетом пространственной невязки схема CROS имеет аппроксимацию $O(\tau^2 + h^2)$ на произвольных сетках по времени и равномерных или квазиравномерных сетках по пространству. Она безусловно устойчива (это можно также проверить методом гармоник, но соответствующее доказательство довольно громоздко).

Единственным недостатком схемы CROS является ее пространственная немонотонность. Наличие немонотонности можно пояснить следующим образом. Общее решение однородного уравнения акустики (6.2) имеет вид двух бегущих навстречу волн. Тем самым уравнение подобно двум уравнениям переноса во встречных направлениях, а для уравнения переноса была доказана теорема 3.2 о немонотонности. Схема CROS, имеющая второй порядок аппроксимации, подпадает под эту теорему.

Эта немонотонность может заметно проявляться при расчетах решений с крутыми фронтами. На решениях с пологими пространственными профилями она проявляется слабо или вообще незаметна. При этом наблюдаемая немонотонность обычно слабее, чем в аналогичных ситуациях для уравнения переноса.

Отметим интересные свойства наблюдаемой немонотонности: провалы и всплески численного решения имеют вид пологих волн. Каждая полуволна по длине равна нескольким (обычно довольно многим) интервалам пространственной сетки. Амплитуды полуволн невелики и быстро убывают по мере удаления от

крутого профиля. Именно поэтому во многих расчетах немонотонность практически незаметна.

По совокупности описанных свойств первое, что рекомендуется попробовать в расчетах, — это схема CROS (в пакетах программ она должна выбираться по умолчанию).

Чисто неявная схема. Пусть пространственные профили решения настолько круты (быть может, разрывны), что при расчетах по схеме CROS возникает неприемлемая немонотонность. Тогда рекомендуется использовать чисто неявную схему ($\alpha = 1$).

Эта схема также безусловно устойчива. Она является строго монотонной, что позволяет рассчитывать фронты любой крутизны. Однако эта схема имеет аппроксимацию лишь $O(\tau)$ и на решениях с пологими фронтами существенно уступает схеме CROS по точности. Поэтому чисто неявную схему рекомендуется использовать как резервную схему пакета программ.

Схема «с полусуммой» ($\alpha = 1/2$). Эта схема безусловно устойчива, а ее аппроксимация есть $O(\tau^2)$. Поэтому до появления схемы CROS схема «с полусуммой» часто использовалась в расчетах. Однако эта схема имеет сильную немонотонность как по пространству, так и по времени. Это подробно описано для обыкновенных дифференциальных уравнений в п. 1.2.3, а для уравнения переноса — в п. 3.1.4. Вдобавок, эта немонотонность имеет резко выраженный пилообразный характер. Полуволна захватывает один-два интервала сетки, а амплитуды полуволн велики. Таким образом, по надежности схема «с полусуммой» существенно уступает схеме CROS.

Однако у схемы «с полусуммой» есть два небольших преимущества перед схемой CROS. Во-первых, коэффициент в члене невязки $O(\tau^2)$ вдвое меньше (пространственные члены невязки одинаковы во всех схемах); это дает небольшой выигрыш в точности на решениях с пологими профилями. Во-вторых, для линейной задачи акустики схема «с полусуммой» строго симметрична по времени, так что невязка раскладывается в ряд по степеням τ^2; схема CROS несимметрична по времени и ее невязка раскладывается в ряд по всем степеням τ. Это играет роль при расчетах со сгущением сеток и исключением погрешности по Ричардсону.

По совокупности свойств схему «с полусуммой» целесообразно ставить на третье место в пакете программ.

Явная схема. Схема Розенброка (6.34) — (6.35) с $\alpha = 0$ оказалась безусловно неустойчивой. Однако можно построить явную схему, не входящую в семейство Розенброка. Она имеет следующий вид:

$$\hat{u}_n = u_n + \tau \left(v_{n+1/2} - v_{n-1/2}\right)/h_{n+1/2},\ 1 \le n \le N-1; \qquad (6.40)$$

$$\begin{gathered} \hat{u}_0 = \mu_1(\hat{t}),\ \hat{u}_N = \mu_2(\hat{t}); \\ \hat{v}_{n-1/2} = v_{n-1/2} + \tau c^2 \left(\hat{u}_n - \hat{u}_{n-1}\right)/h_n + \\ + \tau F_{n-1/2}(t + \tau/2),\ 1 \le n \le N. \end{gathered} \qquad (6.41)$$

Видно, что схема является явной. Переход со слоя на слой происходит следующим образом. Сначала по формулам (6.40) и данным с исходного слоя явно вычисляются $\hat{u}_n$. Затем по формулам (6.41) и уже найденным $\hat{u}_n$ явно вычисляются $\hat{v}_{n-1/2}$.

Устойчивость схемы (6.40) — (6.41) нетрудно исследовать методом гармоник. Делая стандартную подстановку (6.36) и проводя аналогичные выкладки, получим квадратное уравнение для множителя роста:

$$\rho_q^2 - 2\left(1 - s_q/2\right)\rho_q + 1 = 0,\ s_q = 4\left(c\tau/h\right)^2 \sin^2\left(qh/2\right). \qquad (6.42)$$

Оно совпадает с уравнением (6.14) для схемы «крест». Поэтому явная двуслойная схема (6.40) — (6.41) устойчива при выполнении условия Куранта $c\tau \le h$.

Для одномерных задач явная схема не имеет преимуществ по сравнению с описанными выше неявными схемами. Однако эта идея является полезной для многомерных обобщений.

6.2.4. Неограниченная область

Простейшее гиперболическое уравнение (6.1) описывает не только струну, но и распространение акустических возмущений в газе или жидкости. Пусть источник звука расположен в некоторой небольшой части области $[0, a]$. В этом случае звуковые колебания распространяются от источника, достигают границ и отражаются от них; характер этого отражения зависит от вида граничного условия. Затем отраженный звук движется через область в обратном направлении, достигает другой границы и отражается от нее. Процесс повторяется неограниченное число раз. В области возникает система звуковых волн, двигающихся

навстречу друг другу. В точном решении эти волны не затухают, поскольку уравнение (6.1) не содержит диссипативных членов.

Отрезок $[0, a]$ может оказаться очень большим, хотя источник звука расположен в небольшой его части. В этом случае можно считать что $a = +\infty$, т. е. правая граница отсутствует (бесконечно удалена). Тогда идущая вправо часть звуковой волны не достигнет этой границы и не отразится от нее. В системе останется только однократное отражение волны, идущей влево, от левой границы $x = 0$.

Такие задачи можно рассчитывать, предполагая область неограниченной ($a = \infty$) и строя квазиравномерную пространственную сетку с бесконечно удаленной точкой (см. п. 4.1.5). Напомним, что для этого надо ввести некоторое строго монотонное преобразование координат с полюсом, например,

$$x(\xi) = b\xi / \left(1 - \xi^2\right)^r, \ 0 \le \xi \le 1, \ b > 0, \ q > 0. \tag{6.43}$$

По переменной ξ строится равномерная сетка $\xi_n = n/N$, $0 \le n \le N$; полуцелыми узлами вспомогательной сетки будет $\xi_{n-1/2} = (n - 1/2)/N$. Целые и полуцелые узлы пространственной сетки строятся с помощью преобразования (6.43): $x_n = x(\xi_n)$, $x_{n-1/2} = x(\xi_{n-1/2})$. Последняя точка такой сетки оказывается бесконечно удаленной $x_N = \infty$.

В формулы п. 6.2.3 и 6.2.4 входят пространственные шаги h_n и $h_{n-1/2}$. В конечной области их определяют как расстояния между соседними целыми или полуцелыми узлами сетки по x. Однако в неограниченной области их надо переопределять следующим образом:

$$h_n = \frac{1}{N}\left(\frac{dx}{d\xi}\right)_{n-1/2}, \ h_{n+1/2} = \frac{1}{N}\left(\frac{dx}{d\xi}\right)_{n}; \tag{6.44}$$

индексы при производных означают, что их надо брать при значениях $\xi_{n-1/2}$ и ξ_n. При этом длина последнего (неограниченного) интервала сетки остается конечной.

В преобразовании (6.43) чаще всего полагают $r = 1/2$; это обеспечивает особенно простой вид производной $dx/d\xi$. Значение b подбирают так, чтобы та часть области, где разыгрываются значимые события, содержала 50 — 75 % расчетных точек.

Прозрачная граница. Существует другой способ избежать отраженной волны в расчетах. Поставим правую границу на до-

вольно большом, но конечном расстоянии a. Вспомним, что общее решение (6.2) однородного уравнения (6.1) состоит из двух волн произвольного профиля; волна $\phi_1(x+ct)$ бежит влево, а волна $\phi_2(x-ct)$ бежит вправо. Нам нужно, чтобы вблизи правой границы $x=a$ не было отраженной волны, т. е. оставалась только бегущая вправо волна $\phi_2(x-ct)$. Но для такой волны выполняется соотношение $\partial\phi_2/\partial t + c\partial\phi_2/\partial x = 0$. Это условие и надо выбрать в качестве правого граничного условия:

$$(u_t + cu_x)_{x=a} = 0. \tag{6.45}$$

С учетом первого соотношения (6.22) его можно записать в следующем виде:

$$(cu_x + v_x)_{x=a} = 0. \tag{6.46}$$

Любое из граничных условий (6.45) — (6.46) означает прозрачную границу. Условием (6.45) можно пользоваться для трехслойных схем подраздела 6.2, а условием (6.46) — удобнее для двухслойных схем подраздела 6.3.

Условие (6.46) реализуется проще. Опишем его разностную аппроксимацию, обеспечивающую точность $O\left(h^2\right)$; аппроксимация по τ в нем является точной. Введем фиктивные точки $x_{N+1/2}$ и x_{N+1} за правой границей. При этом во всех системах разностных уравнений этого раздела надо увеличить правую границу изменения индекса n. Разностный аналог условия (6.46) запишем так:

$$\begin{aligned} &c\left(\hat{u}_{N+1} - \hat{u}_{N-1}\right) / \left(h_N + h_{N+1}\right) + \\ &+ \left(\hat{v}_{N+1/2} - \hat{v}_{N-1/2}\right) / h_{N+1/2} = 0. \end{aligned} \tag{6.47}$$

Из симметрии видно, что на квазиравномерных сетках обеспечена аппроксимация $O\left(h^2\right)$.

Заметим, что введение фиктивных точек позволяет пользоваться сгущением сеток и нахождением апостериорной асимптотически точной оценки погрешности методом Ричардсона.

6.3. МНОГОМЕРНОЕ УРАВНЕНИЕ

6.3.1. Явная схема

Постановка задачи. Волновое уравнение в p-мерной изотропной среде (либо в неизотропной среде, если у тензора упругости отличны от нуля только диагональные компоненты) имеет следующий вид:

$$\frac{\partial^2 u}{\partial t^2} = \sum_{\alpha=1}^{p} A_\alpha u + f(\mathbf{x}, t),$$
$$A_\alpha u = \frac{\partial}{\partial x_\alpha}\left(c_\alpha^2(\mathbf{x}, t)\frac{\partial u}{\partial x_\alpha}\right), \ \mathbf{x} = \{x_1, x_2, \ldots, x_p\} \in G. \tag{6.48}$$

Величины c_α имеют смысл скорости распространения звука по соответствующим направлениям. Для общности они взяты неодинаковыми по разным направлениям. Физически это соответствует распространению звука в монокристалле, главные оси которого совпадают с осями координат. В поликристаллическом или аморфном твердом веществе, в жидкостях и газах скорости звука по всем направлениям одинаковы.

Уравнение (6.48) задано в области G с границей Г. Рассмотрим задачу нахождения решения уравнения (6.48) с начальными условиями и с краевыми условиями первого рода:

$$u(\mathbf{x}, 0) = \mu_1(\mathbf{x}), \ u_t(\mathbf{x}, 0) = \mu_2(\mathbf{x}), \ \mathbf{x} \in G;$$
$$u(\mathbf{x}, t) = \mu_3(\mathbf{x}, t), \ \mathbf{x} \in \Gamma(G). \tag{6.49}$$

Далее ограничимся случаем, когда G есть прямоугольный параллелепипед, и будем вводить в нем прямоугольные пространственные сетки с шагами h_α по переменным x_α.

Уравнение (6.48) можно заменить системой уравнений первого порядка, однако ограничимся рассмотрением приведенной записи.

Схема «крест». Она строится аналогично одномерной схеме (6.5) на шаблоне, вид которого для двух измерений показан на рис. 6.3. При произвольном числе измерений эта схема для уравнения (6.48) имеет вид:

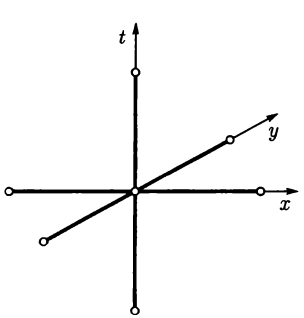


Рис. 6.3. Шаблон схемы «крест»

$$\frac{1}{\tau^2}(\hat{u} - 2u + \check{u}) = \sum_{\alpha=1}^{p} \Lambda_\alpha u + f. \tag{6.50}$$

Одномерные трехточечные разностные операторы Λ_α записываются так же, как в подразделах 6.1, 6.2.

