



МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ  
В ТЕХНИКЕ И В ТЕХНОЛОГИИ

М.П. Галанин, Е.Б. Савенков

# Методы численного анализа математических моделей

Издательство МГТУ  
им. Н.Э. Баумана

М.П. Галанин, Е.Б. Савенков

# Методы численного анализа математических моделей



Москва 2010

*Рецензенты: член-кор. РАН М.А. Гузев;  
проф. А.В. Гулин*

**Галанин М. П.**

- Г15      Методы численного анализа математических моделей / М. П. Галанин, Е. Б. Савенков. – М. : Изд-во МГТУ им. Н. Э. Баумана, 2010. – 591, [1] с. : ил. (Математическое моделирование в технике и в технологии).

ISBN 978-5-7038-3252-3

Книга отражает современный уровень развития численных методов и алгоритмов, ориентированных на применение современной вычислительной техники и позволяющих проводить количественный анализ математических моделей широкого класса реальных природных, социальных и технических объектов.

Изложены методы решения задач линейной алгебры, систем нелинейных алгебраических уравнений, интерполяция функций, методы численного интегрирования и дифференцирования, численные методы решения задач Коши и краевых задач для систем обыкновенных дифференциальных уравнений. Приведены основы общей теории разностных схем и ее применение к построению и анализу методов численного решения эллиптических, параболических и гиперболических уравнений, а также численные методы решения интегральных уравнений. Представлены методы генерации сеток для многомерных задач математической физики, многосеточные методы решения, численные методы для решения уравнения переноса и уравнений газовой динамики, алгоритмические основы метода конечных элементов.

Для студентов старших курсов технических университетов, аспирантов и инженеров. Может быть полезна преподавателям и научным работникам.

УДК 519.6  
ББК 22.193

ISBN 978-5-7038-3252-3

© Галанин М. П., Савенков Е. Б., 2010  
© Оформление. Издательство  
МГТУ им. Н. Э. Баумана, 2010

**180-летию МГТУ  
им. Н.Э. Баумана  
посвящается**

## **ПРЕДИСЛОВИЕ**

В настоящее время методы численного анализа широко применяются в самых разнообразных областях научной и технической деятельности.

Данная книга выходит в серии «Математическое моделирование в технике и в технологии». В ней приведены фундаментальные сведения о методах и приемах численного моделирования, а также рассмотрены области применения прикладной вычислительной науки и пути ее развития.

Книга состоит из двух частей: первая часть представляет интерес для начинающих изучать методику численного моделирования и технологию вычислительного эксперимента, вторая — полезна для опытных специалистов.

Нумерация глав по всей книге сквозная. Параграфы имеют двойную нумерацию (например, 2.1 — первый параграф в главе 2). Ссылки в тексте на параграфы и главы набраны полужирным шрифтом (например, см. 2.1 или см. 2). Аналогично пронумерованы формулы, рисунки, определения, леммы и теоремы (например, (2.3) — третья формула в главе 2, рис. 3.1 — первый рисунок в главе 3). В квадратные скобки заключены номера библиографических источников из помещенного в конце книги списка литературы.

Каждый термин, выделенный в тексте *полужирным курсивом*, представлен в предметном указателе (в алфавитном порядке по существительному в именительном падеже) с указанием страницы, на которой он определен или описан. В начале каждой главы и параграфа *светлым курсивом* выделены термины, отнесенные к ключевым словам, т. е. для понимания излагаемого материала читатель должен знать значение этих терминов. Читатель может уточнить это значение, найдя с помощью предметного указателя необходимую страницу.

После предисловия помещен список основных обозначений, где наряду с краткой расшифровкой указаны параграфы, в которых можно найти более подробное их объяснение. Принятая структура справочного аппарата книги позволяет читателю знакомиться с интересующим его материалом отдельно взятого параграфа.

Надеемся, что данная книга будет интересна как студентам и аспирантам, так и опытным исследователям.

Авторы благодарны главному редактору серии «Математическое моделирование в технике и в технологиях» И.Б. Федорову и членам редакционного совета за предоставленную возможность издания книги, а также рецензентам за благосклонное отношение к нашему труду.

# ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

- $x \in X$  — элемент  $x$  принадлежит множеству  $X$
- $\forall$  — квантор всеобщности
- $\exists$  — квантор существования
- $n = \overline{1, N}$ ,  $n = 1, 2, \dots, N$  — число  $n \in \mathbb{N}$  принимает последовательно все значения из множества  $\mathbb{N}$  натуральных чисел от 1 до  $N$  включительно
- $X \setminus Y$  — разность множеств  $X$  и  $Y$
- $\cup$  — символ операции объединения множеств
- $\cap$  — символ операции пересечения множеств
- $X \subset Y$  — подмножество  $X$  включено в множество  $Y$  ( $Y$  включает  $X$ )
- $X \subseteq Y$  — подмножество  $X$  включено в множество  $Y$  или совпадает с ним
- $|\cdot|$  — абсолютное значение числа или модуль вектора
- $(\cdot)^T$  — символ транспонирования матрицы
- $\nabla$  — градиент вектора или функции (дифференциальный оператор Гамильтона)
- $\Delta, \nabla^2$  — дифференциальный оператор Лапласа
- $\Delta^2$  — бигармонический дифференциальный оператор
- $\text{grad} = \nabla$  — символ дифференциальной операции вычисления градиента
- $\text{div} = \nabla \cdot$  — символ дифференциальной операции вычисления дивергенции
- $A : X \rightarrow Y$  — оператор  $A$  отображает множество  $X$  на (или в) множество  $Y$
- $A : x \mapsto y$  — оператор  $A$  отображает элемент  $x$  в элемент  $y$
- $X \times Y$  — декартово произведение множеств  $X$  и  $Y$
- $D(A)$  — область определения оператора  $A$  **1.1.2**
- $\text{im}(A)$  — область значений оператора  $A$  **1.1.2**
- $\emptyset$  — пустое множество

$\mathbb{R}$  — множество действительных чисел (числовая прямая)

$[a, b], (a, b)$  — отрезок и интервал с концами в точках  $a, b \in \mathbb{R}$

$\inf_{x \in X} f(x)$  — точная нижняя грань множества  $\{f(x), x \in X\} \subset \mathbb{R}$  числовой оси

$\sup_{x \in X} f(x)$  — точная верхняя грань множества  $\{f(x), x \in X\} \subset \mathbb{R}$  числовой оси

$a^*, y^*$  — приближенное значение величины  $a$  и приближенное значение функции  $y^* = y(x^*)$  **B.3.1, B.3.3**

$\Delta(a^*), \Delta(y^*)$  — абсолютная погрешность приближенного значения величины  $a$  и линейная абсолютная оценка погрешности функции  $y$  **B.3.1, B.3.3**

$\delta(a^*)$  — относительная погрешность приближенного значения величины  $a$  **B.3.1**

$X \oplus Y$  — прямая сумма подпространств  $X$  и  $Y$  **1.1.1**

$\|x\|, \|A\|$  — норма элемента  $x$  некоторого линейного нормированного пространства и норма оператора либо матрицы  $A$  **1.1.1, 1.1.2**

$\|x\|_V$  — норма элемента  $x$  линейного нормированного пространства  $V$  **1.1.1**

$\|x\|_A = (Ax, x)^{1/2}$  — энергетическая норма ( $A$ -норма) элемента  $x$  линейного нормированного пространства, соответствующая симметричному положительно определенному оператору  $A$  **1.1.3, 16.3.1**

$\|x\|_\infty, \|A\|_\infty$  — кубическая норма вектора  $x \in \mathbb{R}^n$  и матрицы  $A = A_{n \times n}$  **1.1.5**

$\|x\|_1, \|A\|_1$  — октаэдрическая норма (1-норма) вектора  $x \in \mathbb{R}^n$  и матрицы  $A = A_{n \times n}$  **1.1.5**

$\|x\|_2, \|A\|_2$  — евклидова (сферическая, шаровая) норма (2-норма) вектора  $x \in \mathbb{R}^n$  и матрицы  $A = A_{n \times n}$  **1.1.5**

$\|A\|_s$  — спектральная норма матрицы  $A$  **1.1.5**

$\|A\|_M$  — максимальная норма матрицы  $A$  **1.1.5**

$(x, y)$  — скалярное произведение элементов  $x$  и  $y$  унитарного (евклидова) пространства **1.1.1**

$(x, y)_A = (Ax, y)$  — энергетическое скалярное произведение элементов  $x$  и  $y$  унитарного (евклидова) пространства, соответствующее самосопряженному положительно определенному оператору  $A$  **1.1.3**

$x \perp y, x \perp X$  — элемент  $x$  унитарного (евклидова) пространства ортогонален элементу  $y$  и элемент  $x$  унитарного (евклидова) пространства ортогонален подпространству  $X$  **1.1.1**

$\ker A$	— ядро оператора $A$ <b>1.1.2</b>
$\rho(A)$	— спектральный радиус оператора $A$ <b>1.1.2</b>
$\bar{\rho}(A)$	— числовой радиус оператора $A$ <b>1.1.2</b>
$A^*$	— оператор, сопряженный оператору $A$ <b>1.1.3</b>
$P_n(x)$	— алгебраический полином степени $n$ <b>3.2</b>
$\tilde{f}$	— тот или иной интерполянт заданной функции $f$ <b>3.1</b>
$L_n(x)$	— алгебраический интерполяционный полином степени $n$ <b>3.2.1</b>
$H_n(x)$	— алгебраический интерполяционный полином Эрмита степени $n$ <b>3.2.3</b>
$T_n(x)$	— полином Тейлора степени $n$ <b>3.2.3</b>
$Q_n(x)$	— тригонометрический полином порядка $n$ <b>3.3.5</b>
$S_n(x)$	— интерполяционный сплайн степени $n$ <b>3.4</b>
$I_h$	— квадратурная формула для вычисления интеграла $I$ <b>4.1</b>
$y_{\bar{x}}$	— разностная производная сеточной функции назад (левая) <b>4.7</b>
$y_x$	— разностная производная сеточной функции вперед (правая) <b>4.7</b>
$y_x^o$	— центральная разностная производная сеточной функции <b>4.7</b>
$y_{\bar{xx}}$	— вторая разностная производная сеточной функции <b>4.7</b>
$\Phi[u]$	— значение функционала $\Phi$ на элементе $u$ <b>6.3.1</b>
$y_i = y(x_i)$	— значение сеточной функции $y$ в пространственной точке $x_i$ <b>7.1.2</b>
$y^j = y(t_j)$	— значение сеточной функции $y$ во временной точке $t_j$ <b>7.1.2</b>
$y_i^j = y(x_i, t_j)$	— значение сеточной функции $y$ в точке $(x_i, t_j)$ <b>7.1.2</b>
$\Omega_h$	— сеточная область, соответствующая пространственно-временной области $\Omega$ <b>7.1.2</b>
$A_h$	— разностная аппроксимация оператора $A$ <b>7.1.2</b>
$p_h$	— оператор проектирования на сетку <b>7.1.2</b>
$u_h = p_h u$	— проекция функции $u$ на сетку <b>7.1.2</b>
$\hat{y}$	— значения сеточной функции $y$ на следующем временном слое: $\hat{y}(x_i) = y(x_i, t_{j+1})$ <b>7.2</b>

- $\check{y}$  — значения сеточной функции  $y$  на предыдущем временном слое:  $\check{y}(x_i) = y(x_i, t_{j-1})$  **7.2**
- $\psi_h$  — погрешность аппроксимации разностной схемы и численного алгоритма **7.2**
- $h_i = x_{i+1} - x_i$  — расстояние (шаг сетки) между узлами  $x_{i+1}$  и  $x_i$  **7.3.2**
- $x_{i+1/2} = (x_i + x_{i+1})/2$  — координаты грани разностных ячеек с центрами в узлах  $x_{i+1}$  и  $x_i$  **7.3.2**
- $\check{h}_i = (x_{i+1/2} - x_{i-1/2})/2$  — расстояние между гранями ячейки разностной сетки с центром в узле  $x_i$  **7.3.2**
- $y^{(\sigma)}$  — взвешенное с весом  $\sigma$  по двум временным слоям значение сеточной функции  $y$ ,  $y^{(\sigma)} = \sigma\hat{y} + (1 - \sigma)y$  **8.1**
- # — признак окончания примера или доказательства **B.2**

# **ВВЕДЕНИЕ**

Представлены материалы по истории прикладной математики и по технологии вычислительного эксперимента. Рассмотрены источники ошибок при проведении вычислений, особенности машинной арифметики и ее результаты. Вычислены ошибки арифметических операций. Приведены примеры устойчивых и неустойчивых алгоритмов. Даны библиографические ссылки.

## **В.1. Предмет и содержание книги**

Решение и исследование подавляющего большинства современных задач науки и техники невозможно представить без математического моделирования, под которым обычно понимают набор подходов и методов для построения, исследования и анализа математических моделей тех или иных физических либо технических объектов или процессов.

Математическая модель при этом в большинстве случаев представляет собой систему уравнений в частных производных, отражающую существенные для рассматриваемого приложения особенности поведения исходного исследуемого объекта. Система уравнений должна быть дополнена указанием области, в которой разыскивается решение, а также необходимыми начальными и граничными данными, обеспечивающими единственность решения.

Непосредственное аналитическое исследование таких моделей возможно лишь в крайне ограниченном ряде случаев. При этом чаще всего оно позволяет получать лишь качественные особенности поведения объекта, знания о которых важны, но недостаточны. Это касается многих инженерных и технических приложений, в которых получение точных количественных результатов является одной из самых важных задач.

Таким образом, при становлении и развитии математического моделирования появилась потребность в численных методах, применение которых позволяло получать достаточно точное приближенное численное решение, поскольку даже хорошо, с теоретической точки зрения, понятные модели редко давали количественный или численный результат без дополнительных усилий. Дальнейшее развитие математического моделирования, связанное, прежде всего, со все возрастающей степенью подробности математических моделей, вызванной запросами научной и инженерной практики, привело к необходимости решать многомерные

нестационарные уравнения с переменными коэффициентами, а также нелинейные задачи. Как правило, такие задачи не имеют простых аналитических решений (а зачастую не имеют аналитических решений вообще). Такие решения известны лишь для весьма ограниченного набора частных случаев.

Для приложений и более простых задач аналитические решения могут быть построены, но получение по этим решениям соответствующих численных результатов настолько неэффективно, что не может быть применено на практике несмотря на уровень развития современной вычислительной техники. В качестве примера такой ситуации можно привести задачу нахождения решения системы линейных алгебраических уравнений (СЛАУ), которое всегда может быть выписано в конечном и явном виде, например, с помощью формулы Крамера. Однако получить численное решение для задач типичной размерности, возникающих в инженерной вычислительной практике (от  $\sim 10^4$  и до  $\sim 10^7$  уравнений), по формуле Крамера с точки зрения временных затрат невозможно. В итоге снова возникает проблема построения качественных вычислительных алгоритмов для решения тех или иных задач.

Такое положение вызвало развитие современных методов численного моделирования и вычислительного эксперимента. Основным содержанием настоящей книги и является изложение таких методов.

Отметим, что в книге не рассматриваются вопросы, связанные с самим построением математических моделей исходя из технической либо физической постановки задачи. Считается, что к тому моменту, когда инженер-исследователь или математик-прикладник приступает к разработке или выбору численного метода для решения имеющейся задачи, сама модель, т. е. система соответствующих уравнений с заданными условиями, ему уже известна. На практике вычислитель обычно получает такую модель самостоятельно или от своих коллег — физиков, механиков, инженеров или других специалистов в данной предметной области.

Тем не менее для выбора эффективного численного алгоритма и, в конечном итоге, успешного решения задачи, вычислитель должен знать прямую и обратную связь между моделью и имеющимися в его распоряжении численными алгоритмами и вычислительными ресурсами. Другими словами, успех решения задачи зависит не только от эффективного численного алгоритма, но и от корректного и тщательно обдуманного выбора математической модели. Авторы надеются, что данная книга поможет исследователю оценивать используемые модели

не только с точки зрения их детализации, но и с позиций их дальнейшего численного исследования.

Книга состоит из двух частей.

Первая часть — «Теоретические основы численных методов» — представляет собой переработанный и расширенный вариант лекций по курсу «Методы вычислений», которые авторы читают студентам третьего курса факультета «Фундаментальные науки» МГТУ им. Н.Э. Баумана.

Курс «Методы вычислений» занимает одно из основных мест в цикле естественно-научных дисциплин, определяющих уровень профессиональной подготовки специалистов в области прикладной математики.

Содержанием курса являются изучение методов решения задач линейной алгебры, нахождения решения систем линейных и нелинейных алгебраических уравнений, методов интерполяции функций одного и нескольких пространственных переменных, численного интегрирования и дифференцирования, методов численного решения задач Коши и краевых задач для обыкновенных дифференциальных уравнений (ОДУ), изучение элементов теории разностных схем и метода конечных элементов и их применение к решению параболических, гиперболических и эллиптических уравнений, изучение методов решения интегральных уравнений.

Данный курс основан на знании ряда общематематических дисциплин: математический анализ, линейная алгебра, уравнения математической физики, функциональный анализ, а также прикладных дисциплин: математические модели в механике сплошной среды и др. Наиболее полно материал по этим дисциплинам представлен в серии книг «Математика в техническом университете», состоящей из 21 тома.

Цель курса «Методы вычислений» — установление общей фундаментальной базы для дальнейшего изучения прикладных математических дисциплин, связанных с численным решением задач прикладной математики (конечно-разностные методы, методы конечных элементов, методы граничных элементов и другие методы нахождения приближенного решения краевых задач, численное решение задач теории упругости и пластичности, газо- и гидродинамики, других разделов механики сплошной среды и физики), а также для самостоятельного решения прикладных задач. Общность и неразрывная связь перечисленных дисциплин заставляют рассматривать их как единое целое. Особенностью курса является обучение строгости математических формулировок изучаемого материала с развитием у студентов необходимых для инженера-математика навыков решения конкретных прикладных задач.

Курс основан на материале учебников и монографий по численным методам, существующим в настоящее время, и на личном опыте авторов. С методической точки зрения материал данной книги подобран так, чтобы читатель мог освоить программу курса без обращения к каким-либо иным пособиям. Однако это не означает, что авторы исчерпали всю тематику методов вычислений. Библиография, приведенная в конце книги, дает представление об объеме материала, не вошедшего в книгу. Она включает в себя основные издания по численным методам.

Авторы учились численным методам в основном по лекциям и книгам В.И. Агошкова, В.Б. Андреева, В.Я. Арсенина, Н.Н. Бахвалова, В.В. Воеводина, С.К. Годунова, А.В. Гулина, Ю.Н. Днестровского, Н.Н. Калиткина, Ю.Н. Карамзина, Д.П. Костомарова, Г.И. Марчука, А.Ф. Никифорова, Е.С. Николаева, Ю.П. Попова, Б.Л. Рождественского, В.С. Рябенького, А.А. Самарского, А.Г. Свешникова, А.Н. Тихонова, В.Б. Уварова, Р.П. Федоренко, Н.Н. Яненко. Эти ученые оказали значительное влияние на авторов, что нашло отражение в содержании и стиле книги.

В связи с тем, что материал первой части составлен на основе лекционного курса, в тексте нет прямых библиографических ссылок. Они вынесены в специальные разделы — «Библиографические комментарии», — помещенные в конце каждой главы.

Книга не содержит списка задач или вопросов для самостоятельного изучения. Однако в ней много примеров, демонстрирующих возможные варианты решения задач.

Вторая часть книги — «Избранные вопросы теории и практики численных методов» — содержит материал, который редко включается в учебную литературу и даже в монографии. В этой части книги рассмотрены вопросы, связанные с методами генерации сеток для решения многомерных (в том числе трехмерных) задач, а также численного решения эллиптических уравнений с помощью многосеточных методов. Две главы посвящены численному решению гиперболических уравнений: решению простейшего гиперболического уравнения — уравнения переноса (линейного и квазилинейного) и решению уравнений газовой динамики. В последней главе представлены основные теоретические и алгоритмические положения метода конечных элементов с элементами теории и практики граничных конечных элементов.

Материал второй части основан на результатах работ различных авторов. Так, глава о генерации сеток подготовлена в сотрудничестве с И.А. Щегловым. Материал главы о многосеточных методах представлен С.И. Мартыненко. Результаты численного моделирования уравнения переноса получены совместно с Т.А. Александриковой,

Т.Г. Елениной, С.А. Токаревой. Глава о методах численного решения уравнений газовой динамики написана по материалам О.А. Кузнецова, предоставленным им для публикации.

Во время подготовки этой книги О.А. Кузнецов умер, не успев сделать и крохотной доли того, на что был способен. Авторы надеются, что публикация его результатов послужит делу памяти о замечательном ученом и человеке.

Авторы выражают глубокую благодарность своим коллегам и ученикам за сотрудничество и совместный труд над книгой. Особенную благодарность авторы выражают Ю.П. Попову и Н.А. Тихонову, чьими учениками они являются.

## **В.2. История вычислений**

### **В.2.1. Исторические сведения**

Известно, что математика возникла как прикладная наука. Ее результат — число. Интерес представляли, например, продолжительность паводков и наводнений на Ниле, количество мер зерна, засыпаемого в хранилища, размер пирамиды фараона данной династии. Измерение площадей, ориентация в пространстве (география и навигация) и многое другое требовали знания конкретного числа. По мере углубления знаний об окружающем мире появились абстрактные разделы математики, которые до сих пор черпают постановки задач из природы.

Историю вычислений можно условно разбить на три периода:

- 1) от начала существования человека до Ньютона, когда господствовал прямой счет;
- 2) от Ньютона до Второй мировой войны, когда появились собственно **математические модели**, но считали по ним медленно и мало (в конце этого периода существовали целые цеха расчетчиков на счетных машинках типа «Феликс» или «Мерседес»);
- 3) от Второй мировой войны до наших дней, когда военные задачи заставили резко увеличить объем вычислений и потребовали создания ЭВМ, которые продолжают интенсивно совершенствоваться.

Хорошо известно, что прогресс в развитии персональных компьютеров за последние годы таков, что если бы так развивалось автомобилестроение, то мы бы сейчас имели «Роллс-Ройс» по цене велосипеда. Приведем, в частности, известный закон Гордона Мура, который гла-

сит, что число транзисторов на кристалле удваивается каждые два года. Значит, и возможности компьютеров с центральным процессором на таком кристалле растут соответствующим образом.

### B.2.2. Вычислительный эксперимент

Усложнение современной жизни, рост дороговизны и увеличение опасности современных технических устройств и социальных явлений привели к тому, что появился новый метод теоретического исследования окружающего нас мира. Его называют **вычислительным экспериментом**. В нем исследуют не само исходное явление или **технический объект**, а их **математические модели** (ММ). Результаты изучения ММ должны сформировать новое представление об исходном явлении или объекте. В натурном же эксперименте ведется изучение самого физического (технического, социального) процесса. Схема вычислительного эксперимента приведена на рис. В.1. Ее часто называют схемой модель — алгоритм — программа.

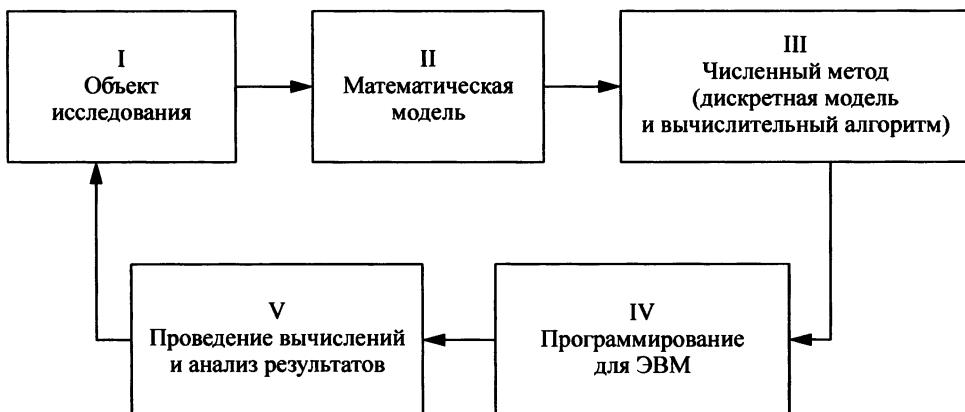


Рис. В.1

**Пример В.1.** Пусть при броске камня с заданной скоростью  $v$  под некоторым углом  $\vartheta$  к горизонту требуется найти такой угол  $\vartheta^*$ , при котором камень улетает на максимальное расстояние  $L$  (рис. В.2). Математическая модель в данном случае имеет следующий вид:  $\dot{x} = v \cos \vartheta$ ,  $\dot{y} = -g$  при  $t > 0$  и  $x = y = 0$ ,  $\dot{y} = v \sin \vartheta$  при  $t = 0$ , где точка над символом означает дифференцирование по времени  $t$ , а  $g$  — ускорение свободного падения. После интегрирования получим

$$x(t) = vt \cos \vartheta, \quad y(t) = -\frac{gt^2}{2} + vt \sin \vartheta.$$

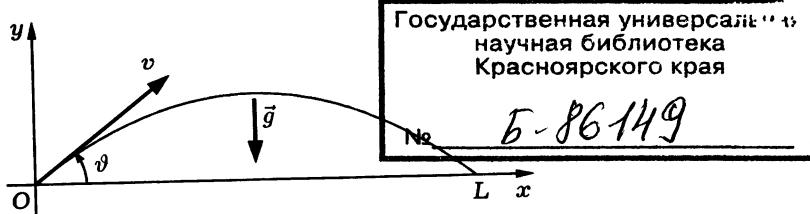


Рис. В.2

В конце полета камня левая часть второго равенства равна нулю, так что время полета  $t_* = 2(v/g)\sin\vartheta$  и соответствующее ему расстояние  $x(t_*, \vartheta) = (v^2/g)\sin 2\vartheta$ . Из условия максимума абсциссы точки падения:

$$\frac{\partial x(t_*, \vartheta)}{\partial \vartheta} \Big|_{\vartheta=\vartheta^*} = 0,$$

получаем  $\vartheta^* = \pi/4$ . Представленное решение имеет смысл при ряде ограничений: масса камня постоянна,  $g = \text{const}$ , нет сопротивления воздуха. Поэтому и ответ удается получить в простой аналитической форме.

Несколько усложним задачу: потребуем, чтобы  $y(t_*) = -h$ , как это показано на рис. В.3.

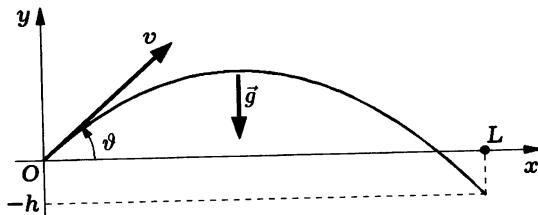


Рис. В.3

Тогда из решения системы уравнений для  $y$  следует  $-h = -gt_*^2/2 + vt_* \sin \vartheta$ , откуда  $t_* = (v/g)\sin \vartheta \pm \sqrt{(v/g)^2 \sin^2 \vartheta + 2h/g}$ . Выбирая положительный корень и подставляя его в решение системы для  $x$ , имеем

$$x(t_*, \vartheta) = \left( \frac{v}{g} \sin \vartheta + \sqrt{\left(\frac{v}{g}\right)^2 \sin^2 \vartheta + 2\frac{h}{g}} \right) v \cos \vartheta$$

и из условия максимума расстояния получаем

$$v \cos 2\vartheta^* + \frac{v^2 \cos 2\vartheta^* - 2gh}{\sqrt{v^2 \sin^2 \vartheta^* + 2gh}} \sin \vartheta^* = 0.$$

Это нелинейное уравнение не имеет аналитического решения в простом виде, поэтому для нахождения значения  $\vartheta^*$  необходимо численными

методами с применением ЭВМ либо решать это уравнение, либо непосредственно искать максимум функции  $x(t_*, \vartheta)$  при выбранном значении  $t_*$ . #

Предметом этой книги является содержание прямоугольника III на рис. В.1, хотя важны все составляющие схемы. Необходимо следить за тем, чтобы не получилось так, как показано на схеме, приведенной на рис. В.4. Здесь представлена кошка в различных видах: как физический объект, как математическая модель, как дискретная модель.

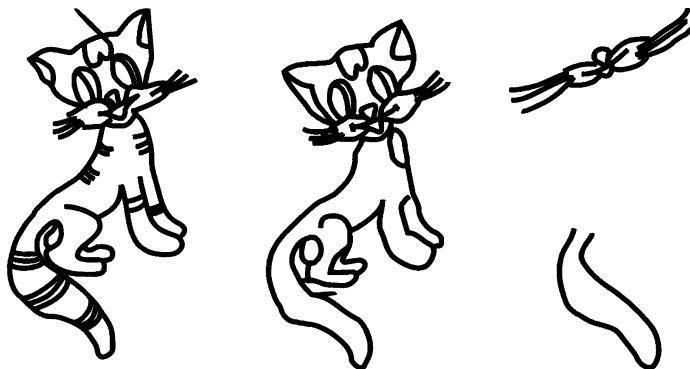


Рис. В.4

На каждом этапе преобразований при переходе от реального объекта к результатам вычислительного эксперимента следует сохранять существенные для рассматриваемой задачи особенности. Только в этом случае есть надежда провести математическое моделирование качественно.

Очевидно, что натурный эксперимент в целом никогда не будет вытеснен вычислительным. В то же время без вычислительного эксперимента анализ математических моделей и соответствующих явлений ныне уже немыслим.

### B.3. Ошибки при вычислениях

Наша задача состоит в определении некоторой величины  $y$  по заданной величине  $x$ . Пусть между ними существует связь  $y = A(x)$ .

Задача выглядит простой и ясной, но в действительности вычислить  $y$ , за исключением тривиальных случаев, невозможно. Причин этому несколько.

1. В реальности входные данные  $x$  измеряются приближенно и никогда (за исключением тривиальной целочисленной информации) точно неизвестны, известно лишь  $\tilde{x}$  — некоторое приближение  $x$ .

2. Ввиду неполноты наших знаний об окружающем мире вместо истинной зависимости известна только некоторая приближенная зависимость  $\tilde{A}$ . В результате наличия двух указанных факторов может быть вычислена лишь величина  $\tilde{y} = \tilde{A}(\tilde{x})$ .

3. При численном решении  $\tilde{A}$  (оператор или функция) заменяется на приближение  $\tilde{A}_h$ , поэтому в результате проведения вычислений может быть найдена величина  $\tilde{y}_h = \tilde{A}_h(\tilde{x})$ .

4. Проведение вычислений на ЭВМ в силу особенностей хранения чисел (причины этого обсуждаются далее) вносит дополнительную ошибку. В результате получается величина  $\tilde{y}_h^* = \tilde{A}_h^*(\tilde{x})$ .

Следовательно, возникают погрешности:

$$\rho_1 = \tilde{y} - y = \tilde{A}(\tilde{x}) - A(x) \quad —$$

**неустранимая погрешность**, которая в рамках данного подхода вызвана неточностью модели и входных данных;

$$\rho_2 = \tilde{y}_h - \tilde{y} = \tilde{A}_h(\tilde{x}) - \tilde{A}(\tilde{x}) \quad —$$

**погрешность численного метода**;

$$\rho_3 = \tilde{y}_h^* - \tilde{y}_h = \tilde{A}_h^*(\tilde{x}) - \tilde{A}_h(\tilde{x}) \quad —$$

**погрешность вычислений**. В итоге погрешность результата

$$\rho = \tilde{y}_h^* - y = \tilde{y}_h^* - \tilde{y}_h + \tilde{y}_h - \tilde{y} + \tilde{y} - y = \rho_3 + \rho_2 + \rho_1.$$

Мы не будем заниматься изучением неустранимой погрешности  $\rho_1$ , модель и входные данные будем считать заданными. Описание алгоритма, приводящего к появлению величины  $\rho_2$ , будет приведено далее. Рассмотрим погрешность вычислений  $\rho_3$ .

В процессе вычислений необходимо обеспечить получение величины  $(\rho_2 + \rho_3)$ , которая была бы в несколько раз меньше  $\rho_1$ .

Заметим, что ранее мы обсуждали только ошибки, носящие объективный по отношению к вычислителю характер.

### B.3.1. Хранение чисел на ЭВМ и ошибки округления

В любой ЭВМ для хранения чисел используется конечное число разрядов. Числа представляются в виде числа либо с фиксированной, либо с плавающей запятой. Чаще всего число хранится в виде числа с плавающей запятой:  $a = Mr^p$ , где  $M$  — число с фиксированной запятой, такое, что  $r^{-1} \leq |M| < 1$ , оно называется **мантиссой**;  $r$  — **основание**

**системы счисления;**  $r$  — целое число — порядок числа  $a$ . Типичный вариант — 32-разрядное число, занимающее в памяти ЭВМ 4 байта. Часть разрядов отводится под знаки числа и порядка и самого порядка. В результате мантисса действительного числа имеет семь значащих цифр, а действительные числа (по абсолютной величине) вообще лежат в диапазоне от  $5,4 \cdot 10^{-79}$  до  $7,2 \cdot 10^{75}$ . В результате в такой ЭВМ точно нельзя представить не только трансцендентные числа (бесконечную непериодическую десятичную дробь), но и рациональные числа (например,  $1/3$ ), а также слишком большие или малые по абсолютной величине.

Пусть под порядок  $p$  отведено  $m$  разрядов, а под мантиссе  $M - l$  разрядов. Тогда числа  $\pm r^{r^m}$  определяют левую и правую границы допустимого числового диапазона.

При этом число  $r^{-r^m}$  представляет собой машинный нуль, а  $r^{r^m-1}$  — машинную бесконечность. Эти названия возникают потому, что числа, меньшие машинного нуля, ЭВМ полагает равными нулю. С числами за пределами машинной бесконечности ЭВМ без специальных процедур работать не может.

Отметим, что в настоящее время возможно использование так называемой «точной арифметики», модули для реализации которой уже разработаны для многих популярных прикладных языков программирования (C, C++, Фортран и др.). В точной арифметике иррациональные числа представляются как набор предикатов (арифметических и алгебраических операторов) от рациональных/целых чисел и установленных констант (таких, как  $e$ ,  $\pi$  и им аналогичных), т. е. в точной арифметике  $\sqrt{3}$  — это действительно  $\sqrt{3}$ , а не  $1,73205080756887729\dots$  Недостатком такого подхода является весьма существенное снижение скорости вычислений, так как описанная точная арифметика пока аппаратно не поддерживается процессорами современных персональных компьютеров. Поэтому все эти операции приходится реализовывать на уровне подпрограмм, что приводит к дополнительным затратам ресурсов.

При появлении в процессе счета мантиссы с числом знаков, большим допускаемого системой, происходит округление. Оно может производиться либо простым отбрасыванием первого лишнего разряда (правило усечения), либо в соответствии со специальными правилами (правило дополнения), либо как-то иначе (например, случайным образом и т. д.).

В результате чаще всего вместо точного числа  $a$  известно лишь его приближение  $a^*$ .

**Определение В.1.** Величина  $\Delta(a^*)$ , про которую известно, что  $|a - a^*| \leq \Delta(a^*)$ , называется **абсолютной погрешностью** (ошибкой) приближенного значения  $a^*$ .

Таким образом, точным выражением является  $a = a^* + \Delta^*(a^*)$ , где  $\Delta^*(a^*)$  — «истинная» ошибка приближенного значения  $a^*$ . Здесь  $|\Delta^*(a^*)| \leq \Delta(a^*)$ . Очевидна неоднозначность такой ошибки  $\Delta(a^*)$ , поэтому она должна быть выбрана наименьшей из всех возможных значений, удовлетворяющих последнему неравенству.

**Определение В.2.** Величина  $\delta(a^*)$ , про которую известно, что  $\left| \frac{a - a^*}{a^*} \right| \leq \delta(a^*)$ , называется **относительной погрешностью** приближенного значения  $a^*$ :

$$\delta(a^*) = \frac{\Delta(a^*)}{|a^*|}.$$

Но обычно используют другую форму записи:

$$a = a^* \pm \Delta(a^*) = a^*(1 \pm \delta(a^*)),$$

показывая, что  $a^*$  есть приближенное значение  $a$ .

Как и в случае абсолютной погрешности,  $\delta(a^*)$  должна быть выбрана минимально возможной из всех значений, удовлетворяющих данному определению.

Далее мы будем использовать истинную абсолютную погрешность  $\Delta^*(a^*)$ , аналогичную истинной относительной погрешности

$$\delta^*(a^*) = \frac{\Delta^*(a^*)}{|a^*|}.$$

Ясно, что относительная погрешность играет главную роль в характеристике точности числа  $a^*$ .

Пусть приближенное число  $a^*$  задано в виде конечной десятичной дроби.

**Определение В.3.** *Значащими цифрами* числа  $a^*$  называются все цифры в его записи начиная с первой ненулевой слева.

**Определение В.4.** Значащую цифру числа  $a^*$  называют *верной*, если абсолютная погрешность числа не превосходит единицы разряда, соответствующего этой цифре.

Очевидно, что данное определение ориентировано на использование наиболее распространенного способа округления — правила дополнения. Обычно в этом случае сохраняемые цифры оставляют неизменными.

ми, если первая из отбрасываемых цифр слева меньше пяти. Если же она равна пяти или более, то в младший сохраняемый разряд добавляется единица.

При округлении по дополнению абсолютная величина погрешности округления не превышает половины единицы разряда, соответствующего последней оставляемой цифре, в случае округления по усечению — единицы того же разряда.

В силу распространенности описанного округления по дополнению чаще всего запись вида  $a = a^*$  означает наличие абсолютной погрешности  $\Delta(a^*)$ , равной половине единицы разряда последней цифры в записи числа  $a^*$ .

Нетрудно заметить, что в случае если число  $a^*$  содержит  $N$  верных значащих цифр, выполнено приближенное равенство  $\delta(a^*) \simeq 10^{-N}$ .

### B.3.2. Ошибки арифметических операций

Подсчитаем ошибки результатов четырех арифметических операций с двумя приближенно заданными числами  $a$  и  $b$ .

1. Сложение. Имеем следующую цепочку соотношений:

$$a + b = a^* + \Delta^*(a^*) + b^* + \Delta^*(b^*) = a^* + b^* + \Delta^*(a^* + b^*).$$

Отсюда

$$\begin{aligned} \Delta^*(a^* + b^*) &= \Delta^*(a^*) + \Delta^*(b^*), \\ \delta^*(a^* + b^*) &= \frac{|a^*|}{|a^* + b^*|} \delta^*(a^*) + \frac{|b^*|}{|a^* + b^*|} \delta^*(b^*). \end{aligned}$$

В результате получаем  $\Delta(a^* + b^*) = \Delta(a^*) + \Delta(b^*)$  и

$$\delta(a^* + b^*) = \frac{|a^*|}{|a^* + b^*|} \delta(a^*) + \frac{|b^*|}{|a^* + b^*|} \delta(b^*).$$

Если  $a^*, b^*$  имеют один знак, то

$$\min(\delta(a^*), \delta(b^*)) \leq \delta(a^* + b^*) \leq \max(\delta(a^*), \delta(b^*)).$$

2. Вычитание. Совершенно аналогично сложению получим

$$\begin{aligned} \Delta(a^* - b^*) &= \Delta(a^*) + \Delta(b^*), \\ \delta(a^* - b^*) &= \frac{|a^*|}{|a^* - b^*|} \delta(a^*) + \frac{|b^*|}{|a^* - b^*|} \delta(b^*). \end{aligned}$$

В случае если  $a^*$  и  $b^*$  близки, видно сильное возрастание относительной ошибки ввиду малости знаменателя в последнем соотношении.

3. Умножение. Выполним умножение, опустив в результате слагаемые второго порядка малости:

$$ab = (a^* + \Delta^*(a^*)) (b^* + \Delta^*(b^*)) \approx a^*b^* + a^*\Delta^*(b^*) + b^*\Delta^*(a^*).$$

Отсюда

$$\Delta(a^*b^*) = |a^*|\Delta(b^*) + |b^*|\Delta(a^*)$$

и

$$\delta(a^*b^*) = \delta(a^*) + \delta(b^*).$$

4. Деление. При вычислении частного также ограничимся главными слагаемыми:

$$\frac{a}{b} = \frac{a^* + \Delta^*(a^*)}{b^* + \Delta^*(b^*)} = \frac{a^*/b^* + \Delta^*(a^*)/b^*}{1 + \Delta^*(b^*)/b^*} \approx \frac{a^*}{b^*} + \frac{\Delta^*(a^*)}{b^*} - \frac{a^*\Delta^*(b^*)}{(b^*)^2}.$$

Отсюда

$$\Delta\left(\frac{a^*}{b^*}\right) = \frac{\Delta(a^*)}{|b^*|} + \frac{|a^*|\Delta(b^*)}{(b^*)^2}$$

и

$$\delta\left(\frac{a^*}{b^*}\right) = \delta(a^*) + \delta(b^*).$$

В этих выражениях не учтены возникающие при вычислении ошибки округления.

Легко видеть, что если не пренебречь слагаемыми второго порядка малости, то в случае умножения

$$\delta(a^*b^*) = \delta(a^*) + \delta(b^*) + \delta(a^*)\delta(b^*).$$

Если поступить аналогично и с делением, то в результате

$$\delta\left(\frac{a^*}{b^*}\right) = \frac{\delta(a^*) + \delta(b^*)}{1 - \delta(b^*)}.$$

В дальнейшем наличие ошибок округлений и операций в явном виде мы будем учитывать редко. Однако необходимо постоянно помнить о них и бороться с результатами их проявления.

**Пример В.2.** Пусть имеются три приближенно заданных числа  $a$ ,  $b$ ,  $c$ , для которых известны значения  $a^*$ ,  $b^*$ ,  $c^*$  и их ошибки. Требуется вычислить величины  $u = (a - b)/c$  и  $v = a/c - b/c$  и выяснить, в каком случае ошибка будет больше.

Очевидно, что в рамках точной арифметики  $u = v$ . В случае приближенно заданных чисел это не так.

Воспользовавшись приближенными формулами для относительной ошибки арифметических операций, получим

$$\begin{aligned}\delta(u^*) &= \delta(a^* - b^*) + \delta(c^*) = \frac{|a^*|}{|a^* - b^*|} \delta(a^*) + \frac{|b^*|}{|a^* - b^*|} \delta(b^*) + \delta(c^*), \\ \delta(v^*) &= \frac{\left| \frac{a^*}{c^*} \right| (\delta(a^*) + \delta(c^*)) + \left| \frac{b^*}{c^*} \right| (\delta(b^*) + \delta(c^*))}{\left| \frac{a^*}{c^*} - \frac{b^*}{c^*} \right|} = \\ &= \frac{|a^*|}{|a^* - b^*|} \delta(a^*) + \frac{|b^*|}{|a^* - b^*|} \delta(b^*) + \frac{|a^*| + |b^*|}{|a^* - b^*|} \delta(c^*).\end{aligned}$$

Здесь

$$u^* = \frac{a^* - b^*}{c^*}, \quad v^* = \frac{a^*}{c^*} - \frac{b^*}{c^*}.$$

В результате оказывается, что относительная ошибка  $v^*$  больше, чем относительная ошибка  $u^*$ . Итог в данном случае не является неожиданным, так как для вычисления  $u^*$  требуется выполнить два действия, а для вычисления  $v^*$  — три. Каждое лишнее действие будет привносить дополнительную ошибку. Однако это только предварительное соображение. Можно привести примеры разных ошибок и в случае равного количества выполняемых операций.

### B.3.3. Погрешность алгоритма

**Определение B.5.** Вычислительный *алгоритм* называется *устойчивым*, если в результате его выполнения ошибки округления возрастают незначительно, и неустойчивым — в противном случае.

С учетом обозначений величин  $\tilde{y}_h^* = \tilde{A}_h^*(\tilde{x})$  определение может быть сформулировано более конкретно: алгоритм  $\tilde{A}_h^*$  является устойчивым, если существует константа  $M$ , такая, что  $\Delta(\tilde{y}_h^*) \leq M\Delta(\tilde{x})$ . При этом она не должна зависеть от  $\Delta(\tilde{x})$ .

Однако вследствие конечности хранимых чисел остается еще вопрос о значении этой константы.

**Пример B.3.** Рассмотрим уравнение

$$x^2 + 2px + q = 0, \quad |q| \ll p^2.$$

Запишем его решение:

$$x_{1,2} = -p \pm \sqrt{p^2 - q}.$$

Пусть  $p, q > 0$ . Тогда при вычислении величины

$$x_1 = \sqrt{p^2 - q} - p$$

используется худшая операция — вычитание близких чисел. Но ту же величину можно записать иначе:

$$x_1 = -\frac{q}{\sqrt{p^2 - q} + p}.$$

При таком вычислении все устойчиво.

**Пример В.4.** Пусть необходимо найти частичную сумму гармонического ряда

$$S = S_{10^7} = \sum_{i=1}^{10^7} \frac{1}{i}.$$

Величина

$$S_n = \sum_{i=1}^n \frac{1}{i} \rightarrow +\infty$$

при  $n \rightarrow \infty$ .

### 1. Вычисление по алгоритму

$$S_1 = 1, \quad S_i = S_{i-1} + \frac{1}{i}, \quad i = 2, 3, \dots, 10^7.$$

Поскольку  $S_n \sim \ln n$  и  $\ln 10 \approx 2,3$ , то  $S_{999999} \approx \ln 10^6 \approx 13,8$ . Если  $i = 10^6$ , то  $S_{10^6} \approx 13,8 + 10^{-6} = 13,8$  при семи значащих цифрах. При сложении мантисса меньшего числа сдвигается вправо и добавляется к мантиссе большего, следовательно, результат не изменится. Таким образом, в результате выполнения алгоритма будет получено совершенно неверное значение суммы.

### 2. Вычисление по алгоритму

$$S_{10^7+1}^* = 0, \quad S_{i-1}^* = S_i^* + \frac{1}{i-1}, \quad i = 10^7 + 1, 10^7, \dots, 2, \quad S = S_1^*.$$

Такой порядок вычислений не содержит операций с резко различными числами. Следовательно, пропадания знаков меньшего числа в этом алгоритме не происходит.

**Пример В.5.** Пусть требуется вычислить последовательность  $\{I_n\}$ , где

$$I_0 = e^{-1} \int_0^1 e^x dx = 1 - e^{-1},$$

$$I_n = e^{-1} \int_0^1 x^n e^x dx = 1 - nI_{n-1}, \quad n = 1, 2, \dots$$

1. Вычислим  $\{I_n\}$  в соответствии с представленным выражением. В результате получим, что  $\Delta(I_n^*) = n! \Delta(I_0^*)$ . Здесь звездочки указывают на приближенный характер используемых величин. Поэтому использование прямой рекуррентной формулы даст катастрофическую потерю точности.

2. Рассмотрим обратное рекуррентное соотношение

$$I_{n-1} = \frac{1 - I_n}{n}.$$

Очевидно, что даже если положить заведомо неверно в качестве «начального» значения  $I_N = 0$  для достаточно большого  $N$  и спуститься по параметру от  $N$  до заданного, то полученная ошибка будет намного меньше, чем в первом алгоритме. #

Приведенные примеры показывают, что часто для вычисления некоторой величины можно построить неустойчивый алгоритм и при наличии устойчивого. Обратное, к сожалению, возможно не всегда. Таким образом, существуют устойчивые и неустойчивые алгоритмы для задач, устойчивость решения которых и не вызывает особых сомнений. В то же время существуют так называемые некорректно поставленные задачи, для решения которых создание устойчивого алгоритма является весьма сложной проблемой. Такие задачи будут рассмотрены далее.

Рассмотрим отдельно **погрешность вычисления функции**. Пусть  $y = f(x)$ ,  $f$  — дифференцируемая функция,  $dy = f' dx$ .

При вычислении величины  $y = f(x^* + \Delta^*(x^*))$  приближенное значение функции  $y^* = f(x^*)$ . Тогда по формуле Лагранжа конечных приращений  $\Delta(y^*) = |f'_x(x^* + \vartheta \Delta^*(x^*))| \Delta(x^*)$ , где  $-1 < \vartheta < +1$ .

**Определение В.6.** Величина

$$\Delta(y^*) = |f'_x(x^*)| \Delta(x^*)$$

называется **линейной абсолютной погрешностью функции**.

### Определение В.7. Величина

$$\Delta(y^*) = \sup_{-1 < \vartheta < +1} |f'_x(x^* + \vartheta \Delta(x^*))| \cdot \Delta(x^*)$$

называется *пределной абсолютной погрешностью вычисления функции*.

На практике чаще всего пользуются линейной оценкой погрешности. Тогда легко решается обратная задача: найти оценку погрешности аргумента, исходя из заданной погрешности функции.

Например,

$$\begin{aligned} y &= \ln x, \quad \Delta(y^*) = \delta(x^*); \\ y &= \operatorname{tg} x, \quad \Delta(y^*) = (1 + \operatorname{tg}^2(x^*))\Delta(x^*) > \Delta(x^*). \end{aligned}$$

Аналогично по формуле Лагранжа конечных приращений оцениваются ошибки вычисления функций многих переменных. Но обратная задача в этом случае может быть решена лишь при дополнительных предположениях — при равном вкладе ошибок каждой переменной в результирующую ошибку функции либо при равенстве ошибок каждого из аргументов, либо при каком-то ином условии. В противном случае найти несколько неизвестных из одного уравнения единственным образом не представляется возможным.

## В.4. Библиографические комментарии

Математические модели, для нахождения численного решения которых разрабатываются численные алгоритмы, описаны во множестве книг, основные из которых [27, 70, 99, 100, 102, 108, 170]. Этот список можно значительно расширить.

Книг по собственно математическому моделированию с описанием различных этапов работы от анализа явления до получения числа сравнительно мало. Наиболее цельными с этой точки зрения являются [69] и [153]. В них подробно описаны разные этапы математического моделирования природных, технических и социальных объектов, а также этапы математического и численного моделирования различных физико-механических явлений (см. также [70, 80, 144, 146, 155]).

Рассмотрение особенностей представления чисел на ЭВМ является необходимым элементом большинства книг, посвященных численному моделированию. Обсуждение этого вопроса можно найти практически во всей основной литературе, приведенной в конце книги. Наиболее

подробно машинная арифметика рассматривается в книгах [17, 61, 62, 83, 90], ориентированных в том числе и на ручной счет. Помимо указанных, рассмотрение способов работы с числами в ЭВМ подробно проведено в [6, 25, 84] и др.

Во многих книгах приведены численные результаты работы алгоритмов, связанные с особенностями машинной арифметики, например в [11, 20].

Многие вопросы, рассмотренные в данной книге, могут быть сопровождены задачами, приведенными в [15, 65, 89, 90, 107, 143, 156].

## **Часть I**

# **Теоретические основы численных методов**

# 1. ЗАДАЧИ ЛИНЕЙНОЙ АЛГЕБРЫ. РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Приведены основные сведения из линейной алгебры и теории *линейных операторов* в конечномерных *пространствах*. Рассмотрены способы решения систем линейных алгебраических уравнений (СЛАУ) как прямые, так и итерационные. В частности, описаны алгоритмы *метода Гаусса* в различных вариантах, включая *метод Холецкого*. Представлены варианты *метода прогонки*: *правая*, *левая*, *встречная*, *циклическая*, *матричная*, *потоковая*. Рассмотрены *стационарные* и *нестационарные итерационные методы* решения СЛАУ, включая *двуслойные* и *трехслойные*. Введено понятие *обусловленности* матрицы. Описан простейший алгоритм решения частичной проблемы *собственных значений*. Представлен один из способов решения плохо обусловленных СЛАУ. Приведены способы хранения больших матриц.

## 1.1. Элементы функционального анализа и линейной алгебры

Основные определения и другие общие сведения из функционального анализа и линейной алгебры, приведенные в этом параграфе, необходимы для облегчения восприятия дальнейшего материала.

Основная наша задача — численное решение уравнения

$$Ay = f,$$

где  $f$  — входная информация;  $A$  — известный оператор;  $y$  — неизвестное, подлежащее определению. Элемент  $y$  ищется среди элементов некоторого пространства  $H$ .

### 1.1.1. Линейные пространства

**Определение 1.1.** Множество  $H$  называется *линейным пространством* над полем  $K$  действительных или комплексных чисел, если выполнены следующие условия:

- 1) задано сложение элементов, т. е. закон, согласно которому любым элементам  $x, y \in H$  поставлен в соответствие элемент  $z = x + y \in H$ , называемый их суммой;

2) задано умножение элемента на число, т. е. закон, согласно которому любым  $x \in H$  и  $\lambda \in K$  поставлен в соответствие элемент  $z = \lambda x \in H$ , называемый произведением элемента  $x$  на число  $\lambda$ ;

3) указанные законы (операции) подчиняются следующим аксиомам:

а)  $x + y = y + x$ ,  $x + (y + z) = (x + y) + z$ ,  $x, y, z \in H$  (коммутативность и ассоциативность сложения);

б)  $\lambda(\mu x) = (\lambda\mu)x$ ,  $x \in H$ ,  $\lambda, \mu \in K$  (ассоциативность умножения на число);

в)  $\lambda(x + y) = \lambda x + \lambda y$ ,  $(\lambda + \mu)x = \lambda x + \mu x$ ,  $x, y \in H$ ,  $\lambda \in K$  (дистрибутивность умножения относительно сложения);

г) в  $H$  существует элемент 0 (нуль), для которого  $x + 0 = x$ ,  $x \in H$ ;

д) для любого  $x \in H$  существует элемент  $(-x) \in H$ , такой, что  $x + (-x) = 0$ ;

е)  $1 \cdot x = x$ ,  $x \in H$ .

Элементы линейного пространства принято называть векторами. Элемент 0 (см. п. «г») называют нулевым вектором, а элемент  $(-x)$  (см. п. «д») называют вектором противоположным вектору  $x$ .

**Определение 1.2.** Элементы  $\{x_i\}$ ,  $i = 1, 2, \dots, n$ ,  $x_i \in H$ , называются **линейно независимыми**, если из равенства

$$\sum_{i=1}^n \lambda_i x_i = 0$$

следует, что  $\lambda_i = 0$ ,  $i = 1, 2, \dots, n$ . Если же равенство нулю последней суммы возможно хотя бы при одном  $\lambda_i \neq 0$ , то исходные элементы называются линейно зависимыми.

Очевидно, что любая совокупность элементов линейно зависима, если содержит нуль.

**Определение 1.3.** Линейное пространство  $H$  называется  **$n$ -мерным**, если в нем существует  $n$  линейно независимых элементов, а любые  $n + 1$  элементов являются линейно зависимыми.

**Определение 1.4.** Непустое множество  $H_1$  элементов линейного пространства  $H$  называется **линейным многообразием**, если вместе с элементами  $x_1, x_2, \dots, x_n \in H_1$  множество  $H_1$  содержит и любую их линейную комбинацию

$$\sum_{i=1}^n \lambda_i x_i.$$

Замкнутое линейное многообразие называется **подпространством**.

**Определение 1.5.** Сумма  $H_1 + H_2 + \dots + H_n$  конечного числа подпространств  $H_1, H_2, \dots, H_n$  есть множество элементов вида  $x = x_1 + x_2 + \dots + x_n$ , где  $x_i \in H_i$ . Если для любого  $x \in H$  существует однозначное представление в таком виде, то говорят, что  $H$  — прямая сумма подпространств  $H_i$ , и пишут

$$H = H_1 \oplus H_2 \oplus \dots \oplus H_n.$$

**Лемма 1.1.** Если  $H = H_1 \oplus H_2$ , то  $H_1 \cap H_2 = \{0\}$ . Обратно, если  $\forall x \in H$  есть  $x = x_1 + x_2$ , где  $x_1 \in H_1, x_2 \in H_2, H_1 \cap H_2 = \{0\}$ , то  $H = H_1 \oplus H_2$ .

Без доказательства.

**Определение 1.6.** Линейное пространство  $H$  называется **нормированным**, если для каждого  $x \in H$  определено вещественное число  $\|x\|$ , называемое **нормой**, которое удовлетворяет условиям:

- 1)  $\|x\| \geq 0$ , причем  $\|x\| = 0$  тогда и только тогда, когда  $x = 0$ ;
- 2)  $\forall x, y \in H \quad \|x + y\| \leq \|x\| + \|y\|$  (неравенство треугольника);
- 3)  $\forall x \in H \quad \forall \lambda \in K \quad \|\lambda x\| = |\lambda| \|x\|$  (однородность нормы).

**Определение 1.7.** Последовательность  $\{x_n\} \subset H$  сходится к  $x \in H$ , т. е.  $x_n \rightarrow x$ , если  $\|x - x_n\| \rightarrow 0$  при  $n \rightarrow \infty$ . Если  $\|x_m - x_n\| \rightarrow 0$  при  $m, n \rightarrow \infty$ , т. е.  $\forall \varepsilon > 0 \exists N: \|x_n - x_m\| < \varepsilon$  при  $n, m > N$ , то последовательность называется **фундаментальной**.

**Определение 1.8.** Линейное нормированное пространство  $H$  называется **полным**, если любая фундаментальная последовательность элементов  $H$  сходится к некоторому элементу из  $H$ .

Полное линейное нормированное пространство называется **банаховым**. Любое конечномерное линейное нормированное пространство является полным.

**Определение 1.9. Нормы**  $\|\cdot\|_1$  и  $\|\cdot\|_2$ , определенные на  $H$ , называются **эквивалентными**, если существуют положительные постоянные  $t$  и  $M$ , не зависящие от  $x$ , такие, что для любого  $x \in H$  выполняются неравенства

$$m\|x\|_1 \leq \|x\|_2 \leq M\|x\|_1.$$

В конечномерном пространстве любые две нормы являются эквивалентными. Однако постоянные  $t$  и  $M$  из определения эквивалентности чаще всего зависят от выбора норм и размерности пространства.

В случае двух эквивалентных норм из сходимости элементов в одной норме следует сходимость в другой.

**Определение 1.10.** Пусть каждой паре  $x, y \in H$ , где  $H$  — линейное действительное (комплексное) пространство, сопоставлено действительное (комплексное) число  $(x, y)$ , такое, что

- 1)  $(x, y) = \overline{(y, x)}$  (симметричность);
- 2)  $(x + y, z) = (x, z) + (y, z)$  (дистрибутивность);
- 3)  $(\lambda x, y) = \lambda(x, y)$  (однородность);
- 4)  $(x, x) \geq 0$ ; причем  $(x, x) = 0$  тогда и только тогда, когда  $x = 0$ .

В этом случае число  $(x, y)$  называют *скалярным произведением*.

Черта над  $(y, x)$  здесь означает комплексное сопряжение.

**Определение 1.11.** Комплексное линейное пространство  $H$  с заданным скалярным произведением называется *унитарным пространством*. Действительное линейное пространство со скалярным произведением называется *евклидовым пространством*. Бесконечномерное унитарное пространство, полное относительно нормы  $\|x\| = \sqrt{(x, x)}$ , называется *гильбертовым пространством*.

В унитарном (евклидовом) пространстве используют, как правило, *евклидову норму*  $\|x\| = \sqrt{(x, x)}$ .

**Неравенство Коши — Буняковского** (или Коши — Буняковского — Шварца) имеет вид  $|(x, y)| \leq \|x\| \|y\|$ .

**Определение 1.12.** Элементы  $x, y \in H$ , где  $H$  — унитарное (евклидово) пространство, называют взаимно ортогональными и пишут  $x \perp y$ , если  $(x, y) = 0$ . Если элемент  $x$  ортогонален любому элементу  $y$  подпространства  $H_1$ , то говорят, что  $x$  ортогонален  $H_1$ :  $x \perp H_1$ . Множество  $H_2 = \{x \in H : x \perp H_1\}$  есть подпространство, называемое *ортогональным дополнением*  $H_1$ . Если  $H_2$  — ортогональное дополнение  $H_1$ , то  $H = H_1 \oplus H_2$ .

**Определение 1.13.** *Система*  $\{x_i\}$ ,  $i = 1, 2, \dots$ , элементов унитарного пространства  $H$  называется *ортонормированной*, если для любых индексов  $i$  и  $j$  верно равенство  $(x_i, x_j) = \delta_{ij}$  ( $\delta_{ij}$  — символ Кронекера,  $\delta_{ii} = 1$ ,  $\delta_{ij} = 0$ , если  $i \neq j$ ). Если в  $H$  не существует элемента, отличного от 0 и ортогонального всем  $x_i$ , то такая *система* называется *полной*.

### 1.1.2. Операторы в линейных нормированных пространствах

Пусть  $X, Y$  — линейные нормированные пространства.

**Определение 1.14.** Говорят, что на множестве  $D \subset X$  задан оператор  $A$  со значениями в  $Y$ , если любому элементу  $x \in D$  сопоставлен

элемент  $y = Ax \in Y$ . При этом  $D = D(A)$  — **область определения оператора**  $A$ ,  $\text{im } A = \{y: y = Ax, x \in D(A)\}$  — **область значений оператора**  $A$ . Если  $D(A) = X$ ,  $\text{im } A \subset X$ , то  $A$  отображает линейное нормированное пространство  $X$  на себя,  $A$  — оператор (действующий) на  $X$ .

Будем через  $E$  обозначать единичный оператор, а через  $0$  — нулевой.

**Определение 1.15.** *Оператор*  $A$  называется **линейным**, если  $D(A)$  — линейное многообразие в  $X$  и

$$A(\lambda x + \mu y) = \lambda Ax + \mu Ay, \quad x, y \in D(A), \quad \lambda, \mu \in K.$$

**Определение 1.16.** *Линейный оператор*  $A$  называется **ограниченным**, если существует такая константа  $M > 0$ , что

$$\|Ax\|_2 \leq M\|x\|_1, \quad x \in D(A), \tag{1.1}$$

где  $\|\cdot\|_1, \|\cdot\|_2$  — нормы в  $X$  и  $Y$  соответственно,

**Определение 1.17.** Наименьшая из постоянных  $M$  из неравенства (1.1) называется **нормой оператора**  $A$ , **подчиненной** (или порожденной, индуцированной) норме  $\|\cdot\|_1$ :

$$\|A\| = \sup_{\substack{x \in D(A) \\ x \neq 0}} \frac{\|Ax\|_2}{\|x\|_1} = \sup_{\substack{x \in D(A) \\ \|x\|_1=1}} \|Ax\|_2.$$

**Определение 1.18.** *Оператор*  $A$  называется **непрерывным** в точке  $x$ , если из  $\|x_n - x\|_1 \rightarrow 0$  следует  $\|Ax_n - Ax\|_2 \rightarrow 0$ .

Линейный ограниченный оператор всегда является непрерывным и, наоборот, линейный непрерывный оператор ограничен.

Все возможные линейные ограниченные операторы из  $X$  в  $Y$  образуют линейное нормированное пространство.

На множестве линейных ограниченных операторов, действующих из  $X$  в  $X$  ( $D(A) = D(B) = X$ ), можно ввести и операцию умножения:

$$AB(x) = A(B(x)).$$

При этом  $AB$  — линейный ограниченный оператор, причем

$$\|AB\| \leq \|A\|\|B\|.$$

В самом деле,

$$\|AB\| = \sup_{x \neq 0} \frac{\|ABx\|_1}{\|x\|_1} \leq \sup_{x \neq 0} \frac{\|A\|\|Bx\|_1}{\|x\|_1} \leq \|A\|\|B\|.$$

**Определение 1.19.** Если  $(AB)x = (BA)x$ ,  $x \in X$ , то операторы  $A$  и  $B$  называются перестановочными или коммутирующими.

**Определение 1.20.** Пусть  $A$  — оператор из  $X$  на  $Y$  ( $D(A) = X$ ,  $\text{im } A = Y$ ). Если для любого  $y \in Y$  существует единственный элемент  $x \in X$ , для которого  $Ax = y$ , то таким соотношением определяется **оператор  $A^{-1}$** , называемый **обратным** для  $A$  и имеющий область определения  $Y$ , а область значений  $X$ , при этом

$$A^{-1}Ax = x, \quad AA^{-1}y = y.$$

**Определение 1.21.** **Ядром** линейного **оператора**  $A$  называется множество тех элементов  $x \in X$ , для которых  $Ax = 0$ . Ядро линейного оператора обозначается  $\ker A$ .

Условие  $\ker A = \{0\}$  является необходимым и достаточным условием того, чтобы линейный оператор  $A$  имел обратный оператор.

**Определение 1.22.** Число

$$\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k}$$

называется **спектральным радиусом** оператора  $A$ . При этом  $\rho(A)$  не зависит от определения нормы и

$$\rho(A) = \inf_{\|\cdot\|} \|A\|.$$

Для линейного ограниченного оператора

$$\rho(A) \leq \|A\|.$$

### 1.1.3. Операторы в гильбертовом пространстве

Из неравенства Коши — Буняковского следует, что

$$|(Ax, x)| \leq \|A\| \cdot \|x\|^2.$$

**Определение 1.23.** Ограниченный **оператор**  $A^*$  называется **сопряженным** ограниченному оператору  $A$ , если

$$(Ax, y) = (x, A^*y), \quad x, y \in H.$$

Если  $A = A^*$ , то  $A$  — **самосопряженный оператор**.

Для любого линейного ограниченного оператора  $\|A^*\| = \|A\|$ . Если  $A$  — линейный ограниченный оператор в  $H$ , то

$$H = \ker A \oplus \text{im } A^* = \ker A^* \oplus \text{im } A.$$

**Определение 1.24.** Оператор  $A$  называется **нормальным**, если  $AA^* = A^*A$ , и **кососимметричным**, если  $A^* = -A$ .

Любой оператор  $A$  можно представить в виде суммы самосопряженного  $A_1$  и кососимметричного  $A_2$  операторов:

$$A = A_1 + A_2, \quad A_1 = \frac{1}{2}(A + A^*), \quad A_2 = \frac{1}{2}(A - A^*).$$

Если  $H$  — действительное пространство, то имеют место соотношения

$$(Ax, x) = (A_1x, x), \quad (A_2x, x) = 0.$$

Для нормального (в частности, самосопряженного) оператора  $\rho(A) = \|A\|$ , так как для таких операторов справедливо  $\|A^k\| = \|A\|^k$ . Здесь используется определенная выше подчиненная норма.

**Определение 1.25.** Числовым радиусом оператора  $A$  называется число

$$\bar{\rho}(A) = \sup_{\|x\|=1} |(Ax, x)|, \quad x \in H.$$

**Определение 1.26.** Линейный оператор  $A$ , действующий в гильбертовом пространстве  $H$ , называется **положительным** ( $A > 0$ ), если

$$(Ax, x) > 0, \quad x \in H, \quad x \neq 0;$$

**неотрицательным** ( $A \geq 0$ ), если

$$(Ax, x) \geq 0, \quad x \in H;$$

**положительно определенным**, если для некоторого положительного числа  $\delta$

$$(Ax, x) \geq \delta(x, x), \quad x \in H.$$

Заметим, что свойство знаковой определенности полностью зависит от самосопряженной части оператора.

Понятие положительного (неотрицательного) оператора вводит в пространстве линейных операторов отношение порядка: неравенство  $A > B$  ( $A \geq B$ ) означает, что  $A - B > 0$  ( $A - B \geq 0$ ).

В случае комплексного линейного пространства  $H$  определение 1.26 распространяется только на самосопряженные операторы.

Пусть  $D$  — самосопряженный положительный оператор в  $H$ . Тогда можно ввести **энергетическое пространство**  $H_D$ , состоящее из элементов  $H$  со скалярным произведением  $(x, y)_D = (Dx, y)$  и нормой  $\|x\|_D = (Dx, x)^{1/2}$ . Если  $D$  — самосопряженный ограниченный положительно определенный в  $H$  оператор, то обычная  $\|\cdot\|$  и энергетическая  $\|\cdot\|_D$  нормы эквивалентны.

**Определение 1.27.** Числа  $\delta = \inf_{\|x\|=1} (Ax, x)$ ,  $\Delta = \sup_{\|x\|=1} (Ax, x)$  называются границами оператора  $A$ , при этом

$$\delta E \leq A \leq \Delta E,$$

где  $E$  — единичный оператор.

**Определение 1.28.** Оператор  $B$  называется квадратным корнем из оператора  $A$ , если выполнено равенство  $B^2 = A$ . Тогда пишут  $B = A^{1/2}$ .

Можно показать, что существует единственный неотрицательный самосопряженный квадратный корень из любого неотрицательного самосопряженного оператора  $A$ , перестановочный со всяkim оператором, перестановочным с  $A$ .

#### 1.1.4. Операторы в конечномерном пространстве

Рассмотрим  $n$ -мерное унитарное (или евклидово) пространство  $H$ ,  $x_1, x_2, \dots, x_n$  — его ортонормированный базис,

$$x = \sum_{k=1}^n c_k x_k, \quad c_k = (x, x_k).$$

Пусть в  $H$  действует линейный оператор  $A$ . В базисе  $x_1, x_2, \dots, x_n$  ему соответствует матрица  $\tilde{A} = (a_{ij})_{n \times n}$ , где  $a_{ij} = (Ax_j, x_i)$ . В самом деле,

$$Ax = \sum_{k=1}^n c_k Ax_k = \sum_{l=1}^n d_l x_l,$$

откуда

$$d_l = \sum_{k=1}^n c_k (Ax_k, x_l) = \sum_{k=1}^n a_{lk} c_k, \quad a_{lk} = (Ax_k, x_l).$$

Любому элементу  $x \in H$  соответствует вектор его координат:

$$x = (c_1, c_2, \dots, c_n).$$

Если оператор  $A$  является самосопряженным в  $H$ , то соответствующая ему матрица  $\tilde{A}$  симметрична в любом ортонормированном базисе.

**Определение 1.29.** Число  $\lambda$  называется **собственным значением** оператора  $A$ , если уравнение  $Ax = \lambda x$  имеет ненулевые решения. При этом  $x$  — **собственный элемент** оператора  $A$ , соответствующий собственному значению  $\lambda$ . Если  $\lambda$  — собственное значение, то  $\ker(A - \lambda E) \neq \{0\}$ .

**Определение 1.30.** Множество  $\sigma(A)$  собственных значений оператора  $A$  называется **спектром оператора**  $A$ .

Укажем некоторые свойства операторов, имеющие отношение к его спектру.

1. Самосопряженный оператор  $A$  имеет  $n$  ортонормированных собственных элементов  $x_1, x_2, \dots, x_n$ , соответствующих вещественным собственным значениям  $\lambda_1, \lambda_2, \dots, \lambda_n$ .

2. Для самосопряженного оператора  $A$

$$\|A\| = \rho(A) = \max_{1 \leq k \leq n} |\lambda_k|.$$

3. Если  $A = A^* \geq 0$ , то все собственные значения оператора  $A$  неотрицательны и, кроме того,

$$\delta(x, x) \leq (Ax, x) \leq \Delta(x, x),$$

где  $\delta = \min_k \lambda_k$ ,  $\Delta = \max_k \lambda_k$ .

Самосопряженная матрица неотрицательна в том и только том случае, если все ее собственные значения неотрицательны. Она положительна в том и только том случае, если все ее собственные значения положительны.

Очевидно, что в последнем случае свойства положительности и положительной определенности совпадают.

Отметим, что в силу однозначности представления матрицы в виде суммы самосопряженной и кососимметричной частей свойства знакоопределенности полностью определяются самосопряженной частью. Следовательно, эквивалентность положительности и положительной определенности в действительном конечномерном случае распространяется и на случай матриц общего вида.

**Определение 1.31.** Число  $\lambda$  называется *собственным значением* оператора  $A$  относительно оператора  $B$ , если уравнение  $Ax = \lambda Bx$  имеет ненулевые решения. При этом  $x$  — *собственный элемент* оператора  $A$  относительно оператора  $B$ , соответствующий собственному значению  $\lambda$ .