Схема (6.50) — явная трехслойная; организация вычислений по ней одинаково проста при любом числе изме-

рений. Нетрудно проверить, что на равномерных сетках она имеет аппроксимацию $O\left(\tau^2 + \sum_{\alpha=1}^{p} h_\alpha^2\right)$. Ее устойчивость можно исследовать методом разделения переменных, подставляя в (6.50) многомерную гармонику:

$$u = \exp\left(\mathtt{i}\sum_{\alpha=1}^{p} q_\alpha x_\alpha\right),\ \hat{u} = \rho u,\ u = \rho\check{u}. \tag{6.51}$$

Учитывая, что

$$\Lambda_\alpha u \to -4\left(\frac{c_\alpha}{h_\alpha}\sin\frac{q_\alpha h_\alpha}{2}\right)^2, \tag{6.52}$$

получим для множителя роста квадратное уравнение

$$\rho^2 - 2(1-2\gamma)\rho + 1 = 0,\ \gamma = \tau^2\sum_{\alpha=1}^{p}\left(\frac{c_\alpha}{h_\alpha}\sin\frac{q_\alpha h_\alpha}{2}\right)^2. \tag{6.53}$$

Это уравнение аналогично одномерному уравнению (6.14). Анализ его корней показывает, что схема (6.50) устойчива при выполнении условия

$$\tau < \left(\sum_{\alpha=1}^{p}\frac{c_\alpha^2}{h_\alpha^2}\right)^{-1/2} \sim \frac{h}{c\sqrt{p}}, \tag{6.54}$$

являющегося обобщением условия Куранта. Это естественное условие, а точность схемы неплохая. Поэтому схему «крест» используют в расчетах, если не требуется особенно высокой надежности вычислений.

Кроме того, важным преимуществом схемы «крест» является то, что она удобна для многопроцессорных компьютеров.

Таким образом, численный расчет многомерных задач акустики не вызывает принципиальных затруднений.

6.3.2. Факторизованные схемы

В явных схемах локальное нарушение устойчивости легко приводит к непредсказуемым ошибкам. Поэтому если требуется высокая надежность расчета, то необходимо строить безусловно

устойчивые неявные схемы. Они несколько сложнее в реализации и плохо приспособлены для многопроцессорных компьютеров. Однако эти схемы экономичны и по трудоемкости сопоставимы со схемой «крест».

Исходная схема. Напишем многомерный аналог одномерной неявной схемы с весами (6.18):

$$\frac{1}{\tau^2}(\hat{u} - 2u + \check{u}) = \sum_{\alpha=1}^{p} \Lambda_\alpha [\sigma\hat{u} + (1 - 2\sigma)u + \sigma\check{u}] + f. \qquad (6.55)$$

Одномерные трехточечные операторы Λ_α записываем так, чтобы они обеспечивали аппроксимацию $O\left(h_\alpha^2\right)$ даже в случае неравномерной сетки. Схема (6.55) симметрична относительно среднего временного слоя, что обеспечивает аппроксимацию $O\left(\tau^2\right)$. Поэтому ее полная аппроксимация есть $O\left(\tau^2 + \sum h_\alpha^2\right)$. Можно показать, что эта схема безусловно устойчива при $1/4 \le \sigma \le 1/2$, как и в одномерном случае.

Однако эта схема содержит на новом слое выражение $\sum \Lambda_\alpha \hat{u}$. Этот оператор уже встречался в многомерном параболическом уравнении (см. подразд. 4.2.). Там было показано, что прямое обращение этого оператора не экономично, а его факторизация дает экономичную схему.

От факторизации требуется, чтобы она не ухудшала аппроксимации $O\left(\tau^2\right)$. Для гиперболического уравнения это можно сделать разными способами. Но предпочтительным является следующий метод.

Эволюционная факторизация. Точно преобразуем схему с весами (6.55) к следующей форме:

$$\left(E - \tau^2\sigma \sum_{\alpha=1}^{p} \Lambda_\alpha\right) \frac{\hat{u} - 2u + \check{u}}{\tau^2} = \sum_{\alpha=1}^{p} \Lambda_\alpha u + f. \qquad (6.56)$$

Аналогично параболическому уравнению приближенно расщепляем оператор в (6.56) на множители и получаем эволюционно-факторизованную схему:

$$\prod_{\alpha=1}^{p} \left(E - \tau^2\sigma\Lambda_\alpha\right) u = \sum_{\alpha=1}^{p} \Lambda_\alpha u + f. \qquad (6.57)$$

Исследуем полученную схему.

Аппроксимация. Разложим факторизованный оператор в (6.57) в ряд по степеням τ и получим

$$E - \tau^2\sigma\sum_\alpha \Lambda_\alpha + \tau^4\sigma^2\sum_\alpha\sum_\beta \Lambda_\alpha\Lambda_\beta + \dots \tag{6.58}$$

Заметим, что перестановочности операторов Λ_α нигде не требуется. Это выражение отличается от нефакторизованного оператора (6.56) членом $O\left(\tau^4\right)$. Это не только сохраняет общую аппроксимацию $O\left(\tau^2\right)$, но даже не меняет коэффициента в соответствующем члене невязки! Напомним, что для параболического уравнения эволюционная факторизация сохраняла аппроксимацию $O\left(\tau^2\right)$, но вносила дополнительный вклад в этот член невязки. Поэтому для гиперболического уравнения эволюционная факторизация оказывается лучше, чем для параболического уравнения.

Устойчивость. Исследование проведем методом гармоник. В каждом операторе Λ_α все переменные, кроме x_α, имеют сохраняющиеся номера сеточных узлов; у переменной x_α присутствуют индексы $n_{\alpha-1}, n_\alpha, n_{\alpha+1}$. Поэтому стандартное введение гармоник приводит к следующим заменам:

$$\Lambda_\alpha \to -4\frac{c_\alpha^2}{h_\alpha^2}\sin^2\left(\frac{q_\alpha h_\alpha}{2}\right),\ \hat{u} \to \rho u,\ u \to \rho\check{u}.$$

Подставляя их в факторизованную схему (6.58), получаем уравнение для множителя роста:

$$\prod_{\alpha=1}^{p}\left(1+2\sigma\gamma_\alpha\right)\left(\rho - 2 + 1/\rho\right) = -2\sum_{\alpha=1}^{p}\gamma_\alpha,\ \gamma_\alpha = \frac{2\tau^2c_\alpha^2}{h_\alpha^2}\sin^2\left(\frac{q_\alpha h_\alpha}{2}\right).$$

Оно преобразуется в квадратное уравнение:

$$\rho^2 - 2\left[1 - \frac{\sum\limits_{\alpha=1}^{p}\gamma_\alpha}{\prod\limits_{\alpha=1}^{p}\left(1+2\sigma\gamma_\alpha\right)}\right]\rho + 1 = 0. \tag{6.59}$$

По теореме Виета произведение корней равно 1; поэтому для устойчивости необходимо и достаточно, чтобы они образовывали

комплексно-сопряженную пару. Для этого дискриминант уравнения (6.59) должен быть не положительным для всех гармоник. Это выполняется при

$$-1 \le 1 - \frac{\sum\limits_{\alpha=1}^{p} \gamma_\alpha}{\prod\limits_{\alpha=1}^{p} (1 + 2\sigma\gamma_\alpha)} \le 1. \tag{6.60}$$

Правое неравенство (6.60) всегда удовлетворяется, поскольку все $\gamma_\alpha > 0$. Левое неравенство (6.60) эквивалентно условию

$$\prod_{\alpha=1}^{p} (1 + 2\sigma\gamma_\alpha) \ge \frac{1}{2} \sum_{\alpha=1}^{p} \gamma_\alpha. \tag{6.61}$$

Видно, что условие (6.61) выполняется для всех гармоник, если $\sigma \ge 1/4$. Таким образом, при $\sigma \ge 1/4$ факторизованная схема (6.57) безусловно устойчива, как и в одномерном случае.

Брать $\sigma > 1/2$ нецелесообразно: при этом вес среднего слоя становится отрицательным. Поэтому в расчетах следует полагать $1/4 \le \sigma \le 1/2$. При выполнении этого условия схема (6.57) безусловно сходится с точностью $O\left(\tau^2 + \sum h_\alpha^2\right)$.

Алгоритм разрешения факторизованной схемы сводит ее к последовательности одномерных прогонок, т. е. является экономичным. Он аналогичен случаю теплопроводности (п. 4.2.2). Опишем его для случая трехмерной задачи, обозначая пространственные операторы через $\Lambda_x, \Lambda_y, \Lambda_z$. Заменим факторизованную схему (6.57) на эквивалентную цепочку трех разностных схем:

$$\left(E - \sigma\tau^2\Lambda_z\right) w = \left(\Lambda_x + \Lambda_y + \Lambda_z\right) u + f; \tag{6.62}$$

$$\left(E - \sigma\tau^2\Lambda_y\right) v = w; \tag{6.63}$$

$$\left(E - \sigma\tau^2\Lambda_x\right) \frac{\hat{u} - 2u + \check{u}}{\tau^2} = v. \tag{6.64}$$

Здесь на трехмерной сетке дополнительно введены функции w и v (не надо их путать с такими же буквами, введенными в предыдущих пунктах!).

В уравнении (6.62) правая часть вычисляется на среднем слое, т. е. является известной. На неизвестную функцию w действует только одномерный трехточечный оператор по переменной z. Поэтому при фиксированных сеточных значениях переменных x и y значения w с разными индексами по переменной z вычисляются одномерной прогонкой по направлению z. Таких прогонок столько, сколько имеется пар сеточных значений x, y.

Вычислив w, получаем правую часть в уравнении (6.63). После этого неизвестная функция v вычисляется одномерными прогонками по направлению y. Найденное v подставляется в правую часть (6.64), после чего величина $(\hat{u} - 2u + \check{u})/\tau^2$ вычисляется одномерными прогонками по направлению x. Из нее находят $\hat{u}$.

Граничные условия. В эволюционно-факторизованной схеме для уравнения теплопроводности для получения точности $O(\tau^2)$ требовалось довольно сложным образом вносить поправки в естественные граничные условия (см. п. 4.2.2). Это было связано с тем, что в каждом множителе схемы оператор Λ_α умножался на τ.

В эволюционно-факторизованной схеме для гиперболического уравнения операторы Λ_α умножаются на τ^2. Поэтому поправочные члены в естественные граничные условия можно не вводить; аппроксимация $O(\tau^2)$ при этом сохранится. Тем самым для последнего уравнения (6.64) в качестве граничных значений для прогонки берутся величины

$$(\hat{u} - 2u + \check{u})/\tau^2|_{\text{гран}}, \tag{6.65}$$

непосредственно вычисленные из краевых условий первого рода по значениям u на всех трех слоях на соответствующих гранях прямоугольного параллелепипеда G.

Поскольку поправки в граничные значения вносить не надо, для функций v и w также берутся граничные значения (6.65), вычисленные на других гранях параллелепипеда.

Такая простота задания граничных условий точности $O(\tau^2)$ получается только при эволюционной факторизации. В этом преимущество эволюционной факторизации перед другими методами факторизации многомерных гиперболических уравнений.

6.4. СИСТЕМЫ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ

6.4.1. Задачи со многими процессами

Простейшая прикладная задача может содержать только один физический процесс. Однако многие задачи включают ряд одновременно протекающих процессов. Даже задачи, при первом взгляде воспринимающиеся как однопроцессные, на самом деле могут состоять из нескольких процессов. При этом каждый отдельный процесс обычно оказывается законом сохранения некоторой физической величины. Приведем некоторые примеры.

- Простейшее уравнение переноса описывает один физический процесс — перенос частиц, летящих с заданной скоростью. Его можно записать в интегральной форме, которая является законом сохранения числа частиц в некотором объеме (изменение числа частиц равно интегралу от потока через границу этого объема).
- Параболическое уравнение (уравнение теплопроводности) является следствием интегрального закона сохранения энергии: изменение энергии в объеме равно интегралу от теплового потока через поверхность. В п. 4.1.5 показано, что это уравнение распадается на два уравнения: определение теплового потока через градиент температуры и уравнение баланса энергии. Фактически это два различных процесса.
- Решение одномерного уравнения акустики (6.2) включает два переноса: вправо и влево. Неудивительно, что оно в п. 6.2.1 заменено на систему двух уравнений, описывающих два разные процесса: перенос амплитуды возмущения и перенос скорости изменения этой амплитуды.
- Одномерные уравнения газодинамики включают законы сохранения трех величин: массы, импульса и энергии. К этим трем уравнениям добавляют еще закон изменения координат, отражающий движение системы как целого.
- В уравнения магнитной газодинамики входят, помимо обычных уравнений газодинамики, еще уравнения изменения электрических и магнитных полей.

Таким образом, в прикладных расчетах приходится решать системы, содержащие несколько уравнений в частных производных для функций $u(x,t), v(x,t), w(x,t), \ldots$. Их можно символически записать следующим образом:

$$\frac{\partial u}{\partial t} = \mathbf{A}\,(u, v, w, \ldots)\,,$$
$$\frac{\partial v}{\partial t} = \mathbf{B}\,(u, v, w, \ldots)\,,\ \frac{\partial w}{\partial t} = \mathbf{C}\,(u, v, w, \ldots)\,,\ \ldots. \tag{6.66}$$

Здесь жирными буквами $\mathbf{A}, \mathbf{B}, \mathbf{C}$ обозначены дифференциальные операторы, содержащие частные производные по пространственным переменным. Как правило, эти операторы содержат только первые и/или вторые производные; более высокие производные встречаются крайне редко. Все задачи должны содержать краевые условия. Будем предполагать, что они включены в соответствующие пространственные операторы.

В исходных задачах могут возникать уравнения, содержащие вторые производные по времени. На примере уравнения акустики мы видели, как эти уравнения превращаются в систему уравнений первого порядка по времени, т. е. приводится к виду (6.66). На примере уравнения теплопроводности было показано, что исходное уравнение может быть приведено к дифференциально-алгебраической системе. В этом случае в системе (6.66) будут возникать алгебраические уравнения, где вместо производной по времени будет стоять 0.