Если операторы  $A$  и  $B$  самосопряженные в  $H$  и оператор  $B$  положительно определен, то существует  $n$  собственных элементов, ортонормированных в энергетическом пространстве  $H_B$ .

### 1.1.5. Нормы векторов и матриц

Рассмотрим конечномерное пространство  $H$  размерности  $n$  и действующий в нем линейный ограниченный оператор  $A$ .

Будем отождествлять оператор  $A$  с его матрицей  $\tilde{A}$ , а вектор — с его координатами. Волну у матрицы  $\tilde{A}$  в дальнейшем опустим.

Найдем выражение для подчиненных норм оператора  $A$  при различном выборе норм  $x$ . Будем считать, что оператор  $A$  действует из  $X$  в  $X$ , способы вычисления норм  $\|x\|$  и  $\|Ax\|$  совпадают.

**1. Кубическая норма**  $\|x\|_\infty = \max_k |x_k|$ . Множество  $\|x\|_\infty \leq 1$  представляет собой куб со стороной длины 2 (рис. 1.1).

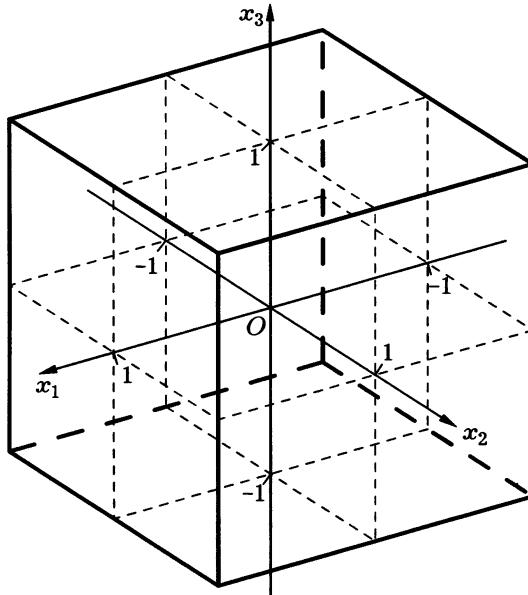


Рис. 1.1

**Предложение 1.1.** Справедливо соотношение

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|.$$

◀ По определению,

$$\begin{aligned} \|Ax\|_\infty &= \max_i \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \max_i \sum_{j=1}^n |a_{ij}| |x_j| \leq \\ &\leq \max_i \sum_{j=1}^n |a_{ij}| \max_j |x_j| \leq \|x\|_\infty \cdot \max_i \sum_{j=1}^n |a_{ij}|. \end{aligned}$$

Найдем вектор  $x \neq 0$ , для которого выполняется равенство. Пусть

$$b_i = \sum_{j=1}^n |a_{ij}| \quad \text{и} \quad b_m = \max_{1 \leq i \leq n} b_i.$$

Выберем  $x_j = \text{sign } a_{mj}$ ,  $\|x\|_\infty = 1$ . Тогда

$$|(Ax)_m| = \sum_{j=1}^n |a_{mj}| = b_m.$$

Для всех остальных компонент

$$|(Ax)_i| = \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \sum_{j=1}^n |a_{ij}| \cdot 1 \leq b_m.$$

Поэтому

$$\|Ax\|_\infty = \max_i \sum_{j=1}^n |a_{ij}| \cdot 1 = \max_i \sum_{j=1}^n |a_{ij}| = \|A\|_\infty. \quad \blacktriangleright$$

**2. Октаэдрическая норма**  $\|x\|_1 = \sum_{i=1}^n |x_i|$ . Множество  $\|x\|_1 \leq 1$  представляет собой октаэдр (рис. 1.2).

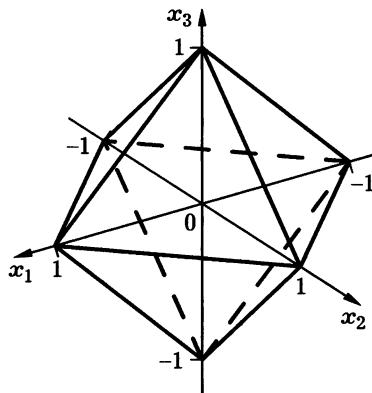


Рис. 1.2

**Предложение 1.2.** Справедливо соотношение

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|.$$

◀ По определению,

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^n \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| |x_j| = \\ &= \sum_{j=1}^n |x_j| \sum_{i=1}^n |a_{ij}| \leq \max_j \sum_{i=1}^n |a_{ij}| \cdot \|x\|_1. \end{aligned}$$

Найдем элемент  $x$ , для которого  $\|Ax\|_1 = \max_j \sum_{i=1}^n |a_{ij}| \cdot \|x\|_1$ . Пусть  $c_j = \sum_{i=1}^n |a_{ij}|$ ,  $c_p = \max_j c_j$ . Выберем вектор  $x$  с компонентами  $x_j = \delta_{jp}$ , т. е. с одной ненулевой единичной  $p$ -й компонентой. Тогда  $\|x\|_1 = 1$ ,

$$(Ax)_i = \sum_{j=1}^n a_{ij} x_j = a_{ip}.$$

Следовательно,

$$\|Ax\|_1 = \sum_{i=1}^n |a_{ip}| = c_p = \max_j \sum_{i=1}^n |a_{ij}| = \|A\|_1. \quad \blacktriangleright$$

**3. Евклидова норма**  $\|x\|_2 = \left( \sum_{i=1}^n x_i^2 \right)^{1/2}$  (иногда — **гильбертова, сферическая или шаровая**). Множество  $\|x\|_2 \leq 1$  — обычный шар радиусом 1 в декартовых координатах (на рис. 1.3 приведен привычный шар радиусом 1 в евклидовой (сферической) норме).

**Предложение 1.3.** Величина  $\|A\|_2 = \left( \sum_{i=1}^n a_{ij}^2 \right)^{1/2}$  не является нормой матрицы, подчиненной сферической норме вектора.

◀ По определению,

$$\|Ax\|_2 = \left( \sum_{i=1}^n \left( \sum_{j=1}^n a_{ij} x_j \right)^2 \right)^{1/2}.$$

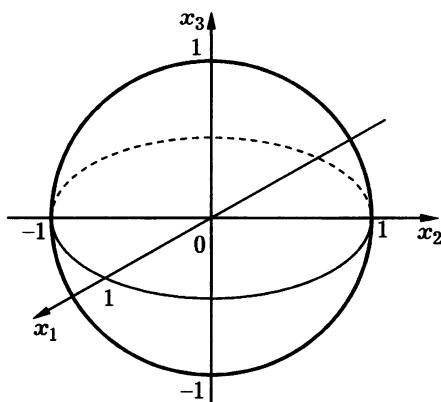


Рис. 1.3

Вследствие неравенства Коши — Буняковского справедлива оценка

$$\left| \sum_{j=1}^n a_{ij} x_j \right| \leq \left( \sum_{j=1}^n a_{ij}^2 \right)^{1/2} \left( \sum_{j=1}^n x_j^2 \right)^{1/2}$$

и тогда

$$\|Ax\|_2 \leq \left( \sum_{i,j=1}^n a_{ij}^2 \right)^{1/2} \|x\|_2.$$

Если пытаться доказать подчиненность рассматриваемой нормы, то необходимо указать вектор  $x$ , для которого выполнено равенство. Но это невозможно, так как, например, для единичного оператора

$$\|E\|_2 = \left( \sum_{i,j=1}^n \delta_{ij}^2 \right)^{1/2} = \sqrt{n}, \quad \|Ex\|_2 = \|x\|_2 \neq \sqrt{n} \cdot \|x\|_2, \quad x \neq 0,$$

при  $n \neq 1$ . Вычислим подчиненную норму  $E$ :

$$\|E\| = \sup_{x \neq 0} \frac{\|Ex\|_2}{\|x\|_2} = 1.$$

В то же время  $\|E\|_2 = \sqrt{n}$ . Следовательно,  $\|A\|_2$  не является нормой матрицы, подчиненной норме вектора  $\|\cdot\|_2$ . Однако  $\|A\|_2$  есть величина, для которой выполнены все три условия нормы, поэтому  $\|A\|_2$  — норма. ►

**Определение 1.32.** *Норма матрицы  $\|A\|$  называется согласованной с нормой вектора  $\|x\|_X$ , если*

$$\|Ax\|_Y \leq \|A\| \cdot \|x\|_X, \quad x \in H.$$

Подчиненные нормы являются согласованными с данной нормой вектора. Рассмотренные выше нормы матрицы  $\|\cdot\|_1$ ,  $\|\cdot\|_\infty$ ,  $\|\cdot\|_2$  являются согласованными с соответствующей нормой вектора.

Для подчиненной нормы матрицы, определяемой соотношением

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

заведомо выполнено неравенство  $\|AB\| \leq \|A\| \cdot \|B\|$  (оно выполнено в том числе и для  $\|A\|_2$ ).

Однако это неравенство не выполняется для других кандидатов на роль нормы, удовлетворяющих свойствам нормы в определении 1.6.

**Предложение 1.4.** Для величины

$$\tilde{N}(A) = \max_{i,j} |a_{ij}|$$

аксиомы нормы справедливы, но неравенство

$$\tilde{N}(AB) \leq \tilde{N}(A)\tilde{N}(B),$$

вообще говоря, не выполнено.

◀ По построению  $\tilde{N}(A) \geq 0$ . Если  $\tilde{N}(A) = 0$ , то  $a_{ij} = 0$  для любых  $i$  и  $j$ , и, следовательно,  $A = 0$ . Проверим остальные свойства нормы:

$$\tilde{N}(\alpha A) = \max_{i,j} |\alpha a_{ij}| = |\alpha| \max_{i,j} |a_{ij}| = |\alpha| \tilde{N}(A).$$

Очевидно также, что  $\tilde{N}(A + B) \leq \tilde{N}(A) + \tilde{N}(B)$ . Тем не менее при

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad A^2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

имеем  $\tilde{N}(A) = 1$ ,  $\tilde{N}(A^2) = 2$  и  $\tilde{N}(A^2) > (\tilde{N}(A))^2$ . ►

В связи с важностью неравенства  $\|AB\| \leq \|A\|\|B\|$  его в явном виде включают в определение нормы матрицы.

**Определение 1.33.** Число  $\|A\|$  называют **нормой матрицы**  $A$ , если

- 1)  $\|A\| \geq 0$  для всех  $A$ ;  $\|A\| = 0$  тогда и только тогда, когда  $A = 0$ ;
- 2)  $\|A + B\| \leq \|A\| + \|B\|$  для любых  $A$  и  $B$  (неравенство треугольника);
- 3)  $\|\lambda A\| = |\lambda| \|A\|$  для любых  $A$  и  $\lambda \in K$  (однородность нормы);
- 4)  $\|AB\| \leq \|A\|\|B\|$  для любых  $A$  и  $B$ .

Далее будем иметь дело именно с такими нормами. Следовательно, величина  $\tilde{N}(A)$ , определенная выше, не является нормой матрицы.

Используемые далее нормы будут, как правило, подчиненными или, как минимум, согласованными.

**Предложение 1.5.** Величина  $\|A\|_M = n \max_{i,j} |a_{ij}|$  согласована с  $\|x\|_1$ ,  $\|x\|_2$ ,  $\|x\|_\infty$  и является нормой.

Без доказательства.

Данную **норму матрицы** часто называют **максимальной**.

**Предложение 1.6.** Величина  $\tilde{N}(A) = 1/\sqrt{n}\|A\|_2$  не является нормой в смысле приведенного определения.

Без доказательства.

**Предложение 1.7.** Пусть  $\|A\|_s = (\max_j \mu_j)^{1/2}$ , где  $\mu_j$  — собственные значения матрицы  $A^*A$  ( $\det(A^*A - \mu_j E) = 0$ ). Величина  $\|\cdot\|_s$  подчинена  $\|\cdot\|_2$  и является нормой.

Без доказательства.

Данная **норма матрицы**  $\|\cdot\|_s$  называется *спектральной*.

Если  $A^* = A$ , то  $\mu_j$  равны квадратам собственных чисел  $A$ , т. е.  $\|A\|_s = \max_j |\lambda_j|$ , где  $\lambda_j$  — собственные значения  $A$ , следовательно,  $\|A\|_s = \rho(A)$  — спектральный радиус оператора  $A$ .

### 1.1.6. Другие нормированные пространства

Приведем примеры других нормированных пространств.

1. Множество действительных чисел с нормой  $\|x\| = |x|$ .
2. Пространство  $C[a, b]$  непрерывных на отрезке  $[a, b]$  функций с нормой  $\|x\|_C = \max_{[a, b]} |x|$ .
3. Пространство  $L_p([a, b])$ ,  $1 < p < +\infty$ , состоящее из функций  $x(t)$ , определенных на отрезке  $[a, b]$  и удовлетворяющих условию  $\int_a^b |x|^p dt < \infty$  (здесь интеграл понимается в смысле Лебега). При этом считаются одинаковыми функции, различающиеся лишь на множестве меры нуль. Нормой в этом пространстве является

$$\|x\|_p = \left( \int_a^b |x|^p dt \right)^{1/p}.$$

4. Бесконечномерные пространства  $c$  и  $l_p$  последовательностей вида  $x = \{x_n\}$  с нормами:

$$\|x\|_c = \sup |x_i|, \quad \|x\|_{l_p} = \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{i=1}^n |x_i^p| \right)^{1/p}.$$

5. **Пространство Соболева**  $W_2^1([a, b])$  представляет собой пополнение<sup>\*</sup> нормированного пространства бесконечно дифференцируе-

\*Процедуру пополнения нормированного пространства можно рассматривать как расширение этого линейного пространства добавлением формальных пределов фундаментальных последовательностей. Детали этой процедуры выходят за рамки данной книги и здесь не рассматриваются.

мых на отрезке  $[a, b]$  функций с нормой

$$\|y\|_{W_2^1} = \left( \int_a^b (y^2(x) + (y'(x))^2) dx \right)^{1/2}.$$

Пополнение нормированного пространства финитных бесконечно дифференцируемых функций с той же нормой приводит к полному нормированному пространству  $W_2^{1,0}$ . В этом нормированном пространстве можно упростить выражение для нормы, опустив под знаком интеграла первое слагаемое.

### 1.1.7. Критерий Адамара и лемма Гершгорина

В заключение приведем два результата, полезных для дальнейшего изложения.

**Определение 1.34.** *Матрица  $A$  называется невырожденной* (неособенной), если ее определитель не равен нулю, и *вырожденной* (особенной) в противном случае.

Решение системы линейных алгебраических уравнений (СЛАУ) с невырожденной матрицей существует и единствено, с вырожденной может не существовать или быть неединственным. Однородная система с квадратной матрицей имеет ненулевое решение тогда и только тогда, когда матрица системы вырожденная.

**Лемма 1.2 (критерий Адамара невырожденности матрицы).** Пусть матрица  $A$  обладает свойством строгого диагонального преобразования:  $|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|$ ,  $i = 1, 2, \dots, n$ . Тогда матрица  $A$  является невырожденной.

◀ Проведем доказательство от противного. Допустим, что матрица с указанным свойством является вырожденной. Тогда СЛАУ  $Ax = 0$  имеет ненулевое решение  $x = (x_1, x_2, \dots, x_n)$ , т. е.  $\sum_{j=1}^n a_{ij}x_j = 0$ ,  $i = 1, 2, \dots, n$ . Выберем максимальный по абсолютной величине элемент решения  $x_k$ :  $|x_k| = \max_{i=1, \dots, n} |x_i| > 0$ . Тогда для строки системы с номером  $k$

$$a_{kk}x_k = - \sum_{j=1, j \neq k}^n a_{kj}x_j.$$

Поэтому

$$|a_{kk}| |x_k| \leq \sum_{j=1, j \neq k}^n |a_{kj}| |x_j| \leq \sum_{j=1, j \neq k}^n |a_{kj}| |x_k|.$$

Отсюда получаем  $|a_{kk}| \leq \sum_{j=1, j \neq k}^n |a_{kj}|$ , что противоречит условию строгого диагонального преобладания. Следовательно, матрица  $A$  является невырожденной. ►

В дальнейшем мы увидим, что многие критерии устойчивости или существования решений, по существу, являются иными формулировками критерия Адамара применительно к конкретной решаемой задаче. Так, например, непосредственно из критерия Адамара вытекает следующая лемма.

**Лемма 1.3 (лемма Гершгорина).** Все собственные значения матрицы  $A$  лежат в объединении **кругов Гершгорина**

$$|z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, 2, \dots, n.$$

◀ Пусть  $\lambda$  — собственное значение матрицы  $A$ . Тогда матрица  $A - \lambda E$  является вырожденной. Критерий Адамара для нее несправедлив. Отсюда  $|\lambda - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|$  для некоторого  $i$ . Объединяя весь спектр, получим нужные круги. ►

Лемма Гершгорина дает самый простой способ нахождения границ спектра, но, к сожалению, он и самый грубый.

## 1.2. Прямые методы решения СЛАУ

Рассмотрим систему линейных алгебраических уравнений (СЛАУ)  $Ax = f$  с квадратной невырожденной матрицей  $A$  порядка  $n$ . Здесь  $x$  — неизвестный  $n$ -мерный вектор,  $f$  — известный  $n$ -мерный вектор. Вектор  $x$  можно записать в виде  $x = A^{-1}f$ . По сути, проблема заключается в вычислении обратной матрицы. Начнем изучать методы решения СЛАУ с **прямых методов**, в которых решение может быть найдено за конечное число действий.

### 1.2.1. Схема метода Гаусса

Идея **метода Гаусса** состоит в последовательном исключении переменных. При  $i = 1$  имеем

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = f_1.$$

Пусть  $a_{11} \neq 0$ . Положим

$$c_{1j} = \frac{a_{1j}}{a_{11}}, \quad j = 2, 3, \dots, n; \quad y_1 = \frac{f_1}{a_{11}}.$$

Тогда исходная система принимает вид

$$\begin{aligned} x_1 + c_{12}x_2 + \dots + c_{1n}x_n &= y_1, \\ \sum_{j=1}^n a_{ij} x_j &= f_i, \quad i = 2, 3, \dots, n. \end{aligned}$$

Умножаем первое уравнение на  $a_{i1}$  и вычитаем из  $i$ -го ( $i = 2, 3, \dots, n$ ), тогда

$$a_{ij}^{(1)} = a_{ij} - a_{i1}c_{1j}, \quad j = 2, 3, \dots, n, \quad f_i^{(1)} = f_i - a_{i1}y_1, \quad a_{i1}^{(1)} = \delta_{i1}.$$

В результате получим

$$\left\{ \begin{array}{l} x_1 + c_{12}x_2 + \dots + c_{1n}x_n = y_1, \\ a_{22}^{(1)}x_2 + \dots + a_{2n}^{(1)}x_n = f_2^{(1)}, \\ \dots \dots \dots \dots \dots \dots \\ a_{n2}^{(1)}x_2 + \dots + a_{nn}^{(1)}x_n = f_n^{(1)}. \end{array} \right.$$

Теперь  $x_1$  содержится только в первой строке и мы можем рассмотреть  $n - 1$  последние строки. Если  $a_{22}^{(1)} \neq 0$ , то процедуру повторяем. Проделывая это  $n$  раз, получаем систему треугольного вида:

$$\left\{ \begin{array}{l} x_1 + c_{12}x_2 + \dots + c_{1n}x_n = y_1, \\ x_2 + \dots + c_{2n}x_n = y_2, \\ \dots \dots \dots \dots \dots \\ x_n = y_n \end{array} \right.$$

с матрицей  $C$ , т. е.  $Cx = y$ , где

$$C = \begin{pmatrix} 1 & c_{12} & \dots & c_{1,n-1} & c_{1n} \\ 0 & 1 & \dots & c_{2,n-1} & c_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & c_{n-1,n} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \quad —$$

верхняя треугольная матрица. В ней элементы на главной диагонали и выше нее отличны от нуля. Нижняя треугольная матрица — это матрица, у которой все элементы выше главной диагонали равны нулю.

Далее выполняем обратный ход:

$$x_n = y_n,$$

$$x_i = y_i - \sum_{j=i+1}^n c_{ij}x_j, \quad i = n-1, n-2, \dots, 1.$$

### 1.2.2. Расчетные формулы метода Гаусса

Цикл по  $i$  от 1 до  $(n - 1)$ :  $y_i = f_i/a_{ii}$ .

Цикл по  $j$  от  $i + 1$  до  $n$ :  $c_{ij} = a_{ij}/a_{ii}$ .

Цикл = по  $i'$  от  $i + 1$  до  $n$ :  $f_{i'} = f_{i'} - a_{i'i}y_i$ .

Цикл по  $j'$  от  $i + 1$  до  $n$ :  $a_{i'j'} = a_{i'j'} - a_{i'i}c_{ij'}$ .

Для  $i = n$  получим  $y_n = f_n/a_{nn}$ .

Считаем, что значения переприсваиваются переменным в памяти ЭВМ, поэтому верхние индексы опущены.

Далее выполняем обратный ход метода Гаусса.

Формулы обратного хода приведены выше.

Главное ограничение метода Гаусса — требование отличия от нуля величины  $a_{ii}^{(i-1)}$ , называемой ведущим элементом на  $i$ -м шаге исключения.

### 1.2.3. Число действий в методе Гаусса

Ограничимся подсчетом числа операций умножения и деления, так как их выполнение требует намного больше времени, чем выполнение операций сложения и вычитания.

1. Для вычисления коэффициентов  $c_{ij}$  необходимое число операций составляет

$$\sum_{i=1}^n (n-i) = 1 + \dots + n - 1 = \frac{n(n-1)}{2}.$$

2. При вычислении коэффициентов  $a_{ij}$  для каждого значения  $i$  требуется  $(n-i)^2$  умножений. Для вычисления всех  $a_{ij}$  необходимо выполнить следующее число операций:

$$\begin{aligned} S_n &= \sum_{i=1}^{n-1} (n-i)^2 = \sum_{i=1}^{n-1} i^2 = \frac{1}{3} \sum_{i=1}^{n-1} ((i+1)^3 - i^3 - 3i - 1) = \\ &= \frac{1}{3} \left( n^3 - 1 - \frac{3}{2}(n-1)n - (n-1) \right) = \frac{(n-1)n(2n-1)}{6}. \end{aligned}$$

Отметим полезную оценку величины  $S_n$ , идея получения которой может быть использована при оценке многих сумм. Из геометрического смысла определенного интеграла вытекают неравенства

$$S_n < \int_1^n x^2 dx = \frac{n^3 - 1}{3}, \quad S_n > \int_1^n (x-1)^2 dx = \frac{(n-1)^3}{3}.$$

Взяв полусумму этих оценок, получим приближенную формулу для суммы последовательных квадратов:

$$S_n \approx \frac{(n^3 - 1) + (n - 1)^3}{6} = \frac{(n - 1)(2n^2 - n + 2)}{6},$$

что очень близко к точному значению величины  $S_n$ .

3. Вычисление коэффициентов  $y_i$  требует  $n$  делений.

4. Количество делений при вычислении коэффициентов  $f_{i'}$  составляет

$$\sum_{i=1}^n (n - i) = \frac{n(n - 1)}{2}.$$

5. Таким образом, для проведения прямого хода метода Гаусса необходимое количество операций составляет

$$\frac{n(n - 1)}{2} + \frac{(n - 1)n(2n - 1)}{6} + n + \frac{n(n - 1)}{2} = \frac{n(n + 1)(2n + 1)}{6}.$$

6. Для проведения обратного хода метода Гаусса нужно

$$\sum_{i=1}^{n-1} (n - i) = \frac{n(n - 1)}{2}$$

действий.

7. Суммируя, получаем общее количество делений и умножений в методе Гаусса:

$$\Sigma = \frac{1}{3}n(n^2 - 1) + n^2 = \frac{1}{3}n(n^2 + 3n - 1) \sim \frac{1}{3}n^3.$$

Приведем пример оценки затрат машинного времени. При  $n = 10^3$  следует выполнить  $1/3 \cdot 10^9$  операций. Если ЭВМ делает  $10^6$  операций деления или умножения в секунду, то необходимо  $1/3 \cdot 10^3$  секунд, т. е. примерно 5 минут машинного времени. При  $n = 10^4$  требуется уже 5000 минут, т. е. примерно 83 часа.

Таким образом, очевидны ограничения на применимость метода Гаусса по числу действий.

#### 1.2.4. Выбор главного элемента

Рассмотрим некоторые ограничения в использовании метода Гаусса.

1. Основным ограничением является условие  $a_{ii}^{(i-1)} \neq 0$ . Исключение элемента  $x_i$  нельзя проводить, если получен нуль на главной диагонали. Но в первом столбце промежуточной системы все элементы не могут

оказаться нулями, так как это означало бы равенство нулю определителя матрицы  $A$ , что не соответствует условию. Следовательно, путем перестановок строк всегда можно добиться появления в  $i$ -й строке элемента  $a_{ii}^{(i-1)} \neq 0$  и продолжить процедуру.

Если элемент  $|a_{ii}^{(i-1)}|$  мал, то это тоже плохо, так как значения  $|c_{ij}|$  велики и будут приводить к значительным ошибкам в вычислениях. Поэтому на практике, как правило, реализуют **метод Гаусса с выбором главного элемента**.

В этом случае после исключения  $(i-1)$ -й переменной в  $i$ -м столбце ищут значение  $\max_{i \leq i' \leq n} |a_{i'i}|$ . Выбрав главный элемент, например  $a_{i^*i}$ , строки  $i$  и  $i^*$  меняют местами и проводят следующий шаг метода Гаусса. После следующего исключения процедуру повторяют.

В ряде задач главный элемент разыскивают в  $i$ -й строке. В этом случае использование элемента  $a_{ii}^*$  требует перестановки столбцов, что эквивалентно перенумерации неизвестных. Данный способ алгоритмически существенно сложнее, чем предыдущий.

2. Отметим применимость описанного алгоритма для нахождения обратной матрицы. Такая задача сводится к решению матричного уравнения  $A \cdot A^{-1} = E$ , где неизвестной является  $A^{-1}$ . Процедура ее нахождения состоит в  $n$ -кратном решении описанной задачи с различными правыми частями. Неизвестными при этом являются столбцы искомой матрицы  $A^{-1}$ . Приведение матрицы  $A$  к треугольному виду делается один раз. Далее  $n$  раз придется обработать правые части и выполнить обратный ход метода Гаусса. Применение метода Гаусса для обращения матрицы влечет за собой в несколько раз большее число действий, чем обычное решение системы.

3. В методе Гаусса никак не используется структура матрицы, поэтому его можно применять для матриц произвольного вида. Наличие определенной структуры позволяет существенно ускорить процесс решения. По расположению ненулевых элементов различают матрицы следующих видов: ленточные, ящичные, квазитреугольные, треугольные, пяти- и трехдиагональные и др. Если заранее знать расположение нулей в матрице, то расчет можно организовать так, чтобы не включать в процесс вычисления нулевые элементы. Тогда скорость работы заметно увеличится. Мы ограничимся рассмотрением трехдиагональных матриц и родственных им (см. 1.4).

Существуют и другие прямые методы: метод Жордана, метод оптимального исключения, метод окаймления, метод отражений, метод ортогонализации, QR-алгоритм и многие другие. Мы ограничимся

рассмотрением метода прогонки с рядом его вариантов и метода квадратного корня.

За неимением места не будем рассматривать теорию и практику решения систем линейных разностных уравнений с постоянными коэффициентами. Свойства таких задач во многом напоминают (или прямо совпадают) свойства линейных обыкновенных дифференциальных уравнений (ОДУ) с постоянными коэффициентами. Благодаря этому соответствуанию решение таких задач иногда удается найти аналитически, что дает специализированный прямой метод нахождения решения (см. 1.13).

### 1.3. Обусловленность СЛАУ

**Определение 1.35.** Система линейных алгебраических уравнений (СЛАУ)  $Ax = f$  устойчива по правой части, если существует такая постоянная  $M \geq 0$ , что для любого возмущения правой части  $\delta f$  справедлива следующая оценка:

$$\|\delta x\| \leq M \|\delta f\|.$$

При этом постоянная  $M$  не должна зависеть от правой части  $f$  и решения  $x$ .

Определение 1.35 имеет смысл как в конечномерном, так и бесконечномерном случаях. В конечномерном случае при невырожденной матрице существование постоянной  $M$  — тривиальный факт.

Если  $\det A \neq 0$ , то  $x = A^{-1}f$ ,  $\delta x = A^{-1}\delta f$  в силу линейности  $A$  и  $A^{-1}$ . Отсюда имеем оценку  $\|\delta x\| \leq \|A^{-1}\| \|\delta f\|$ , т. е. чем меньше определитель  $A$ , тем больше определитель  $A^{-1}$ , больше  $M_1 = \|A^{-1}\|$  и больше влияние ошибок правой части на ошибки решения.

Обычно именно малость определителя исходной матрицы  $A$  (и большое значение определителя обратной матрицы) считают признаком плохой устойчивости решаемой системы. Однако это не так, в чем убеждает простой пример матрицы  $A = \varepsilon E$  с произвольным малым  $\varepsilon$ . Ее определитель очевидно мал, но проблем в решении систем с такой матрицей нет (при  $\varepsilon \neq 0$ ).

Истинный критерий устойчивости дает переход к относительным ошибкам  $x$  и  $f$ :

$$\|f\| \leq \|A\| \|x\|,$$

откуда получаем

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|}{\|A\|^{-1}} \cdot \frac{\|\delta f\|}{\|f\|} = \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\delta f\|}{\|f\|}.$$

**Определение 1.36.** Число  $M_A = \|A^{-1}\| \|A\|$  называется **числом обусловленности** матрицы  $A$  (и  $A^{-1}$  в силу симметрии).

Отметим, что умножение (или деление) матрицы на число ее обусловленность не меняет.

Матрицы с большим числом  $M_A$  называются плохо обусловленными, при решении СЛАУ с такими матрицами идет резкое накопление погрешностей. При этом характеристики типа «большой» и «малый» чаще всего имеют относительный характер в зависимости от возможностей (машинных и алгоритмических), имеющихся у вычислителя.

**Лемма 1.4.** Число обусловленности  $M_A$  матрицы  $A$  обладает следующими свойствами:

$$1) \ M_A \geqslant 1; \quad 2) \ M_A \geqslant \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}; \quad 3) \ M_{AB} \leqslant M_A \cdot M_B.$$

◀ Рассмотрим свойство 2. Легко показать, что  $\|A\| \geqslant |\lambda_{\max}|$ . Пусть  $\|A\|$  — подчиненная норма, тогда  $\|A\| = \sup_{x \neq 0} \|Ax\|/\|x\|$ ,  $\|A\| \geqslant |\lambda_{\max}|$ , так как в качестве вектора  $x$  под знаком sup может стоять собственный вектор, соответствующий максимальному (по модулю) собственному числу. Для него  $\|Ax\| = |\lambda_{\max}| \|x\|$ . Если же  $\|A\|$  — согласованная норма, то для того же вектора  $Ax = \lambda_{\max}x$  из условия согласованности следует  $\|Ax\| \leqslant \|A\| \|x\|$ , откуда  $\|A\| \geqslant |\lambda_{\max}|$ .

Для обратной матрицы максимальным по модулю является число  $|\lambda_{\min}^{-1}|$ , обратное к минимальному по модулю собственному числу, откуда  $\|A^{-1}\| \geqslant |\lambda_{\min}^{-1}|$ . Следовательно,

$$M_A = \|A^{-1}\| \|A\| \geqslant \left| \frac{\lambda_{\max}}{\lambda_{\min}} \right|.$$

Отсюда сразу следует свойство 1, а свойство 3 следует из неравенства  $\|AB\| \leqslant \|A\| \|B\|$  для матриц. ►

**Замечание 1.1.** Рассмотрим пространство  $H$  с евклидовой нормой  $\|y\|^2 = (y, y)$ . Пусть матрица  $A$  симметрична, т. е. оператор  $A$  самосопряженный. Тогда он имеет  $n$  собственных векторов, образующих ортонормированный базис, и произвольный вектор  $x$  можно представить в виде

$$x = \sum_{k=1}^n c_k x_k.$$

В результате

$$Ax = \sum_{k=1}^n c_k \lambda_k x_k, \quad \|Ax\|^2 = \sum_{k=1}^n c_k^2 \lambda_k^2, \quad \|x\|^2 = \sum_{k=1}^n c_k^2.$$

Отсюда  $\|A\| = |\lambda_{\max}|$ . Аналогично  $\|A^m\| = |\lambda_{\max}^m|$ , откуда  $\rho(A) = |\lambda_{\max}|$ . Таким же образом  $\|A^{-1}\| = |\lambda_{\min}|^{-1}$ , откуда  $M_A = |\lambda_{\max}|/|\lambda_{\min}|$ .

Исследуем теперь устойчивость решения СЛАУ относительно возмущений матрицы.

При проведении вычислений неизбежны ошибки округления, вследствие чего матрица  $A$  искажается. В результате приходится решать систему  $\tilde{A}\tilde{x} = \tilde{f}$  вместо  $Ax = f$ .

Оценим  $\delta x = \tilde{x} - x$  через  $\delta A = \tilde{A} - A$ . Пусть для простоты  $f = \tilde{f}$ . Тогда

$$(A + \delta A)(x + \delta x) = f = Ax.$$

Отсюда  $(A + \delta A)\delta x = -\delta Ax$  и

$$\delta x = -(E + A^{-1}\delta A)^{-1}A^{-1}\delta Ax.$$

Следовательно,

$$\begin{aligned} \|\delta x\| &\leq \|A^{-1}\| \|\delta A\| \|(E + A^{-1}\delta A)^{-1}\| \|x\|, \\ \frac{\|\delta x\|}{\|x\|} &\leq M_A \frac{\|\delta A\|}{\|A\|} \|(E + A^{-1}\delta A)^{-1}\|. \end{aligned}$$

Ограничимся полученным неравенством, считая последнюю норму в правой части величиной  $O(1)$ , т. е.

$$\|(E + A^{-1}\delta A)^{-1}\| \leq C = \text{const.}$$

Приведем без доказательства оценку погрешности вычисления матрицы в алгоритме метода Гаусса. Возмущение (ошибка) матрицы  $A$  в нем является следствием представления  $A$  в ходе вычислений в виде  $A = LU$ , где  $L$  — нижняя треугольная матрица;  $U$  — верхняя треугольная матрица ( $U = C$  в наших обозначениях при описании метода). Поэтому в результате фактически получаем  $\tilde{A} = \tilde{L}\tilde{U}$ . Ограничимся оценкой возмущения матрицы  $A$ . Для метода Гаусса

$$\frac{\|\delta A\|}{\|A\|} = O(n2^{-t}),$$

где  $n$  — размерность матрицы;  $t$  — число разрядов мантиссы в двоичном представлении числа на ЭВМ, используемой для вычислений.

Следовательно, в методе Гаусса

$$\frac{\|\delta x\|}{\|x\|} = O(M_A n 2^{-t}).$$

На практике найти значение числа обусловленности конкретной матрицы чаще всего бывает сложно, если только не решается спектральная задача об определении границ спектра. При этом чаще всего используется необходимое условие плохой обусловленности — малость определителя исходной матрицы  $A$  — с учетом его ограниченности.

**Пример 1.1. Матрица Гильберта**, служащая классическим образцом плохой обусловленности, имеет вид  $H_n = \{h_{ij}\}$ ,  $i, j = 1, 2, \dots, n$ , с элементами  $h_{ij} = (i + j - 1)^{-1}$ . Норма обратной к ней матрицы экспоненциально растет с ростом  $n$ . В результате число ее обусловленности при  $n = 8$  превышает  $10^{10}$ . Вследствие этого решение СЛАУ с такой матрицей может не содержать ни одного верного знака.

Матрица Гильберта появляется естественным образом при попытке приблизить некоторый набор данных полиномом. Как будет показано далее (см. 3), норма интерполяционного полинома растет с повышением его степени. Это свойство проявляется и в плохой обусловленности матрицы возникающей системы уравнений.

## 1.4. Метод прогонки решения СЛАУ с трехдиагональной матрицей

### 1.4.1. Метод правой прогонки

Рассмотрим случай трехдиагональной матрицы  $A$ , т. е. систему линейных алгебраических уравнений (СЛАУ) вида

$$a_i x_{i-1} - b_i x_i + c_i x_{i+1} = -d_i, \quad 1 \leq i \leq n; \quad a_1 = c_n = 0.$$

Будем решать эту систему методом исключения Гаусса. Воспользуемся структурой решаемой системы: выразим первый элемент  $x_1$  через  $x_2$  и подставим во второе уравнение системы. Тогда снова получим уравнение, связывающее между собой два неизвестных, но теперь только  $x_2$  и  $x_3$ . Далее процесс можно продолжить вплоть до выражения  $x_{n-1}$  через  $x_n$  после обработки предпоследнего уравнения. Итогом описанных действий является обнуление нижней диагонали исходной матрицы и приведение матрицы к треугольному (верхнему) виду. В результате последнее уравнение позволит определить  $x_n$ . После этого необходимо

проводить обратный ход метода Гаусса, циклически определив все компоненты неизвестного вектора. Описанный словесно алгоритм можно реализовать, положив в основу поиска решения линейную функциональную связь одной компоненты неизвестного вектора с последующей. Таким образом, получим **метод прогонки**, который представляет собой специализированную реализацию метода исключения Гаусса, существенным образом использующую структуру матрицы системы.

Будем искать решение в виде

$$x_i = \alpha_{i+1}x_{i+1} + \beta_{i+1}, \quad i = 1, 2, \dots, n-1,$$

где  $\alpha_{i+1}$ ,  $\beta_{i+1}$  — прогоночные коэффициенты. Для первого уравнения получим

$$x_1 = \frac{c_1}{b_1}x_2 + \frac{d_1}{b_1} = \alpha_2x_2 + \beta_2,$$

где

$$\alpha_2 = \frac{c_1}{b_1}; \quad \beta_2 = \frac{d_1}{b_1} \quad (b_1 \neq 0).$$

Далее для строки с номером  $i \neq 1$  проводим исключение переменной  $x_{i-1}$ :

$$a_i(\alpha_i x_i + \beta_i) - b_i x_i + c_i x_{i+1} = -d_i.$$

Отсюда

$$x_i = \frac{c_i}{b_i - a_i \alpha_i} x_{i+1} + \frac{d_i + a_i \beta_i}{b_i - a_i \alpha_i} = \alpha_{i+1} x_{i+1} + \beta_{i+1},$$

где

$$\alpha_{i+1} = \frac{c_i}{b_i - a_i \alpha_i}; \quad \beta_{i+1} = \frac{d_i + a_i \beta_i}{b_i - a_i \alpha_i}, \quad i = 2, 3, \dots, n-1.$$

При  $i = n$  имеем

$$a_n(\alpha_n x_n + \beta_n) - b_n x_n = -d_n,$$

откуда

$$x_n = \frac{d_n + a_n \beta_n}{b_n - a_n \alpha_n} = \beta_{n+1}.$$

Окончательно получаем следующие формулы прямого хода:

$$\alpha_2 = \frac{c_1}{b_1}, \quad \beta_2 = \frac{d_1}{b_1},$$

$$\alpha_{i+1} = \frac{c_i}{b_i - a_i \alpha_i}, \quad \beta_{i+1} = \frac{d_i + a_i \beta_i}{b_i - a_i \alpha_i}, \quad i = 2, 3, \dots, n-1,$$

и формулы обратного хода:

$$x_n = \beta_{n+1} = \frac{d_n + a_n \beta_n}{b_n - a_n \alpha_n}; \quad x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad i = n-1, n-2, \dots, 1.$$

Число операций деления и умножения при такой прогонке равно  $5n$ . В этой оценке учтены лишь старшие по параметру  $n$  составляющие числа действий. Вывод оценки предполагает, что знаменатель в прогоночных коэффициентах вычисляется только один раз. Меньшее по порядку  $n$  число действий при решении системы обеспечить уже невозможно (сравните с числом  $\frac{1}{3}n^3$  действий в методе Гаусса).

Для успешной реализации алгоритма необходимо, чтобы в процессе расчета знаменатели не обращались в нуль, а коэффициент умножения в обратном ходе (по абсолютной величине) не превышал единицы, т. е. ошибки округления не накапливались при обратном ходе прогонки. Следовательно, можно сформулировать определение.

**Определение 1.37.** Будем называть прогонку корректной, если знаменатели в формулах для прогоночных коэффициентов не обращаются в нуль, и устойчивой, если все  $|\alpha_i| \leq 1$ .

**Лемма 1.5.** Если в трехдиагональной матрице выполнено *условие диагонального преобладания*:

$$|b_i| \geq |a_i| + |c_i|,$$

где хотя бы для одного  $i$  выполнено строгое неравенство, то исходная система уравнений имеет решение, которое может быть получено с помощью метода прогонки. Алгоритм прогонки в указанных условиях является корректным и устойчивым.

◀ Так как  $a_1 = 0$ , то  $|\alpha_2| \leq 1$  и  $|\alpha_2| > 0$ . Отсюда

$$|b_i - a_i\alpha_i| \geq |b_i| - |a_i||\alpha_i| \geq |c_i| + |a_i|(1 - |\alpha_i|) \geq |c_i|,$$

если  $|\alpha_i| \leq 1$  для всех  $i = 2, 3, \dots, n - 1$ . Тогда

$$|\alpha_{i+1}| = \frac{|c_i|}{|b_i - a_i\alpha_i|} \leq 1.$$

Осталось показать, что  $b_n - a_n\alpha_n \neq 0$ . Единственная сложность доказательства вызвана тем, что  $c_n = 0$ . По условию леммы хотя бы в одной строке есть строгое диагональное преобладание. Если оно достигается при  $i = n$ , то

$$|b_n - a_n\alpha_n| \geq |b_n| - |a_n||\alpha_n| > |a_n|(1 - |\alpha_n|) \geq 0.$$

Таким образом, знаменатель в выражении для  $x_n$  отличен от нуля. Если же строгое неравенство достигается в строке  $i = i_0$ , то при обработке данной строки в соответствии с проведенными выше выкладками будет получено  $|\alpha_{i_0+1}| < 1$ , откуда следует, что все  $\alpha$  с большими номерами также меньше единицы (по модулю). ►

**Замечание 1.2.** Условия леммы носят достаточный характер. Заметим, что они во многом соответствуют *критерию Адамара* невырожденности матрицы. Описанный алгоритм является весьма надежным. Он часто хорошо работает и без выполнения условий диагонального преобладания.

**Замечание 1.3.** Описанный вариант метода Гаусса иногда называют методом *правой прогонки*, так как в нем непосредственное определение неизвестных происходит справа налево.

**Замечание 1.4.** Далее мы рассмотрим методы *левой, встречной* и *матричной прогонок*. Кроме того, существуют методы *потоковой, циклической* и *пятидиагональной прогонок*. Опишем их коротко без специального анализа.

1. Потоковая прогонка чаще всего применяется в задачах с резко различающимися коэффициентами, в которых помимо самих неизвестных  $x_i$ ,  $i = 1, 2, \dots, n$ , необходимо найти еще и так называемые потоки  $w_i = -a_i(x_i - x_{i-1})$ ,  $i = 2, 3, \dots, n$ . При этом верны соотношения  $c_i = a_{i+1}$ ,  $i = 1, 2, \dots, n-1$ ,  $a_1 = 0$ ,  $c_n = 0$ .

Как правило, при буквальном вычислении потоков по приведенной формуле используется вычитание близких чисел, что ведет к большим ошибкам. Поэтому для получения устойчивых результатов применяют метод потоковой прогонки, в котором такое действие исключено. Исключения можно добиться путем расширения решаемой системы уравнений, добавив к ней  $n-1$  уравнений — определений соответствующих потоков. Затем формулы прогонки (например, правой) можно переписать так, чтобы в результате были получены потоки и решения исходной системы без непосредственного вычисления разностей.

2. Метод циклической прогонки применяется при решении задач с модифицированной матрицей, имеющей с формальной точки зрения пять ненулевых диагоналей, т. е. решаемая в этом случае система уравнений полностью совпадает с рассмотренной, кроме первого и последнего уравнений. Они имеют вид

$$a_1x_n - b_1x_1 + c_1x_2 = -d_1, \quad a_nx_{n-1} - b_nx_n + c_nx_1 = -d_n.$$

В рассматриваемой матрице главная и две примыкающие к ней слева и справа диагонали отличны от нуля. Кроме того, ненулевые элементы находятся в правом верхнем и левом нижнем углах. Подобные задачи возникают, например, при конечномерной дискретизации дифференциальных задач с периодическими граничными условиями.

Нетрудно видеть, что прямой ход исключения метода Гаусса приведет к появлению ненулевых значений в столбце с номером  $n$  при обработке каждой строки. Следовательно, постулируемая линейная зависимость должна иметь вид

$$x_i = \alpha_{i+1}x_{i+1} + \beta_{i+1} + \gamma_{i+1}x_n, \quad i = 1, 2, \dots, n-2,$$

где  $\alpha_{i+1}$ ,  $\beta_{i+1}$ ,  $\gamma_{i+1}$  — прогоночные коэффициенты. Аналогично правой прогонке из первого уравнения получим значения  $\alpha_2$ ,  $\beta_2$ ,  $\gamma_2$ . Обработка следующих уравнений системы с номерами  $i = 2, 3, \dots, n-2$  обычным образом позволит получить рекуррентные соотношения для прогоночных коэффициентов.

Подстановка линейной зависимости с уже известными прогоночными коэффициентами в уравнение с номером  $n-1$  дает линейную связь  $x_{n-1}$  и  $x_n$ . В результате оказывается возможным искать решение в виде

$$x_i = \delta_{i+1}x_n + \varepsilon_{i+1}, \quad i = 1, 2, \dots, n-1,$$

где  $\delta_{i+1}$ ,  $\varepsilon_{i+1}$  — новые прогоночные коэффициенты. Значения коэффициентов  $\delta_n$ ,  $\varepsilon_n$  находятся из уравнения исходной системы с номером  $n-1$ . После этого предыдущая связь значений трех неизвестных позволяет найти весь набор новых прогоночных коэффициентов. Далее остается лишь подставить полученные соотношения в последнее уравнение исходной системы, найти  $x_n$  и значения всех остальных неизвестных.

3. При решении СЛАУ с пятидиагональной матрицей, в которой отличные от нуля элементы стоят на главной диагонали и примыкающим к ней слева и справа диагоналям (по две с каждой стороны), также можно использовать соответствующее обобщение прогонки. Ее формулы записываются сравнительно просто по аналогии, например, с обычной правой прогонкой.

#### 1.4.2. Методы левой и встречной прогонок

Рассмотрим ту же СЛАУ с трехдиагональной матрицей  $A$ , что и в 1.4.1. Ее решение можно найти методом правой прогонки. Этот алгоритм весьма работоспособен и практичен. Однако бывают ситуации, когда условия устойчивости этого алгоритма не выполнены и на практике алгоритм неустойчив. В этом случае может оказаться устойчивой левая прогонка. Рассмотрим ее подробнее.

В правой прогонке исключение неизвестных происходило слева направо: сначала  $x_1$ , потом  $x_2$  и т. д. Однако ничто не мешает сделать

исключение в обратном порядке:  $x_n$ , потом  $x_{n-1}$  и т. д. Основанный на такой процедуре алгоритм носит название левой прогонки.

Очевидность процедуры позволяет сразу выписать формулы данной прогонки:

$$x_{i+1} = \xi_{i+1}x_i + \eta_{i+1}, \quad i = 1, 2, \dots, n-1,$$

где  $\xi_{i+1}, \eta_{i+1}$  — прогоночные коэффициенты, которые вычисляются по следующим формулам:

$$\begin{aligned} \xi_n &= \frac{a_n}{b_n}; & \eta_n &= \frac{d_n}{b_n} \quad (b_n \neq 0); \\ \xi_i &= \frac{a_i}{b_i - c_i \xi_{i+1}}; & \eta_i &= \frac{d_i + c_i \eta_{i+1}}{b_i - c_i \xi_{i+1}}, \quad i = 2, 3, \dots, n-1; \\ x_1 &= \frac{d_1 + c_1 \eta_2}{b_1 - c_1 \xi_2} = \eta_1. \end{aligned}$$

Далее ходом слева направо можно найти все остальные неизвестные.

Встречаются задачи, в которых требуется найти либо одну величину  $x_m$ , либо группу подряд идущих значений компонентов неизвестного вектора, включающих  $x_m$ . В этом случае применим метод **встречных прогонок**. Состоит он в последовательном исключении неизвестных с номерами, большими и меньшими  $m$ . Неизвестные с большими номерами исключаются методом левой прогонки, а с меньшими — методом правой прогонки. Для этого необходимо найти прогоночные коэффициенты  $\alpha_i, \beta_i, i = 2, 3, \dots, m$ , и  $\xi_i, \eta_i, i = m+1, m+2, \dots, n$ .

Далее в строке с номером  $m$  решаемой СЛАУ получим

$$a_m(\alpha_m x_m + \beta_m) - b_m x_m + c_m(\xi_{m+1} x_m + \eta_{m+1}) = -d_m.$$

Отсюда можно найти неизвестную величину  $x_m$ . Если необходимо вычислить компоненты вектора с номерами, меньшими или большими  $m$ , то это делается с помощью соответствующих формул правой или левой прогонок с уже известными прогоночными коэффициентами.

#### 1.4.3. Метод матричной прогонки

Рассмотрим специальный случай решения следующей СЛАУ вида

$$\begin{aligned} A_i X_{i-1} - B_i X_i + C_i X_{i+1} &= -D_i, \quad 1 \leq i \leq n; \\ A_1 &= C_n = 0. \end{aligned}$$

В отличие от рассмотренных выше случаев здесь каждый элемент  $X_i$ , как и  $D_i$ , представляет собой  $m$ -мерный вектор, а коэффициенты  $A_i$ ,

$B_i, C_i$  являются квадратными матрицами размером  $m \times m$ . Подобные системы возникают при решении многомерных задач математической физики или при решении систем дифференциальных уравнений в одномерном случае.

Очевидно, что рассматриваемая СЛАУ не является системой с трехдиагональной матрицей, если рассматривать все компоненты неизвестных. Однако ее специальная структура позволяет практически без изменений использовать алгоритм прогонки. Представленный далее алгоритм называется методом **матричной прогонки**.

Будем искать решение в виде

$$X_i = \alpha_{i+1} X_{i+1} + \beta_{i+1}, \quad i = 1, 2, \dots, n-1,$$

где  $\alpha_{i+1}, \beta_{i+1}$  — прогоночные коэффициенты, в данном случае это квадратные матрицы размером  $m \times m$ . Как и ранее, из первого уравнения получим

$$\alpha_2 = B_1^{-1} C_1, \quad \beta_2 = B_1^{-1} D_1$$

(матрица  $B_1$  невырождена). Далее после обработки строки с номером  $i \neq 1$  имеем

$$\alpha_{i+1} = (B_i - A_i \alpha_i)^{-1} C_i, \quad \beta_{i+1} = (B_i - A_i \alpha_i)^{-1} (D_i + A_i \beta_i), \quad i = 2, 3, \dots, n-1.$$

Обработка строки  $i = n$  дает

$$X_n = (B_n - A_n \alpha_n)^{-1} (D_n + A_n \beta_n) = \beta_{n+1}.$$

Приведенные формулы позволяют провести как прямой, так и обратный ход прогонки и полностью найти решение задачи.

Как видно, в отличие от простой прогонки теперь деление заменяется на обращение матриц. Поэтому описанный алгоритм является достаточно трудоемким по числу операций: каждое обращение матриц требует порядка  $m^3$  действий. В результате для решения задач с большим  $m$  матричную прогонку используют редко.

## 1.5. Метод квадратного корня

Необходимо решить следующую систему линейных алгебраических уравнений:

$$Ax = f.$$

Рассмотрим случай эрмитовой (в комплексном случае) или симметричной (в действительном случае) матрицы. Элементы такой матрицы удовлетворяют условию  $a_{ij} = \bar{a}_{ji}$ ,  $i, j = 1, 2, \dots, n$ , где черта означает комплексное сопряжение.

Для решения поставленной задачи существует специальная реализация *метода исключения Гаусса*, называемая **методом квадратного корня** или **методом Холецкого**. Метод основан на использовании структуры матрицы  $A$ , позволяющей представить матрицу в виде  $A = S^*DS$ , где  $S$  — верхняя треугольная матрица с положительными элементами на диагонали;  $S^*$  — ее транспонированная, т. е. нижняя треугольная, матрица;  $D$  — диагональная матрица, на диагонали которой находятся  $\pm 1$ .

Метод Гаусса, по существу, представляет собой разложение матрицы  $A$  в произведение  $A = LU$  левой треугольной и правой треугольной матриц с последующим обращением матрицы  $U$ . Для симметричной матрицы  $LU$ -разложение легко преобразовать в разложение  $S^*DS$ , на котором основан метод квадратного корня.

Пусть  $S = (s_{ij})$ ,  $d_{ii}$  — диагональные элементы  $D$ :

$$D = \text{diag}(d_{11}, \dots, d_{nn}),$$

тогда

$$(DS)_{ij} = \sum_{k=1}^n d_{ik}s_{kj} = d_{ii}s_{ij}.$$

Пусть матрица  $S$  (как и  $A$ ) действительная, тогда  $(S^*)_{ij} = s_{ji}$  и

$$(S^*DS)_{ij} = \sum_{k=1}^n s_{ki}d_{kk}s_{kj} = a_{ij}, \quad i, j = 1, 2, \dots, n.$$

Так как  $a_{ij} = a_{ji}$ , то можно рассмотреть лишь случай  $i \leq j$ :

$$a_{ij} = \sum_{k=1}^{i-1} s_{ki}d_{kk}s_{kj} + s_{ii}d_{ii}s_{ij} + \sum_{k=i+1}^n s_{ki}d_{kk}s_{kj}.$$

Но  $s_{ki} = 0$  при  $k \geq i+1$ , поэтому

$$a_{ij} = \sum_{k=1}^{i-1} s_{ki}d_{kk}s_{kj} + s_{ii}d_{ii}s_{ij}, \quad i \leq j.$$

Так, при  $i = j$

$$a_{ii} = \sum_{k=1}^{i-1} d_{kk}s_{ki}^2 + s_{ii}^2d_{ii}, \quad i = 1, 2, \dots, n.$$

Следовательно,

$$d_{ii} = \operatorname{sign}\left(\frac{1}{s_{ii}^2} \left(a_{ii} - \sum_{k=1}^{i-1} d_{kk} s_{ki}^2\right)\right),$$

$$s_{ii} = \left|a_{ii} - \sum_{k=1}^{i-1} d_{kk} s_{ki}^2\right|^{1/2}, \quad i = 1, 2, \dots, n.$$

При  $i < j$

$$s_{ij} = \frac{1}{s_{ii} d_{ii}} \left( a_{ij} - \sum_{k=1}^{i-1} s_{ki} d_{kk} s_{kj} \right).$$

Расчет осуществляется по следующему алгоритму:

$$s_{11} = |a_{11}|^{1/2}, \quad d_{11} = \operatorname{sign}\left(\frac{a_{11}}{s_{11}^2}\right), \quad s_{1j} = \frac{a_{1j}}{s_{11} d_{11}}, \quad j = 2, 3, \dots, n;$$

$$s_{22} = |a_{22} - d_{11} s_{12}^2|^{1/2}, \quad d_{22} = \operatorname{sign}\left(\frac{a_{22} - d_{11} s_{22}}{s_{22}^2}\right),$$

$$s_{2j} = \frac{a_{2j} - s_{12} d_{22} s_{1j}}{s_{22} d_{22}}, \quad j = 3, 4, \dots, n.$$

Далее проводим расчет при  $i = 3, 4, \dots, n$ .

После разложения, т. е. представления  $A$  в виде  $A = S^* DS$ , решение системы  $Ax = f$  сводится к решению трех систем — двух треугольных и одной диагональной:

$$S^* z = f, \quad Dy = z, \quad Sx = y \quad (S^* D y = f).$$

После исключения  $z$  имеем:

$$y_1 = \frac{f_1}{s_{11} d_{11}},$$

$$y_i = \frac{1}{s_{ii} d_{ii}} \left( f_i - \sum_{j=1}^{i-1} s_{ji} d_{jj} y_j \right), \quad i = 2, 3, \dots, n;$$

$$x_n = \frac{y_n}{s_{nn}},$$

$$x_i = \frac{1}{s_{ii}} \left( y_i - \sum_{j=i+1}^n s_{ij} x_j \right), \quad i = n-1, n-2, \dots, 1.$$

Число действий в методе квадратного корня примерно в два раза меньше числа действий в методе Гаусса за счет использования данных о структуре матрицы  $A$ , т. е. ее симметрии.

## 1.6. Итерационные методы решения СЛАУ

Рассмотрим систему  $Ax = f$ , где  $A$  — матрица размером  $n \times n$ ;  $x, f$  —  $n$ -векторы.

Как уже отмечалось ранее, практически все числа в памяти ЭВМ хранятся в приближенном виде. Проводимые вычисления ведут к появлению новых ошибок. При этом ошибка результата дополнительно зависит еще и от свойств матрицы системы уравнений — от числа ее обусловленности. Все это приводит к тому, что результат решения системы практически никогда не бывает совершенно точным. Поэтому есть смысл в рассмотрении таких методов решения данного уравнения, которые, вообще говоря, даже без учета ошибок округления дают лишь некоторое приближение  $\tilde{x}$  к точному значению решения. При определенных условиях это приближение можно сделать весьма близким к точному.

### 1.6.1. Каноническая форма одношаговых итерационных методов

*Одношаговые* (или *двухслойные* — нужно помнить результат только одной последней итерации) *итерационные методы* характеризуются следующей канонической формой записи:

$$B_{k+1} \frac{x_{k+1} - x_k}{\tau_{k+1}} + Ax_k = f, \quad k = 0, 1, \dots$$

Здесь  $B_{k+1}$  — обратимая матрица, задающая метод;  $x_k$  —  $k$ -е приближение решения;  $k$  — номер итерации;  $\tau_{k+1}$  — итерационный параметр. Считается известным начальное приближение  $x_0$ . Оператор  $B_{k+1}$  часто называют *предобуславливателем* данной системы уравнений. Оператор  $B_{k+1}^{-1}$  должен вычисляться сравнительно просто. Тогда решение ищется по следующим формулам:

$$x_{k+1} = x_k + \tau_{k+1} B_{k+1}^{-1} (f - Ax_k).$$

Такое определение задает целое семейство методов. Выбор матрицы  $B_{k+1}$ , задающей метод, и итерационного параметра  $\tau_{k+1}$  позволяет варьировать свойства итерационного процесса. Ясно, что с точки зрения простоты реализации наилучшим является выбор единичного оператора  $B_{k+1} = E$  и постоянного параметра. Но с точки зрения скорости наилучшим является выбор  $B_{k+1} = A$ . В этом случае при  $\tau_{k+1} = 1$  точное решение будет найдено за одну итерацию. Очевидно, однако, что

последний вариант является бессмысленным в рассматриваемом контексте: оператор  $B_{k+1}$  должен легко обращатьсяся. Если это так, то использование итерационного процесса для обращения просто обращающейся матрицы  $A$  является излишним.

Из сказанного очевидна важность решения задачи выбора (или расчета) как оператора  $B_{k+1}$ , так и итерационных параметров, что принципиально может позволить оптимизировать процесс численного решения исходной системы. Разнообразие критерииов оптимальности при этом добавляет степеней свободы.

**Определение 1.38.** *Итерационный метод* называется **явным**, если  $B_{k+1} = E$ , и **неявным** в противном случае.

**Определение 1.39.** *Итерационный метод* называется **стационарным**, если операторы  $B_{k+1} = B$  и параметры  $\tau_{k+1} = \tau$  не зависят от  $k$ , и **нестационарным** в противном случае.

### 1.6.2. Примеры одношаговых итерационных методов

**Пример 1.2.** *Метод простой итерации:*

$$\frac{x_{k+1} - x_k}{\tau} + Ax_k = f.$$

Данный алгоритм соответствует случаю  $B = E$  и постоянному  $\tau$ .

**Пример 1.3.** Метод

$$\frac{x_{k+1} - x_k}{\tau_{k+1}} + Ax_k = f,$$

соответствующий специальному выбору параметров  $\tau_{k+1}$ , таких, что норма разности  $x_m - x$  ( $m$  — заранее заданный номер итерации,  $x$  — точное решение) является минимальной, называется **методом Ричардсона с чебышевскими параметрами**.

Отсюда следует, что данный метод является оптимальным среди всех итерационных методов, делающих  $m$  итерационных шагов. Рассмотрим его подробнее и найдем набор итерационных параметров, обеспечивающих оптимальность. Пусть  $z_k = x_k - x$  — погрешность  $k$ -й итерации. Поскольку  $f = Ax$ , имеем уравнение для погрешности

$$\frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0.$$

Отсюда получаем

$$z_{k+1} = (E - \tau_{k+1}A)z_k, \quad z_m = \prod_{k=1}^m (E - \tau_k A)z_0.$$

Минимальность нормы  $z_m$  означает, что итерационные параметры выбираются такими, чтобы норма оператора перехода от начальной ошибки  $z_0$  к конечной была минимальна:

$$\|z_m\| = \min_{\tau_1, \dots, \tau_m} \left\| \prod_{k=1}^m (E - \tau_k A) z_0 \right\| \leq \min_{\tau_1, \dots, \tau_m} \left\| \prod_{k=1}^m (E - \tau_k A) \right\| \|z_0\|.$$

Пока не делалось никаких предположений относительно свойств оператора  $A$ . Представим окончательное решение задачи выбора итерационных параметров для простейшего случая самосопряженного оператора  $A$ , удовлетворяющего условию

$$\gamma_1 E \leq A \leq \gamma_2 E, \quad \gamma_1 > 0.$$

Здесь  $\gamma_1, \gamma_2$  — постоянные энергетической эквивалентности операторов  $A$  и  $E$ . В данном случае величины  $\gamma_1, \gamma_2$  — просто границы  $\lambda_{\min}, \lambda_{\max}$  спектра оператора  $A$ . В условиях самосопряженности норма оператора равна его максимальному собственному числу. В результате имеем

$$\|z_m\| \leq \min_{\tau_1, \dots, \tau_m} \max_{\gamma_1 \leq t \leq \gamma_2} \left| \prod_{k=1}^m (1 - \tau_k t) \right| \|z_0\|.$$

Следовательно, задача поиска набора итерационных параметров свелась к нахождению полинома  $P_m(t)$  порядка  $m$ , равного единице при  $t = 0$  и наименее уклоняющегося от нуля, а именно:

$$\min_{\tau_1, \dots, \tau_m} \max_{\gamma_1 \leq t \leq \gamma_2} \left| \prod_{k=1}^m (1 - \tau_k t) \right| = \max_{\gamma_1 \leq t \leq \gamma_2} |P_m(t)| = q_m, \quad \|z_m\| \leq q_m \|z_0\|.$$

Решение данной задачи получено В.А. Марковым в 1892 г. Искомый полином имеет вид

$$P_m(t) = q_m T_m \left( \frac{1 - \tau_0 t}{\rho_0} \right),$$

где

$$q_m = \left( T_m \left( \frac{1}{\rho_0} \right) \right)^{-1} = \frac{2\rho_1^m}{1 + \rho_1^{2m}}; \quad T_m(x) = \cos(m \arccos x) \quad (|x| \leq 1);$$

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}; \quad \rho_0 = \frac{1 - \eta}{1 + \eta}, \quad \rho_1 = \frac{1 - \eta^{1/2}}{1 + \eta^{1/2}}, \quad \eta = \frac{\gamma_1}{\gamma_2}.$$

Здесь  $T_m(x)$  — **полином Чебышева** первого рода степени  $m$ .

Из условия совпадения корней искомого полинома и полинома Чебышева получаем формулу для итерационных параметров:

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k = \cos\left(\frac{2k-1}{2m}\pi\right), \quad k = 1, 2, \dots, m.$$

где  $\mu_k$  — корни полинома Чебышева. Поэтому полученный набор итерационных параметров называется чебышевским.

К сожалению, метод Ричардсона с чебышевскими параметрами является неустойчивым при выполнении действий с произвольно упорядоченным набором параметров. Для получения устойчивого алгоритма расчет необходимо выполнять при специально выбранном упорядочении данного набора (см. 1.13).

Укажем на неочевидно решаемый вопрос о выборе числа итераций в методе Ричардсона. Обычно на практике его решают путем проведения расчетов циклическим образом: задавшись некоторым числом, выполняют вычисления циклами до момента достижения заданной точности по некоторому критерию.

Отметим, что существуют неявные варианты метода Ричардсона и более общие варианты его обоснования.

**Пример 1.4.** Представим  $A = A_1 + D + A_2$ , где  $A_1$  — нижняя треугольная матрица;  $D = \text{diag}(a_{ii})$ ;  $A_2$  — верхняя треугольная матрица. Запишем решаемое уравнение в виде

$$A_1x + Dx + A_2x = f.$$

Алгоритм **метода Якоби** задается следующим образом:

$$x_{k+1} = D^{-1}(f - A_1x_k - A_2x_k),$$

или

$$Dx_{k+1} + (A - D)x_k = f, \quad D(x_{k+1} - x_k) + Ax_k = f.$$

Это соответствует случаю  $B = D$ ,  $\tau = 1$ .

**Пример 1.5.** Алгоритм **метода Зейделя** определен соотношениями

$$(A_1 + D)x_{k+1} + A_2x_k = f,$$

или

$$(A_1 + D)(x_{k+1} - x_k) + Ax_k = f.$$

Это соответствует случаю  $B = A_1 + D$ ,  $\tau = 1$ .

**Пример 1.6. Метод верхней релаксации:**

$$(D + \omega A_1) \frac{x_{k+1} - x_k}{\omega} + Ax_k = f,$$

т. е. в данном случае  $B = D + \omega A_1$ ,  $\tau = \omega$ . Этот метод является обобщением метода Зейделя. #

Итерации проводятся, как правило, до выполнения одного из условий:

$$\|x_{k+1} - x_k\| \leq \varepsilon, \quad \|x_{k+1} - x_k\| \leq \varepsilon \|x_k\| + \varepsilon_0, \quad \left\| \frac{x_{k+1} - x_k}{\|x_k\| + \varepsilon_0} \right\| \leq \varepsilon.$$

Последнее условие является самым строгим и жестким.

Указанные условия прерывания итерационного процесса оперируют не нормой погрешности численного решения, а нормами его изменения за одну итерацию. Иногда это приводит к неверному заключению о сходимости метода, если, например, метод очень медленно сходится.

Тогда может оказаться успешным применение другого условия:

$$\|Ax_{k+1} - f\| \leq \varepsilon,$$

которое связано с некоторой операторной нормой погрешности, а именно, нормой невязки. Данный критерий часто называют критерием по невязке. При этом необходимо отметить, что в случае малости нормы оператора  $A$  данный критерий также может оказаться неприемлемым.

Так как сама погрешность численного решения является неизвестной, идеальный критерий прерывания процесса указать невозможно. Следовательно, его выбор чаще всего определяется умением и навыками вычислителя. На практике обычно используют некоторый набор критериев, проверяя сходимость несколькими способами одновременно.

### 1.6.3. Условия сходимости стационарных итерационных методов

**Теорема 1.1.** Пусть  $A$  — симметричная положительно определенная матрица,  $\tau > 0$  и выполнено неравенство

$$B - 0,5\tau A > 0.$$

Тогда стационарный итерационный метод

$$B \frac{x_{k+1} - x_k}{\tau} + Ax_k = f$$

сходится.

◀ Пусть  $z_k = x_k - x$  — погрешность на  $k$ -й итерации. Так как  $f = Ax$ , то

$$B \frac{z_{k+1} - z_k}{\tau} + Az_k = 0.$$

Нужно показать, что норма погрешности стремится к нулю при  $k \rightarrow \infty$ . Проделаем ряд преобразований:

$$\begin{aligned} z_{k+1} &= (E - \tau B^{-1} A) z_k, \quad Az_{k+1} = (A - \tau AB^{-1} A) z_k, \\ (Az_{k+1}, z_{k+1}) &= ((A - \tau AB^{-1} A) z_k, (E - \tau B^{-1} A) z_k) = \\ &= (Az_k, z_k) - \tau (Az_k, B^{-1} Az_k) - \tau (AB^{-1} Az_k, z_k) + \\ &\quad + \tau^2 (AB^{-1} Az_k, B^{-1} Az_k). \end{aligned}$$

В силу симметрии  $A$

$$(AB^{-1} Az_k, z_k) = (B^{-1} Az_k, Az_k).$$

Следовательно,

$$\begin{aligned} J_{k+1} &= (Az_{k+1}, z_{k+1}) = J_k - 2\tau (Az_k, B^{-1} Az_k) + \\ &\quad + \tau^2 (AB^{-1} Az_k, B^{-1} Az_k) = J_k - 2\tau ((B - 0,5\tau A)B^{-1} Az_k, B^{-1} Az_k), \end{aligned}$$

где  $J_k = (Az_k, z_k)$ .

Если  $B - 0,5\tau A > 0$ , то  $J_{k+1} \leq J_k$ ,  $J_k \geq 0$ , так как  $A > 0$ . Значит, последовательность  $J_k$  монотонно не возрастает и ограничена нулем снизу. Поэтому существует

$$\lim_{k \rightarrow \infty} J_k = J.$$

Положительная определенность матрицы  $B - 0,5\tau A$  означает, что существует такое  $\delta > 0$ , при котором

$$((B - 0,5\tau A)y, y) \geq \delta \|y\|^2.$$

Следовательно,

$$J_{k+1} - J_k + 2\tau \delta \|B^{-1} Az_k\|^2 \leq 0,$$

откуда при  $k \rightarrow \infty$  получим

$$\lim_{k \rightarrow \infty} \|B^{-1} Az_k\|^2 = 0.$$

Обозначим  $\omega_k = B^{-1}Az_k$ , тогда  $z_k = A^{-1}B\omega_k$ . Отсюда

$$\|z_k\| \leq \|A^{-1}B\|\|\omega_k\|$$

и  $\lim_{k \rightarrow \infty} \|z_k\| = 0$ . ▶

**Следствие 1.1.** Пусть  $A$  — симметричная положительно определенная матрица с диагональным преобладанием, т. е.

$$a_{ii} > \sum_{j \neq i} |a_{ij}|, \quad i = 1, 2, \dots, n.$$

Тогда метод Якоби сходится.

◀ Условие сходимости имеет в данном случае вид

$$D - 0,5A > 0,$$

т. е.  $A < 2D$ . Рассмотрим положительно определенную форму

$$(Ax, x) = \sum_{i,j} a_{ij}x_i x_j.$$

Для нее имеем оценку

$$\begin{aligned} (Ax, x) &\leq \frac{1}{2} \sum_{i,j} |a_{ij}|x_i^2 + \frac{1}{2} \sum_{i,j} |a_{ij}|x_j^2 = \frac{1}{2} \sum_{i,j} |a_{ij}|x_i^2 + \frac{1}{2} \sum_{i,j} |a_{ji}|x_i^2 = \\ &= \frac{1}{2} \sum_{i,j} (|a_{ij}| + |a_{ji}|)x_i^2 = \sum_{i,j} |a_{ij}|x_i^2. \end{aligned}$$

Последнее равенство верно в силу симметричности  $A$ . Отсюда

$$(Ax, x) \leq \sum_{i,j} |a_{ij}|x_i^2 = \sum_{i=1}^n x_i^2 \left( |a_{ii}| + \sum_{j \neq i} |a_{ij}| \right).$$

Но вследствие положительной определенности  $a_{ii} > 0$ ,  $i = 1, 2, \dots, n$ . Используя далее условие диагонального преобладания, имеем

$$(Ax, x) < \sum_{i=1}^n x_i^2 (a_{ii} + a_{ii}) = 2(Dx, x),$$

т. е.  $A < 2D$ . ▶

**Следствие 1.2.** Пусть  $A$  — симметричная положительно определенная матрица. Тогда метод верхней релаксации сходится при  $0 < \omega < 2$ . В частности, метод Зейделя ( $\omega = 1$ ) сходится. Рассматривается действительный случай.