Для решения задачи (6.66) вводится пространственная сетка. Значения решения в узлах x_n этой сетки рассматривают как вектор-функции от времени: $u_n(t) = u(x_n, t)$, $v_n(t) = v(x_n, t)$, $w_n(t) = w(x_n, t)$, $\mathbf{u} = \{u_n\}$, $\mathbf{v} = \{v_n\}$, $\mathbf{w} = \{w_n\}$. Дифференциальные операторы заменяем разностными: $\mathbf{A} \to A$, $\mathbf{B} \to B$, $\mathbf{C} \to C$, ...; для этого следует использовать интегро-интерполяционный метод и строить консервативные разностные схемы, желательно бикомпактные. Применяя метод прямых, заменяем (6.66) системой обыкновенных дифференциальных уравнений очень большого числа неизвестных:

$$\frac{d\mathbf{u}}{dt} = A(\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots), \frac{d\mathbf{v}}{dt} = B(\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots), \frac{d\mathbf{w}}{dt} = C(\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots); \tag{6.67}$$

эта система может быть также дифференциально-алгебраической. Рассмотрим альтернативные алгоритмы решения этой системы.

6.4.2. Расщепление по процессам

Идеология этого метода сложилась в конце 1940-х годов. Тогда передовой научной проблемой была задача расчета мощно-

сти ядерного взрыва. Она сводилась к решению системы большого числа уравнений в частных производных, описывающих процессы газодинамики с теплопроводностью и переноса нейтронов с их размножением при делении урана или плутония. Для решения этой задачи создавались компьютеры первого поколения. Их вычислительные возможности (быстродействие и оперативная память) были ничтожны по современным критериям. Например, в оперативную память с трудом помещались данные, необходимые для решения только одного уравнения системы.

Поэтому сложилась следующая процедура решения системы (6.67). Эта система имеет первый порядок по времени. Поэтому схема выбиралась двуслойной. Для перехода от исходного слоя t к новому слою $\hat{t}$ каждое уравнение вызывалось в оперативную память по отдельности.

Уравнение для каждой отдельной функции имеет первый порядок по времени. Поэтому для данной функции можно строить двуслойную схему интегрирования. Все остальные функции при этом «замораживаются», т. е. не меняются в ходе данного интегрирования. Подобные расчеты не претендуют на высокую точность, поэтому для них выбирают двуслойные схемы точности $O(\tau^2)$. При этом схема для каждой функции может быть своя. Простейший расчет, т. е. переход $t \to \hat{t}$, выглядит следующим образом.

Сначала решаем первое уравнение системы (6.67) для функции $\mathbf{u}(t)$. При этом значения функций $v, w, \ldots$ нам известны только на исходном слое. Поэтому результат решения обозначим через $\tilde{\mathbf{u}}$, а не через $\hat{\mathbf{u}}$: $\mathbf{u} \to \tilde{\mathbf{u}}$. Поскольку прочие функции взяты с исходного слоя, по отношению к ним схема является явной. При такой процедуре общая аппроксимация системы (6.67) есть только $O(\tau)$, а не $O(\tau^2)$.

Затем решаем второе уравнение системы (6.67) для функции $\mathbf{v}(t)$. Функции $\mathbf{w}(t)$ и последующие мы также вынуждены брать с исходного слоя. Однако функцию $\mathbf{u}(t)$ мы уже знаем на новом слое. Поэтому во второе уравнение мы можем подставлять вместо $\mathbf{u}$ значения $(\mathbf{u} + \tilde{\mathbf{u}})/2$ или другое выражение, обеспечивающее аппроксимацию $O(\tau^2)$ для этого члена. Разумеется, влияние неучтенных функций $\mathbf{w}$ и других приведет к ошибкам аппроксимации $O(\tau)$ в соответствующих членах. Таким образом, переход $\mathbf{v} \to \tilde{\mathbf{v}}$ будет иметь несколько меньший коэффициент в остаточном члене $O(\tau)$, чем в первом уравнении.

Решая третье уравнение (6.67) для $\mathbf{w}(t)$, мы можем использовать улучшенные выражения уже для двух первых функций: $(\mathbf{u}+\tilde{\mathbf{u}})/2$ и $(\mathbf{v}+\tilde{\mathbf{v}})/2$.

Таким образом последовательно получим переход на новый слой для всей системы (6.67). Разумеется, итоговый порядок аппроксимации будет при этом лишь $O(\tau)$.

Отдельные схемы. В сложных задачах на первый план выходит надежность схем. Поэтому для каждого отдельного процесса выбирают, как правило, безусловно устойчивую схему. Такая схема может быть только неявной; метод интегрирования по времени должен быть A-устойчивым. Соответствующие схемы для отдельных процессов рассматривались выше в этой и предыдущих главах. Напомним, что для диссипативных процессов (теплопроводность, диффузия, квазилинейный перенос с диссипативным членом в правой части) целесообразно использовать L-устойчивые методы интегрирования по времени (например, схему CROS).

Для линейных одномерных простейших процессов даже неявные схемы приводят к безытерационному алгоритму (например, прогонке). Если задача нелинейная, то неявная схема требует итераций для нахождения разностного решения. Эти итерации надо выполнять до сходимости итерационного процесса с достаточно высокой точностью. Только в этом случае реализуется безусловная устойчивость схемы. Если в нелинейной неявной схеме принудительно ограничить число итераций, безусловной устойчивости может не быть.

Ранее отмечалось, что для отдельных процессов обычно выбирают схемы точности $O\left(\tau^2\right)$. Однако на особенно трудных задачах, вроде квазилинейного уравнения теплопроводности с быстро бегущей тепловой волной, схемы точности $O\left(\tau^2\right)$ могут оказаться недостаточно надежными: итерационный процесс нахождения разностного решения при не слишком малом шаге τ перестает сходиться. В этом случае переходят на чисто неявные схемы точности $O(\tau)$: их надежность выше.

Если задача достаточно проста или если скорость компьютера позволяет брать достаточно малый шаг τ, допустимо использование явных схем для недиссипативных процессов (переноса или акустики).

Второй порядок точности. Если для отдельных процессов используются схемы точности $O\left(\tau^2\right)$, то желательно полу-

чить для полной схемы с расщеплением по процессам также точность $O\left(\tau^2\right)$. Этого можно добиться, если повторить процесс перехода $t \to \hat{t}$. Назовем этот переход для всех процессов ***кругом***.

Однократное выполнение круга дает точность $O(\tau)$. Выполним второй круг, выбирая для каждого процесса замороженные функции следующим образом. Для большей наглядности будем считать, что мы делаем переход для функции $\mathbf{v}$. Для следующих за ней функций подставляем $(\mathbf{w} + \tilde{\mathbf{w}})/2$, используя значения $\tilde{\mathbf{w}}$ с первого круга. Для ранее выполняемых процессов вместо $\tilde{\mathbf{u}}$ подставляем в полусумму то значение, которое уже вычислено на втором круге. Выполнение второго круга обеспечивает суммарную точность $O\left(\tau^2\right)$.

Можно аналогичным образом вычислять третий, четвертый и последующие круги. При этом общий порядок аппроксимации по времени останется $O(\tau^2)$, но коэффициент в остаточном члене при этом обычно несколько уменьшается.

Безусловная устойчивость. Для хорошей надежности нужна безусловная устойчивость не просто схем для отдельных процессов, а полной схемы перехода $t \to \hat{t}$ по всем процессам. Исследование устойчивости таких схем можно выполнять методом гармоник, как это делалось в п. 6.2.3. Обычно такую устойчивость удается доказать только в том случае, если предполагать выполнение многих итерационных кругов до сходимости итерационного процесса, т. е. до тех пор, пока решение на новом слое не перестанет меняться от круга к кругу. Если это не сделано, то безусловная устойчивость схем для отдельных процессов еще не гарантирует безусловную устойчивость схемы в целом. Поэтому на практике зачастую выполняют много кругов ради устойчивости, хотя трудоемкость при этом существенно возрастает, а точность улучшается лишь незначительно по сравнению с двумя кругами.

Заметим, что при недостаточно малом шаге τ этот итерационный процесс может и не сходиться.

Многомерность. Описанные выше процедуры первоначально широко использовались для одномерных задач на компьютерах малой мощности. Когда мощность компьютеров значительно возрасла, для одномерных задач появился более хороший метод, описанный далее. Однако многомерные задачи требуют огромных компьютерных ресурсов. Поэтому в ближайшие

годы для многомерных задач метод расщепления по процессам будет оставаться основным даже для расчетов на суперкомпьютерах.

6.4.3. Жесткий метод прямых (Stiff Method of Lines)

Этот метод оказывается выгодным для одномерных задач. Он позволяет выполнять вычисления даже на современных персональных компьютерах. Метод является обобщением способа, использованного в подразделе 6.2 для системы двух уравнений. Опишем его.

Из всех сеточных векторов $\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots$ составим единый сеточный вектор $\mathbf{z}$ по следующему правилу. Сначала сгруппируем значения всех функций в нулевой (граничной) точке, затем в первой точке и так далее вплоть до N-й (граничной) точки:

$$\mathbf{z} = \{u_0, v_0, w_0, \ldots; u_1, v_1, w_1, \ldots; \ldots; u_N, v_N, w_N, \ldots\}. \quad (6.68)$$

Размерность каждой отдельной сеточной функции есть $N+1$, а размерность суммарного вектора $\mathbf{z}$ есть $K(N+1)$, где K — число отдельных процессов. Тогда полную разностную схему символически можно записать в следующем виде:

$$\frac{d\mathbf{z}}{dt} = \mathbf{F}(\mathbf{z}); \quad (6.69)$$

здесь разностный оператор $\mathbf{F}$ составлен из разностных операторов $A, B, C, \ldots$, в которых узловые значения отдельных сеточных функций заменены на значения единой сеточной функции $\mathbf{z}$ с соответствующими индексами. Ранее отмечалось, что операторы $A, B, C, \ldots$ являются двухточечными или трехточечными по пространству. Это означает, что в правых частях (6.69) могут присутствовать одновременно не любые компоненты вектора $\mathbf{z}$, а только с номерами, сравнительно близкими к номеру уравнения. В самом деле, пусть некоторая компонента z_j обозначает значение одной из функций в точке x_n. Тогда в j-м уравнении (6.69) могут присутствовать только значения всех других функций в той же точке x_n и соседних точках $x_{n\pm1}$.

Отдельные процессы нередко бывают диссипативными. Тогда вся система (6.69) оказывается жесткой. Поэтому ее целесообразно интегрировать с помощью $L2$-устойчивой схемы CROS, имеющей аппроксимацию $O(\tau^2)$. Формулы имеют следующий вид:

$$\hat{\mathbf{z}} = \mathbf{z} + \tau \mathbf{R} \mathbf{e} \mathbf{y}, \quad \left(E - \frac{1+\mathrm{i}}{2} \tau \mathbf{F}_{\mathbf{z}} \right) \mathbf{y} = \mathbf{F}(\mathbf{z}). \tag{6.70}$$

Рассмотрим структуру матрицы Якоби $\mathbf{F}_{\mathbf{z}}$; она изображена на рис. 6.4. Поскольку вектор $\mathbf{z}$ имеет размерность $K(N+1)$, порядок этой матрицы таков же. Эту матрицу можно разбить на квадратные клетки размерности K, равной числу процессов. Число клеток по каждой стороне матрицы есть число пространственных узлов $N+1$. В предыдущем абзаце говорилось, какие сеточные значения реально присутствуют в правых частях. Отсюда видно, что в матрице $\mathbf{F}_{\mathbf{z}}$ диагональные клетки могут быть плотно заполненными. Кодиагональные клетки также могут быть плотно заполненными, хотя обычно они будут заполнены довольно слабо. Все прочие клетки остаются незаполненными.

Таким образом, матрица $\mathbf{F}_{\mathbf{z}}$ имеет ленточную структуру. Для хорошей точности расчета пространственные сетки берут подробными: $N \sim 100 - 1000$. Число процессов обычно бывает не слишком большим, $K \leq 10$, поэтому ширина ленты оказывается много меньше порядка матрицы Якоби. Тем самым линейную систему (6.70) можно экономично решать методом исключения Гаусса для ленточной матрицы.

Мы получим результаты с наименьшими погрешностями, если будем использовать аналитические выражения для матрицы Якоби. Однако при практической реализации этого метода нахождение аналитических выражений часто приводит к чрезмерно громоздким алгоритмам и программам. Поэтому нередко используют разностную аппроксимацию частных производных,

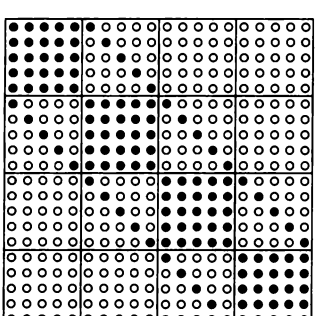


Рис. 6.4. Типичная структура матрицы Якоби для жесткого метода прямых

варьируя значения аргументов $\mathbf{z}$ (см. п. 1.2.3). Это приводит к некоторому ухудшению точности, однако на 64-разрядном компьютере результаты обычно остаются хорошими (но вычисления с 32-разрядными числами недопустимы).

Преимущества данного метода следующие. *1.* Схема $L2$-устойчива, т. е. безусловно устойчива (на линейных задачах) и имеет хорошие диссипативные свойства: хорошо сглаживает возможные счетные немонотонности. *2.* Она имеет хорошую точность $O\left(\tau^2\right)$. *3.* Схема безытерационна, так что не возникает вопроса о сходимости итераций и связанных с этим осложнений, в отличие от расщепления по процессам. *4.* Объем расчетов не слишком велик.