◀ Для данного метода  $B = D + \omega A_1$ ,  $\tau = \omega$ ,  $A = A_1 + D + A_2$ . В случае симметричной матрицы имеем  $A_1^* = A_2$ . Нужно показать, что

$$(D + \omega A_1) - 0,5\omega(A_1 + D + A_2) > 0.$$

При  $0 < \omega < 2$

$$\begin{aligned} ((B - 0,5\tau A)x, x) &= (1 - 0,5\omega)(Dx, x) + \\ &\quad + 0,5\omega(A_1x, x) - 0,5\omega(A_2x, x) = (1 - 0,5\omega)(Dx, x) > 0, \end{aligned}$$

так как  $D$  — положительно определенная матрица. Случай  $\omega \leq 0$  не рассматривается, поскольку по условию теоремы 1.1 параметр  $\tau > 0$ . ►

**Следствие 1.3.** Метод простой итерации сходится при  $\tau < \frac{2}{\lambda_{\max}}$ , где  $\lambda_{\max}$  — максимальное собственное значение симметричной положительно определенной матрицы  $A$ .

◀ Условие  $B - 0,5\tau A > 0$  в данном случае есть  $E - 0,5\tau A > 0$ , что эквивалентно условию положительности минимального собственного значения матрицы  $E - 0,5\tau A$ , т. е.  $1 - 0,5\tau\lambda_{\min} > 0$ . Отсюда  $\tau < \frac{2}{\lambda_{\min}}$ . ►

## 1.7. Итерационные методы решения СЛАУ вариационного типа

Описанные в 1.6 итерационные методы сходятся медленно. Лучший из них (*метод Ричардсона*) требует знания границ оператора  $A$ : минимального  $\lambda_{\min}$  и максимального  $\lambda_{\max}$  собственных значений. Даже *метод простой итерации* требует знания  $\lambda_{\max}$ . Рассмотрим *итерационные методы* так называемого *вариационного типа* решения системы линейных алгебраических уравнений (СЛАУ), которые не требуют знания границ оператора.

Каноническая форма методов такова:

$$B \frac{x_{k+1} - x_k}{\tau_{k+1}} + Ax_k = f, \quad k = 0, 1, \dots$$

Рассмотрим итерационные методы канонического вида с постоянными  $B_{k+1} = B$  и переменными итерационными параметрами  $\tau_{k+1}$ . Выберем их так, чтобы при переходе от  $k$ -го слоя к  $(k+1)$ -му стала минимальной норма

$$\|z_{k+1}\|_D = (Dz_{k+1}, z_{k+1})^{1/2},$$

где  $z_{k+1} = x_{k+1} - x$  — погрешность решения на  $(k+1)$ -й итерации;  $D$  — некоторый строго положительно определенный самосопряженный линейный оператор, задающий (кроме  $B$ ) рассматриваемый метод.

Выше рассмотрен метод Ричардсона с чебышевскими параметрами, обеспечивающий минимальность погрешности через заданное число итераций  $t$ . На первый взгляд кажется, что проведение вычислений таким образом, чтобы на каждом шаге получалась минимальная погрешность, является лучшим способом нахождения решения. Однако это не так. Для пояснения приведем следующий пример: пусть шарик, находящийся в некотором начальном положении 0 (рис. 1.4), необходимо перевести в положение с минимальной энергией, т. е. опустить вниз).

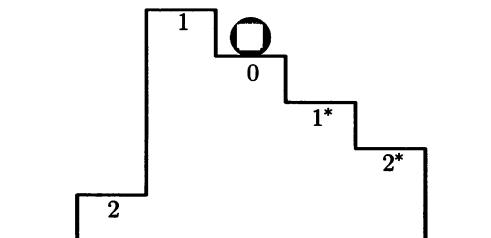


Рис. 1.4

Сравним результаты двух шагов в соответствии с приведенным рисунком. Если требовать минимальности энергии на каждом шаге, то шарик должен двигаться по пути  $0-1^*-2^*$ . Если же потребовать минимальности через два шага, то получим маршрут  $0-1-2$ . Итог окажется заметно лучше, чем в первом варианте, что вполне объяснимо: во втором варианте шаг  $0-1$  является заведомо плохим с точки зрения локальной (одношаговой) оптимальности. Однако после него шарик занимает очень удобную позицию, стартуя с которой достигает минимума энергии.

Если требовать минимальности энергии на каждом шаге, то шарик должен двигаться по пути  $0-1^*-2^*$ . Если же потребовать минимальности через два шага, то получим маршрут  $0-1-2$ . Итог окажется заметно лучше, чем в первом варианте, что вполне объяснимо: во втором варианте шаг  $0-1$  является заведомо плохим с точки зрения локальной (одношаговой) оптимальности. Однако после него шарик занимает очень удобную позицию, стартуя с которой достигает минимума энергии.

### 1.7.1. Расчетные формулы

Рассмотрим действительный случай. Если перейти к погрешности  $z_k$ , то получим

$$B \frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0,$$

откуда

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A) z_k$$

и

$$Dz_{k+1} = D(E - \tau_{k+1} B^{-1} A) z_k.$$

В результате

$$\begin{aligned} \|z_{k+1}\|_D^2 &= (Dz_k - \tau_{k+1} DB^{-1} Az_k, z_k - \tau_{k+1} B^{-1} Az_k) = \\ &= \|z_k\|_0^2 - \tau_{k+1} (DB^{-1} Az_k, z_k) - \tau_{k+1} (Dz_k, B^{-1} Az_k) + \\ &\quad + \tau_{k+1}^2 (DB^{-1} Az_k, B^{-1} Az_k). \end{aligned}$$

Поскольку  $D$  — самосопряженный оператор, то второй и третий члены в правой части полученного равенства совпадают. Квадратный относительно итерационного параметра трехчлен дает условие минимума для  $\|z_{k+1}\|_D$ , из которого получаем следующее значение итерационного параметра:

$$\tau_{k+1} = \frac{(B^{-1}Az_k, Dz_k)}{(DB^{-1}Az_k, B^{-1}Az_k)}.$$

Следовательно,

$$\|z_{k+1}\|_D^2 = \|z_k\|_D^2 - \frac{(Dz_k, B^{-1}Az_k)^2}{(DB^{-1}Az_k, B^{-1}Az_k)} \leq \|z_k\|_D^2.$$

Введем обозначения:  $r_k = Az_k = Ax_k - f$  — **невязка** решения,  $w_k = B^{-1}r_k = B^{-1}(Ax_k - f)$  — **поправка** решения на  $k$ -й итерации, тогда

$$\tau_{k+1} = \frac{(w_k, Dz_k)}{(Dw_k, w_k)}, \quad x_{k+1} = x_k - \tau_{k+1}w_k.$$

Алгоритм решения выглядит следующим образом: по заданному  $x_k$  вычисляется невязка  $r_k$ , по ней — поправка  $w_k = B^{-1}r_k$ , далее вычисляются  $\tau_{k+1}$  и  $x_{k+1}$ . Потом цикл повторяется.

Словесное описание является ясным и понятным. Однако анализ полученных формул показывает, что данный алгоритм реализуем далеко не всегда: по заданному  $x_k$  без особого труда определяется невязка  $r_k$ , а по ней — поправка  $w_k$ . Для произвольного же оператора  $D$ , удовлетворяющего описанным выше условиям, вычислить  $Dz_k$  нельзя, так как погрешность  $z_k$  неизвестна. В результате найти параметр  $\tau_{k+1}$  невозможно. Таким образом, практически реализуемыми являются варианты метода с такими операторами  $D$ , для которых можно вычислить  $Dz_k$ . Как будет видно далее, обычно эти операторы сами представляют собой произведение некоторых операторов и имеют вид  $D = \dots A$ . В этом случае  $Dz_k = \dots r_k$ . В результате новый итерационный параметр может быть вычислен.

### 1.7.2. Оценка скорости сходимости

Введем обозначение  $y_k = D^{1/2}z_k$ , тогда  $\|y_k\|^2 = \|z_k\|_D^2$  ( $D$  и  $D^{1/2}$  — самосопряженные операторы). В этих обозначениях

$$z_{k+1} = (E - \tau_{k+1}B^{-1}A)z_k.$$

Следовательно,

$$\begin{aligned} y_{k+1} &= (E - \tau_{k+1}D^{1/2}B^{-1}AD^{-1/2})y_k = (E - \tau_{k+1}C)y_k, \\ C &= D^{1/2}B^{-1}AD^{-1/2} = D^{-1/2}(DB^{-1}A)D^{-1/2}. \end{aligned}$$

В результате

$$\|z_{k+1}\|_D^2 = \|(E - \tau_{k+1}C)y_k\|^2 = \|y_{k+1}\|^2.$$

Поскольку  $\tau_{k+1}$  выбирается из условия минимума величины  $\|z_{k+1}\|_D$ ,

$$\begin{aligned} \|z_{k+1}\|_D &= \min_{\tau_{k+1}} \|(E - \tau_{k+1}C)y_k\| \leq \\ &\leq \min_{\tau_{k+1}} \|E - \tau_{k+1}C\| \|y_k\| = \min_{\tau_{k+1}} \|E - \tau_{k+1}C\| \|z_k\|_D. \end{aligned}$$

В итоге

$$\|z_m\|_D \leq \rho^m \|z_0\|_D, \quad \rho = \min_{\tau} \|E - \tau C\|.$$

**Теорема 1.2.** Пусть оператор  $DB^{-1}A$  является самосопряженным и выполнены условия

$$\gamma_1 D \leq DB^{-1}A \leq \gamma_2 D,$$

где  $\gamma_1, \gamma_2$  — положительные постоянные. Тогда

$$\rho = \frac{1 - \eta}{1 + \eta}, \quad \eta = \frac{\gamma_1}{\gamma_2}, \quad \|z_m\|_D \leq \left( \frac{1 - \eta}{1 + \eta} \right)^m \|z_0\|_D.$$

◀ Так как

$$\gamma_1(Dx, x) \leq (DB^{-1}Ax, x) \leq \gamma_2(Dx, x), \quad x = D^{-1/2}y,$$

то получим

$$\gamma_1(y, y) \leq (D^{-1/2}DB^{-1}AD^{-1/2}y, y) \leq \gamma_2(y, y),$$

откуда  $\gamma_1 E \leq C \leq \gamma_2 E$ . Следовательно,  $-\gamma_2 \tau E \leq -\tau C \leq -\gamma_1 \tau E$  при  $\tau \geq 0$ . Далее при  $\tau \geq 0$  получаем

$$(1 - \gamma_2 \tau)E \leq E - \tau C \leq (1 - \gamma_1 \tau)E.$$

Поскольку

$$\tau_{k+1} = \frac{(Cy_k, y_k)}{(Cy_k, Cy_k)},$$

то  $\tau_{k+1} \geq 0$ . Поэтому вычисление нормы проводится при  $\tau \geq 0$ . Легко видеть, что оператор  $C$  в силу сделанных предположений является самосопряженным. Поэтому оценка оператора  $E - \tau C$  позволяет получить оценку его спектра: спектр лежит на отрезке  $[1 - \gamma_2 \tau, 1 - \gamma_1 \tau]$ .

Следовательно, так как в самосопряженном случае норма есть максимальное по модулю собственное число, то

$$\|E - \tau C\| \leq \max \{|1 - \gamma_1 \tau|, |1 - \gamma_2 \tau|\} = N(\tau).$$

Найдем параметр  $\tau$ , минимизирующий  $N(\tau)$  (искомая оценка есть минимум указанной величины по  $\tau$ ). В точке  $A$  (рис. 1.5) выполнено условие  $1 - \gamma_1 \tau = -1 + \gamma_2 \tau$ , т. е.  $\tau = 2/(\gamma_1 + \gamma_2)$ . Отсюда получаем, что

$$\rho \leq \min_{\tau} N(\tau) = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \frac{1 - \eta}{1 + \eta}. \quad \blacktriangleright$$

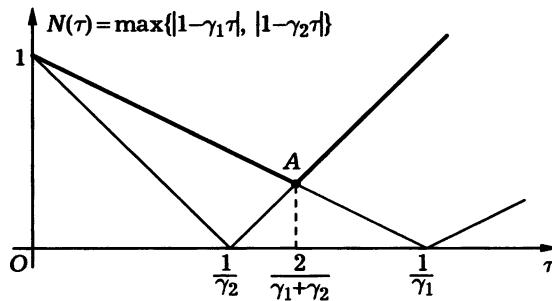


Рис. 1.5

**Следствие 1.4.** Оптимальным параметром  $\tau$  в обобщенном методе простых итераций

$$B \frac{x_{k+1} - x_k}{\tau} + Ax_k = f,$$

минимизирующим норму  $\|z_k\|_D$  с операторами  $A$ ,  $B$ ,  $D$ , которые удовлетворяют условиям теоремы 1.2, является  $\tau = 2/(\gamma_1 + \gamma_2)$ .

### 1.7.3. Частные случаи методов

В 1.7.1 при выводе общих формул уже обсуждалась реализуемость итерационных методов вариационного типа. Ниже приведены примеры методов, для которых все необходимые величины можно вычислить практически.

1. Если  $D = A^*A$ , то  $\|z_k\|_D^2 = (A^*Az_k, z_k) = (Az_k, Az_k) = (r_k, r_k)$ . В результате получаем **метод минимальных невязок** ( $A$  — невырожденный оператор). Параметры метода:

$$\tau_{k+1} = \frac{(w_k, A^*r_k)}{(Aw_k, Aw_k)} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}.$$

2. Если  $D = A = A^* > 0$ , то  $\|z_k\|_D^2 = (Az_k, z_k)$  — норма погрешности в энергетическом пространстве. При этом оператор  $A$  должен быть самосопряженным положительно определенным. В результате имеем **метод скорейшего спуска**. Параметры метода:

$$\tau_{k+1} = \frac{(w_k, r_k)}{(Aw_k, w_k)}.$$

3. Если

$$\begin{aligned}\|z_k\|_D^2 &= (Bw_k, w_k) = (BB^{-1}Az_k, B^{-1}Az_k) = \\ &= (Az_k, B^{-1}Az_k) = (A^*B^{-1}Az_k, z_k),\end{aligned}$$

то получим **метод минимальных поправок**. В этом случае  $D = A^*B^{-1}A$ ,  $Dz_k = A^*B^{-1}r_k = A^*w_k$ , где  $B$  — самосопряженный положительный оператор. Тогда

$$\tau_{k+1} = \frac{(w_k, A^*w_k)}{(B^{-1}Aw_k, Aw_k)}.$$

4. **Методу минимальных погрешностей** соответствует  $D = B_0 = A^*B$ ,  $B_0 = B_0^*$ . Тогда  $B = (A^*)^{-1}B_0$ ,

$$\|z_k\|_D^2 = \|z_k\|_{B_0}^2 = (A^*Bz_k, z_k)$$

и

$$Dz_k = B_0z_k, \quad w_k = B^{-1}r_k = B_0^{-1}A^*r_k.$$

Следовательно,

$$\begin{aligned}(w_k, Dz_k) &= (B_0^{-1}A^*r_k, B_0z_k) = (r_k, Az_k) = (r_k, r_k), \\ Dw_k &= A^*BB^{-1}Az_k = A^*r_k, \quad (Dw_k, w_k) = (r_k, Aw_k),\end{aligned}$$

откуда

$$\tau_{k+1} = \frac{(r_k, r_k)}{(r_k, Aw_k)}.$$

Во всех четырех случаях считается, что операторы  $A$ ,  $B$  ( $B_0$ ) таковы, что  $D$  — самосопряженный положительный оператор,  $DB^{-1}A$  — самосопряженный оператор и что выполнено приведенное выше неравенство с  $\gamma_1, \gamma_2 > 0$ . Тогда для уменьшения начальной погрешности (в норме  $D$ ) в  $\varepsilon^{-1}$  раз требуется выполнить  $n_0(\varepsilon) = \frac{\ln(\varepsilon^{-1})}{\ln(\rho^{-1})}$  итераций.

Проблема заключается в том, что в реальных задачах коэффициент  $\rho$  очень близок к единице (например, на практике отношение  $\gamma_1/\gamma_2 \sim 10^{-4}$  считается достаточно большим).

## 1.8. Методы сопряженных направлений

Приведем расчетные формулы с их кратким пояснением для **двухшаговых (трехслойных) итерационных методов — методов сопряженных направлений**. При их реализации необходимо знать решения на двух предыдущих итерациях. Рассматриваемые методы принадлежат к следующему классу итерационных методов:

$$B \frac{(x_{k+1} - x_k) + (1 - \alpha_{k+1})(x_k - x_{k-1})}{\tau_{k+1} \alpha_{k+1}} + Ax_k = f, \quad k = 1, 2, \dots$$

В отличие от одношаговых методов здесь имеется дополнительный итерационный параметр  $\alpha_{k+1}$ . Кроме того, начать расчеты можно только при наличии  $x_0, x_1$ , т. е. двух начальных приближений. Первое из них обычно берется произвольным, а второе вычисляется путем выполнения одной итерации *однослоиного итерационного метода* вариационного типа того же вида, но при  $k = 0, \alpha_1 = 1$ . Если два предыдущих приближения известны и параметры найдены, то новое итерационное приближение находится без особого труда (оператор  $B^{-1}$  должен быть сравнительно легко обратим).

Итерационные параметры выберем так, чтобы при переходе от нулевого слоя к произвольному  $m$ -му операторная норма погрешности  $\|z_m\|_D = (Dz_m, z_m)^{1/2}$  стала минимальной из всех возможных. Здесь  $z_k = x_k - x$  — погрешность решения на  $k$ -й итерации;  $D$  — некоторый *строго положительно определенный оператор*, задающий (кроме  $B$ ) рассматриваемый метод;  $D$  — линейный *самосопряженный оператор*.

Постановка задачи о минимизации погрешности соответствует постановке задачи о построении *метода Ричардсона с чебышевскими параметрами*. Из этого следует, что оценка нормы погрешности после  $m$  итераций в методе сопряженных направлений будет не хуже, чем оценка ошибки метода Ричардсона, приведенная выше.

Решение задачи минимизации дает следующие формулы для итерационных параметров:

$$\begin{aligned} \tau_{k+1} &= \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \dots, \\ \alpha_1 &= 1, \quad \alpha_{k+1} = \left( 1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Dw_k, z_k)}{(Dw_k, w_k)} \frac{1}{\alpha_k} \right)^{-1}, \quad k = 1, 2, \dots \end{aligned}$$

Здесь используется тот же набор обозначений, что и в случае одношаговых (двухслойных) итерационных методов вариационного типа. Их родство состоит не только в этом. Из расчетных формул следует, что выражения для итерационных параметров  $\tau_{k+1}$  одношаговых и

двуухшаговых методов совпадают, а вычисление  $\alpha_{k+1}$  никаких новых скалярных произведений не требует. Таким образом, для нахождения итерационных параметров в методах сопряженных направлений используется практически то же самое количество действий.

Вычисление нового итерационного приближения несколько более трудоемко, чем в случае однослойных методов. Однако это вполне компенсируется значительно большей скоростью сходимости. Как указывалось выше, она не ниже, чем у метода Ричардсона с чебышевскими параметрами.

Более того, в конечномерном пространстве методы сопряженных направлений (при естественных ограничениях, накладываемых на применяемые операторы) сходятся за число итераций, не превышающее размерности пространства.

Без излишней детализации приведем частные случаи методов сопряженных направлений. Каждый из них имеет свой одношаговый аналог.

1. Если  $D = A^*A$ , имеем **метод сопряженных невязок**.
2. Если  $D = A = A^* > 0$ , имеем **метод сопряженных градиентов**.
3. Если  $D = A^*B^{-1}A$ , получается **метод сопряженных поправок**.
4. Если  $D = B_0 = A^*B$ , имеем **метод сопряженных погрешностей**.

Во всех случаях считается, что операторы  $A$ ,  $B$  ( $B_0$ ) таковы, что  $D$  — самосопряженный положительный оператор,  $DB^{-1}A$  — самосопряженный оператор и что выполнено приведенное в 1.7 неравенство с  $\gamma_1, \gamma_2 > 0$ .

Описанные методы обладают рядом замечательных свойств, часть из которых уже описана выше.

Отметим, что при решении плохо обусловленных задач вследствие наличия ошибок округления метод может и не сойтись за число итераций, равное порядку системы. В этом случае необходимо предпринимать специальные меры.

Заметим также, что двухшаговые методы часто проявляют свойство «насыщения» погрешности: норма ошибки численного решения быстро убывает на первых итерациях, но затем выходит на практический постоянное ненулевое значение («полочку»). В этом случае иногда помогает прерывание итерационного цикла и его повтор с полученного приближения. При этом снова делается один шаг по формулам одношагового метода, после чего повторяются итерации двухшагового.

## 1.9. Итерационное уточнение решения

При численном решении системы линейных алгебраических уравнений (СЛАУ)  $Ax = f$  *прямым* или *итерационным методом* получаемое решение является только приближением к точному. При этом чем «хуже» матрица системы (больше ее *число обусловленности*), тем дальнее полученное решение от точного. Полученная погрешность зависит, конечно же, и от разрядности ЭВМ, качества программной реализации алгоритма и т. д. Но в любом случае решение является неточным.

Повысить точность приближенного решения можно способом *итерационного уточнения*. Выполним следующие этапы до сходимости (по некоторому критерию), положив для начала  $x_0 = 0$ . Итак, имеем цикл по  $k$  от 0 с указанным начальным приближением.

1. Находим невязку  $r_k = f - Ax_k$ .
2. Нормируем невязку, вычисляя  $b_k = \frac{r_k}{\|r_k\|}$ . В результате получаем  $\|b_k\| = 1$ .
3. Решаем систему уравнений  $Ay_k = b_k$  и находим поправку  $y_k$ .
4. Находим новое приближение  $x_{k+1} = x_k + y_k \|r_k\|$ .
5. Проверяем критерий прекращения итераций.
6. Если критерий не выполнен, возвращаемся в начало цикла.

Заметим, что в результате нормирования правой части системы (невязки) на этапе 2 всегда решается одна и та же (по норме правой части) система (этап 3). Поэтому никаких дополнительных проблем малость правой части не привносит. Так как решается одна и та же система, то обращение матрицы может быть выполнено один раз (при использовании прямого метода) либо может быть выбран хороший предобуславливатель, полученный на предыдущей итерации.

Наиболее важными являются этапы расчета невязки и пересчета приближения. На них возможны как вычитание близких чисел, так и сложение чисел, порядки которых далеки друг от друга.

Если число обусловленности решаемой системы уравнений не слишком велико, то описанный итерационный метод сходится сравнительно быстро. При этом под сходимостью чаще всего понимается установление значащих цифр в приближенном решении.

## 1.10. Решение проблемы собственных значений

Исследуем решение задачи на собственные значения

$$Ax_k = \lambda_k x_k.$$

**Определение 1.40.** Полной проблемой собственных значений называется задача отыскания всех собственных значений и соб-

ственных векторов матрицы  $A$ , *частичной* (ограниченной) — лишь их части.

Ограничимся задачей нахождения минимального и максимального собственных значений *строго положительно определенной матрицы*  $A$ , имеющей ортонормированный базис из собственных векторов.

Пусть

$$0 < \alpha(A) = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_{n-1} < \lambda_n = \beta(A) \quad —$$

собственные значения оператора  $A$ , а  $e_i$ ,  $i = 1, 2, \dots, n$ , — собственные векторы.

**Предложение 1.8.** Построим следующий итерационный процесс:

$$\varphi_0 = g, \quad \varphi_{k+1} = \frac{A\varphi_k}{\|\varphi_k\|}.$$

Тогда  $\beta(A) = \lim_{k \rightarrow \infty} \|\varphi_k\|$ .

◀ Разложим  $\varphi_0$  по собственным векторам матрицы  $A$ :

$$\varphi_0 = g = \sum_{i=1}^n c_i e_i.$$

Тогда

$$A^k \varphi_0 = \sum_{i=1}^n c_i \lambda_i^k e_i, \quad \|A^k \varphi_0\|^2 = \sum_{i=1}^n c_i^2 \lambda_i^{2k}.$$

Следовательно,

$$\varphi_{k+1} = \frac{A\varphi_k}{\|\varphi_k\|} = \frac{A^2 \varphi_{k-1}}{\|\varphi_{k-1}\|} \frac{\|\varphi_{k-1}\|}{\|A\varphi_{k-1}\|} = \frac{A^3 \varphi_{k-2}}{\|\varphi_{k-2}\|} \frac{\|\varphi_{k-2}\|}{\|A\varphi_{k-2}\|} = \dots = \frac{A^{k+1} \varphi_0}{\|A^k \varphi_0\|},$$

где

$$\frac{\|A^{k+1} \varphi_0\|^2}{\|A^k \varphi_0\|^2} = \lambda_n^2 + O\left(\frac{\lambda_{n-1}^{2k}}{\lambda_n^{2k}}\right),$$

откуда

$$\frac{\|A^{k+1} \varphi_0\|}{\|A^k \varphi_0\|} = \beta(A) + O\left(\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k}\right).$$

Значит,

$$\beta(A) = \lim_{n \rightarrow \infty} \|\varphi_n\|. \quad \blacktriangleright$$

**Следствие 1.5.** Для любой строго положительно определенной матрицы  $A$  имеет место равенство

$$\alpha(A) = \beta(A) - \beta(\beta(A)E - A).$$

◀ Рассмотрим оператор  $B = \beta(A)E - A$ . Его максимальное собственное число  $\beta(B) = \beta(A) - \alpha(A)$ . Отсюда  $\alpha(A) = \beta(A) - \beta(B)$ . Для нахождения  $\beta(B)$  используется тот же алгоритм, что и в предложении 1.8. ►

**Замечание 1.5.** Для поиска  $(n - 1)$ -го (второго по величине) собственного значения необходимо исключить из получаемых в процессе итераций векторов часть, соответствующую  $n$ -му собственному вектору. Это позволит проделать ту же процедуру, но уже в подпространстве, ортогональном этому вектору. В частности, можно использовать следующие соотношения:

$$A^{k+1}\varphi_0 - \lambda_n A^k \varphi_0 = \sum_{i=1}^{n-1} c_i (\lambda_i^{k+1} - \lambda_i^k \lambda_n) e_i,$$

$$A^k \varphi_0 - \lambda_n A^{k-1} \varphi_0 = \sum_{i=1}^{n-1} c_i (\lambda_i^k - \lambda_i^{k-1} \lambda_n) e_i.$$

Если  $\lambda_{n-1}$  отделено от  $\lambda_{n-2}$ , то отношение норм двух последних векторов должно стремиться к  $\lambda_{n-1}$  при  $k \rightarrow \infty$ .

**Замечание 1.6.** Лемма 1.3 о кругах Гершгорина позволяет тривиально оценить границы собственных значений. Довольно часто верхняя граница может быть оценена достаточно надежно. Нижняя же, как правило, всегда оценивается плохо.

## 1.11. О регуляризации плохо обусловленных СЛАУ

Рассмотрим систему

$$Ax = f.$$

Пусть *число обусловленности матрицы  $A$*  велико. В этом случае формально полученное решение задачи несет в себе столько погрешностей округления, что теряет всякий смысл. Необходима иная постановка исходной задачи.

Следствием точной постановки является равенство

$$\|Ax - f\|^2 = 0.$$

Но в случае плохо обусловленной системы вполне разумно считать, что

$$\|Ax - f\|^2 \approx 0.$$

Для определенности, т. е. выделения единственного решения, требуется добавить условие, которого не было в исходной постановке задачи. Такое условие может быть взято в виде минимума отклонения от заданного вектора  $x_0$  и минимума нормы разности  $(Ax - f)$  в виде

$$\Phi(x, x_0) = \|Ax - f\|^2 + \alpha\|x - x_0\|^2 \rightarrow \min.$$

Здесь  $\alpha > 0$  — управляющий параметр, называемый *параметром регуляризации*. Рассматриваемый метод называется *методом регуляризации СЛАУ*.

Преобразуем последнее выражение:

$$\begin{aligned} \Phi(x, x_0) = & (x, A^*Ax) - 2(Ax, f) + (f, f) + \\ & + \alpha(x, x) - 2\alpha(x, x_0) + (x_0, x_0) \rightarrow \min. \end{aligned}$$

Нахождение экстремума этого функционала (варьирование  $x$ ) дает следующее уравнение для  $x$ :

$$A^*Ax - A^*f + \alpha Ex - \alpha Ex_0 = 0,$$

или

$$(A^*A + \alpha E)x = A^*f + \alpha Ex_0.$$

Здесь матрица  $A^*A$  уже положительно определена. Если  $\alpha > 0$ , то при его достаточно большом значении нахождение решения задачи, например, методом Гаусса не составит труда. Но  $x = x(x_0, \alpha)$ . Откуда выбираются  $x_0, \alpha$ ?

Если относительно  $x_0$  нет соображений физического или технического характера, то обычно полагают  $x_0 = 0$ . Если есть, то берут нужное.

Рассмотрим выбор  $\alpha$ . Его нельзя брать слишком малым (иначе опять получится плохо обусловленная задача) и нельзя брать слишком большим (получится большая норма невязки  $\|Ax - f\|$ , т. е. плохое решение исходной СЛАУ). Обычно  $\alpha$  выбирается по принципу невязки: ищут такое  $\alpha$ , чтобы было выполнено условие

$$\|Ax - f\| \simeq \|\delta f\| + \|\delta Ax\|,$$

где  $\delta f$  и  $\delta A$  — априорно заданные погрешности правой части и оператора.

Изложенное выше является одним из простейших примеров *метода регуляризации* некорректно поставленных задач *А.Н. Тихонова*.

## 1.12. Хранение больших разреженных матриц

При дискретизации задач математической физики возникают системы линейных алгебраических уравнений (СЛАУ) с большим числом неизвестных  $n$ . В современных многомерных задачах значение  $n$  может достигать многих тысяч и миллионов. Матрица  $A$  возникшей СЛАУ формально имеет  $n^2$  элементов. Даже для современных ЭВМ объем этой информации может оказаться недопустимо велик. Однако чаще всего **матрица** редко заполнена (*разрежена*), т. е. содержит большое количество нулей. Общее число ненулевых элементов  $N$  в таких задачах удовлетворяет условию  $N \ll n^2$ . Ясно, что хранить большое количество нулей бессмысленно. Опишем различные способы хранения таких матриц.

1. Хранение матрицы  $A$  в виде двумерного массива элементов  $\{a_{ij}\}$ ,  $i, j = 1, 2, \dots, n$ , требует запоминания  $n^2$  элементов, большинство из которых нули. Вариант неприемлем.

2. Простейшим вариантом является хранение только ненулевых элементов матрицы  $A$  в виде линейного массива  $\{a_k\}$ ,  $k = 1, 2, \dots, N$ , состоящего чаще всего из действительных чисел. Помимо ненулевых элементов упорядочить по строкам слева направо. При этом для каждого элемента необходимо указать номера строки и столбца, которые он занимает в исходной матрице:  $\{ia_k\}$ ,  $k = 1, 2, \dots, N$ , и  $\{ja_k\}$ ,  $k = 1, 2, \dots, N$ . Данные массивы целочисленные. В этом варианте требуется хранить  $3N$  значений.

3. Минимальным по требуемым ресурсам памяти является вариант хранения только ненулевых элементов матрицы  $A$  в виде линейного массива  $\{a_k\}$ ,  $k = 1, 2, \dots, N$ , и одного целочисленного массива  $\{iia_k\}$ ,  $k = 1, 2, \dots, N$ ; последний массив содержит номер соответствующего элемента в исходной двумерной матрице при ее хранении, например в виде одномерного массива по столбцам. В этом случае номер  $k$ -го элемента  $iia_k = i_k + (j_k - 1)n$ . Здесь  $i_k$ ,  $j_k$  — номера строки и столбца рассматриваемого элемента. Этот вариант требует хранить  $2N$  значений. Очевидно, что при работе с матрицей, хранящейся в таком виде, придется постоянно пересчитывать номера строки и столбца каждого элемента.

4. Оптимальным по требуемым ресурсам памяти и удобству работы является так называемый «разреженный строчный» формат хранения данных. При этом необходимо хранить только упорядоченные по строкам слева направо ненулевые элементы матрицы  $A$  в виде линейного массива  $\{a_k\}$ ,  $k = 1, 2, \dots, N$ , массив  $\{ja_k\}$ ,  $k = 1, 2, \dots, N$ , с номерами столбцов элементов в исходной матрице и массив  $\{iiak\}$ ,

$k = 1, 2, \dots, n + 1$ . Два последних массива целочисленные. В массиве  $\{iii a_k\}$  хранятся номера первого элемента  $k$ -й строки при сплошной нумерации линейного одномерного массива. Таким образом, данный массив указывает, что элементы с номерами от  $iii a_k$  до  $iii a_{k+1} - 1$  (включительно) при сплошной нумерации одномерного массива соответствуют  $k$ -й строке исходной матрицы. Отсюда следует, что необходимо положить  $iii a_1 = 1$ ,  $iii a_{n+1} = N + 1$ . В этом варианте требуется хранить  $2N + n + 1$  значений.

### 1.13. Библиографические комментарии

Материал, приведенный в данной главе, широко применяется при численном решении практических задач математической физики. Почти в каждой из них требуется решить систему линейных (или нелинейных) уравнений. Широко распространенное выражение «данный метод не требует решения систем линейных алгебраических уравнений» означает чаще всего лишь то, что матрица решаемой системы имеет либо диагональный вид, либо треугольный. Поиск решения таких систем не составляет проблем.

Важность решения задач линейной алгебры привела к тому, что существует огромное число публикаций, посвященных этой проблеме.

Базовыми дисциплинами в данной области являются функциональный анализ в конечномерных пространствах и теория линейных операторов [8], [17, 18, 87, 103, 105, 154]. В [8], в частности, исследована матрица Гильберта. Наше изложение в значительной степени следует указанным книгам. Представляет интерес литература по классической линейной алгебре (например, [82]), особенно важна специализированная литература по матричному анализу и решению задач вычислительной алгебры [29, 31, 46, 78, 175, 188].

Наиболее интересна книга [154], посвященная численному решению систем уравнений, возникающих при конечномерной дискретизации задач математической физики. В ней, в частности, содержится решение задачи о построении полинома, наименее отклоняющегося от нуля, теория решения линейных разностных уравнений с постоянными коэффициентами, а также общая теория итерационных методов решения СЛАУ, как одношаговых, так и двухшаговых. В [154] подробно описаны различные прямые методы решения, такие как метод редукции и метод быстрого преобразования Фурье. Они являются специализированными для решения некоторых классов задач и достигают рекордных по числу действий характеристик.

Необходимо отметить, что задачи вычислительной алгебры рассматриваются во всех книгах учебного характера (например, [92]), указанных в библиографии.

Мы лишь коснулись проблемы собственных значений и ее решения. Однако ряд задач науки и техники состоит в решении данной проблемы в конкретном случае. Укажем специализированные по данному вопросу книги [88] и [174]. Много полезного можно найти в [17, 18, 80].

Теория решения некорректно поставленных задач уже превратилась в самостоятельный раздел прикладной математики и математической физики. Представленное в данной главе описание метода регуляризации может быть существенно расширено с помощью работ [118, 145, 168, 169].