Если исходная система (6.67) является дифференциально-алгебраической, то система (6.70) также должна быть дифференциально-алгебраической. В этом случае перед производной $\mathbf{z}_t$ должна стоять сингулярная матрица G (см. подразд. 1.3). Тогда в схеме (6.70) вместо матрицы E также ставится G.

Жесткий метод прямых отлично работает на линейных одномерных задачах, а также на большинстве нелинейных задач. Только на задачах с очень сильной нелинейностью (например, очень быстро бегущей тепловой волне) он может оказаться недостаточно надежным: возникает ограничение на величину шага τ.

6.4.4. Пример

Рассмотрим затвердевание эпоксидной смолы. Она является вязкой жидкостью и состоит из молекул мономера. Эти молекулы полимеризуются (соединяются) под действием катализатора, который также является вязкой жидкостью. Для начала процесса обе жидкости тщательно смешивают. При этом жидкости разбиваются на мелкие капли. Капли настолько мелкие, что мономер и катализатор хорошо проникают друг в друга за счет молекулярной диффузии. Тогда химические процессы идут не только на поверхности, но и в объеме, т. е. намного быстрее. Составим математическую модель процесса.

Рассмотрим одномерную задачу, вводя координату x перпендикулярно поверхности соприкасновения капель. Обозначим через $u_k(x,t)$, $1 \le k \le K$, концентрацию k-го химического вещества: катализатора, мономера и полимеров со всевозможными разумными длинами цепей. Тогда процессы диффузии компо-

нент и химических реакций будут описываться следующей системой уравнений:

$$\frac{\partial u_k}{\partial t} = D_k \frac{\partial^2 u_k}{\partial x^2} + f_k\left(u_1, u_2, \ldots, u_K\right), \; 1 \le k \le K. \tag{6.71}$$

Здесь D_k — коэффициенты диффузии, которые для простоты берем постоянными, а f_k — скорости химических реакций. Припишем границе капель значение $x = 0$. Если радиус левой капли есть a, а правой — b, то из симметрии процесса граничные условия можно взять следующим образом:

$$\frac{\partial u_k}{\partial x} = 0 \text{ при } x = -a \text{ и } x = b. \tag{6.72}$$

Начальные условия при $t = 0$ будут разрывными: при $x < 0$ имеется только мономер, при $x > 0$ —только катализатор, а полимерные молекулы отсутствуют.

Подробно рассмотрим решение жестким методом прямых.

Разностная схема. Введем сетку x_n, $-M \le n \le N$, с достаточно малым шагом h, который для простоты берем постоянным. Поскольку интервалов будет много, возможной несоизмеримостью радиусов a и b можно пренебречь и полагать $a \approx Mh$, $b \approx Nh$. Формально среда является слоистой. Однако коэффициенты диффузии взяты постоянными. Поэтому у этой задачи есть только классическое (не обобщенное) решение, так что необходимости в построении бикомпактных схем нет, а разрывные начальные данные в параболических уравнениях не представляют опасности для L-устойчивых схем. Это позволяет приписать концентрации не узлам, а серединам интервалов.

Обозначим $u_k\left(x_{n-1/2}, t\right) = u_{k,n-1/2}$. Заменяя пространственную производную второй разностью, получим следующие уравнения:

$$\begin{gathered}\frac{du_{k,n-1/2}}{dt} = \frac{D_k}{h^2}\left(u_{k,n-3/2} - 2u_{k,n-1/2} + u_{k,n+1/2}\right) + f_k, \\ \left(u_{1,n-1/2}, u_{2,n-1/2}, \ldots, u_{K,n-1/2}\right), \; 1 \le k \le K;\end{gathered} \tag{6.73}$$

границы пространственных индексов обсудим позже. Схема имеет аппроксимацию $O\left(h^2\right)$. Чтобы получить такую же аппроксимацию граничных условий, воспользуемся методом фиктивных точек и введем концентрации на пол-интервала левее и правее центров капель. Тогда получим

$$u_{k,-M-1/2} = u_{k,-M+1/2} \text{ и } u_{k,N-1/2} = u_{k,N+1/2},\ 1 \le k \le K. \quad (6.74)$$

Отсюда видно, что для (6.73) надо полагать $-M+1 \le n \le N$. Таким образом, в схеме (6.73) использованы значения концентраций в $N+M$ серединах интервалов.

Краевые условия (6.74) являются алгебраическими уравнениями. Тем самым система (6.73) — (6.74) является дифференциально-алгебраической системой. Вместо ее непосредственного решения удобнее исключить «заграничные» значения концентраций из (6.73) с помощью краевых условий (6.74). Тогда полученная система будет чисто дифференциальной.

Видно, что вектор всех переменных $\mathbf{z}$ содержит $K(N+M)$ компонент. При этом в матрице Якоби диагональные клетки будут полностью заполнены за счет производных от f_k, которые могут зависеть от всех химических компонент в данной точке. Зато в кодиагональных клетках будут присутствовать только диагональные члены, обусловленные второй пространственной разностью. Именно этот случай изображен на рис. 6.4. Поэтому в данном примере ширина ленты будет составлять всего $2K+1$. Тем самым применение схемы CROS (6.70) приводит к достаточно простым вычислениям.

Заметим, что при вычислении матрицы Якоби производные по концентрациям, входящие в пространственную разность, следует брать точно, а при вычислении частных производных от f_k допустимы разностные замены.

Глава 7

ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ

7.1. КОРРЕКТНО ПОСТАВЛЕННЫЕ ЗАДАЧИ

7.1.1. Элементы теории

Интегральным называют уравнение, в котором неизвестная функция $u(x)$ стоит под знаком интеграла. Одномерное нелинейное интегральное уравнение имеет вид

$$\int_a^b K\left(x, \xi, u\left(\xi\right)\right) d\xi = F\left(x, u\left(x\right)\right), \; a \leq x \leq b, \tag{7.1}$$

где ядро $K\left(x, \xi, u\right)$ и правая часть $F\left(x, u\right)$ — заданные функции, а $u(x)$ — неизвестная функция.

К интегральным уравнениям приводят многие физические задачи. Так, задача восстановления переданного радиосигнала $u(t)$ по принятому сигналу $f(t)$ сводится к решению интегрального уравнения типа свертки:

$$\int_0^t K(t-\tau)u(\tau)d\tau = f(t), \tag{7.2}$$

где ядро $K(\xi)$ зависит от свойств приемной аппаратуры и среды, через которую проходил сигнал.

Заметим, что даже для задач, записанных в терминах уравнений в частных производных, первичной обычно является формулировка в виде интегральных законов сохранения, т. е. интегральных уравнений. В предыдущих главах такие формулировки использовались для построения консервативных разностных схем.

Интегральные уравнения в некоторых отношениях удобнее дифференциальных. Во-первых, интегральное уравнение содержит в себе полную постановку задачи. Например, интегральное уравнение

$$u(x) = u_0 + \int_{x_0}^{x} f(\xi, u(\xi))\, d\xi \tag{7.3}$$

эквивалентно задаче Коши для дифференциального уравнения

$$\frac{du(x)}{dx} = f(x, u), \; u(x_0) = u_0. \tag{7.4}$$

Тем самым для уравнения (7.3) не требуется задавать никаких дополнительных условий, начальных или граничных.

Во-вторых, в интегральных уравнениях переход от одной переменной ко многим является естественным. Так, многомерным аналогом (7.1) является уравнение

$$\begin{gathered} \int_G K(\mathbf{x}, \xi, u(\xi))\, d\xi = F(\mathbf{x}, u(\mathbf{x})), \\ \mathbf{x} = \{x_1, x_2, \ldots, x_p\} \in G(\mathbf{x}), \end{gathered} \tag{7.5}$$

отличающееся от (7.1) только тем, что интегрирование проводится по многомерной области G. Поскольку оба уравнения не требуют дополнительных условий и полностью определяют задачу, аналогия является полной. Тем самым теоретическое обоснование постановок и методов решения одномерных задач непосредственно обобщается на случай многих измерений.

Наоборот, в дифференциальных уравнениях переход от одной переменной к нескольким, т. е. от обыкновенных дифференциальных уравнений к уравнениям в частных производных, является принципиальным усложнением, приводит к новым постановкам задач и требует новых методов для их обоснования.

Далее ограничимся рассмотрением одномерного уравнения (7.1) и некоторых его частных случаев.

Линейные задачи. Лучше всего изучены уравнения, в которые неизвестная функция $u(x)$ входит линейно. Их традиционно записывают в следующем виде:

$$u(x) - \lambda \int_a^b K(x, \xi)\, u(\xi)\, d\xi = f(x), \; a \le x \le b. \tag{7.6}$$

Это уравнение называют уравнением Фредгольма второго рода; ядро $K(x, \xi)$ этого уравнения определено на квадрате $a \le x \le b$, $a \le \xi \le b$.

В ряде задач ядро $K(x,\xi)$ отлично от нуля только на треугольнике $a \le \xi \le x \le b$ (т. е. $K(\xi,x) = 0$ при $x < \xi$). Тогда уравнение (7.6) приобретает следующий вид:

$$u(x) - \lambda \int_a^x K(x,\xi)\, u(\xi)\, d\xi = f(x),\ a \le x \le b. \qquad (7.7)$$

Его называют уравнением Вольтерра второго рода. Это уравнение теоретически исследовать или численно решить много проще, чем уравнение Фредгольма.

Если в уравнениях (7.6) и (7.7) отбросить член $u(x)$ вне интеграла, то получим уравнения Фредгольма и Вольтерра первого рода. Задачи для уравнений первого рода являются некорректно поставленными и будут рассмотрены в подразделе 7.2. Для уравнений второго рода задачи корректно поставлены; остановимся на этих задачах.

Для ***однородного уравнения*** Фредгольма второго рода (7.6) (т. е. при $f(x) \equiv 0$) ставится задача на собственные значения:

$$u(x) = \lambda \int_a^b K(x,\xi)\, u(\xi) d\xi,\ a \le x \le b. \qquad (7.8)$$

Требуется найти такие значения параметра $\lambda = \lambda_s$, при которых уравнение (7.8) имеет нетривиальные решения $u(x) \not\equiv 0$; будем обозначать эти решения как $u_s(x)$. Значения λ_s называют собственными значениями ядра $K(x,\xi)$, а $u_s(x)$ — собственными функциями.

Если ядро вещественное и симметричное, $K^*(x,\xi) = K(x,\xi) = K(\xi,x)$, то оно имеет по меньшей мере одно собственное значение. Все собственные значения такого ядра вещественны, а его собственные функции ортогональны друг другу. Заметим, что система собственных функций $u_s(x)$ может быть неполной и даже конечной.

Неоднородное уравнение Фредгольма (7.6) при значении параметра λ, не равном ни одному из собственных значений λ_s ядра, имеет решение $u(x)$, притом единственное.

Если ядро $K(x,\xi)$ и правая часть $f(x)$ непрерывны вместе со своими p-ми производными, то решение также p раз непрерывно дифференцируемо. В этом легко убедиться, продифференцировав (7.6) p раз:

$$u^{(p)}(x) = f^{(p)}(x) + \lambda \int_a^b \frac{\partial^p K(x,\xi)}{\partial x^p} u(\xi) d\xi.$$

При сделанных предположениях правая часть этого равенства непрерывно зависит от x, что доказывает наше утверждение.

Для симметричного ядра решение неоднородного уравнения (7.6) представляется в виде разложения Шмидта:

$$u(x) = f(x) + \sum_s \frac{\lambda}{\lambda_s - \lambda} u_s(x) \int_a^b f(\xi) u_s(\xi) d\xi; \qquad (7.9)$$

если ядро $K(x,\xi)$ и правая часть $f(x)$ интегрируемы с квадратом, то этот ряд сходится абсолютно и равномерно. В данном случае из формулы (7.9) непосредственно видно, что при $\lambda \neq \lambda_s$ решение $u(x)$ существует, единственно и непрерывно зависит от $f(x)$, что означает корректность задачи (7.6).

Пусть параметр λ равен одному из собственных значений λ_s ядра $K(x,\xi)$. Тогда неоднородное уравнение Фредгольма (7.6) при произвольной правой части $f(x)$, вообще говоря, не имеет решения. Однако при некоторых правых частях $f(x)$ оно может иметь решение, притом не единственное. Таким образом, при $\lambda = \lambda_s$ в классе непрерывных или даже достаточно гладких правых частей $f(x)$ задача (7.6) является некорректно поставленной.

Уравнение Вольтерра не имеет собственных значений: если в уравнении (7.7) положить $f(x) = 0$, то оно будет иметь только тривиальное решение $u(x) = 0$. Поэтому неоднородное уравнение (7.7) всегда имеет решение, притом единственное.

7.1.2. Сеточный метод

Это несложный и универсальный численный метод, позволяющий получать решение одномерных, даже нелинейных задач с высокой точностью, а двумерных — с удовлетворительной. При этом расчеты на последовательности сгущающихся сеток позволяют попутно получить апостериорные асимптотически точные оценки погрешности.

Общий случай. Рассмотрим одномерное нелинейное уравнение (7.1), предполагая у ядра $K(x,\xi,u)$ и правой части $F(x,u)$

наличие непрерывных p-х производных по всем аргументам (заметим, что на высокую точность сеточного расчета можно надеяться лишь при достаточно больших p). Возьмем на $[a,b]$ какую-нибудь квадратурную формулу, например линейную формулу с узлами x_n и весами c_n $(1 \le n \le N)$:

$$\int_a^b \phi(\xi)d\xi \approx \sum_{n=1}^{N} c_n \phi(x_n). \tag{7.10}$$

Напомним, что использовать квадратуры порядка точности выше гладкости решения p нецелесообразно: реальный порядок точности все равно не может быть выше p (см. кн. 1). Введем в квадрате $[a \le x \le b, a \le \xi \le b]$ сетку x_n, ξ_m, где как x_n, так и ξ_m являются узлами формулы (7.10). В уравнении (7.1) положим $x = x_n$, а интеграл по ξ аппроксимируем суммой (7.10). Обозначая через $u_n = u(x_n)$ значения решения в узлах, получим соотношения

$$\sum_{m=1}^{N} c_m K(x_n, x_m, u_m) = F(x_n, u_n), \ 1 \le n \le N. \tag{7.11}$$

Это система N нелинейных алгебраических уравнений для определения такого же количества неизвестных (приближенных) значений u_n.

Решать эту систему целесообразно методом Ньютона с применением усечения (см. кн. 1). Напомним, что этот метод для произвольных нелинейных систем удовлетворительно работает лишь для систем небольших порядков $N <\sim 10 - 20$. Для бóльших N трудно найти такое нулевое приближение, которое обеспечивало бы сходимость метода Ньютона.

Многосеточный метод. В качестве квадратуры (7.10) возьмем формулу средних или трапеций на равномерной или квазиравномерной сетке с числом интервалов N. Эти формулы имеют точность $O\left(N^{-2}\right)$, т. е. для высокой точности необходимо большое N, при котором сходимость итерационного процесса Ньютона проблематична. Поэтому построим последовательность сгущающихся сеток. Число интервалов первой сетки возьмем небольшим: $N_1 \sim 10$. Последующие сетки получим удвоением числа интервалов.

На первой сетке N_1 невелико. Поэтому обычно сравнительно легко удается найти такое нулевое приближение, которое обес-

печивает сходимость итерационного процесса Ньютона. Вблизи корня итерационный процесс Ньютона сходится очень быстро (квадратично), поэтому нетрудно получить сеточное решение $u_n^{(1)}$ почти с точностью до ошибок компьютерного округления.

Разумеется, отличие сеточного решения от точного будет заметным, поскольку шаг сетки велик, но не очень большой.

Рассмотрим расчет на второй сетке. Поскольку $u_n^{(1)}$ довольно близко к точному решению, оно должно быть довольно близким и к неизвестному пока решению на второй сетке $u_n^{(2)}$. Поэтому $u_n^{(1)}$ можно взять в качестве нулевого приближения для второй сетки; разумеется, надо при этом провести интерполяцию с узлов первой сетки на узлы второй сетки. Практика расчетов показывает, что при таком выборе нулевого приближения итерационный процесс Ньютона на второй сетке сходится обычно за $2-5$ итераций.

Аналогично проводятся расчеты на всех последующих сетках. При этом нетрудно довести расчеты до сеток с очень большим N, обеспечивающим высокую точность сеточного решения. При этом основной объем расчетов будет приходиться на вычисление матрицы производных на последней сетке, а общая трудоемкость расчетов будет приемлемой.

Точность расчетов необходимо контролировать, используя метод Ричардсона и сравнивая сеточные решения на соседних сетках при одинаковых значениях x. При этом получается апостериорная асимптотически точная оценка погрешности. Если использовать эту оценку для уточнения сеточных значений, т. е. применять рекуррентный метод Ричардсона, то можно получить результат повышенного порядка точности. При этом требуемая точность достигается с меньшими N.

Отсюда видно следующее. Пусть начальная сетка имеет число интервалов N_1, а всего используется k сеток. Тогда при применении рекуррентного метода Ричардсона последняя сетка будет иметь $N_1 2^{k-1}$ интервалов, а окончательный уточненный результат будет соответствовать квадратуре точности $O\left(N_k^{-2k}\right)$.

Сделаем полезное замечание. В приложениях нередко возникают задачи, когда ядро имеет особенность на диагонали квадрата $x = \xi$, а вне этой диагонали его гладкость высока. Поскольку сетки по x и ξ равномерны и одинаковы, то на каждой линии $x = x_n$ диагональ проходит через узел сетки по ξ. Поэтому ко-

гда особенность является разрывом самого ядра, следует применять формулу средних; при этом квадратура сохраняет точность $O\left(N^{-2}\right)$ и каждое рекуррентное уточнение методом Ричардсона повышает порядок точности на 2. Применять формулу трапеций в этом случае нельзя. Если же особенность является разрывом лишь производных ядра, то можно применять как формулу средних, так и формулу трапеций.

Квадратуры Гаусса. Если ядро имеет достаточно много непрерывных производных, можно использовать в качестве (7.10) квадратурную формулу Гаусса с небольшим числом узлов $N \sim 5 - 10$. Напомним, что погрешность формулы Гаусса эквивалентна квадратурам точности $O\left(N^{-2N}\right)$. Поэтому они могут обеспечить хороший результат при небольших N. Поскольку число узлов невелико, то обеспечить сходимость ньютоновских итераций для решения системы нелинейных алгебраических уравнений (7.11) сравнительно легко. По экономичности этот способ существенно превосходит многосеточный метод.

Если ньютоновские итерации сходятся плохо, то можно применить прием, аналогичный многосеточному методу. Сначала следует провести расчет с $N \sim 5$: при таком малом числе узлов легче добиться сходимости итераций. Когда итерации сойдутся, полученное сеточное решение надо интерполировать на сетку с $N \sim 10$ и взять его в качестве нулевого приближения для ньютоновских итераций.

Напомним, что для реализации высокой точности квадратур Гаусса значения узлов x_n и весов c_n следует задавать с компьютерной точностью.

Отметим недостаток квадратур Гаусса. Их нельзя применять, если ядро или его требуемые производные имеют какие-либо особенности (даже особенности на диагонали квадрата): интегрирование по ξ вдоль линий $x = \texttt{const}$ проходит через эти особенности и не позволяет получить высокий порядок точности.

Линейные уравнения. Описанные выше методы рассчитаны на нелинейные интегральные уравнения. Для линейных интегральных уравнений квадратуры приводят к системам линейных алгебраических уравнений относительно сеточной функции. Поэтому алгоритм существенно упрощается. Рассмотрим основные виды линейных задач.

1. Для однородного уравнения Фредгольма второго рода (7.8) ставится задача на собственные значения.

Алгебраическая система (7.11) принимает следующий вид:

$$\sum_{m=1}^{N} c_m K_{nm} u_m = \frac{1}{\lambda} u_n,\ 1 \le n \le N;\ K_{nm} = K(x_n, x_m). \qquad (7.12)$$

Система (7.12) представляет собой задачу на определение собственных значений матрицы K' порядка N с элементами $K'_{nm} = c_m K_{nm}$. Эта матрица имеет N собственных значений $\lambda_s^{(N)}$, $1 \le s \le N$. Они являются приближением к первым собственным значениям λ_s ядра $K(x, \xi)$. Точность этих приближений обычно хороша только для индексов $s \ll N$.

Спектр и сеточное решение (7.12) определяют методами, описанными в кн. 1. Матрица K' является, вообще говоря, плотно заполненной. Скорости даже персональных компьютеров позволяют проводить вычисления с $N \sim 100 - 1000$ для неэрмитовых матриц и существенно больших N для эрмитовых матриц. В этом случае хорошую точность можно получить по крайней мере для первых 10—30 собственных значений.

Если ядро $K(x, \xi)$ вещественно и симметрично, то его точные собственные значения вещественны. Однако сеточное ядро K' в системе (7.12) при этом может оказаться несимметричным за счет коэффициентов c_m. Докажем, что несмотря на это, сеточные собственные значения и собственные функции останутся вещественными. Мы используем квадратуры с $c_n > 0$. Введем новую сеточную функцию $v_n = u_n / \sqrt{c_n}$. Тогда система (7.12) примет следующий вид:

$$\sum_{m=1}^{N} \sqrt{c_n c_m} k_{nm} v_m = \frac{1}{\lambda} v_n,\ 1 \le n \le N. \qquad (7.13)$$

Видно, что матрица системы (7.13) вещественна и симметрична. Следовательно, ее собственные значения и собственные векторы вещественны.

Заметим, что для нелинейных задач применять квадратуры Гаусса бессмысленно: в них используются небольшие N, а из сказанного выше видно, что для нахождения даже первых собственных значений необходимо $N \gg 1$. Поэтому целесообразно использовать квадратуры средних или трапеций и обязательно применять сгущение сеток и метод Ричардсона.

2. Неоднородное уравнение Фредгольма (7.6) приводит к линейной неоднородной алгебраической системе:

$$u_n - \lambda \sum_{m=1}^{N} c_m K_{nm} u_m = f_n,\ 1 \le n \le N,\ f_n = f(x_n). \qquad (7.14)$$

Матрица системы будет плотно заполненной. Сеточное решение легко вычисляется методом исключения Гаусса (см. кн. 1). Даже персональные компьютеры легко решают такие системы с $N > 1000$. Таким образом, в этой задаче нетрудно получить более высокую точность расчета, чем в задаче на собственные значения. Как и для задач на собственные значения, здесь необходимо использовать большие N, т. е. квадратуры средних или трапеций, но не квадратуры Гаусса. Все расчеты необходимо проводить на последовательности сгущающихся вдвое сеток, применять метод Ричардсона и контролировать сходимость сеточного решения к предельной функции при увеличении N.

Линейная система (7.14) имеет единственное решение, если λ отлична от всех собственных значений матрицы $\lambda_s^{(N)}$. Но при больших N величины $\lambda_s^{(N)}$ близки к точным собственным значениям ядра λ_s. Следовательно, описанный алгоритм хорошо обусловлен, если параметр λ не лежит в малой окрестности одного из собственных значений λ_s ядра. Если при расчетах со сгущением сеток наблюдается четкое стремление сеточных функций u_n к предельной функции $u(x)$, то мы имеем дело именно с данным случаем.

Пусть при сгущении сеток устойчивого предельного перехода нет. Это означает, что λ близко к одному из собственных значений λ_s ядра. В этом случае как точная задача (7.6), так и сеточная задача (7.14) плохо обусловлены. Для получения хорошей точности в этом случае требуются очень большие N.

3. Уравнение Вольтерра (7.7) получают из уравнения Фредгольма (7.6), полагая $K(x, \xi) = 0$ при $x < \xi$. Алгебраическая система (7.14) имеет при этом треугольную матрицу:

$$u_n - \lambda \sum_{m=1}^{n} c_m K_{nm} u_m = f_n,\ 1 \le n \le N. \qquad (7.15)$$

Система (7.15) является хорошо обусловленной при любом λ и решается обратным ходом метода исключения Гаусса всего за $3N^2/2$ действий. Поэтому здесь объем вычислений невелик даже при огромных N. Это позволяет находить решение с большой точностью.

Повышение гладкости. Если ядро или правая часть недостаточно гладки, то бесполезно применять квадратурные формулы высокой точности или многократно сгущать сетку при квадратурах невысокой точности. В этих случаях иногда удается улучшить гладкость, преобразуя исходную задачу. Например, пусть ядро $K(x,\xi)$ уравнения (7.6) непрерывно, а $f(x)$ лишь кусочно-непрерывна. Решение $u(x)$ в этом случае также лишь кусочно-непрерывно. Введем новую неизвестную функцию $v(x) = u(x) - f(x)$. Тогда (7.6) преобразуется к следующему виду:

$$v(x) - \lambda \int_a^b K(x,\xi)v(\xi)d\xi = \phi(x),$$
$$\phi(x) = \lambda \int_a^b K(x,\xi)f(\xi)d\xi. \tag{7.16}$$

Функция $\phi(x)$ непрерывна; следовательно, $v(x)$ также будет непрерывной. Поэтому решать уравнение (7.16) проще, чем исходное уравнение (7.6).

Многомерные задачи допускают применение сеточного метода. Надо лишь вместо одномерных квадратур использовать многомерные кубатурные формулы. Однако получить удовлетворительную точность можно только для достаточно гладких ядер и правых частей, причем в областях простейшей формы (прямоугольный параллелепипед). Для этого в качестве кубатурных формул нужно использовать произведение одномерных квадратур Гаусса. Получить апостериорные оценки точности в этом случае не удается.

7.1.3. Метод Галёркина

Для интегральных уравнений метод обычно называют ***методом моментов***. Изложим его на примере уравнения Фредгольма второго рода. Будем искать решение в виде разложения по полной системе функций $\psi_k(x)$:

$$u(x) \approx f(x) + \lambda \sum_{k=1}^{N} c_k \psi_k(x). \tag{7.17}$$

В отличие от краевых задач для обыкновенных дифференциальных уравнений $u(x)$ не должно удовлетворять никаким краевым

условиям. Поэтому от системы $\psi_k(x)$ не надо требовать ничего, кроме полноты.

Подставим разложение (7.17) в уравнение (7.6) и потребуем ортогональности невязки ко всем функциям $\psi_m(x)$, $1 \le m \le N$. Получим линейную алгебраическую систему уравнений для нахождения c_k:

$$\sum_{k=1}^{N} a_{mk} c_k = b_m, \ 1 \le m \le N,$$

$$a_{mk} = \int_a^b \psi_m(x)\psi_k(x)dx - \lambda \int_a^b \int_a^b K(x,\xi)\psi_m(x)\psi_k(\xi)dxd\xi, \quad (7.18)$$

$$b_m = \int_a^b \int_a^b K(x,\xi)\psi_m(x)f(\xi)dxd\xi.$$

В случае задачи на собственные значения (7.8) надо полагать $f(x) = 0$ в (7.17), что дает $b_m = 0$ в (7.18).