Развитие алгоритмов вычислительной линейной алгебры происходит как по линии создания новых, так и по линии модернизации классических алгоритмов в целях их приспособления к практическим нуждам: например, метода Гаусса для решения больших разреженных СЛАУ со специальными алгоритмами сжатия полосы, содержащей ненулевые элементы [63].

Укажем также литературу, посвященную технологии работы с матрицами специального вида, например [77, 78, 131, 173, 186, 191].

Современные алгоритмы вычислительной линейной алгебры представлены также в [57]. Развитие ЭВМ с параллельной архитектурой привело к появлению алгоритмов, предназначенных для использования именно на таких машинах [30, 127].

Во второй части данной книги представлен многосеточный метод решения СЛАУ, возникающих при дискретизации задач математической физики. Максимальное использование информации об исходной задаче позволяет получить наилучшую реализацию метода и достичь неулучшаемых характеристик. Пионерскими работами в этой области являются [7, 13, 179, 180].

## 2. РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Представлены наиболее распространенные методы решения скалярного нелинейного уравнения и систем нелинейных уравнений. Обоснованы *метод деления отрезка пополам*, *метод хорд* и *метод Ньютона*. Доказана сходимость *стационарных методов*, сводимых к поиску неподвижной точки отображения. Рассмотрены *внутренние* и *внешние итерации*, применяемые для нахождения решения систем нелинейных уравнений. Приведены примеры *гибридных методов* для решения систем.

### 2.1. Решение скалярных уравнений

Рассмотрим задачу о нахождении корней уравнения  $f(x) = 0$ , где  $f(x)$  — некоторая заданная функция. Существуют различные задачи, связанные с поиском корней:

- 1) нахождение области локализации корней;
- 2) нахождение корня;
- 3) определение его кратности;
- 4) уточнение значений найденных корней и оценка их точности.

Ограничимся рассмотрением лишь одной задачи: нахождение простого (некратного) корня уравнения  $f(x) = 0$ , расположенного на отрезке  $[a, b]$  непрерывности функции  $f(x)$ , причем справедливо неравенство  $f(a)f(b) < 0$ . Для определенности будем считать, что  $f(a) < 0$ ,  $f(b) > 0$  (если это не так, то всегда можно вместо  $f$  принять  $-f$ , что ничего не меняет). При этих условиях, как известно из курса математического анализа (теорема Коши о значениях непрерывной на отрезке функции), на отрезке  $[a, b]$  присутствует корень уравнения  $f(x) = 0$ .

**Замечание 2.1.** Для корректного решения поставленной задачи нужно быть уверенным в наличии не более одного корня на рассматриваемом отрезке  $[a, b]$ . Разные знаки функции на границах отрезка гарантируют только нечетность числа действительных корней на данном отрезке. Для решения вопроса о единственности корня в одномерном случае может быть использован аппарат классического математического анализа по качественному исследованию (нахождение областей монотонности, выпуклости и т. п.) графика функции. Этот же аппарат можно использовать для нахождения области локализации корней.

Технически же проще вычислить значения функции в ряде точек и найти участки, на границах которых функция принимает значения с разными знаками. Это можно сделать, например, методом деления отрезка пополам, приведенным в 2.1.1.

**Замечание 2.2.** При поиске всех корней решение задачи приходится вести последовательно, находя их поочередно. При этом после нахождения простого корня  $x_k$  необходимо исходную функцию разделить на  $x - x_k$ . Далее нужно искать решение модифицированной задачи. В случае кратного корня алгоритм потребуется модифицировать.

Отметим, что значения корней по-разному зависят от ошибок коэффициентов исходной функции. Известно, что наибольшие по абсолютной величине корни наименее устойчивы. Поэтому поиск всех корней необходимо начинать с меньших корней, после чего делить функцию и продолжать процесс.

### 2.1.1. Метод деления отрезка пополам (метод «вилки»)

Пусть

$$a_0 = a, \quad b_0 = b, \quad c_1 = \frac{a_0 + b_0}{2}.$$

Если  $f(c_1) = 0$ , то решение задачи найдено. Если  $f(c_1) > 0$ , то  $a_1 = a_0$ ,  $b_1 = c_1$ , если же наоборот, то  $a_1 = c_1$ ,  $b_1 = b_0$ , и т. д. В результате на  $n$ -м шаге получим отрезок, содержащий решение, длиной

$$b_n - a_n = \frac{b - a}{2^n}, \quad n = 0, 1, \dots$$

Если в качестве приближения к решению  $x$  взять  $x_* = \frac{a_n + b_n}{2}$ , то

$$|x_* - x| \leq \frac{1}{2}(b - a)2^{-n},$$

где  $x$  — точное значение корня, которое заведомо находится на отрезке  $[a_n, b_n]$ , если процесс не завершился раньше.

Следовательно, для нахождения решения с точностью  $\varepsilon$  требуется выполнить следующее число итераций:

$$n = \left\lceil \frac{\ln \frac{b-a}{2\varepsilon}}{\ln 2} \right\rceil + 1.$$

Здесь  $[x]$  — целая часть числа  $x$ .

### 2.1.2. Итерационные методы решения типа простой итерации

Уравнение  $f(x) = 0$  заменим на уравнение  $x = F(x)$ , эквивалентное исходному. Как правило,  $F(x) = x + \tau(x)f(x)$ , где  $\tau(x)$  — знакопостоянная на  $[a, b]$  функция. Организуем итерационный процесс

$$x_{k+1} = F(x_k), \quad k = 0, 1, \dots$$

**Теорема 2.1 (сходимость метода простой итерации).** Пусть функция  $F(x)$  липшиц-непрерывна с постоянной  $q \in (0, 1)$  на отрезке  $[c - \delta, c + \delta] = \tilde{O}_\delta(c)$ , т. е. для любых  $x', x'' \in \tilde{O}_\delta(c)$  справедливо неравенство

$$|F(x') - F(x'')| \leq q|x' - x''|,$$

причем

$$|F(c) - c| \leq (1 - q)\delta.$$

Тогда уравнение  $x = F(x)$  имеет единственное решение  $x_*$  на отрезке  $\tilde{O}_\delta(c)$ , которое может быть найдено в результате описанного итерационного процесса при любом  $x_0 \in \tilde{O}_\delta(c)$ . Для погрешности справедлива оценка

$$|x_k - x_*| \leq q^k |x_0 - x_*|, \quad k = 0, 1, \dots,$$

или

$$|x_k - x_*| \leq \frac{q^k}{1 - q} |F(x_0) - x_0|, \quad k = 0, 1, \dots$$

◀ Пусть  $x_0 \in \tilde{O}_\delta(c)$ . Допустим, что и  $x_k \in \tilde{O}_\delta(c)$ . Докажем, что  $x_{k+1} \in \tilde{O}_\delta(c)$ . В самом деле,

$$x_{k+1} - c = F(x_k) - F(c) + F(c) - c,$$

откуда

$$|x_{k+1} - c| \leq q|x_k - c| + (1 - q)\delta \leq q\delta + (1 - q)\delta = \delta.$$

Следовательно,  $x_{k+1} \in \tilde{O}_\delta(c)$ . Так как  $x_{k+1} - x_k = F(x_k) - F(x_{k-1})$ , то

$$|x_{k+1} - x_k| \leq q|x_k - x_{k-1}| \leq q^k |x_1 - x_0| = q^k |F(x_0) - x_0|, \quad k = 1, 2, \dots$$

Тогда

$$x_{k+p} - x_k = \sum_{l=1}^p (x_{k+l} - x_{k+l-1}),$$

и

$$\begin{aligned}|x_{k+p} - x_k| &\leq |F(x_0) - x_0| \sum_{l=1}^p q^{k+l-1} = \\&= |F(x_0) - x_0| q^k \frac{1-q^p}{1-q} \leq \frac{q^k}{1-q} |F(x_0) - x_0|.\end{aligned}$$

Отсюда заключаем, что  $\{x_n\}$  — фундаментальная последовательность. Она имеет предел, находящийся на отрезке  $\tilde{O}_\delta(c)$ . Так как  $F(x)$  — непрерывная функция, то, переходя к пределу в соотношении  $x_{k+1} = F(x_k)$  при  $k \rightarrow \infty$ , получим

$$\lim_{k \rightarrow \infty} x_{k+1} = \lim_{k \rightarrow \infty} x_k = F\left(\lim_{k \rightarrow \infty} x_k\right),$$

откуда

$$x_* = \lim_{k \rightarrow \infty} x_k, \quad x_* \in \tilde{O}_\delta(c).$$

Устремив  $p \rightarrow \infty$  в неравенстве для  $|x_{k+p} - x_k|$ , получим оценку точности решения через известные величины:

$$|x_k - x_*| \leq \frac{q^k}{1-q} |F(x_0) - x_0|.$$

Поскольку  $x_{k+1} = F(x_k)$ ,  $x_* = F(x_*)$ , то

$$|x_{k+1} - x_*| = |F(x_k) - F(x_*)| \leq q|x_k - x_*| \leq q^{k+1}|x_0 - x_*|,$$

или

$$|x_k - x_*| \leq q^k |x_0 - x_*|.$$

Данный корень является единственным на  $\tilde{O}_\delta(c)$ , так как, предположив наличие двух корней  $x_*^1$  и  $x_*^2$ , получим

$$|x_*^1 - x_*^2| = |F(x_*^1) - F(x_*^2)| \leq q|x_*^1 - x_*^2|,$$

откуда  $x_*^1 = x_*^2$ , поскольку  $q \in (0, 1)$ . ►

**Следствие 2.1.** Если верно не условие липшиц-непрерывности, а  $|F'(x)| \leq q < 1$  в  $\tilde{O}_\delta(c)$ , то все условия теоремы 2.1 выполнены, и ее выводы справедливы.

**Следствие 2.2.** Пусть функция  $F(x)$  непрерывно дифференцируема в окрестности точки  $x_*$ ,  $F(x_*) = x_*$ , и  $|F'(x_*)| < 1$ . Тогда существует такое  $\delta > 0$ , что в  $\delta$ -окрестности  $\tilde{O}_\delta(x_*)$  уравнение  $x = F(x)$  имеет единственный корень, итерационный процесс в  $\tilde{O}_\delta(x_*)$  сходится, если  $x_0 \in \tilde{O}_\delta(x_*)$ .

◀ В силу непрерывности функции  $F'(x)$  и условия  $|F'(x_*)| < 1$  существует  $\delta$ -окрестность  $\tilde{O}_\delta(x_*)$  точки  $x_*$ , в которой  $|F'(x)| \leq q^* < 1$ . Тогда выполнены все условия теоремы:  $c = x_*$ ,  $|F(c) - c| = 0 < (1 - q^*)\delta$ . Поэтому утверждение следствия справедливо. ►

**Определение 2.1.** Если погрешность метода удовлетворяет оценке

$$|x_k - x_*| \leq Lq^k|x_0 - x_*|,$$

то говорят, что имеет место **линейная сходимость со скоростью геометрической прогрессии** с показателем  $q$ .

В рассмотренном нами случае  $L = 1$  и на каждой итерации выполнено неравенство  $|x_k - x_*| \leq q|x_{k-1} - x_*|$ .

### 2.1.3. Варианты метода простой итерации

**Метод релаксации.** В этом случае

$$F(x) = x + \tau f(x).$$

Тогда итерационный процесс имеет вид

$$\frac{x_{k+1} - x_k}{\tau} = f(x_k)$$

и  $F'_x = 1 + \tau f'_x$ . По теореме 2.1 метод сходится при  $|1 + \tau f'_x| < 1$ , т. е. при  $-2 < \tau f'(x) < 0$ . Если в некоторой окрестности корня  $f'_x < 0$  и  $0 < m < |f'_x| < M$ , то метод сходится при  $\tau < 2/M$ .

Найдем оптимальное значение параметра. Пусть  $z_k = x_k - x_*$  — погрешность. Тогда

$$\frac{z_{k+1} - z_k}{\tau} = f(x_* + z_k) = f(x_* + z_k) - f(x_*) = f'(x_* + \vartheta z_k)z_k$$

и

$$z_{k+1} = z_k(1 + \tau f'(x_* + \vartheta z_k)).$$

Отсюда

$$|z_{k+1}| \leq \max_y |1 + \tau f'(y)| |z_k| \leq \left(\max_y |1 + \tau f'(y)|\right)^{k+1} |z_0|.$$

Но

$$\max_y |1 + \tau f'(y)| \leq \max(|1 - \tau m|, |1 - \tau M|).$$

Выберем  $\tau$ , которое минимизирует эту величину (определение оптимального  $\tau$  проиллюстрировано на рис. 2.1):

$$\tau = \frac{2}{m+M}.$$

Отсюда получим

$$|z_k| \leq \left(\frac{M-m}{M+m}\right)^k |z_0|,$$

или

$$|z_k| \leq \left(\frac{1-\xi}{1+\xi}\right)^k |z_0|, \quad \xi = \frac{m}{M}.$$

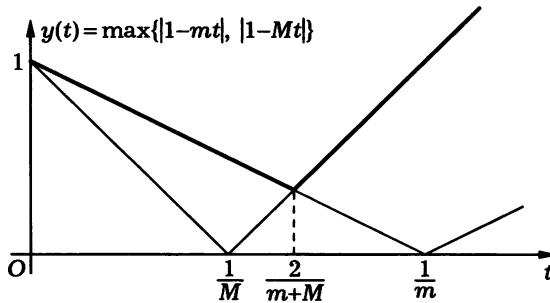


Рис. 2.1

Аналогичная задача о «минимаксе» встречалась при исследовании итерационных методов вариационного типа.

**Метод Ньютона (метод касательных).** Для данного метода

$$F(x) = x - \frac{f(x)}{f'(x)}.$$

Тогда

$$F'(x) = 1 - \frac{1 + \frac{ff''}{(f')^2}}{\frac{ff''}{(f')^2}} = \frac{ff''}{(f')^2}.$$

Пусть всюду на  $[a, b]$

$$|f'(x)| \geq m > 0, \quad |f''(x)| \leq M.$$

Тогда существует такая  $\varepsilon$ -окрестность корня  $x_*$ , что если  $x_0 \in \tilde{O}_\varepsilon(x_*)$ , то итерационный процесс сходится к корню. Действительно, всюду на отрезке  $[a, b]$

$$|F'(x)| = \left| \frac{ff''}{(f')^2} \right| \leq \frac{|f|}{m^2} M.$$

Из непрерывности  $f(x)$  следует, что в некоторой  $\varepsilon$ -окрестности корня  $x_*$  выполнено неравенство

$$|f(x)| \leq q \frac{m^2}{M}, \quad 0 < q < 1.$$

Таким образом, в этой окрестности справедливы условия теоремы и ее выводы.

Следовательно, речь идет о сходимости метода Ньютона в малом, т. е. лишь при удачном попадании начального приближения в окрестность корня.

Оценим погрешность. Имеем

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Разложим  $f(x_*)$  в точке  $x_k$ :

$$f(x_*) = f(x_k) + f'(x_k)(x_* - x_k) + \frac{1}{2}f''(\xi)(x_k - x_*)^2 = 0,$$

где  $\xi \in (x_k, x_*)$  или  $\xi \in (x_*, x_k)$  в зависимости от взаимного расположения корня и его приближения. Отсюда

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{f(x_k) - f(x_*)}{f'(x_k)} = x_* + \frac{1}{2} \frac{f''(\xi)}{f'(x_k)} (x_k - x_*)^2.$$

Следовательно,

$$|x_{k+1} - x_*| \leq \frac{M}{2m} |x_k - x_*|^2 \leq \left( \frac{M}{2m} \right)^{2^{k+1}-1} |x_0 - x_*|^{2^{k+1}}.$$

Можно получить более сильные результаты, если предположить наличие у функции  $f(x)$ , например, монотонной производной  $f'(x)$  определенного знака на отрезке  $[a, b]$ .

**Определение 2.2.** Если погрешность метода удовлетворяет оценке

$$|x_k - x_*| \leq L |x_{k-1} - x_*|^p,$$

то говорят, что метод сходится с  $p$ -м порядком.

В соответствии с определением метод Ньютона имеет квадратичную скорость сходимости.

**Замечание 2.3.** Существуют многочисленные варианты метода Ньютона. Самый простой из них заключается в использовании вычислений в одной выбранной точке производной на всех итерациях. При этом квадратичная скорость сходимости теряется.

**Замечание 2.4.** Рассмотрим следующую модификацию метода Ньютона:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f(x_k - f(x_k)f'(x_k)^{-1})}{f'(x_k)}.$$

Нетрудно видеть, что скорость сходимости данного метода является кубической.

**Замечание 2.5.** Если вычисление второй производной рассматриваемой функции не вызывает проблем, то можно использовать другую модификацию метода Ньютона:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f''(x_k)(f(x_k))^2}{2(f'(x_k))^3},$$

также дающую кубическую скорость сходимости.

**Метод хорд (метод секущих).** В этом случае

$$F(x) = x - f(x) \frac{b-x}{f(b)-f(x)}.$$

При  $x = b$  полагаем

$$F(b) = b - \frac{f(b)}{f'(b)}.$$

Исследуем сходимость метода хорд для случая дважды непрерывно дифференцируемой на  $[a, b]$  функции  $f(x)$ , такой, что  $f'(x) > 0$  и  $f''(x) \geq 0$ .

Пусть, как и ранее,  $x_*$  — искомый корень ( $f(x_*) = 0$ ). Допустим, что на некоторой  $k$ -й итерации выполнено неравенство  $x_k < x_* < b$ . Исследуем значение  $x_{k+1}$ . В соответствии с используемым итерационным процессом

$$\begin{aligned} x_{k+1} - x_k &= -f(x_k) \frac{b-x_k}{f(b)-f(x_k)} = \\ &= -(f(x_k) - f(x_*)) \frac{b-x_k}{f(b)-f(x_*) + f(x_*) - f(x_k)} = \\ &= -f'(\xi)(x_k - x_*) \frac{b-x_k}{f'(\eta)(b-x_*) + f'(\xi)(x_* - x_k)}. \end{aligned}$$

Преобразования выполнены с помощью формулы Лагранжа конечных приращений. В соответствии со сделанными предположениями фигурирующие в ней средние точки таковы:  $\xi \in (x_k, x_*)$ ,  $\eta \in (x_*, b)$ ,  $\xi < \eta$ . Тогда в соответствии со знаком второй производной  $f'(\eta) \geq f'(\xi)$  и  $x_{k+1} - x_k \leq x_* - x_k$ . Кроме того, так как точка  $x_k$  лежит левее корня, справедливо неравенство  $x_{k+1} - x_k > 0$ .

В результате получаем, что последовательность итерационных приближений  $\{x_k\}$  монотонно возрастает и ограничена сверху значением корня. Следовательно, по теореме Вейерштрасса она сходится. Легко видеть, что ее пределом является искомый корень уравнения.

Совершенно так же работает метод хорд в случае функции  $f(x)$ , такой, что  $f'(x) < 0$  и  $f''(x) \leq 0$ .

Если же имеется одна из двух нерассмотренных пар сочетаний знаков первой и второй производной, то метод хорд нужно использовать в том же виде с заменой  $b$  на  $a$ .

**Интерполяционные методы.** Данные методы состоят в замене функции  $f(x)$  на ее приближение  $\tilde{f}(x)$ , сконструированное путем интерполяции (иногда — экстраполяции) по нескольким точкам, и выборе в качестве приближенного корня уравнения  $f(x) = 0$  корня уравнения  $\tilde{f}(x) = 0$ .

Рассмотрим разные варианты.

1. Замена функции прямой, проходящей через две точки  $x_k, x_{k+1}$ :

$$\tilde{f}(x) = f(x_k) + (f(x_{k+1}) - f(x_k)) \frac{x - x_k}{x_{k+1} - x_k}.$$

Тогда получаем **метод секущих** (его некоторый вариант):

$$x_{k+2} = x_k + f(x_k) \frac{x_{k+1} - x_k}{f(x_k) - f(x_{k+1})}.$$

Без доказательства отметим, что скорость сходимости метода секущих находится примерно посередине между линейной и квадратичной.

2. Замена функции прямой

$$\tilde{f}(x) = f(x_k) + f'(x_k)(x - x_k),$$

проходящей через точку  $x_k$ , в которой известны сама функция и ее производная. Тогда получаем **метод Ньютона**, имеющий квадратичную скорость сходимости:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

3. Замена функции параболой, проходящей через три точки:

$$\tilde{f}(x) = ax^2 + bx + c,$$

где  $a, b, c$  находятся из системы уравнений

$$\tilde{f}(x_i) = ax_i^2 + bx_i + c = f(x_i), \quad i = k - 2, k - 1, k.$$

Как будет показано в 3, эта парабола может быть выписана без непосредственного решения последней системы в виде

$$\begin{aligned}\tilde{f}(x) = & f(x_{k-2}) \frac{(x - x_{k-1})(x - x_k)}{(x_{k-2} - x_{k-1})(x_{k-2} - x_k)} + \\ & + f(x_{k-1}) \frac{(x - x_{k-2})(x - x_k)}{(x_{k-1} - x_{k-2})(x_{k-1} - x_k)} + f(x_k) \frac{(x - x_{k-2})(x - x_{k-1})}{(x_k - x_{k-2})(x_k - x_{k-1})}.\end{aligned}$$

Тогда корень  $x_{k+1}$  принимается равным одному из корней уравнения  $\tilde{f}(x_{k+1}) = 0$ .

Полученный метод называется **методом парабол**.

Несмотря на повышение порядка приближающей функции, что, казалось бы, должно привести к повышению скорости сходимости, выигрыш здесь очень небольшой. Скорость сходимости данного метода ниже квадратичной, но выше скорости сходимости метода секущих.

Метод парабол принципиально отличается от методов, основанных на замене исходной функции ее линейным приближением, тем, что может дать комплексные корни при действительных предыдущих приближениях.

Отметим также, что метод парабол оказывается весьма эффективным средством нахождения корней алгебраических многочленов. Итерационные приближения данного метода практически всегда быстро сходятся к корню уравнения.

4. Замена функции параболой (полиномом Тейлора второго порядка)

$$\tilde{f}(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2,$$

проходящей через точку  $x_k$ , в которой известны сама функция, ее первая и вторая производные. Тогда

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \left( 1 \pm \left( 1 - 2 \frac{f(x_k)f''(x_k)}{(f'(x_k))^2} \right)^{1/2} \right).$$

Разложив подкоренное выражение с точностью до квадратичных слагаемых и выбрав нужный знак, получим представленную выше модификацию метода Ньютона.

Можно этого не делать и использовать выписанное выражение для корней непосредственно. В этом случае будем иметь еще один вариант метода парабол.

Методы построения интерполянтов будут подробно изучаться в 3.

## 2.2. Решение систем нелинейных уравнений

Пусть теперь необходимо решить систему уравнений

$$f_i(x_1, x_2, \dots, x_n) = 0, \quad i = 1, 2, \dots, n,$$

и найти  $n$  неизвестных  $x_1, x_2, \dots, x_n$ . Будем считать, что  $x$  есть  $n$ -мерный вектор, компоненты которого равны  $x_1, x_2, \dots, x_n$ , а  $F$  —  $n$ -мерный вектор с компонентами  $f_1, f_2, \dots, f_n$ . Тогда необходимо найти решение векторного уравнения

$$F(x) = 0.$$

Рассмотрим *итерационные одношаговые методы* решения такого уравнения вида

$$B_{k+1} \frac{x^{k+1} - x^k}{\tau_{k+1}} + F(x^k) = 0, \quad k = 0, 1, \dots,$$

где  $x^k$  — значение приближенного решения на  $k$ -й итерации.

Существуют и нелинейные (относительно  $x^{k+1}$ ) итерационные методы. Но мы ограничимся пока методами указанного вида. Будем считать, что  $B_{k+1}$  — линейный оператор (матрица размером  $n \times n$ ), имеющий обратный.

Для нахождения  $x^{k+1}$  необходимо решить операторное уравнение

$$B_{k+1}x^{k+1} = g(x^k) = B_{k+1}x^k - \tau_{k+1}F(x^k).$$

Как и ранее, метод называется явным при  $B_{k+1} = E$  и неявным в противном случае, стационарным при  $B_{k+1} = B$ ,  $\tau_{k+1} = \tau$  и нестационарным в противном случае.

**Определение 2.3.** При решении уравнения

$$B_{k+1}x^{k+1} = g(x^k)$$

итерационным способом *итерации* последнего называются *внутренними*, а *итерации*  $x^k$  — *внешними*.

### 2.2.1. Сходимость стационарного метода

Если  $B_{k+1} = B$ ,  $\tau_{k+1} = \tau$ , то

$$x^{k+1} = x^k - \tau B^{-1}F(x^k) = S(x^k).$$

Исходное уравнение  $F(x) = 0$  можно переписать в виде  $x = S(x)$ , т. е.  $\tau \neq 0$ ,  $B$  — невырожден. Таким образом, искомое решение есть неподвижная точка оператора  $S$ .

**Определение 2.4.** Говорят, что *оператор*  $S$  является *сжимающим* на множестве  $K$  с коэффициентом сжатия  $q$ , если существует такое число  $q \in (0, 1)$ , что

$$\|S(x') - S(x'')\| \leq q\|x' - x''\|, \quad x', x'' \in K.$$

**Теорема 2.2 (принцип сжимающих отображений).** Пусть оператор  $S$  определен в шаре

$$\bar{U}_r(a) = \{x: \|x - a\| \leq r\}$$

и является сжимающим в нем с коэффициентом  $q$ , причем

$$\|S(a) - a\| \leq (1 - q)r, \quad q \in (0, 1).$$

Тогда в  $\bar{U}_r(a)$  оператор  $S$  имеет единственную неподвижную точку  $x_*$  и итерации  $x^{k+1} = S(x^k)$  сходятся к  $x_*$  для любого  $x^0 \in \bar{U}_r(a)$ . Для погрешности справедливы оценки

$$\|x^k - x_*\| \leq q^k \|x^0 - x_*\|, \quad \|x^k - x_*\| \leq \frac{q^k}{1-q} \|S(x^0) - x^0\|.$$

◀ Пусть  $x^0 \in \bar{U}_r(a)$  и все  $x^k \in \bar{U}_r(a)$ . Докажем, что и  $x^{k+1} \in \bar{U}_r(a)$ . В самом деле,

$$x^{k+1} - a = S(x^k) - a = S(x^k) - S(a) + S(a) - a.$$

Тогда

$$\begin{aligned} \|x^{k+1} - a\| &\leq \|S(x^k) - S(a)\| + \|S(a) - a\| \leq \\ &\leq q\|x^k - a\| + (1 - q)r \leq qr + (1 - q)r = r, \end{aligned}$$

и, следовательно,  $x^{k+1} \in \bar{U}_r(a)$ . Оценим  $x^{k+1} - x^k$ :

$$\|x^{k+1} - x^k\| = \|S(x^k) - S(x^{k-1})\| \leq q\|x^k - x^{k-1}\| \leq q^k \|S(x^0) - x^0\|.$$

Отсюда следует, что последовательность  $\{x^k\}$  является фундаментальной, а именно:

$$\begin{aligned} \|x^{k+p} - x^k\| &= \left\| \sum_{l=1}^p (x^{k+l} - x^{k+l-1}) \right\| \leq \|S(x^0) - x^0\| \sum_{l=1}^p q^{k+l-1} = \\ &= \frac{q^k - q^{k+p}}{1-q} \|S(x^0) - x^0\| \leq \frac{q^k}{1-q} \|S(x^0) - x^0\|. \end{aligned}$$

Так как множество  $\bar{U}_r(a)$  замкнуто, то  $\{x^k\} \rightarrow x_* \in \bar{U}_r(a)$ . Оператор  $S$  является непрерывным в  $\bar{U}_r(a)$  в силу сжимаемости, поэтому переходя к пределу при  $k \rightarrow \infty$  в  $x^{k+1} = S(x^k)$ , получим  $x_* = S(x_*)$ , т. е.  $x_*$  — решение исходного уравнения. Следовательно, решение в  $\bar{U}_r(a)$  существует. Оно единственno, так как допустив существование второго решения  $x_{**}$ , имеем

$$\|x_* - x_{**}\| = \|S(x_*) - S(x_{**})\| \leq q\|x_* - x_{**}\|.$$

Учитывая, что  $q \in (0, 1)$ , приходим к заключению, что  $\|x_* - x_{**}\| = 0$ , т. е.  $x_* = x_{**}$ .

Переходя к пределу при  $p \rightarrow \infty$  в неравенстве для  $\|x^{k+p} - x^k\|$ , получаем

$$\begin{aligned} \|x_* - x^k\| &\leq \frac{q^k}{1-q} \|S(x^0) - x^0\|, \\ \|x^{k+1} - x_*\| &= \|S(x^k) - S(x_*)\| \leq \|x^k - x_*\| \leq q^{k+1} \|x^0 - x_*\|, \end{aligned}$$

т. е. вторую оценку погрешности решения. ►

**Замечание 2.6.** По существу, теорема 2.1 есть частный случай теоремы 2.2.

### 2.2.2. Примеры итерационных методов

**Метод релаксации.** В этом случае  $B_{k+1} = E$ ,  $\tau_{k+1} = \tau$ ,  $S(x) = x - \tau F(x)$ . Метод сходится, если  $\|S'\| < 1$ , где  $S' = E - \tau F'$ , а

$$F' = \begin{pmatrix} (f_1)'_{x_1} & (f_1)'_{x_2} & \cdots & (f_1)'_{x_n} \\ \vdots & \vdots & \ddots & \vdots \\ (f_n)'_{x_1} & (f_n)'_{x_2} & \cdots & (f_n)'_{x_n} \end{pmatrix}.$$

**Метод Пикара.** Пусть  $F(x) = Ax + G(x)$ , где  $A$  — линейный оператор. Тогда итерации можно определить следующим образом:

$$Ax^{k+1} + G(x^k) = 0,$$

т. е.

$$B_{k+1} = A, \quad \tau_{k+1} = \tau = 1, \quad S(x) = x - \tau B^{-1}F = x - A^{-1}F.$$

Метод сходится при  $\|S'\| = \|E - A^{-1}F'\| < 1$ .

Можно провести **модификацию метода Пикара**: вместо

$$A(x^{k+1} - x^k) + F(x^k) = 0$$

принять

$$A \frac{x^{k+1} - x^k}{\tau} + F(x^k) = 0,$$

т. е. ввести параметр  $\tau$ , который управляет скоростью сходимости.

**Метод Ньютона.** В этом случае  $B_{k+1} = F'(x^k)$ ,  $\tau_{k+1} = 1$ , т. е.

$$F'(x^k)(x^{k+1} - x^k) + F(x^k) = 0.$$

Для реализации метода необходимо существование матрицы, обратной матрице  $F'(x^k)$ .

Как и в скалярном случае, метод имеет квадратичную сходимость, если начальное приближение выбрано удачно. Доказательство сходимости опустим.

**Модифицированный метод Ньютона.** В этом случае  $B_{k+1} = F'(x^0)$ ,  $\tau_{k+1} = 1$ . Тогда обращать  $B_{k+1}$  в отличие от исходного варианта метода нужно лишь один раз.

**Метод Ньютона с параметром.** Этот вариант метода имеет вид

$$F'(x^k) \frac{x^{k+1} - x^k}{\tau_{k+1}} + F(x^k) = 0.$$

Дополнительное исследование проводить не будем.

**Нелинейный метод Якоби.** Ранее рассмотренные методы являются линейными. Рассмотрим **нелинейный метод Якоби**, относящийся к нелинейным методам:

$$f_i(x_1^k, x_2^k, \dots, x_{i-1}^k, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k) = 0, \quad i = 1, 2, \dots, n.$$

При этом необходимо решить  $n$  скалярных уравнений, независимых друг от друга. Порядок решения уравнений произволен.

**Нелинейный метод Зейделя.** В этом методе новое итерационное приближение находится из уравнений

$$f_i(x_1^{k+1}, x_2^{k+1}, \dots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k) = 0, \quad i = 1, 2, \dots, n.$$

При реализации метода необходимо решить  $n$  скалярных уравнений, но при решении  $k$ -го уравнения используется информация, полученная при решении предыдущих  $k - 1$  уравнений.

**Гибридный метод: внешние итерации по Зейделю, а внутренние — по Ньютону.** Метод соответствует ситуации, когда внешние итерации выполняются методом Зейделя, а внутренние — методом Ньютона:

$$\begin{aligned} \frac{\partial f_i}{\partial x_i}(x_1^{k+1}, \dots, x_{i-1}^{k+1}, (x_i^{k+1})^m, x_{i+1}^k, \dots, x_n^k)((x_i^{k+1})^{(m+1)} - (x_i^{k+1})^m) + \\ + f_i(x_1^{k+1}, \dots, x_{i-1}^{k+1}, (x_i^{k+1})^m, x_{i+1}^k, \dots, x_n^k) = 0. \end{aligned}$$

Если сделать всего одну внутреннюю итерацию, приняв ее результат за  $x_i^{k+1}$ , то получим некий новый явный метод. В случае  $n = 2$  он имеет вид

$$\begin{aligned} \frac{\partial f_1}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1} - x_1^k) + f_1(x_1^k, x_2^k) = 0, \\ \frac{\partial f_2}{\partial x_2}(x_1^{k+1}, x_2^k)(x_2^{k+1} - x_2^k) + f_2(x_1^{k+1}, x_2^k) = 0. \end{aligned}$$

**Гибридный метод: внешние итерации по Ньютону, внутренние — по Зейделю.** Метод соответствует ситуации, когда внешние итерации выполняются методом Ньютона, а внутренние — методом Зейделя. Внешние итерации метода Ньютона имеют вид

$$F'(x^k)(x^{k+1} - x^k) + F(x^k) = 0,$$

где

$$F' = A_- + D + A_+,$$

$A_-$ ,  $A_+$  — нижняя и верхняя треугольные матрицы,  $D$  — диагональная матрица. Тогда внутренние итерации имеют вид:

$$(A_- + D)(x^{(k+1)})^{(m+1)} + A_+(x^{(k+1)})^m - F'(x^{(k)})(x^k) + F(x^k) = 0.$$

Если опять совершить одну внутреннюю итерацию, то получим новый явный метод. Для случая  $n = 2$  он имеет вид

$$\begin{aligned} \frac{\partial f_1}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1} - x_1^k) + f_1(x_1^k, x_2^k) = 0, \\ \frac{\partial f_2}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1} - x_1^k) + \frac{\partial f_2}{\partial x_2}(x_1^k, x_2^k)(x_2^{k+1} - x_2^k) + f_2(x_1^k, x_2^k) = 0. \end{aligned}$$

Этот метод отличается от предыдущего.

Два последних метода демонстрируют сравнительную легкость конструирования новых итерационных методов решения систем нелинейных уравнений.

### 2.3. Библиографические комментарии

Современные задачи математической физики, как правило, являются нелинейными. Поэтому решать нелинейные уравнения и системы таких уравнений приходится очень часто. В связи с этим вопросы решения таких уравнений обсуждаются практически во всей учебной литературе, приведенной в библиографии.

Отметим особо книгу [129], целиком посвященную решению больших систем нелинейных уравнений. В ней содержится как теоретический материал, так и ценные практические указания.

Во времена низкой производительности ЭВМ особое внимание уделялось построению методов повышенной скорости сходимости для решения нелинейных уравнений, например в классическом руководстве [17, 18], в частности, подробно обсужден алгоритм решения проблемы собственных значений путем нахождения нулей определителя матрицы  $A - \lambda E$ .

Отметим специально книги [25] и [80], содержащие большой материал по методам решения нелинейных уравнений.

### 3. МЕТОДЫ ИНТЕРПОЛИРОВАНИЯ ФУНКЦИЙ

Представлены варианты одномерной и многомерной интерполяции функций. Рассмотрены глобальная и локальная полиномиальные интерполяции. Описаны *интерполяционные полиномы в форме Лагранжа и Ньютона, полином Эрмита*. Исследованы сходимость интерполяции при увеличении числа точек, устойчивость интерполяции по отношению к ошибкам исходной функции, зависимость ошибки интерполяции от гладкости функции. Введено принципиальное понятие *насыщаемости алгоритма*, в данном случае — интерполяции. Описан алгоритм *сплайн-интерполяции*. Изложены простейшие способы двумерной интерполяции, в том числе с помощью *конечных элементов*.