Основной трудностью, препятствующей применению метода моментов, является сложность вычисления двукратных интегралов, входящих в (7.18). Поэтому обычно ограничиваются малым числом членов суммы (7.17). На хорошую точность при этом можно рассчитывать только в том случае, если удалось удачно подобрать систему функций $\psi_k(x)$.

Метод применим и к нелинейному уравнению (7.1), но тогда он приводит к нелинейной алгебраической системе.

7.2. НЕКОРРЕКТНЫЕ ЗАДАЧИ

7.2.1. Регуляризация

Если в интегральном уравнении (7.1) правая часть $F(x, u(x))$ не зависит от решения, т. е. $u(x)$ входит только под знак интеграла, то задача оказывается некорректно поставленной. Классическим примером некорректных задач является уравнение Фредгольма первого рода:

$$\int_a^b K(x,\xi)\, u(\xi)d\xi = f(x), \ c \le x \le d, \quad (7.19)$$

и уравнение Вольтерра первого рода:

$$\int_a^x K(x,\xi)\,u(\xi)d\xi = f(x),\ c \le x \le d. \tag{7.20}$$

В отличие от уравнений второго рода ядро уравнения Фредгольма (7.19) задано на прямоугольнике $[c \le x \le d, a \le \xi \le b]$, а в уравнении Вольтерра (7.20) — на трапеции $[c \le x \le d, a \le \xi \le x]$ (при $c < a < d$ эта трапеция превращается в два треугольника), причем функции $u(\xi)$ и $f(x)$ определены на разных отрезках и могут принадлежать разным классам функций U и F.

Покажем, что задача (7.19) неустойчива по правой части и, тем самым, некорректна. Для этого рассмотрим высокочастотное возмущение $\delta u(\xi) = \omega \exp(\mathrm{i}\omega^2\xi)$, $\omega \gg 1$; его амплитуда велика из-за множителя ω перед экспонентой. Ему соответствует возмущение правой части

$$\delta f(x) = \int_a^b K(x,\xi)\delta u(\xi)d\xi = \omega \int_a^b K(x,\xi)\exp\left(\mathrm{i}\omega^2\xi\right)d\xi.$$

Интегрируя по частям, получим

$$\delta f(x) = \frac{1}{\mathrm{i}\omega}K(x,\xi)e^{\mathrm{i}\omega^2\xi}\Big|_{\xi=a}^{\xi=b} - \frac{1}{\mathrm{i}\omega}\int_a^b \frac{\partial K(x,\xi)}{\partial\xi}e^{\mathrm{i}\omega^2\xi}d\xi = O\left(\frac{1}{\omega}\right). \tag{7.21}$$

Это означает, что для достаточно больших частот величина $\|\delta f\|_C = O(1/\omega)$ оказывается сколь угодно малой. Следовательно, существуют такие сколь угодно малые возмущения правой части $\delta f(x)$, которым соответствуют большие возмущения решения с нормой $\|\delta u\|_C = O(\omega)$, т. е. задача (7.19) неустойчива.

Для уравнения Вольтерра (7.20) справедливы те же рассуждения. Напомним, что в кн. 1 мы уже сталкивались с некорректностью задачи численного дифференцирования функции $f(x)$. Эта задача сводится к решению уравнения

$$\int_a^x u(\xi)d\xi = f(x), \tag{7.22}$$

т. е. является частным случаем уравнения Вольтерра первого рода с ядром $K(x,\xi) = 1$ (при $\xi \le x$). Здесь также введение в пра-

вую часть малого возмущения $\delta f(x) = \omega^{-1} \exp(\mathrm{i}\omega^2 x)$ приводит к большому возмущению решения $\delta u(\xi) = \mathrm{i}\omega \exp(\mathrm{i}\omega^2 \xi)$.

Кроме того, если задачи (7.19) и (7.20) при некоторой правой части $f(x)$ имеют решения, то при сколь угодно малом изменении $f(x)$ решение может исчезнуть. Например, задача (7.22) имеет решение для дифференцируемых $f(x)$. Прибавим к $f(x)$ какую-нибудь непрерывную, но не дифференцируемую функцию (например, функцию Вейерштрасса, домноженную на $\epsilon \ll 1$). Изменение $f(x)$ в норме C мало́, а решение перестает существовать.

Непосредственно решать некорректные задачи любыми численными методами бессмысленно. Если $f(x)$ задана с погрешностью $\delta f(x)$, то соответствующее решение $u_\delta(\xi)$ может или не существовать, или отличаться от искомого решения на большую величину $\delta u(\xi)$. Даже если $f(x)$ задана точно, но отыскание решения выполняется численными методами, то неизбежно вносятся погрешности метода и округления. Это снова приводит к большой погрешности решения $\delta u(\xi)$.

Регуляризация. Поскольку точное решение задач (7.19) — (7.20) мы не можем найти, будем искать приближенное решение. Для определенности ограничимся задачей (7.19). Введем невязку

$$\psi(x) = \int_a^b K(x,\xi) u(\xi) d\xi - f(x), \quad c \le x \le d. \tag{7.23}$$

Для точного решения $\psi(x) = 0$, а для приближенного решения $\psi(x) \approx 0$. Будем понимать малость невязки в смысле гильбертовой нормы L_2:

$$\Psi[u] = \|\psi\|_{L_2}^2 = \int_c^d \left[\int_a^b K(x,\xi) u(\xi) d\xi - f(x) \right]^2 dx. \tag{7.24}$$

Если потребовать $\Psi[u] = \min$, то снова возвращаемся к исходной некорректной задаче, для которой $\Psi[u] = 0$. Поэтому надо искать какое-то разумное приближенное решение с $\Psi[u] \approx 0$.

Обычно ищут приближенное решение в некотором достаточно узком классе функций U. Ряд таких задач был рассмотрен в кн. 1. Например, для плохо обусловленных систем линейных алгебраических уравнений искалось решение с не слишком большой нормой $\|\|_u$. При численном дифференцировании табулированной функции рекомендовалось аппроксимировать таблицу

некоторой функцией $\phi(x, \mathbf{c})$ с вектором параметров $\mathbf{c}$, размерность которого много меньше числа узлов таблицы. Производная ϕ_x принимается за искомое приближенное решение. Подобные видоизменения исходной задачи называют ее регуляризацией.

Сделаем замечание. Пусть ядро уравнения Фредгольма (7.19) задано на квадрате, т. е. $a = c$, $b = d$. В этом частном случае возможна следующая регуляризация. Добавим в левую часть уравнения слагаемое $\epsilon u(x)$ с $|\epsilon| \ll 1$. Тогда задача перейдет в уравнение Фредгольма второго рода. Если $1/\epsilon$ не совпадает ни с одним из собственных значений ядра, задача будет корректной. В то же время из-за малости ϵ следует ожидать близости ее решения к точному.

Однако для ядра, заданного на прямоугольнике, этот прием неприменим.

7.2.2. Вариационный метод регуляризации

Рассмотрим задачу (7.19). Описанная в п. 7.2.1 неустойчивость по правой части была связана с высокочастотными возмущениями. Она приводила к большим возмущениям $u(\xi)$. Нетрудно видеть, что возмущения производных $u'(\xi)$, $u''(\xi)$ и т. д. при этом будут еще больше: тем больше, чем выше порядок производной. Поэтому целесообразно искусственно ограничить величину самого решения и, возможно, величину его k-х производных вплоть до некоторого $k = K$.

Для этого введем так называемый стабилизатор А. Н. Тихонова порядка K:

$$\Omega[u] = \sum_{k=0}^{K} \int_a^b p_k(\xi) \left[u^{(k)}(\xi)\right]^2 d\xi, \; p_k(\xi) > 0; \qquad (7.25)$$

весовые множители $p_k(\xi)$ введены для общности, и обычно их можно считать константами. Чтобы стабилизатор принимал не слишком большие значения, должны быть не слишком велики все производные $u^{(k)}(x)$ с $0 \le k \le K$. Пространство функций с ограниченным стабилизатором (7.25) называют пространством Соболева W_2^K. Легко видеть, что у этих функций старшая производная $u^{(K)}(x)$ интегрируема с квадратом на $[a, b]$, а все младшие производные непрерывны и ограниченны.

Будем искать приближенное решение задачи (7.19) в классе функций из пространства W_2^K. Для этого заменим исходную задачу на регуляризованную:

$$M_\epsilon[u] \equiv \Psi[u] + \epsilon\Omega[u] = \min,\ 0 \le \epsilon \ll 1; \tag{7.26}$$

величину ϵ называют *параметром регуляризации*. Подставляя в (7.26) значения функционала $\Psi[u]$ и $\Omega[u]$, получим интегральную форму записи регуляризованной задачи:

$$\begin{aligned} M_\epsilon[u] \equiv \int_c^d \left[\int_a^b K(x,\xi)u(\xi)d\xi - f(x)\right]^2 dx + \\ +\epsilon\sum_{k=0}^{K}\int_a^b p_k(\xi)\left[u^{(k)}(\xi)\right]^2 d\xi = \min. \end{aligned} \tag{7.27}$$

Рассмотрим алгоритм решения данной задачи.

Уравнение Эйлера. Для нахождения минимума (7.27) приравняем нулю производную $M_\epsilon[u]$ по $\delta u(\xi)$. Получим следующее выражение:

$$\begin{aligned} \int_c^d dx\left[\int_a^b K(x,\eta)u(\eta)d\eta - f(x)\right]\int_a^b K(x,\xi)\delta u(\xi)d\xi + \\ +\epsilon\sum_{k=0}^{K}\int_a^b p_k(\xi)u^{(k)}(\xi)\delta u^{(k)}(\xi)d\xi = 0. \end{aligned} \tag{7.28}$$

Интегралы под знаком суммы содержат вариации производных $\delta u^{(k)}(\xi)$. Их надо преобразовать к вариациям самой функции $\delta u(\xi)$. Это делается k-кратным интегрированием по частям:

$$\begin{aligned} \int_a^b p_k(\xi)u^{(k)}(\xi)\delta u^{(k)}(\xi)d\xi = \\ = \sum_{r=0}^{k-1}(-1)^r\delta u^{(k-1-r)}(\xi)\frac{d^r}{d\xi^r}\left[p_k(\xi)u^{(k)}(\xi)\right]\Big|_{\xi=a}^{\xi=b} + \\ +(-1)^k\int_a^b \delta u(\xi)\frac{d^k}{d\xi^k}\left[p_k(\xi)u^{(k)}(\xi)\right]d\xi. \end{aligned} \tag{7.29}$$

Выражение (7.29) содержит вариацию $\delta u(\xi)$ во всех точках отрезка под знаком интеграла, а также сумму с вариациями в краевых точках.

Подставим (7.29) в (7.28). Полученное выражение будет содержать интеграл от $\delta u(\xi)$ и суммы с вариациями граничных условий. Поскольку вариации могут быть произвольными, то стоящие перед ними выражения должны обращаться в нуль. Обращение в нуль выражений, стоящих перед $\delta u(\xi)$, приводит к следующему интегро-дифференциальному уравнению:

$$\epsilon \sum_{k=0}^{K} (-1)^k \frac{d^k}{d\xi^k} \left[p_k(\xi) u^{(k)}(\xi) \right] + \int_a^b Q(\xi, \eta) u(\eta) d\eta = \Phi(\xi), \quad a \le \xi \le b, \tag{7.30}$$

с ядром и правой частью

$$Q(\xi, \eta) = \int_c^d K(x, \xi) K(x, \eta) dx, \ \Phi(\xi) = \int_c^d K(x, \xi) f(x) dx. \tag{7.31}$$

Старшая производная в дифференциальном операторе (7.30) имеет порядок $2K$, а новое интегральное ядро $Q(\xi, \eta)$ симметрично и определено на квадрате $[a \le \xi \le b,\ a \le \eta \le b]$.

Заметим, что новая правая часть $\Phi(\xi)$ получена сверткой $f(x)$ с ядром $K(x, \xi)$. Аналогичная операция использовалась ранее для повышения гладкости решения уравнения Фредгольма второго рода (7.16). Отсюда видно, что гладкость $\Phi(\xi)$ выше, чем гладкость $f(x)$. Следовательно, решение интегро-дифференциального уравнения (7.30) ищется для класса более гладких правых частей, чем решение исходной задачи (7.19). В частности, $\Phi(\xi)$ будет непрерывно, если ядро $K(x, \xi)$ непрерывно.

Граничные условия для вариационного уравнения Эйлера (7.30) получаем, приравнивая нулю множители перед вариациями $\delta u^{(k)}(\xi)$ в граничных точках $\xi = a$ и $\xi = b$ в выражении (7.29). Получим по K условий на каждой границе:

$$q_r [u(a)] = 0, \ q_r [u(b)] = 0, \ 1 \le r \le K,$$
$$q_r [u(\xi)] = \sum_{k=r}^{K} (-1)^k \frac{d^{k-r}}{d\xi^{k-r}} \left(p_k(\xi) u^{(k)}(\xi) \right). \tag{7.32}$$

Полное число граничных условий равно порядку дифференциального оператора в (7.30).

О сходимости. Регуляризованное уравнение (7.30) содержит параметр ϵ, поэтому будем обозначать его решение через $u(\xi;\epsilon)$. Это уравнение содержит дифференциальный оператор порядка $2K$, а правая часть $\Phi(\xi)$ непрерывна. Поэтому регуляризованное решение $u(\xi;\epsilon)$ имеет $2K$ непрерывных производных по ξ. Очевидно, регуляризованное решение отличается от нерегуляризованного решения $u(\xi)$ уравнения (7.19). Справедливы следующие теоремы.

Теорема 7.1. Задача (7.30) — (7.32) корректно поставлена при любом $\epsilon > 0$. •

Теорема 7.2. Пусть для некоторого $f(x)$ существует решение $u(\xi)$ нерегуляризованной задачи (7.19). Тогда при $K = 1$ решение $u(\xi;\epsilon)$ регуляризованной задачи (7.30) — (7.32) при положительном $\epsilon \to 0$ сходится к $u(\xi)$, причем равномерно (т. е. в $\|\|_C$). •

Доказательства этих теорем достаточно громоздки. Однако наглядно пояснить действие регуляризации можно следующим образом. Примем для простоты $p_k(\xi) \equiv \mathtt{const}$. Некорректность связана с возмущениями, имеющими вид высокочастотных гармоник. Пусть правая часть $\Phi(\xi)$ получила возмущение $\beta\exp(\mathrm{i}\omega\xi)$ с $\omega \gg 1$. Тогда решение получит возмущение $\delta u(\xi;\epsilon) = \alpha\exp(\mathrm{i}\omega\xi)$. Подставим эти возмущения в (7.30). Каждое дифференцирование приводит к умножению на ω, а интегрирование — к делению на ω (последнее пояснено в п. 7.2.1). Ограничиваясь только порядками членов по ω и не конкретизируя стоящие при них множители, получим следующее оценочное соотношение:

$$\left(\frac{1}{\omega} + \epsilon\sum_{k=0}^{K} p_k\omega^{2k}\right)\alpha \sim \beta. \tag{7.33}$$

Рассмотрим поведение возмущений, предполагая $\omega \gg 1$.

Если $\epsilon = 0$, то $\alpha \sim \omega\beta \gg \beta$. Возмущения решения велики, и расчет неустойчив, т. е. регуляризации нет (см. п. 7.2.1.).

Если $\epsilon > 0$, а $K = 0$, то $\alpha \sim \beta/\epsilon$. При любом фиксированном $\epsilon > 0$ возмущения решения по порядку величины равны возмущениям правой части и расчет становится устойчивым. Чем больше ϵ, тем меньше возмущения решения и «разболтка» в численном расчете. Но сдвиги фаз отдельных гармоник приводят к

тому, что сходимость будет только среднеквадратичной (т. е. в $\|\|\|_{L_2}$). Такую регуляризацию называют ***слабой***.

Если $K = 1$, то $\alpha \sim \beta/\epsilon\omega^2 \ll \beta$. Возмущения высоких частот оказываются малыми. Устойчивость расчета хорошая, и $u(\xi;\epsilon)$ равномерно сходится к $u(\xi)$ (т. е. в $\|\|\|_C$). Такую регуляризацию называют ***сильной***.

Если $K > 1$, то амплитуды $\alpha \sim \beta\omega^{-2K}$ настолько быстро убывают при $\omega \to \infty$, что обеспечивается равномерная сходимость не только регуляризованного решения, но и его $(k-1)$-й производной.

Выбор параметров. Регуляризованная задача содержит ряд параметров: порядок регуляризации K, параметр регуляризации ϵ, весовые множители $p_k(\xi)$ для $0 \le k \le K$. Приведем некоторые рекомендации по их выбору.

При порядке регуляризации $K = 0$ амплитуды высокочастотных возмущений не подавляются, а фазы различных гармоник сдвигаются на неодинаковые величины; это зачастую приводит к появлению заметных немонотонностей в регуляризованном решении. Они напоминают разболтку, возникающую в расчетах уравнений в частных производных по немонотонным разностным схемам. Поэтому значения $K = 0$ обычно считают неудовлетворительными. Наоборот, при $K > 1$ высокие гармоники подавляются слишком сильно и регуляризованное решение плохо передает детальные особенности точного решения, типа узких экстремумов (они оказываются сильно сглаженными). Поэтому на практике чаще всего выбирают $K = 1$.

Весовые множители $p_k(\xi)$ целесообразно выбирать с учетом имеющихся сведений о функции $f(x)$. Пусть известна оценка погрешности $\delta f(x)$. Рассчитаем, какова будет соответствующая погрешность $\delta\Phi(\xi)$, и зададим $p_k(\xi)$ большими в тех участках, где $\delta\Phi(\xi)$ велика. Если данных о погрешности $f(x)$ нет или оценки $\delta\Phi(\xi)$ сделать затруднительно, то полагают $p_k(\xi) \equiv 1$.

Оптимальный выбор параметра ϵ является задачей с противоречивыми критериями. Чем больше ϵ, тем более гладкий вид имеет регуляризованное решение $u(\xi;\epsilon)$. Таким образом, увеличение ϵ уменьшает разболтку, обусловленную некорректностью исходной задачи. Однако при этом одновременно увеличивается отличие $u(\xi;\epsilon)$ от точного решения $u(\xi)$. Удовлетворительный результат можно получить, если известна оценка погрешности $\delta f(x)$ в некоторой норме (обычно в $\|\|\|_{L_2}$). В этом случае про-

водят серию расчетов с разными значениями ϵ. Для каждого значения вычисляют невязку регуляризованного решения:

$$\psi\left[u(\xi;\epsilon)\right] = \int_a^b K(x,\xi)u(\xi;\epsilon)d\xi - f(x). \qquad (7.34)$$

Норма невязки будет монотонно возрастающей функцией ϵ. При слишком малых ϵ будет выполняться $\|\psi\| < \|\delta f\|$, а само регуляризованное решение будет «разболтанным» (что нетрудно наблюдать на мониторе). При слишком большом ϵ регуляризованное решение станет достаточно гладким, но будет выполняться $\|\psi\| > \|\delta f\|$. Оптимальным следует считать такое ϵ, при котором $\|\psi\| \approx \|\delta f\|$.

Если удовлетворительных оценок $\delta f(x)$ нет, поступают следующим образом. Первый расчет проводят с заведомо большим значением ϵ. Далее уменьшают значение ϵ в геометрической прогрессии (например, со знаменателем $\sqrt{10}$). При больших ϵ расчетный профиль будет гладким. Наблюдая за расчетами на мониторе, останавливаются в тот момент, когда в регуляризованном решении появляется заметная разболтка. В качестве оптимального выбирают предыдущее значение ϵ, когда разболтка еще мало заметна.

Описанная процедура напоминает обработку экспериментальных наблюдений методом наименьших квадратов (см. кн. 1).

7.2.3. Некоторые приложения

Некорректные задачи встречаются в практике вычислений довольно часто. К ним относятся сглаживание и дифференцирование экспериментально измеренных функций, суммирование рядов Фурье с неточно заданными коэффициентами, решение плохо обусловленных линейных систем, задачи оптимального управления, аналитическое продолжение функций, линейное программирование (оптимальное планирование), обратные задачи теплопроводности и геологической разведки, восстановление переданного сигнала по принятому при наличии искажений аппаратуры и многие другие.

Некоторые из этих задач встречались в предыдущих главах. Покажем, как они регуляризируются вариационным методом. Для определенности ограничимся сильной регуляризацией, полагая в формулах (7.30) — (7.32) $K = 1$.

Сглаживание функций. Пусть функция $f(x)$, $a \le x \le b$, измерена экспериментально и содержит заметную случайную погрешность. Тогда математическая задача имеет вид $u(x) = f(x)$. Ее можно записать в каноническом виде, взяв в качестве ядра дельта-функцию Дирака: $K(x,\xi) = \delta(x-\xi)$. Это ядро симметрично и определено на квадрате $[a \le x \le b; a \le \xi \le b]$. Нетрудно видеть, что тогда $Q(\xi,\eta) = \delta(\xi-\eta)$ и $\Phi(\xi) = f(\xi)$. Тогда регуляризованная задача (7.30)–(7.32) принимает следующий вид:

$$\epsilon\left[\frac{d}{d\xi}\left(p_1\frac{du}{d\xi}\right) - p_0 u(\xi)\right] - u(\xi) + f(\xi) = 0,\ u'(a) = u'(b) = 0. \tag{7.35}$$

Таким образом, сглаженная функция $u(x)$ удовлетворяет линейному обыкновенному дифференциальному уравнению второго порядка, для которого поставлена вторая краевая задача. Методы численного решения этой задачи подробно разобраны в гл. I.

Замечания. 1. Для численного решения уравнения (7.35) нецелесообразно использовать схемы высокого порядка точности, поскольку $f(x)$ может содержать экспериментальную ошибку.

2. Решение уравнения (7.35) имеет непрерывную $u''(\xi)$. Однако уже первая производная $u'(\xi)$ при $\epsilon \to 0$ не сходится в $\|\|_C$ к точному решению, поскольку граничные условия в (7.35) не могут совпадать со значениями f' на границах.

3. Последнее обстоятельство приводит к значительной погрешности регуляризованного решения $u(x)$ вблизи границ отрезка $[a,b]$. Можно уменьшить погрешность сглаживания вблизи концов отрезка $[a,b]$, если использовать регуляризацию более высокого порядка. Однако в п. 7.2.2 отмечалось, что при этом могут чрезмерно сгладиться детали решения типа узких экстремумов.

4. При выборе $p_0(\xi)$ и $p_1(\xi)$ руководствуются соображениями из п. 7.2.2.

Дифференцирование функций. Задачу дифференцирования $u(x) = f'(x)$, $a \le x \le b$, можно записать в виде уравнения Вольтерра первого рода (7.20) с ядром $K(x,\xi) = 1$ на треугольнике $a \le \xi \le x \le b$. Его формально можно считать уравнением Фредгольма первого рода с разрывным ядром на квадрате:

$$\begin{gathered} K(x,\xi) = 1 \text{ при } a \le \xi \le x \le b, \\ K(x,\xi) = 0 \text{ при } \xi > x. \end{gathered} \tag{7.36}$$

Поскольку требование непрерывности ядра не является существенным, применим к этой задаче алгоритм (7.30) — (7.32). Легко получим

$$Q(\xi,\eta) = b - \max(\xi,\eta),$$
$$\Phi(\xi) = \int_{\xi}^{b} [f(x) - f(a)]\,dx. \tag{7.37}$$

Тогда для сильной регуляризации получаем следующее интегродифференциальное уравнение и краевые условия:

$$-\epsilon\left[\frac{d}{d\xi}\left(p_1(\xi)\frac{du}{d\xi}\right) - p_0(\xi)u(\xi)\right] + \int_{a}^{\xi}(b-\xi)\,u(\eta)d\eta + \\ + \int_{\xi}^{b}(b-\eta)\,u(\eta)d\eta = \int_{\xi}^{b}[f(x)-f(a)]\,dx, \tag{7.38}$$
$$u'(a) = 0,\ u'(b) = 0.$$

К этой задаче также относятся все замечания, данные к примеру (7.35).

Суммирование ряда Фурье. Пусть на отрезке $[a,b]$ задана полная ортонормированная система функций $\phi_s(\xi)$, являющаяся системой собственных функций задачи Штурма — Лиувилля:

$$\frac{d}{d\xi}\left[p_1(\xi)\frac{d\phi}{d\xi}\right] - [p_0(\xi) - \lambda]\,\phi(\xi) = 0, \tag{7.39}$$
$$\phi'(a) = 0,\ \phi'(b) = 0,$$

с положительными весами $p_1(\xi)$, $p_0(\xi)$. Требуется просуммировать ряд Фурье

$$f(x) = \sum_{s=1}^{\infty}\beta_s\phi_s(x), \tag{7.40}$$

коэффициенты которого β_s заданы приближенно (например, они вычислены с ошибками округления).

Эту задачу можно рассматривать как сглаживание неточно заданной функции $f(x)$. Воспользуемся для ее решения уравнением (7.35), где в качестве $p_0(\xi)$ и $p_1(\xi)$ выбраны веса, входящие

в задачу Штурма—Лиувилля (7.39). Будем искать регуляризованное решение также в виде ряда Фурье:

$$u(\xi) = \sum_{s=1}^{\infty} \alpha_s \phi_s(\xi). \tag{7.41}$$

Подставляя (7.41) и (7.40) в (7.35) и учитывая (7.39), получим

$$\alpha_s = \frac{\beta_s}{1 + \epsilon \lambda_s}, \tag{7.42}$$

где $\lambda_s > 0$ — собственные значения задачи Штурма—Лиувилля (7.39). Этот способ, в отличие от обрезания ряда Фурье по некоторому числу членов, позволяет не думать о том, с какой точностью были вычислены β_s.

Заметим, что почленное дифференцирование регуляризованного ряда Фурье можно рассматривать как некоторый способ решения задачи о регуляризации дифференцирования функции.

Простейший частный случай регуляризации получим, если положим $p_1(\xi) \equiv 1$ и $p_0(\xi) \equiv 0$. Тогда задача Штурма—Лиувилля (7.39) превращается в обычное гармоническое уравнение. Для граничных условий (7.39) решениями этой задачи будут только косинусы:

$$\phi_s(\xi) = \cos\left[\pi s\,(\xi - a)\,/\,(b-a)\right],\ \lambda_s = \pi^2 s^2 / (b-a)^2 = O\left(s^2\right). \tag{7.43}$$

Достаточно быстрое увеличение λ_s при возрастании s приводит к соответствующему подавлению амплитуд высших гармоник по формуле (7.42). Такое подавление высокочастотных шумов обеспечивает хорошую регуляризацию.