#### 3.1. Постановка задачи интерполяции. Простейшие варианты интерполирования

Идеальный вариант решения задачи в большинстве случаев состоит в нахождении функциональной зависимости решения от входных данных, т. е. нахождении функции  $y = f(x)$ , где  $x$  — одна или несколько независимых переменных из некоторого множества  $\Omega$ . Однако для передачи решения приближенно такая полная информация бывает излишней. Для представления решения достаточно выдать значения функции  $f(x)$  на некоторой сетке  $\Omega_h$  (дискретное множество значений аргумента  $x$  из множества  $\Omega$ ). Пусть  $x_i$  — элемент сетки.

**Определение 3.1.** Процедура сопоставления функции  $f(x)$  ее сеточных значений  $f_i$  в точках  $x_i$  называется *операцией ограничения функции* на сетку (часто обозначается  $R$  — от restriction), т. е.

$$R: f(x) \rightarrow \{f_i\}.$$

Как правило,  $f_i = f(x_i)$ . Таким образом создаются таблицы, графики, схемы и т. д. Однако потребитель может захотеть вычислить значение функции  $f(x)$  не в точке  $x_i$ , а в другом месте, не совпадающем с точкой сетки. Тогда возникает необходимость по данным  $\{f_i\}$  определить некоторую функцию  $\tilde{f}(x)$ , которая в идеале является некоторым приближением исходной функции  $f(x)$ .

**Определение 3.2.** Процедура восстановления  $\tilde{f}$  по  $\{f_i\}$  называется **интерполяцией** (часто обозначается  $I$  — от interpolation).

В итоге имеем следующую схему:

$$f(x) \xrightarrow{R} \{f_i\} \xrightarrow{I} \tilde{f}(x).$$

При этом необходимо знать, какова норма ошибки  $\|f - \tilde{f}\|$ . Ясно, что наиболее сложно определить ее в случае неизвестной функции  $f(x)$ .

Очевидно, что вариантов построения функции  $\tilde{f}(x)$  бесконечно много даже в простейшем, одномерном, случае, который мы и рассмотрим для начала.

Итак, пусть  $\Omega = \{x: a \leq x \leq b\}$ . В качестве сетки выберем набор точек

$$\Omega_h = \{a = x_0 < x_1 < x_2 \cdots < x_n = b\}.$$

Пусть в точках сетки  $x_i$  известны значения функции  $f_i$ . Необходимо по ним восстановить функцию  $\tilde{f}(x)$ . Рассмотрим разные способы восстановления.

**Определение 3.3.** Термин «интерполяция» в узком смысле обычно употребляют, если значение аргумента восстанавливаемой функции находится между крайними точками задания функции. Если же значение аргумента выходит за границы, то говорят об **экстраполяции**.

### 3.1.1. Кусочно-линейная интерполяция

Потребуем, чтобы  $\tilde{f}(x_i) = f_i$  и  $\tilde{f}(x)$  была линейна на каждом участке  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ . Тогда очевидно, что

$$\tilde{f}(x) = \frac{1}{x_i - x_{i-1}} ((x - x_{i-1})f_i - (x - x_i)f_{i-1}), \quad x \in [x_{i-1}, x_i].$$

Введем функции

$$\varphi_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0}, & x \in [x_0, x_1]; \\ 0, & x \geq x_1; \end{cases} \quad \varphi_n(x) = \begin{cases} \frac{x - x_{n-1}}{x_n - x_{n-1}}, & x \in [x_{n-1}, x_n]; \\ 0, & x \leq x_{n-1}; \end{cases}$$

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i]; \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}]; \\ 0, & x \notin [x_{i-1}, x_{i+1}], \end{cases} \quad i = 1, 2, \dots, n-1.$$

**Определение 3.4.** Носители функций  $\varphi_i(x)$ , т. е. множества, на которых функции отличны от нуля, называются **конечными элементами**. Самы функции  $\varphi_i(x)$  называются **базисными функциями**  $i$ -го конечного элемента или **функциями формы** данного элемента. Иногда их также называют конечными элементами.

Введенные выше функции  $\varphi_i(x)$  являются простейшими **одномерными** кусочно-линейными **базисными функциями** конечных элементов (рис. 3.1).

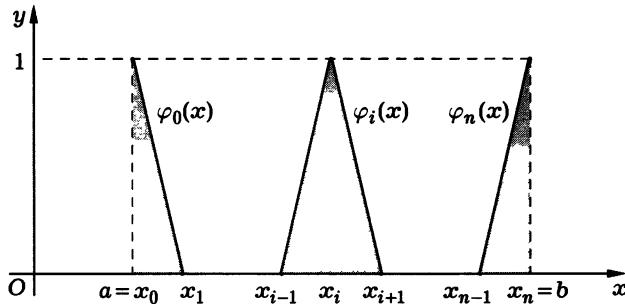


Рис. 3.1

Используя построенные функции формы, запишем

$$\tilde{f}(x) = \sum_{i=0}^n \varphi_i(x) f_i.$$

**Теорема 3.1.** Пусть  $f(x)$  — липшиц-непрерывная функция с постоянной  $q$  на отрезке  $[a, b]$ , т. е.

$$|f(x') - f(x'')| \leq q|x' - x''|, \quad x', x'' \in [a, b].$$

Тогда

$$|f(x) - \tilde{f}(x)| \leq \frac{1}{2}qh, \quad x \in [a, b],$$

где

$$h = \max_{1 \leq i \leq n} |x_i - x_{i-1}|.$$

Эта оценка неулучшаема в данном классе функций.

◀ Пусть  $x \in [x_{i-1}, x_i]$ ,  $h_i = x_i - x_{i-1}$ ,  $x = x_{i-1} + \alpha h_i$ ,  $\alpha = (x - x_{i-1})/h_i$ , где  $0 \leq \alpha \leq 1$ . Тогда  $\tilde{f}(x) = \alpha f_i + (1 - \alpha) f_{i-1}$  и

$$\begin{aligned} |f(x) - \tilde{f}(x)| &= |\alpha f_i + (1 - \alpha) f_{i-1} - \alpha f_i - (1 - \alpha) f_{i-1}| \leq \\ &\leq \alpha|f_i - f_{i-1}| + (1 - \alpha)|f_{i-1} - f_{i-1}| \leq \alpha q|x - x_i| + (1 - \alpha)q|x - x_{i-1}| = \\ &= \alpha q(1 - \alpha)h_i + (1 - \alpha)q\alpha h_i = 2q\alpha(1 - \alpha)h_i \leq \frac{1}{2}qh_i. \end{aligned}$$

Отсюда

$$\|f - \tilde{f}\|_C \leq \frac{1}{2}qh, \quad x \in [a, b].$$

Данная оценка неулучшаема на функциях рассматриваемого класса. Покажем это на примере функции, для которой реализуется строгое равенство. Пусть  $n = 1$ ,  $a = -1$ ,  $b = 1$ ,  $f(x) = |x|$  — липшиц-непрерывная функция с постоянной  $q = 1$ : для  $x', x'' \geq 0$  (или  $x', x'' \leq 0$ ) это очевидно ( $|f'(x)| = 1$ ). При  $x' \geq 0$ ,  $x'' \leq 0$  выполнена цепочка соотношений

$$|f(x') - f(x'')| = |x' - |x''|| = |x' + x''| \leq |x'| + |x''| = x' - x'' = |x' - x''|.$$

Обратная ситуация рассматривается аналогично.

Для данного случая  $h = 2$ ,  $\tilde{f} = 1$ , откуда

$$\max_{[-1,1]} |f - \tilde{f}| = 1 = \frac{qh}{2}$$

достигается в точке  $x = 0$ . ►

**Пример 3.1.** Приведем пример линейной функции, наилучшим образом (в среднеквадратичном смысле) приближающей функцию  $f(x)$ , заданную своими тремя значениями:  $f_0$ ,  $f_1$ ,  $f_2$ , в точках  $x_0$ ,  $x_1$ ,  $x_2$ . Для простоты расстояние между точками примем одинаковым, т. е.  $x_2 - x_1 = x_1 - x_0 = h$ . Будем искать аппроксимирующую функцию в виде  $\varphi(x) = a + b(x - x_1)$ , неизвестные коэффициенты  $a, b$  при этом находим из условия

$$\delta^2 = \sum_{k=0}^2 (\varphi(x_k) - f_k)^2 \sim \min.$$

В результате получаем решение задачи в виде

$$a = \frac{1}{3}(f_0 + f_1 + f_2), \quad b = \frac{1}{2h}(f_2 - f_0).$$

При этом минимальное среднеквадратичное отклонение

$$\delta^2 = \frac{1}{6}h^4(f_{\bar{x}x})^2, \quad f_{\bar{x}x} = \frac{1}{h^2}(f_0 - 2f_1 + f_2).$$

Как мы увидим в 4, величина  $f_{\bar{x}x}$ , называемая разностной производной второго порядка, при условиях определенной гладкости функции аппроксимирует обычную вторую производную и равна ее значению в некоторой средней точке.

**Определение 3.5.** Алгоритм построения аппроксимирующей функции из условия минимума среднеквадратичного отклонения называется *методом наименьших квадратов*.

### 3.1.2. Варианты интерполяции

Очевидно, что их бесконечно много.

**Пример 3.2.** Пусть  $x_1 = -1$ ,  $x_2 = 0$ ,  $x_3 = 1$ , во всех узлах  $f_i = 0$ . Тогда функция

$$\tilde{f}(x) = c(x+1)^\alpha x^\beta (x-1)^\gamma$$

для произвольных (с очевидными оговорками) положительных  $\alpha$ ,  $\beta$ ,  $\gamma$  и любого  $c$  проходит через эти три точки. Если же  $f_i \neq 0$ , то к любой функции, проходящей через эти три точки, можно добавить еще и  $\tilde{f}$ . Полученная кривая также проходит через точки  $(x_i, f_i)$ ,  $i = 1, 2, 3$ . #

Кроме того, часто результирующая функция становится более гладкой, если сделать подходящую замену переменных

$$\xi = \varphi(x), \quad \eta = \psi(y).$$

Ее необходимо либо угадывать, либо выбирать с учетом дополнительных (физических, технических) данных.

Возможны интерполяции с помощью рациональных функций (например, дробно-линейных). Можно функции интерполировать по интервалам, можно требовать точного прохождения интерполяционной функции через заданные точки, а можно и не требовать (метод наименьших квадратов приближения функции, имеющий число параметров, меньшее  $n$ ).

Возможна постановка общей задачи приближения заданной функции  $f(x)$  на некотором участке суммой вида

$$\sum_{i=0}^n c_i \varphi_i(x).$$

Как известно, в случае тригонометрических полиномов в результате минимизации ошибки в  $L_2$  получается частичная сумма ряда Фурье.

На практике наиболее распространены интерполяция полиномиальная, тригонометрическая и сплайн-интерполяция.

## 3.2. Полиномиальная интерполяция

Базой подобных приближений является следующая теорема

**Теорема 3.2 (теорема Вейерштрасса).** Для любой непрерывной на  $[a, b]$  функции  $f(x)$  существует полином  $P_n(x)$ , приближающий  $f(x)$  с любой наперед заданной точностью:

$$\forall \varepsilon > 0 \quad \exists P_n(x): \quad \|f(x) - P_n(x)\|_C < \varepsilon.$$

Для построения аппроксимации поступим следующим образом. Будем искать полином степени  $n$  вида

$$P_n(x) = a_0 + a_1x + \cdots + a_nx^n,$$

проходящий через  $n + 1$  точек  $(x_i, f_i)$ ,  $i = 0, 1, \dots, n$ . Для определения параметров получим систему из  $n + 1$  уравнений с  $n + 1$  неизвестными:

$$a_0 + \sum_{k=1}^n a_k x_i^k = f_i, \quad i = 0, 1, \dots, n.$$

**Определение 3.6.** Полином  $P_n(x)$ , удовлетворяющий описанным условиям, называется *интерполяционным полиномом (интерполянтом)*.

**Теорема 3.3.** Указанный интерполяционный полином существует и является единственным при  $x_i \neq x_j$ ,  $i \neq j$ .

◀ Определитель данной матрицы является *определителем Вандермонда*, который, как известно из курса линейной алгебры, отличен от нуля, если точки сетки не совпадают. Следовательно, решение задачи определения коэффициентов полинома существует и оно единственno.

Приведем другое доказательство единственности. Допустим, что утверждение теоремы неверно. Тогда существуют хотя бы два разных полинома  $P_n^{(1)}(x)$  и  $P_n^{(2)}(x)$ , удовлетворяющих поставленным условиям. Их разность — полином  $P_n(x) = P_n^{(1)}(x) - P_n^{(2)}(x)$  — также является полиномом  $n$ -го порядка и в точках сетки он должен удовлетворять условиям  $P_n(x_i) = 0$ ,  $i = 0, 1, \dots, n$ , т. е. быть равным нулю в  $n + 1$  точках. Такое возможно только в случае  $P_n(x) \equiv 0$ , когда решение задачи построения интерполяционного полинома единственно. ►

**Замечание 3.1.** Заметим, что полиномиальная интерполяция есть частный случай приближения функции, заданной таблично, суммами вида  $\sum_{i=0}^n c_i \varphi_i(x)$  при  $\varphi_i(x) = x^i$ ,  $i = 0, 1, \dots, n$ .

В общей ситуации обычно рассматривается интерполяция по так называемой *чебышевской системе функций*. Так называют систему из  $n + 1$  функций  $\varphi_i(x)$ ,  $i = 0, 1, \dots, n$ , любая линейная комбинация которых не может иметь  $n + 1$  различных корней на участке интерполяции.

Известно, что  $1, x, x^2, \dots, x^n$  образуют систему Чебышева на любом отрезке.

### 3.2.1. Интерполяционный полином в форме Лагранжа

Запишем полином

$$P_n(x) = L_n(x) = \sum_{k=0}^n f_k \prod_{\substack{i=0 \\ i \neq k}}^n \frac{(x - x_i)}{(x_k - x_i)}.$$

В этом соотношении, называемом **интерполяционным полиномом в форме Лагранжа**, внутренний полином  $n$ -го порядка обращается в нуль в точках  $x_i$  при  $i \neq k$  и в единицу в точке  $x_k$ . Следовательно, он реализует интерполяционный полином. Его можно переписать с помощью функции

$$\omega(x) = \prod_{i=0}^n (x - x_i).$$

Тогда **базисный полином**

$$\varphi_k^n(x) = \prod_{\substack{i=0 \\ i \neq k}}^n \frac{(x - x_i)}{(x_k - x_i)} = \frac{\omega(x)}{(x - x_k)\omega'_x(x_k)}.$$

Таким образом,

$$L_n(x) = \sum_{k=0}^n f_k \varphi_k^n(x).$$

Введем **остаточный член интерполяционного полинома**:

$$r_n(x) = f(x) - L_n(x).$$

**Лемма 3.1.** Пусть функция  $f(x)$  имеет  $n + 1$  непрерывных производных на  $[a, b]$ . Тогда для любых  $x \in [a, b]$ ,  $\{x_i\}$ ,  $i = 0, 1, \dots, n$ , и  $f$  существует точка  $\xi \in [a, b]$ , такая, что

$$r_n(x) = \frac{1}{(n+1)!} \omega(x) f^{(n+1)}(\xi).$$

◀ Рассмотрим вспомогательную функцию

$$g(t) = f(t) - L_n(t) - r_n(x) \frac{\omega(t)}{\omega(x)},$$

где  $x$  — параметр. Функция  $g(t)$  имеет  $n + 1$  корней  $t = x_i$ , в которых  $f(t) = L_n(t)$  и  $\omega(t) = 0$ , а также  $(n + 2)$ -й корень  $t = x$ , в котором  $g(t) = g(x) = 0$  в силу определения остаточного члена  $r_n(x)$ . Таким

образом, производная  $g'_t(t)$  имеет не менее  $n + 1$  корней,  $g''_{t^2}(t)$  — не менее  $n$  корней и так далее,  $g_{t^{n+1}}^{(n+1)}(t)$  имеет как минимум один корень  $\xi \in [a, b]$ . В этой точке

$$\begin{aligned} g_{t^{n+1}}^{(n+1)}(\xi) &= f^{(n+1)}(\xi) - L_n^{(n+1)}(\xi) - r_n(x) \frac{(n+1)!}{\omega(x)} = \\ &= f^{(n+1)}(\xi) - r_n(x) \frac{(n+1)!}{\omega(x)} = 0, \end{aligned}$$

откуда

$$r_n(x) = \frac{1}{(n+1)!} \omega(x) f^{(n+1)}(\xi). \quad \blacktriangleright$$

**Замечание 3.2.** Если  $M_{n+1} = \|f^{(n+1)}\|_C$ , то оценку точности приближения можно записать в виде

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |\omega(x)|.$$

Эта оценка справедлива как при  $x \in [a, b]$  (интерполяция), так и при  $x \notin [a, b]$  (экстраполяция). Но в последнем случае точка  $\xi$  может уже находиться вне участка  $[a, b]$  ( $\xi \in [\min(a, x), \max(b, x)]$ ). При этом функция  $f(x)$  должна иметь  $n + 1$  непрерывных производных на указанном расширенном отрезке.

**Теорема 3.4.** Пусть  $x_i - x_{i-1} = h$ ,  $i = 1, 2, \dots, n$ ,  $x \in [a, b]$ ,  $f(x)$  имеет непрерывную и, значит, ограниченную производную  $(n+1)$ -го порядка на  $[a, b]$ . Тогда

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}}{n+1} h^{n+1}.$$

◀ Пусть  $x = x_{i^*} + \alpha h$ ,  $\alpha \neq 0, 1$  (в противном случае  $\omega(x) = 0$ ). Тогда

$$\begin{aligned} \omega(x) &= \prod_{i=0}^n (x - x_i) = h^{n+1} \prod_{i=0}^n (i^* + \alpha - i) = \\ &= h^{n+1} \prod_{i=0}^{i^*} (i^* + \alpha - i) \prod_{i=i^*+1}^n (i^* + \alpha - i). \end{aligned}$$

Отсюда получаем оценку

$$|\omega(x)| = h^{n+1} \prod_{i=0}^{i^*} (i^* + \alpha - i) \prod_{i=i^*+1}^n (i - \alpha - i^*) \leq h^{n+1} (i^* + 1)! (n - i^*)!$$

При этом  $i^* \leq n - 1$ . В результате

$$\frac{|\omega(x)|}{n!} \leq h^{n+1} \frac{(i^* + 1)!}{n(n-1)\cdots(n-i^*+1)}.$$

В числителе  $(i^* + 1)$  сомножителей, в знаменателе  $n - (n - i^*) = i^*$  сомножителей, последний из которых  $n - i^* + 1 \geq 2$ . Отсюда

$$\frac{|\omega(x)|}{n!} \leq h^{n+1}.$$

В результате получим, что

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}}{n+1} h^{n+1}. \quad \blacktriangleright$$

**Следствие 3.1.** Рассмотрим случай  $n = 1$ , т. е. линейную интерполяцию на участках  $[x_{i-1}, x_i]$ . Тогда

$$|f(x) - L_n(x)| \leq \frac{M_2}{8} h^2, \quad x \in [x_{i-1}, x_i].$$

Эта оценка неулучшаема в классе функций, имеющих ограниченную вторую производную.

◀ В самом деле,

$$\|r_1\|_C \leq \frac{M_2}{2!} \|(x - x_{i-1})(x - x_i)\|_C \leq \frac{M_2}{2} \frac{h^2}{4} = \frac{1}{8} M_2 h^2.$$

Для доказательства того, что оценка неулучшаема, достаточно рассмотреть в качестве примера функцию

$$f(x) = \left( x - \frac{x_{i-1} + x_i}{2} \right)^2$$

с параметрами  $x_{i-1} = -1$ ,  $x_i = 1$ , для которой  $M_2 = 2$ ,  $h = 2$ ,  $L_1 = 1$ ,  $\|r_1\|_C = 1$ , т. е. приведенная оценка является точной. ►

Рассмотрим вопрос об ошибке экстраполяции, т. е. вычислении  $L_n(x)$  при  $x \notin [a, b]$ . Тогда анализ, подобный приведенному в теореме 3.4, показывает, что:

при  $x \in [b, b + h]$

$$|f - L_n| = |r_n| \leq h^{n+1} \max_{\xi \in [a, x]} |f^{(n+1)}(\xi)|,$$

при  $x \in [b + h, b + 2h]$

$$|f - L_n| = |r_n| \leq h^{n+1}(n+2) \max_{\xi \in [a, x]} |f^{(n+1)}(\xi)|,$$

при  $x \in [b + 2h, b + 3h]$ :

$$|f - L_n| = |r_n| \leq \frac{1}{2} h^{n+1} (n+2)(n+3) \max_{\xi \in [a, x]} |f^{(n+1)}(\xi)|$$

и т. д. Из полученных оценок следует, что ошибки начинают резко возрастать: вспомним ошибки планов и прогнозов. Формальное пояснение представлено на рис. 3.2, где изображена зависимость функции  $\omega(x)$ , обращающаяся в нуль в четырех точках (т. е.  $n = 3$ ). Вне области интерполяирования данная функция резко возрастает. Этот рост проявляется в виде неустойчивости экстраполяции.

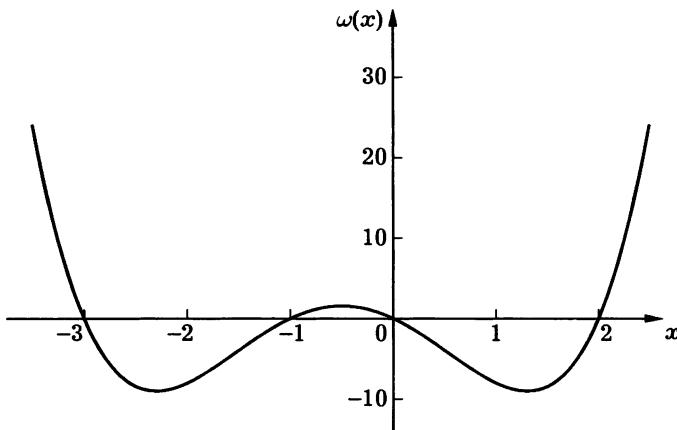


Рис. 3.2

### 3.2.2. Интерполяционный полином в форме Ньютона

**Определение 3.7.** Назовем *разделенными разностями* нулевого порядка значения функции в точке  $x_i$ , первого порядка — значения

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i},$$

второго порядка — значения

$$f(x_i, x_j, x_k) = \frac{f(x_j, x_k) - f(x_i, x_j)}{x_k - x_i}$$

и т. д. Разность  $k$ -го порядка имеет вид

$$f(x_1, x_2, \dots, x_{k+1}) = \frac{f(x_2, x_3, \dots, x_{k+1}) - f(x_1, x_2, \dots, x_k)}{x_{k+1} - x_1}.$$

**Лемма 3.2.** Справедливо равенство

$$f(x_1, x_2, \dots, x_k) = \sum_{j=1}^k \frac{f(x_j)}{\prod_{\substack{i \neq j \\ i=1}}^{k} (x_j - x_i)}.$$

◀ Применим метод математической индукции. Для  $k = 2$  равенство верно. Пусть оно верно и для  $k - 1$ . Тогда для разности порядка  $k$  имеем

$$\begin{aligned} f(x_1, x_2, \dots, x_{k+1}) &= \frac{1}{x_{k+1} - x_1} (f(x_2, x_3, \dots, x_{k+1}) - f(x_1, x_2, \dots, x_k)) = \\ &= \frac{1}{x_{k+1} - x_1} \left( \sum_{j=2}^{k+1} \frac{f(x_j)}{\prod_{\substack{i \neq j \\ 2 \leq i \leq k+1}}^{k+1} (x_j - x_i)} - \sum_{j=1}^k \frac{f(x_j)}{\prod_{\substack{i \neq j \\ 1 \leq i \leq k}}^k (x_j - x_i)} \right). \end{aligned}$$

Слагаемые в сумме при  $j = k + 1$  дают выражение нужного вида, входящее в предполагаемую формулу. Аналогично и при  $j = 1$ . При  $2 \leq j \leq k$  имеем коэффициент перед  $f(x_j)$ , равный

$$\begin{aligned} \frac{1}{\prod_{\substack{i \neq j \\ 2 \leq i \leq k+1}}^{k+1} (x_j - x_i)} - \frac{1}{\prod_{\substack{i \neq j \\ 1 \leq i \leq k}}^k (x_j - x_i)} &= \\ &= \frac{(x_j - x_1) - (x_j - x_{k+1})}{\prod_{\substack{i \neq j \\ 1 \leq i \leq k+1}}^{k+1} (x_j - x_i)} = \frac{x_{k+1} - x_1}{\prod_{\substack{i \neq j \\ 1 \leq i \leq k+1}}^{k+1} (x_j - x_i)}. \end{aligned}$$

В этом выражении числитель сокращается со знаменателем  $x_{k+1} - x_k$ , стоящим перед разностью сумм. В результате получаем справедливость утверждения леммы. ►

**Следствие 3.2.** Значение разделенной разности  $f(x_1, x_2, \dots, x_k)$  не зависит от порядка следования аргументов.

**Теорема 3.5.** *Интерполяционный полином  $P_n(x)$  может быть записан в форме Ньютона:*

$$\begin{aligned} P_n(x) &= f(x_0) + (x - x_0)f(x_0, x_1) + (x - x_0)(x - x_1)f(x_0, x_1, x_2) + \dots \\ &\quad \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1})f(x_0, x_1, \dots, x_n). \end{aligned}$$

◀ Проведем доказательство методом математической индукции. Обозначим через  $P_i(x)$  интерполяционный полином (записанный, например, в форме Лагранжа), проходящий через точки  $x_0, x_1, \dots, x_i$ . Для случая  $i = 1$  видно непосредственно, что

$$P_1(x) = L_1(x) = f(x_0) + (x - x_0)f(x_0, x_1).$$

Допустим, что это верно и при  $i - 1 > 1$ . Тогда многочлен  $P_{i-1}(x)$  может быть записан как в форме Ньютона, так и в форме Лагранжа. Представим многочлен  $P_i$  в виде  $P_i = P_{i-1} + (P_i - P_{i-1})$ . При этом оба многочлена  $P_i, P_{i-1}$  проходят через точки  $x_0, x_1, \dots, x_{i-1}$ , т.е. их разность равна нулю в этих точках. Поскольку  $(P_i - P_{i-1})$  есть многочлен степени  $i$ , то

$$P_i - P_{i-1} = A_i(x - x_0)(x - x_1) \cdots (x - x_{i-1}).$$

Потребуем, чтобы коэффициент  $A_i$  был таким, что

$$P_i(x_i) = f(x_i) = P_{i-1}(x_i) + A_i(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1}).$$

Отсюда

$$\begin{aligned} A_i &= \frac{f(x_i) - \sum_{j=0}^{i-1} f(x_j) \prod_{\substack{k=0 \\ k \neq j}}^{i-1} \frac{(x_i - x_k)}{(x_j - x_k)}}{\prod_{k=0}^{i-1} (x_i - x_k)} = \frac{f(x_i)}{\prod_{k=0}^{i-1} (x_i - x_k)} - \\ &- \sum_{j=0}^{i-1} \frac{f(x_j)}{\prod_{\substack{k=0 \\ k \neq j}}^{i-1} (x_j - x_k)(x_i - x_j)} = \frac{f(x_i)}{\prod_{k=0}^{i-1} (x_i - x_k)} + \sum_{j=0}^{i-1} \frac{f(x_j)}{\prod_{\substack{k=0 \\ k \neq j}}^i (x_j - x_k)} = \\ &= \sum_{j=0}^i \frac{f(x_j)}{\prod_{\substack{k=0 \\ k \neq j}}^i (x_j - x_k)} = f(x_0, x_1, \dots, x_i). \end{aligned}$$

В результате

$$P_i = P_{i-1} + (x - x_0)(x - x_1) \cdots (x - x_{i-1})f(x_0, x_1, \dots, x_i),$$

тогда справедлива формула Ньютона представления интерполяционного полинома. ►

Форма Ньютона интерполяционного полинома более удобна для численных расчетов, в особенности «ручных», чем форма Лагранжа.

### 3.2.3. Интерполяционный полином Эрмита

Рассмотрим несколько иную постановку задачи. Пусть есть  $m + 1$  точек  $x_0 = a < x_1 < \dots < x_m = b$ , в каждой из которых (в  $x_k$ ) известно  $n_k$  величин  $f^{(0)}(x_k), f^{(1)}(x_k), \dots, f^{(n_k-1)}(x_k)$ , где  $n_k$  — кратность  $k$ -го узла. Построим интерполяционный многочлен, проходящий через эти точки так, что в каждой точке  $x_k$  он и его производные порядков  $1, 2, \dots, n_k - 1$  принимают заданные значения. Будем строить многочлен степени

$$n = \sum_{k=0}^m n_k - 1.$$

**Определение 3.8.** Многочлен, удовлетворяющий указанным условиям, называется **интерполяционным полиномом Эрмита**, а сама интерполяция — интерполяцией с кратными узлами.

Можно доказать, что такой многочлен  $H_n(x)$  существует и является единственным. Поскольку  $H_n(x)$  представляет собой полином  $n$ -го порядка и удовлетворяет системе из  $n + 1$  уравнений

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \quad i = 0, 1, \dots, n_k - 1, \quad k = 0, 1, \dots, m,$$

коэффициенты  $H_n(x)$  линейно выражаются через  $f^{(i)}(x_k)$ . Поэтому полином  $H_n(x_k)$  можно искать в виде

$$H_n(x) = \sum_{k=0}^m \sum_{i=0}^{n_k-1} c_{ik}(x) f^{(i)}(x_k),$$

где  $c_{ik}(x)$  — полиномы  $n$ -го порядка.

Коэффициенты полинома Эрмита можно получить в явном виде по аналогии с коэффициентами полинома Лагранжа (Ньютона).

Проще это сделать, воспользовавшись формой Ньютона интерполяционного полинома. А именно, интерполяцию с кратными узлами можно представить как построение интерполяционного полинома по  $n + 1$  точкам, в которых каждая точка  $x_k$  повторяется  $n_k$  раз, т. е. число повторов соответствует ее кратности. Буквальное вычисление разделенных разностей по приведенным выше формулам при этом приведет к неопределенности типа  $0/0$ . Однако возникшие неопределенностии могут быть раскрыты согласно правилу Лопитала. При этом появляются производные.

Этим замечанием относительно общей ситуации мы и ограничимся. Часто оказывается, что в конкретном случае проще решить задачу

исходя из ее конкретной постановки, а не пользуясь готовой формулой. Приведем примеры.

**Пример 3.3.** Рассмотрим следующую задачу: требуется построить полином Эрмита по данным в  $m + 1$  точках  $x_0 = a < x_1 < \dots < x_m = b$ , в каждой из которых, т. е. в  $x_k$ , известны значения функции и ее производной (здесь  $n_k = 2$ ):

$$f(x_k) = f_k, \quad f'(x_k) = f'_k, \quad k = 0, 1, \dots, m.$$

В данном случае  $n = 2(m + 1) - 1 = 2m + 1$ . Построим полином в явном виде. Так как коэффициенты интерполяционного полинома линейно выражаются через заданные значения функции и производной в точках сетки, полином Эрмита можно записать в следующем виде:

$$H_n(x) = \sum_{k=0}^m (a_k(x)f_k + b_k(x)f'_k),$$

где  $a_k(x)$ ,  $b_k(x)$  — полиномы степени  $n$ . Сопоставление полинома выписанного вида и условий, налагаемых на него в точках сетки, дает следующие уравнения для определения коэффициентов полиномов  $a_k(x)$ ,  $b_k(x)$ ,  $k = 0, 1, \dots, m$ :

$$\begin{aligned} a_k(x_i) &= \delta_{ki}, \quad a'_k(x_i) = 0; \\ b_k(x_i) &= 0, \quad b'_k(x_i) = \delta_{ki}. \end{aligned}$$

Здесь  $i = 0, 1, \dots, m$ .

Каждый из полиномов содержит  $n + 1 = 2m + 2$  коэффициентов, которые еще не определены. Число решаемых уравнений равно числу неизвестных, так что задача может быть разрешимой.

Для построения полиномов в явном виде вспомним базисные полиномы Лагранжа  $\varphi_k^m(x)$ , удовлетворяющие условиям  $\varphi_k^m(x_i) = \delta_{ki}$ . Однако всем уравнениям, которым должны удовлетворять коэффициенты искомых полиномов  $a_k(x)$ ,  $b_k(x)$  степени  $2m + 1$ , они удовлетворять не могут, так как являются полиномами только степени  $m$ .

Основные свойства базисного полинома Лагранжа и тот факт, что нуль или единица остаются неизменными при их возведении в целую положительную степень, позволяют записать искомый полином  $b_k(x)$  в виде

$$b_k(x) = (x - x_k)(\varphi_k^m(x))^2.$$

Несколько сложнее предсказать вид второго полинома  $a_k(x)$ . Анализ накладываемых условий показывает, что его нужно искать в форме

$$a_k(x) = (1 + \alpha_k(x - x_k))(\varphi_k^m(x))^2,$$

где  $\alpha_k$  — пока не определенный коэффициент. Легко видеть, что первое уравнение, которому должен удовлетворять искомый полином, выполнено. Второе уравнение позволяет найти единственный неизвестный коэффициент. В результате имеем нужное решение

$$a_k(x) = (1 - 2\varphi_k^{m'}(x_k)(x - x_k))(\varphi_k^m(x))^2.$$

Искомый полином Эрмита построен.

**Пример 3.4.** Рассмотрим «несимметричную» задачу: требуется построить полином Эрмита по заданным в  $m + 1$  точках

$$x_0 = a < x_1 < \cdots < x_m = b$$

значениям функции. При этом в каждой точке известно значение функции и в одной точке  $x_{k_0}$  — значение ее производной:

$$f(x_k) = f_k, \quad k = 0, 1, \dots, m; \quad f'(x_{k_0}) = f'_{k_0},$$

В данном примере  $n_k = 1$  при  $k \neq k_0$ ;  $n_{k_0} = 2$ ;  $n = (m + 1) + 1 - 1 = m + 1$ . Решение задачи также нетрудно получить явно. Полином Эрмита можно записать в следующем виде:

$$H_n(x) = b_{k_0}(x)f'_{k_0} + \sum_{k=0}^m a_k(x)f_k,$$

где  $a_k(x)$ ,  $k = 0, 1, \dots, m$ , и  $b_{k_0}(x)$  — полиномы степени  $n$ . В данном случае имеем следующие уравнения для коэффициентов искомых полиномов  $a_k(x)$ ,  $k = 0, 1, \dots, m$ , и  $b_{k_0}(x)$  в точках сетки  $x_i$ ,  $i = 0, 1, \dots, m$ :

$$a_k(x_i) = \delta_{ki}, \quad b_{k_0}(x_i) = 0,$$

и уравнения в точке  $x_{k_0}$ :

$$a'_k(x_{k_0}) = 0, \quad b'_{k_0}(x_{k_0}) = 1.$$

Опуская рассуждения, аналогичные приведенным при рассмотрении предыдущего примера, запишем окончательное решение:

$$\begin{aligned} b_{k_0}(x) &= (x - x_{k_0})\varphi_{k_0}^m(x), \\ a_k(x) &= (1 + \alpha_k(x - x_k))\varphi_k^m(x), \\ \alpha_k &= -\frac{\varphi_k^{m'}(x_{k_0})}{\delta_{kk_0} + \varphi_k^{m'}(x_{k_0})(x_{k_0} - x_k)}. \end{aligned}$$

Построение закончено.