Плохо обусловленные системы. Рассмотрим плохо обусловленную систему линейных алгебраических уравнений $A\mathbf{u}=\mathbf{f}$, где A — квадратная матрица, а $\mathbf{u}$ и $\mathbf{f}$ — конечномерные векторы. Ее можно регуляризировать, записывая непосредственно в вариационной форме (7.27) и выбирая $K = 0$:

$$\|A\mathbf{u} - \mathbf{f}\|^2 + \epsilon \|\mathbf{u}\|^2 = \min,\ \|\mathbf{u}\|^2 = (\mathbf{u}, \mathbf{u}). \tag{7.44}$$

Формально $K = 0$ соответствует слабой регуляризации. Но в конечномерном пространстве все нормы эквивалентны. Поэтому сходимость регуляризованного решения к точному при $\epsilon \to 0$ является равномерной.

Уравнение (7.44) означает, что среди решений, приближенно удовлетворяющих исходной задаче, ищут вектор наименьшей длины. Часто рассматривают более общую постановку:

$$\|A\mathbf{u} - \mathbf{f}\|^2 + \epsilon \|\mathbf{u} - \mathbf{u_0}\|^2 = \min. \tag{7.45}$$

Она определяет приближенное решение, наименее отличающееся от заданного вектора $\mathbf{u_0}$. Его называют ***нормальным*** решением. Постановку (7.45) используют, например, в задачах линейного программирования.

Постановка (7.45) является задачей на минимум квадратичной формы. Она сводится к решению линейной алгебраической системы

$$\left(A^H A + \epsilon E\right) \mathbf{u} = A^H \mathbf{f} + \epsilon \mathbf{u}_0. \tag{7.46}$$

Благодаря слагаемому ϵE эта система хорошо обусловлена, по крайней мере, при не слишком малых $\epsilon > 0$. Поэтому ее нетрудно решить методом исключения Гаусса.

Описанный алгоритм применяют также для решения систем с вырожденной матрицей A.

7.2.4. Разностные схемы

При вариационном методе регуляризации численно решать приходится либо задачу на минимум функционала (7.27), либо краевую задачу для интегро-дифференциального уравнения Эйлера (7.30)—(7.32). К этим задачам целесообразно применять разностные методы. Дадим пример построения разностной схемы исходя из интегро-дифференциальной системы (7.30)—(7.32).

Чтобы формулы были не слишком громоздкими, сделаем упрощающее предположение. Сетки возьмем равномерные. При этом сетки по переменным ξ и η должны быть совпадающими с одинаковым шагом γ, а сетка по x может иметь другой шаг h:

$$\begin{gathered} \xi_n = \eta_n = a + \gamma_n,\ 0 \le n \le N,\ \gamma = (b-a)/N; \\ x_j = c + h_j,\ 0 \le j \le J,\ h = (d-c)/J. \end{gathered} \tag{7.47}$$

Воспользуемся квадратурными формулами трапеций, поскольку они легко позволяют применить метод Ричардсона с двукратным сгущением сеток. Для реализации второго порядка точности формул трапеций надо потребовать, чтобы ядро $K(x, \xi)$ имело вторые непрерывные производные по каждому аргументу;

заметим, что наличие смешанной второй производной не требуется. При рекуррентном применении метода Ричардсона на последовательности s сгущающихся вдвое сеток надо требовать наличия $2s$ непрерывных производных по каждому аргументу; существования смешанных производных также не требуется.

Ограничимся случаем сильной регуляризации ($K = 1$) с постоянными весовыми множителями $p_0(\xi) = p_1(\xi) = 1$. Для передачи первой и второй производных в граничных точках введем значения u_{-1} и u_{N+1}. Сохраняя для сеточных функций те же обозначения, что и в интегро-дифференциальном уравнении, получим следующую схему:

$$-\frac{\epsilon}{\gamma^2}\left(u_{n-1} - 2u_n + u_{n+1}\right) + \epsilon u_n + \\ +\gamma\left(\frac{1}{2}Q_{n0}u_0 + \sum_{m=1}^{N-1} Q_{nm}u_m + \frac{1}{2}Q_{nN}u_N\right) = \Phi_n,\ 0 \le n \le N; \tag{7.48}$$

ядро и правая часть определяются формулами

$$Q_{nm} = h\left(\frac{1}{2}K_{0n}K_{0m} + \sum_{j=1}^{J-1} K_{jn}K_{jm} + \frac{1}{2}K_{Jn}K_{Jm}\right), \\ K_{jn} = K(x_j, \xi_n); \tag{7.49}$$
$$\Phi_n = h\left(\frac{1}{2}K_{0n}f_0 + \sum_{j=1}^{J-1} K_{jn}f_j + \frac{1}{2}K_{Jn}f_J\right),\ f_j = f(x_j).$$

Краевые условия принимают следующий вид:

$$u_{-1} = u_1,\ u_{N-1} = u_{N+1}. \tag{7.50}$$

Краевые условия позволяют исключить u_{-1} и u_{N+1} из первого и последнего уравнений системы (7.48). Тогда система (7.48) будет линейной алгебраической системой порядка $N+1$ относительно такого же количества неизвестных u_n, $0 \le n \le N$. Матрица этой системы является плотно заполненной, так что систему целесообразно решать методом исключения Гаусса.

Матрица системы (7.48) несимметрична из-за членов с коэффициентами $1/2$ в скобках. Однако если умножить уравнения с $n = 0$ и $n = N$ на $1/2$, то матрица становится симметричной.

Вдобавок, она положительна. В этом случае в методе исключения Гаусса не нужно выбирать главный элемент: он всегда лежит на главной диагонали.

Остановимся на сходимости схемы. При сделанных предположениях схема имеет аппроксимацию $O\left(h^2+\gamma^2\right)$. При сгущениях сеток мы визуально или программно проверяем, сходится ли сеточное решение к предельной функции при $\gamma \to 0$, $h \to 0$. Если сходимость к пределу наблюдается, то схема устойчива. Тогда из теоремы 2.7 следует сходимость разностного решения к точному.

Сгущение сеток по ξ, η и x надо проводить одновременно в одно и то же число раз (вдвое). Погрешность аппроксимации симметричных квадратур разлагается в ряд по степеням γ^2 и h^2. Поэтому в рекуррентном методе Ричардсона каждое сгущение повышает порядок точности на 2.

В прикладных расчетах $f(x)$ может оказаться экспериментальной функцией, измеряемой на некоторой неравномерной сетке $\{x_j\}$. В этом случае ядро и правая часть (7.49) следует вычислять по формуле трапеций для неравномерной сетки (см. кн. 1). Однако сетки по ξ и η можно по-прежнему выбирать равномерными. При этом в методе Ричардсона можно проводить только сгущение сеток по ξ и η, но не по x. Поэтому при сгущениях будет стремиться к нулю та часть погрешности, которая связана с интегрированием по η; погрешность интегрирования по x устранить уже невозможно.

Сглаживание функции. Эта задача была рассмотрена в п. 7.2.3. В ней ядро является дельта-функцией, так что нельзя применять схему (7.48) — (7.50), рассчитанную на гладкие ядра. Следует воспользоваться непосредственно уравнением (7.35). Заменяя производные простейшими разностями на равномерной сетке, получим следующую разностную схему:

$$\frac{\epsilon}{\gamma^2}u_{n-1} - \left(\frac{2\epsilon}{\gamma^2}+\epsilon+1\right)u_n + \frac{\epsilon}{\gamma^2}u_{n+1} = -f_n,\ 0 \le n \le N; \qquad (7.51)$$

$$u_{-1} = u_1,\ u_{N+1} = u_{N-1}.$$

Для нахождения u_n получилась линейная алгебраическая система с трехдиагональной матрицей. Она легко решается методом Гаусса для ленточной матрицы или прогонкой.

Дифференцирование функции сводится к задаче с разрывным ядром $K(x,\xi)$ (см. п. 7.2.3). Здесь также нельзя пользоваться разностными схемами (7.48) — (7.50), рассчитанными на

гладкие ядра. Воспользуемся специальными формулами (7.38), аппроксимируя квадратуры формулами трапеций, а производные — простейшими разностями на равномерных сетках. Получим следующую систему уравнений:

$$-\epsilon\left[\frac{1}{\gamma^2}\left(u_{n-1}-2u_n+u_{n+1}\right)-u_n\right]+$$

$$+\gamma\left[(b-\xi_n)\left(\frac{1}{2}u_0+\sum_{m=1}^{n}u_m\right)+\sum_{m=n+1}^{N-1}(b-\xi_m)\,u_m\right]=$$

$$=\gamma\left[\frac{1}{2}\left(f_n-f_0\right)+\sum_{m=n+1}^{N-1}\left(f_m-f_0\right)+\frac{1}{2}\left(f_N-f_0\right)\right],\ 0\le n\le N;$$

$$u_{-1}=u_1,\ u_{N-1}=u_{N+1}. \tag{7.52}$$

Исключая u_{-1} и u_{N+1} с помощью граничных условий, получим линейную систему порядка $N+1$ с плотно заполненной матрицей.

Заметим, что к задачам сглаживания и дифференцирования функций возможен иной подход. Разложим $f(x)$ в ряд Фурье (см. кн. 1). Проведем регуляризацию полученного ряда по формуле (7.42). Регуляризованный ряд $u(x)$ является сглаживанием функции $f(x)$, а его почленное дифференцирование является регуляризацией численного дифференцирования $f(x)$. Этот способ имеет даже одно преимущество: значения $u(x)$ и $u'(x)$ получаются не в узлах сетки, а в виде явных формул, пригодных для любого x. Этот способ нетрудно применить и в том случае, когда $f(x)$ задана на неравномерной сетке.

СПИСОК ЛИТЕРАТУРЫ

1. *Калиткин Н. Н.* Численные методы : в 2 кн. Кн. 1 / Н. Н. Калиткин, Е. А. Альшина. — М : Издат. центр «Академия», 2012.

2. *Калиткин Н. Н.* Численные методы. — СПб : БХВ-Петербург, 2011.

3. *Самарский А. А.* Численные методы / А. А. Самарский, А. В. Гулин. — М : Наука, 1989.

4. *Самарский А. А.* Введение в численные методы. — М : Наука, 1987.

5. *Калиткин Н. Н.* Вычисления на квазиравномерных сетках / [Н. Н. Калиткин и др.]. — М. : Физматлит, 2005.

6. *Марчук Г. И.* Методы вычислительной математики. — М. : Наука, 1989.

7. *Хайрер Э.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи / Э. Хайрер, С. Нерсетт, Г. Ваннер. — М. : Мир, 1990.

8. *Хайрер Э.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи / Э. Хайрер, Г. Ваннер. — М. : Мир, 1999.

9. *Федоренко Р. П.* Введение в вычислительную физику. — М. : Изд-во МФТИ, 1994.

10. *Рихтмайер Р. Д.* Разностные методы решения краевых задач / Р. Д. Рихтмайер, К. Мортон. — М. : Мир, 1972.

11. *Галанин М. П.* Методы численного анализа математических моделей / М. П. Галанин, Е. Б. Савенков. — М. : Изд-во МГТУ им. Н. Э. Баумана, 2010.

12. *Бахвалов Н. С.* Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г. М. Кобельков. — М. : БИНОМ, Лаборатория знаний, 2004.

13. *Рябенький В. С.* Введение в вычислительную математику. — М. : Наука, 1994.

14. *Плохотников К. Э.* Вычислительные методы. Теория и практика в среде MATLAB, курс лекций. — М. : Горячая линия. — Телеком, 2009.

15. *Плис А. И.* Лабораторный практикум по высшей математике / А. И. Плис, Н. А. Сливина. — М. : Высшая школа, 1994.

16. *Фаддеев Д. К.* Вычислительные методы линейной алгебры / Д. К. Фаддеев, В. Н. Фаддеева. — М. : Физматгиз, 1963.

17. *Самарский А. А.* Разностные методы решения задач газовой динамики / А. А. Самарский, Ю. П. Попов. — М. : Наука, 1992.

18. *Тихонов А. Н.* Методы решения некорректных задач / А. Н. Тихонов, В. Я. Арсенин. — М. : Наука, 1979.

19. *Тихонов А. Н.* Вводные лекции по прикладной математике / А. Н. Тихонов, Д. П. Костомаров. — М. : Наука, 1984.

ОГЛАВЛЕНИЕ

Учебное издание

Калиткин Николай Николаевич,
Корякин Павел Владимирович

ЧИСЛЕННЫЕ МЕТОДЫ

В двух книгах

КНИГА 2

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Учебник

Редактор *Л. В. Честная*
Технический редактор *Н. И. Горбачева*
Компьютерная верстка: *Т. А. Клименко*
Корректор *Г. Н. Петрова*

Изд. № 101114031. Подписано в печать 22.05.2013. Формат 60 × 90/16.
Гарнитура «Ньютон». Бумага офсетная № 1. Печать офсетная. Усл. печ. л. 19,0.
Тираж 1 200 экз. Заказ № 34418.

ООО «Издательский центр «Академия». www.academia-moscow.ru
129085, Москва, пр-т Мира, 101В, стр. 1.
Тел./факс: (495) 648-0507, 616-00-29.
Санитарно-эпидемиологическое заключение № РОСС RU. АЕ51. Н 16476 от 05.04.2013.

Отпечатано в соответствии с качеством предоставленных издательством
электронных носителей в ОАО «Саратовский полиграфкомбинат».
410004, г. Саратов, ул. Чернышевского, 59. www.sarpk.ru

ЧИСЛЕННЫЕ МЕТОДЫ

В двух книгах

Книга 2

МЕТОДЫ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

ISBN 978-5-7695-5091-1

Н. Н. Калиткин
П. В. Корякин

Издательский центр «Академия»
www.academia-moscow.ru

2

ЧИСЛЕННЫЕ МЕТОДЫ