**Пример 3.5.** Простейшим примером полинома Эрмита является **полином Тейлора**, построенный по заданным в одной точке значениям функции и ее первых производных до  $n$ -го порядка включительно. Он имеет вид

$$T_n(x) = \sum_{k=0}^n \frac{1}{k!} f^{(k)}(x_0)(x - x_0)^k.$$

Полином Тейлора является не интерполянтом, а экстраполянтом. Из математического анализа хорошо известно, что разность значений  $n+1$  раз непрерывно дифференцируемой в некоторой окрестности точки  $x_0$  функции  $f(x)$  и ее полинома Тейлора  $T_n(x)$  в указанной окрестности может быть представлена через производную функции  $(n+1)$ -го порядка:

$$f(x) - T_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x - x_0)^{n+1},$$

где  $\xi \in (x_0, x)$  (или  $\xi \in (x, x_0)$ ). Как видно, выражение для ошибки очень похоже на выражение для ошибки интерполяции, полученное выше.

### 3.3. Сходимость и устойчивость полиномиальной интерполяции

При проведении *интерполяции* желательно, чтобы была достигнута минимальная ошибка  $\|f - \tilde{f}\|$  в некоторой норме и чтобы норма ошибки стремилась к нулю с увеличением числа точек интерполяции. В этом случае говорят, что интерполяционный процесс сходится. Для описания его сходимости используют те же характеристики, что и для описания сходимости функциональных последовательностей.

#### 3.3.1. Оптимизация узлов сетки

В случае полиномиальной аппроксимации  $\tilde{f}(x) = L_n(x)$  и

$$\|f - \tilde{f}\|_C = \|f - L_n\|_C = \|r_n\|_C \leq \frac{M_{n+1}}{(n+1)!} \|\omega\|_C.$$

Поставим задачу выбора такого набора узлов, чтобы равномерная (в  $C$ ) норма правой части была минимальна:

$$\min_{x_0, \dots, x_n} \max_{[a, b]} |\omega(x)|, \quad \omega(x) = \prod_{i=0}^n (x - x_i).$$

Это классическая задача о построении **полинома**  $(n+1)$ -го порядка, **наименее уклоняющегося от нуля**. В данном случае коэффициент при старшей степени неизвестного равен единице. Несмотря на это

различие, решение задачи является таким же, что и решение, которое приведено при построении *метода Ричардсона* (см. 1). Данная задача решена В.А. Марковым. Искомый полином является **полиномом Чебышева** первого рода и имеет вид

$$\omega(x) = T_{n+1}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}} \cos\left((n+1) \arccos \frac{2x-(b+a)}{b-a}\right),$$

корни полинома задаются следующим соотношением:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = 0, 1, \dots, n,$$

при этом они являются узлами интерполяции. Для такого набора узлов

$$\|\omega\|_C = \frac{1}{2^{2n+1}} (b-a)^{n+1}$$

и

$$\|f - L_n\|_C \leq \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Полученная оценка называется наилучшей равномерной оценкой погрешности интерполяции и является неулучшаемой. Для доказательства достаточно указать такую функцию, для которой в полученной оценке ошибки реализуется равенство.

Рассмотрим интерполяцию на чебышевской сетке из  $n+1$  узлов функции

$$f(x) = a_{n+1}x^{n+1} + a_nx^n + \dots + a_0$$

с помощью интерполяционного полинома  $n$ -й степени. Тогда  $M_{n+1} = \|f^{(n+1)}\|_C = a_{n+1}(n+1)!$

В результате получим следующее точное представление остаточного члена интерполяционного полинома:

$$f(x) - L_n(x) = \frac{1}{(n+1)!} \omega(x) f^{(n+1)}(\xi) = a_{n+1} \omega(x) = a_{n+1} T_{n+1}(x).$$

Норма выписанного остаточного члена в точности совпадает с наилучшей равномерной оценкой погрешности интерполяции, что доказывает неулучшаемость полученной оценки.

Если вспомнить формулу Стирлинга:  $n! \approx n^n e^{-n} \sqrt{2\pi n}$ , то наилучшая оценка погрешности может быть преобразована к виду

$$\frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}} \approx M_{n+1} \frac{e^{n+1}}{\sqrt{2\pi(n+1)}} \left(\frac{b-a}{n+1}\right)^{n+1} \frac{1}{2 \cdot 4^n}.$$

Если же интерполяция проводится на равномерной сетке, то

$$|f - L_n| \leq \frac{M_{n+1}}{n+1} h^{n+1} = \frac{M_{n+1}}{n+1} \left( \frac{b-a}{n} \right)^{n+1}.$$

Следовательно, оценка в случае интерполяции по узлам многочлена Чебышева отличается в

$$\sqrt{\frac{n+1}{2\pi}} \left(\frac{e}{4}\right)^n \frac{e}{2} \left(1 - \frac{1}{n+1}\right)^{n+1} \simeq \sqrt{\frac{n+1}{2\pi}} \left(\frac{e}{4}\right)^n \frac{1}{2}$$

раз. Для небольших  $n$  эта величина является не очень малой. Но чебышевские сетки обладают и рядом других преимуществ по сравнению с равномерными.

### 3.3.2. Устойчивость интерполяционного полинома относительно погрешностей функции

Пусть значения функции  $f$  известны не точно, а лишь с некоторой погрешностью  $\delta f_i$ :  $f_i = f_i^0 + \delta f_i$ . Возникает вопрос: как сильно исказится при этом интерполяционный полином? Интерполант, построенный по неточным значениям функции в силу линейности полинома по значениям функции равен

$$L_n(x) = L_n(x, f^0 + \delta f) = L_n(x, f^0) + L_n(x, \delta f).$$

В этом выражении второй аргумент у полинома  $L_n$  указывает на зависимость интерполяционного полинома от значений функции.

Для установления влияния ошибки входных данных на построенный полином необходимо оценить величину

$$\max_{\|\delta f\| \leq \delta} \|L_n(x, \delta f)\|_C.$$

Если ввести нормированную на единицу ошибку  $\delta f_i = \delta \cdot \widetilde{\delta f}_i$ , то норма возмущения  $\widetilde{\delta f}_i$  будет не больше единицы. Тогда для оценки влияния ошибки входных данных требуется вычислить величину

$$\eta = \max_{\|\widetilde{\delta f}\| \leq 1} \|L_n(x, \widetilde{\delta f})\|_C.$$

В этом случае оценка возмущения интерполанта по множеству  $\|\delta f\| \leq \delta$  равна  $\delta \eta$  в силу линейности полинома  $L_n$  по значениям функции.

Величина  $\eta$  определяется сеткой и является оценкой чувствительности интерполяционного полинома к погрешностям в задании значений  $f_i$ .

**Определение 3.9.** Величину  $\eta$  называют **нормой интерполяционного полинома** на данной сетке, так как

$$\|L_n\| \leq \eta \|f\|,$$

или **константой Лебега** интерполяционного полинома.

Приведем без доказательства важнейшие результаты: для равномерной сетки

$$\eta = O(2^n),$$

на чебышевской сетке

$$\eta = O(\ln n).$$

Отсюда видна сильная зависимость постоянной Лебега от структуры сетки.

Таким образом, погрешности, связанные с неточностью входной информации, на равномерной сетке растут как  $2^n$ . Поэтому на практике, как правило, равномерные сетки при интерполяции используются лишь при  $n \leq 5$  ( $2^5 = 32$ ).

### 3.3.3. Устойчивость интерполяционного полинома относительно априорной информации

Пусть для функции  $y = f(x)$  построен полином  $L_n(x)$  на некоторой сетке. Что будет, если в действительности  $f(x)$  не имеет  $(n+1)$ -й ограниченной производной?

Еще С. Н. Бернштейн доказал, что полиномы, интерполирующие функцию  $y = f(x) = |x|$  (только липшиц-непрерывную функцию) на равномерной сетке на отрезке  $[-1, 1]$ , таковы, что

$$\|f - L_n\|_C \rightarrow \infty \quad \text{при} \quad n \rightarrow \infty.$$

При этом полином  $L_n(x)$  не сходится к  $f(x)$  ни в одной точке отрезка  $[-1, 1]$  за исключением точек  $-1, 0, 1$ .

В общем случае верна **теорема Фабера**, которая утверждает, что для любой последовательности сеток  $\{\Omega_h\}$ ,  $\Omega_h \subset [a, b]$ , существует непрерывная на  $[a, b]$  функция  $f(x)$ , такая, что

$$\{L_n(x)\} \not\rightarrow f(x)$$

(т. е. не сходится равномерно) на  $[a, b]$ .

Однако есть и обратная теорема — **теорема Марцинкевича**: для любой непрерывной на  $[a, b]$  функции  $f(x)$  существует последовательность сеток  $\{\Omega_h\}$ , для которой

$$\{L_n(x)\} \rightrightarrows f(x)$$

на  $[a, b]$ .

Такие сетки нужно строить для каждой функции отдельно. Они очень сложны и в практических расчетах не используются.

Приведенные выше результаты показывают, что теория интерполяции есть глубокая и хорошо разработанная дисциплина математики.

В частности, доказаны существование и единственность интерполяционного полинома  $P_n$ , реализующего **наилучшее равномерное приближение**

$$\min_{P_n} \|P_n - f\|_C = E_n(f).$$

**Пример 3.6.** Рассмотрим задачу построения наилучшего равномерного приближения функции  $f(x) = x^p$  при  $p > 0$  на отрезке  $[0, 1]$  полиномом нулевого порядка  $P_0(x) = a$ .

Легко видеть, что при  $a \in [0, 1]$  имеем  $\|P_0 - f\|_C = \max(a, 1 - a)$ . Если искомое  $a$  не принадлежит отрезку  $[0, 1]$ , то норма ошибки заведомо больше. Найдем такое  $a$ , на котором достигается минимум рассматриваемой ошибки. Очевидно, что ответом является  $a = 0.5$ . #

Нам в очередной раз встретилась задача о минимаксе. Интересно, что ее ответ является частным случаем более общей ситуации.

**Пример 3.7.** Рассмотрим непрерывную на области задания своих аргументов функцию  $u = f(x_1, x_2, \dots, x_n)$ . Пусть ее аргументы заданы приближенно:  $x_i = x_i^* \pm \Delta(x_i^*)$ ,  $i = 1, 2, \dots, n$ . Как известно из теории ошибок, это означает, что

$$x_i \in [x_i^* - \Delta(x_i^*), x_i^* + \Delta(x_i^*)], \quad i = 1, 2, \dots, n.$$

Следовательно, функция рассматривается на  $n$ -мерном параллелепипеде  $G$ , стороны которого определяются заданием компонент аргумента.

Требуется найти постоянную  $u^*$ , которая бы наилучшим образом в смысле равномерной нормы приближала данную функцию на параллелепипеде  $G$ . Другая формулировка того же задания заключается в нахождении значения функции, наилучшим образом приближающего ее значение в случае неточного задания аргументов. Казалось бы, ответом является значение

$$u^{**} = f(x_1^*, x_2^*, \dots, x_n^*).$$

Однако это не так.

Поскольку функция непрерывна на замкнутом параллелепипеде  $G$ , она достигает на нем своих точных нижней и верхней граней:

$$u_1 = \inf_{(x_1, x_2, \dots, x_n) \in G} f(x_1, x_2, \dots, x_n), \quad u_2 = \sup_{(x_1, x_2, \dots, x_n) \in G} f(x_1, x_2, \dots, x_n).$$

В результате решаемая задача преобразуется следующим образом:

$$\begin{aligned} \min_{u^*} \|u^* - f\|_C &= \min_{u^*} \max_{(x_1, x_2, \dots, x_n) \in G} |u^* - f(x_1, x_2, \dots, x_n)| = \\ &= \min_{u^*} \max_{u \in [u_1, u_2]} |u^* - u|. \end{aligned}$$

Последняя задача уже мало чем отличается от предыдущей (см. пример 3.6). Опуская аналогичные рассуждения, получаем ответ:  $u^* = (u_1 + u_2)/2$ . #

Вернемся к задаче построения наилучшего равномерного приближения в общем случае. Алгоритм построения такого  $P_n$  известен, но он опять-таки очень сложен и на практике не применяется. Однако известно, в частности, что асимптотика  $E_n(f)$  при больших  $n$  однозначно связана с гладкостью функции  $f$ .

Если  $f$  имеет на  $[a, b]$  ограниченную  $l$ -ю производную, то при  $n \rightarrow \infty$

$$E_n(f) \sim O(n^{-l}).$$

Необычность ситуации состоит в том, что стандартная (на равномерной сетке, содержащей  $n+1$  точек) оценка отклонения  $L_n(x)$  от  $f(x)$  зависит от нормы  $(n+1)$ -й производной. Эта оценка вида

$$|r_n| \leq \frac{M_{n+1}}{n+1} h^{n+1}$$

неприменима при фиксированной гладкости функции и  $n \rightarrow \infty$ . Но даже при наличии производных любого порядка сходимость зависит от соотношения между  $M_{n+1}$  и  $h^{n+1}$ . Известны примеры функций  $f \in C^\infty[a, b]$ , для которых  $|r_n| \not\rightarrow 0$  при  $n \rightarrow \infty$ .

**Пример 3.8.** Удивительно простой пример такой функции приведен Рунге — это функция

$$f(x) = \frac{1}{1 + 25x^2}.$$

Использование глобальной полиномиальной интерполяции на равномерной сетке дает расходимость на участках  $|x| \in (0.73, 1)$  при бесконечном увеличении числа точек разбиения.

Причиной этого, очевидно, является рост нормы производной данной функции при увеличении ее порядка. #

Таким образом, лишь для полинома наилучшего приближения заведомо можно получить

$$r_n \sim O(n^{-l}),$$

где  $l$  фиксировано.

**Замечание 3.3.** Отметим, что сходимость интерполянта к исходной функции при наличии у нее нужной гладкости, обеспечивающей оценку ошибки интерполяции  $|r_n|$ , в случае фиксированного числа точек при уменьшении шага сетки гарантируется этой оценкой. Однако этот процесс предполагает возможность измельчать сетку, на которой заданы значения исходной функции, в непосредственной окрестности произвольной точки. Ясно, что это далеко не всегда возможно.

**Замечание 3.4.** Рассмотрим случай фиксированного участка, на котором заданы точки интерполяции, и функцию, имеющую производные любого порядка. Пусть количество точек может неограниченно возрастать, а сами точки могут располагаться произвольным образом. Обеспечивают ли указанные условия сходимость интерполянта? Ответ, вообще говоря, отрицательный. Для пояснения приведем пример функции

$$f(x) = \begin{cases} 0, & x \in [-1, 0]; \\ \exp\left(-\frac{1}{x}\right), & x \in (0, 1]. \end{cases}$$

Она имеет производные всех порядков. Однако если расположить точки интерполяции на левой половине отрезка  $[-1, 1]$ , то интерполянт будет тождественно равен нулю. Ошибка будет фиксирована и никак не будет зависеть от числа точек.

Рассмотрим вопрос о существовании полинома, дающего отклонение от  $f$ , близкое к оптимальному.

**Теорема 3.6.** Пусть  $L_n(x)$  — интерполяционный полином на чебышевской сетке для функции  $f(x)$ . Тогда

$$\|L_n - f\|_C \leq (1 + \tilde{C} \ln n) E_n(f).$$

◀ Запишем соотношение

$$f_i = f(x_i) = P_n(x_i) + f(x_i) - P_n(x_i).$$

В силу линейности  $L_n$  по значениям функции имеем

$$\begin{aligned} L_n(x) &= L_n(x, f_i) = L_n(x, P_n(x_i)) + L_n(x, f_i - P_n(x_i)) = \\ &= P_n(x) + L_n(x, f_i - P_n(x_i)). \end{aligned}$$

Равенство  $L_n(x, P_n(x_i)) = P_n(x)$  справедливо вследствие единственности интерполяционного полинома. Для оценки второго слагаемого используем неравенство

$$\|L_n(x, f_i - P_n(x_i))\|_C \leq \eta \|f - P_n\|_C \leq \tilde{C} \ln n E_n(f).$$

Следовательно,

$$\begin{aligned} \|L_n - f\|_C &= \|P_n - f + L_n(x, f_i - P_n(x_i))\|_C \leq \\ &\leq E_n(f) + \tilde{C} \ln n E_n(f) = (1 + \tilde{C} \ln n) E_n(f). \end{aligned} \quad \blacktriangleright$$

Таким образом, интерполяционный полином на чебышевской сетке вследствие «слабости» функции  $\ln n$  является почти наилучшим. И в случае функции, имеющей  $l$  ограниченных производных ( $l \geq 1$ ), с гарантией выполняется сходимость  $\|L_n - f\|_C \rightarrow 0$  при  $n \rightarrow \infty$ .

### 3.3.4. Наилучшие приближения в гильбертовом пространстве

До сих пор мы рассматривали в основном задачу об интерполяции функции, заданной таблично. Представим более подробно общую задачу об аппроксимации функции  $f(x)$ , являющейся элементом некоторого линейного нормированного пространства  $H$ . Рассмотрим в нем линейно независимую систему функций  $\varphi_i(x)$ ,  $i = 0, 1, \dots, n$ .

**Определение 3.10.** Функция

$$\varphi(x) = \sum_{i=0}^n c_i \varphi_i(x)$$

называется **обобщенным полиномом**, построенным по указанной системе.

Найдем такой обобщенный полином, который дает минимальное отклонение от функции  $f(x)$ .

**Определение 3.11.** Обобщенный полином  $\tilde{\varphi}(x)$ , являющийся решением задачи минимизации  $\min_{\varphi} \|f - \varphi\|_H$ , называется **элементом наилучшего приближения**.

Существование и единственность элемента наилучшего приближения определяются пространством  $H$ , которому принадлежат рассматриваемые функции.

Приведем пример пространства  $L_1[-1, 1]$ , в котором проводится приближение функции  $f(x) \equiv 1$  линейной функцией  $\varphi = cx$ . Тогда имеет место равенство

$$\|f - \varphi\|_{L_1} = \begin{cases} 2, & c \in [-1, 1]; \\ \frac{c^2 + 1}{|c|} > 2, & c \in (-\infty, -1) \cup (1, +\infty). \end{cases}$$

Таким образом, любая функция  $\varphi = cx$  при  $c \in [-1, 1]$  является наилучшим приближением единичной функции в смысле пространства  $L_1[-1, 1]$ .

Рассмотрим случай вещественного гильбертова пространства, в котором норма элемента есть квадратный корень из скалярного произведения элемента на себя. Наиболее известным примером такого пространства является пространство  $L_2$ , в котором скалярное произведение элементов есть интеграл от произведения функций (элементов), взятого в некоторых случаях с заданным положительным весом  $\rho$  по области изменения аргументов.

При минимизации погрешности аппроксимации возникает следующая система линейных алгебраических уравнений (СЛАУ) для коэффициентов элемента наилучшего приближения:

$$\sum_{j=0}^n \tilde{c}_j (\varphi_j, \varphi_i) = (f, \varphi_i).$$

Матрица данной системы является **матрицей Грама** используемой системы функций. В случае линейной независимости системы функций матрица Грама невырождена, поэтому элемент наилучшего приближения существует и является единственным.

**Лемма 3.3.** Если  $\tilde{\varphi}$  — элемент наилучшего приближения в  $H$ , то погрешность  $f - \tilde{\varphi}$  ортогональна  $\tilde{\varphi}$ , т. е.

$$(f - \tilde{\varphi}, \tilde{\varphi})_H = 0.$$

**Следствие 3.3.** Если  $\tilde{\varphi}$  — элемент наилучшего приближения в  $H$ , то

$$\|f - \tilde{\varphi}\|_H^2 = \|f\|_H^2 - \|\tilde{\varphi}\|_H^2.$$

Доказательство леммы и следствия опустим.

Лемма показывает, что погрешность аппроксимации элементом наилучшего приближения ортогональна ему. Это означает, что ошибка лежит в подпространстве, ортогональном линейной оболочке, натянутой на рассматриваемую систему. Следствие леммы является аналогом теоремы Пифагора и позволяет вычислить норму ошибки.

Запишем решение задачи построения полинома наилучшего приближения в случае ортонормированной системы функций  $\varphi_i(x)$ ,  $i = 0, 1, \dots, n$ . Тогда справедливы равенства

$$\tilde{c}_i = (f, \varphi_i), \quad i = 0, 1, \dots, n.$$

**Определение 3.12.** Определенные таким способом числа  $\tilde{c}_i$  называются коэффициентами Фурье элемента  $f \in H$  по ортонормированной системе  $\{\varphi_i(x)\}$ ,  $i = 0, 1, \dots, n$ , а обобщенный полином  $\tilde{\varphi}(x) = \sum_{i=0}^n \tilde{c}_i \varphi_i(x)$  — **многочленом (полиномом) Фурье**.

Если система функций  $\{\varphi_i(x)\}$ ,  $i = 0, 1, \dots$ , является полной, то имеет место равенство Парсеваля

$$\|f\|_H^2 = \sum_{i=0}^{\infty} c_i^2.$$

Следовательно, ряд, составленный из квадратов коэффициентов Фурье, сходится, а его остаток стремится к нулю. В силу этого ошибка приближения функции многочленом Фурье стремится к нулю в смысле пространства  $L_2$ . В результате аппроксимация возможна с любой наперед заданной точностью.

В случае рассмотрения тригонометрической системы функций получаем обычное приближение Фурье.

Успех аппроксимации обобщенным полиномом в конкретной ситуации определяется правильным выбором системы функций, с помощью которых такая аппроксимация производится, и соответствующего пространства, задаваемого в том числе и способом вычисления в нем нормы. При этом заведомо лучше использовать ортонормированную систему функций по сравнению с той исходной системой, ортогонализацией которой получена ортонормированная система. Использование неортогональной системы ведет к быстрому убыванию определителя решаемой системы и получению завышенных ошибок аппроксимации.

Например, матрица Грама самой естественной, казалось бы, системы функций  $\{\varphi_i(x) = x^i\}$ ,  $i = 0, 1, \dots$ , дает самый стандартный пример плохо обусловленной матрицы, называемой **матрицей Гильберта**. Ее использование уже при  $n > 5$  становится невозможным.

Описанное свойство часто трактуется как переполненность степенного базиса.

Если аппроксимируемая функция известна только в точках сетки, то аппроксимация с использованием интегралов становится невозможной. В этом случае часто интеграл заменяется конечной суммой по

точкам сетки, в которой перед каждым слагаемым стоит свой, вообще говоря, вес. Его значение, например, может отражать значимость данной конкретной точки или поведение функции. Элемент наилучшего приближения далее находится с использованием процедуры, практически аналогичной описанной выше. Однако значения коэффициентов элемента аналитически не выписываются.

В случае использования конечномерного аналога пространства  $L_2$  получается ранее обсуждавшийся метод наименьших квадратов, часто реализуемый при обработке экспериментальной информации.

**Замечание 3.5.** Отметим, что при рассмотрении экспериментальной информации, заданной приближенно, часто возникает задача приближенной аппроксимации с заданной точностью  $\varepsilon$ : среди всех обобщенных полиномов требуется найти полином, для которого  $\|f - \varphi\|_H < \varepsilon$ .

Очевидно, что для обеспечения единственности решения такой задачи необходимы дополнительные условия.

### 3.3.5. Насыщаемость алгоритма интерполяции. Тригонометрическая интерполяция

Вернемся еще раз к равномерным сеткам. Как показано выше,

$$|r_n| \leq \frac{M_{n+1}}{n+1} h^{n+1},$$

где  $n + 1$  — число узлов сетки. Возникает вопрос: улучшится ли качество интерполяции на данной сетке, если при фиксированном  $n$  рассмотреть функцию большей гладкости? Данная оценка показывает, что улучшения не произойдет. Ошибка и в этом случае останется величиной  $O(h^{n+1})$ .

**Определение 3.13.** Алгоритм, обладающий свойством независимости погрешности от увеличения гладкости функции, называют **насыщаемым алгоритмом**.

Возникает вопрос: существуют ли **ненасыщаемые алгоритмы** интерполяции? Ответ на него положительный. Примером такого алгоритма является **тригонометрическая интерполяция** функциями вида

$$Q_n(x) = a_0 + \sum_{k=1}^n (a_k \cos k\omega x + b_k \sin k\omega x), \quad \omega = \frac{2\pi}{b-a}.$$

Для таких полиномов, совпадающих с периодической функцией с периодом  $L = b - a$  в точках равномерной сетки  $a = x_0 < x_1 < \dots < x_{2n} < b$ , т. е.

$$Q_n(x_i) = f(x_i), \quad i = 0, 1, \dots, 2n,$$

имеют место следующие свойства:

погрешность

$$\|r_n\|_C = \|f - Q_n\|_C = O\left(\frac{M_{l+1}}{n^{l-1}}\right),$$

где  $M_{l+1}$  — оценка  $(l + 1)$ -й производной  $f$ , т. е. скорость убывания погрешности автоматически учитывает гладкость функции  $f(x)$ ;

константа Лебега  $\eta = O(\ln n)$ , т. е. возрастает с ростом  $n$  существенно медленнее, чем в обычной полиномиальной интерполяции на равномерной сетке.

Сопоставление условий, из которых определяются коэффициенты тригонометрического полинома, и его функционального вида позволяет записать полином в аналитической форме (аналогично полиному Лагранжа, представленному выше):

$$Q_n(x) = \sum_{k=0}^{2n} f(x_k) q_k^n(x),$$

где

$$q_k^n(x) = \prod_{\substack{i=0 \\ i \neq k}}^{2n} \frac{\sin \frac{1}{2}\omega(x - x_i)}{\sin \frac{1}{2}\omega(x_k - x_i)} =$$

*базисный тригонометрический полином.*

Можно показать, что система функций  $1, \cos k\omega x, \sin k\omega x$ ,  $k = 1, 2 \dots, n$ , является базисом в пространстве сеточных функций, заданных на используемой интерполяционной сетке. При этом она ортогональна относительно дискретного аналога обычного скалярного произведения в  $L_2$ . Это позволяет выписать аналитические выражения для коэффициентов  $a_0, a_k, b_k$ ,  $k = 1, 2, \dots, n$ , в виде, являющимся дискретным аналогом обычных выражений для коэффициентов Фурье.

Далее при решении задачи Штурма — Лиувилля для разностного оператора второго порядка мы увидим, что собственные функции такого оператора с точностью до незначительных деталей совпадают с рассматриваемой системой. Следовательно, она является решением такой задачи и обладает свойствами, присущими собственным функциям.

**Замечание 3.6.** Анализ выражений для коэффициентов разложения  $a_0, a_k, b_k$ ,  $k = 1, 2 \dots, n$ , позволяет тривиально получить оценку

константы Лебега вида  $\eta = O(n)$ . Однако более детальный анализ дает возможность установить на равномерной сетке логарифмическую оценку константы Лебега. Это, в частности, свидетельствует об оптимальности равномерной сетки для тригонометрической интерполяции. Расчетные формулы при этом получаются также самыми простыми.

**Замечание 3.7.** Совпадение оценок констант Лебега тригонометрической интерполяции и полиномиальной на чебышевской сетке является неслучайным. Легко видеть, что замена  $x$  на  $\varphi$  по правилу

$$x = \frac{a+b}{2} + \frac{b-a}{2} \cos \varphi$$

приводит к преобразованию равномерной сетки переменной  $\varphi$  в чебышевскую сетку переменной  $x$ . При этом тригонометрический полином переменной  $\varphi$  становится обычным полиномом переменной  $x$ .

### 3.4. Сплайн-интерполяция

От многих недостатков глобальной полиномиальной интерполяции свободна интерполяция кусочно-многочленная, или кусочно-полиномиальная. Пусть имеются точки

$$a = x_0 < x_1 \cdots < x_n = b.$$

Найдем функцию  $S_3(x)$ , которая представляет собой многочлен третьей степени на любом отрезке  $[x_{i-1}, x_i]$  длиной  $h_i = x_i - x_{i-1}$ :

$$S_3(x) = a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3.$$

Потребуем, чтобы на концах отрезка  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ , функция  $S_3$  принимала заданные значения:

в точке  $x = x_{i-1}$

$$S_3(x_{i-1}) = y_{i-1} = a_i;$$

в точке  $x = x_i$

$$S_3(x_i) = y_i = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3.$$

Видно, что число неизвестных параметров в два раза превышает число уравнений.

Для увеличения числа уравнений потребуем еще непрерывности первой и второй производных во внутренних точках сетки:

$$S'_3 = b_i + 2c_i(x - x_{i-1}) + 3d_i(x - x_{i-1})^2,$$

$$S''_3 = 2c_i + 6d_i(x - x_{i-1}).$$

Непрерывность производных означает, что

$$\begin{cases} b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1}, & i = 1, 2, \dots, n-1. \\ 2c_i + 6d_i h_i = 2c_{i+1}, \end{cases}$$

В результате имеем  $2n + 2(n-1) = 4n - 2$  уравнений для  $4n$  неизвестных. Еще два уравнения получим, полагая  $S_3''$  равной нулю в точках  $x = x_0 = a$  и  $x = x_n = b$ :

$$2c_1 = 0, \quad 2c_n + 6d_n h_n = 0 \quad (\text{или } c_{n+1} = 0).$$

Для вывода расчетных соотношений приведем рассматриваемую систему уравнений к удобному виду, исключив  $a_i$ ,  $b_i$ ,  $d_i$ :

$$\begin{aligned} d_n &= -\frac{c_n}{3h_n}, \quad d_i = \frac{c_{i+1} - c_i}{3h_i}, \\ b_i &= \frac{y_i - y_{i-1}}{h_i} - c_i h_i - \frac{(c_{i+1} - c_i)h_i}{3}, \\ b_n &= \frac{y_n - y_{n-1}}{h_n} - c_n h_n + \frac{c_n h_n}{3}, \end{aligned}$$

тогда равенство первых производных дает

$$\begin{aligned} \frac{1}{h_i}(y_i - y_{i-1}) - c_i h_i - \frac{1}{3}(c_{i+1} - c_i)h_i + 2c_i h_i + (c_{i+1} - c_i)h_i &= \\ = \frac{1}{h_{i+1}}(y_{i+1} - y_i) - c_{i+1} h_{i+1} - \frac{1}{3}(c_{i+2} - c_{i+1})h_{i+1}. \end{aligned}$$

Заменим индекс  $i$  на  $i - 1$  для получения уравнений привычного вида:

$$\begin{aligned} c_{i-1} h_{i-1} + (2h_{i-1} + 2h_i)c_i + c_{i+1} h_i &= \\ = 3\left(\frac{1}{h_i}(y_i - y_{i-1}) - \frac{1}{h_{i-1}}(y_{i-1} - y_{i-2})\right). \end{aligned}$$

При этом  $c_1 = c_{n+1} = 0$ . Равенство  $c_{n+1} = 0$  легко получить из сравнения условий для  $S_3''$  в точке  $x_n = b$  и во внутренних точках.

Таким образом, получена трехдиагональная система линейных алгебраических уравнений (СЛАУ) со строгим диагональным преобладанием: разность коэффициентов перед  $c_i$  и  $c_{i+1}$ ,  $c_{i-1}$  равна  $(h_{i-1} + h_i) > 0$ . Следовательно, задача определения  $c_i$  корректно поставлена, решение легко находится методом прогонки. Далее вычисляются остальные коэффициенты.

Построенная функция называется *интерполяционным кубическим сплайном* (от английского слова spline — планка, рейка). Название происходит от чертежного приема. Он состоит в проведении кривой по гибкой металлической линейке, которая проходит через заданные точки. Приложенная линейка приобретает форму, соответствующую минимуму упругой энергии:

$$\int_a^b (u'')^2 dx \rightarrow \min.$$

Отсюда получим уравнение Эйлера  $y^{(4)} = 0$ . Его решением  $y$  является многочлен третьей степени на каждом из интервалов сетки. В узлах сетки должны быть непрерывны решение  $y$  и его две первые производные  $y'$  и  $y''$ .

Описанное обстоятельство служит причиной того, что данный сплайн носит название естественного или чертежного.

Помимо простоты, сплайн-интерполяция замечательна еще и своей сходимостью.

**Теорема 3.7.** Пусть  $y = f(x) \in C^4(a, b)$ ,  $M_4 = \|f^{(4)}\|_C$ ,  $S_3(x)$  — сплайн третьей степени. Тогда

$$\|f - S_3\|_C \leq C_1 M_4 h^4, \quad \|f' - S'_3\|_C \leq C_2 M_4 h^3, \quad \|f'' - S''_3\|_C \leq C_3 M_4 h^2.$$

Доказательство теоремы опустим.

Отсюда следует, что для указанного класса функций не только сплайн  $S_3$  сходится к функции  $f$ , но и первая и вторая производные. Сплайн  $S_3$  можно дважды дифференцировать.

**Замечание 3.8.** Заметим, что рассмотренный выше в начале данной главы случай кусочно-линейной интерполяции является не только простейшим примером интерполянта вообще, но и сплайна в частности.

**Определение 3.14.** *Сплайном степени  $m$*  называется функция  $S_m(x)$ , заданная на отрезке  $[a, b]$  с указанными точками разбиения, которая непрерывна на отрезке вместе со своими производными вплоть до некоторого порядка  $p$  и на каждом отрезке разбиения  $[x_{i-1}, x_i]$  совпадает с некоторым алгебраическим полиномом  $P_{m,i}(x)$  степени  $m$ . Разность  $m - p$  между степенью сплайна и порядком наивысшей непрерывной производной называется *дефектом сплайна*.

Таким образом, кусочно-линейный интерполянт является линейным (первой степени) сплайном с дефектом, равным единице. Рассмотренный выше кубический сплайн имеет тот же дефект.

Достаточно часто используются и кубические сплайны с дефектом, равным двум.

Очевидно, что для задания такого сплайна, у которого, вообще говоря, вторая производная в точках сетки не является непрерывной, необходимо указать дополнительные условия. Такими условиями могут быть значения производной сплайна в точках разбиения  $s_i = S'_m(x_i)$ , называемые наклоном сплайна в точке  $x_i$ .

Нетрудно получить явное выражение для такого интерполянта (на отрезке разбиения  $[x_{i-1}, x_i]$ ) без решения трехдиагональной СЛАУ:

$$\begin{aligned} S_3(x) = & \frac{(x - x_i)^2(2(x - x_{i-1}) + h_i)}{h_i^3} y_{i-1} + \\ & + \frac{(x - x_{i-1})^2(2(x_i - x) + h_i)}{h_i^3} y_i + \frac{(x - x_i)^2(x - x_{i-1})}{h_i^2} s_{i-1} + \\ & + \frac{(x - x_{i-1})^2(x - x_i)}{h_i^2} s_i. \end{aligned}$$

Существуют различные способы задания наклонов сплайна в точках сетки. Самым простым является случай, в котором известны производные исходной функции в точках  $x_i$ ,  $i = 0, 1, \dots, n$ . Тогда  $s_i = y'_i$ . Такой сплайн называется локальным, так как на каждом отрезке разбиения  $[x_{i-1}, x_i]$  он полностью определяется значениями функции и ее производной на границах отрезка. Очевидно, что этот сплайн является также и интерполяционным полиномом Эрмита на каждом таком отрезке.

Построенный выше кубический сплайн с дефектом, равным единице, естественно считать глобальным сплайном, так как его коэффициенты определяются данными на всей сетке сразу.

**Замечание 3.9.** Довольно часто в рассмотрение вводятся так называемые *B-сплайны*. Они представляют собой кусочно-полиномиальные функции, задаваемые полиномом в областях его неотрицательности и нулем в остальной части числовой оси (в одномерном случае). Так, *B-сплайн* нулевой степени представляет собой характеристическую функцию одного полуинтервала между точками сетки, линейный *B-сплайн* ранее рассматривался при описании базисных функций линейного конечного элемента. Такие сплайны широко используются при решении задач математической физики методом конечных элементов.

### 3.5. Двумерная интерполяция

Рассмотрим двумерную сетку

$$\Omega_h = \{(x_i, y_j) : a \leq x_i \leq b, c \leq y_j \leq d\},$$

т. е. набор точек на плоскости. Пусть в каждой точке  $\Omega_h$  задано значение функции  $z_{ij}$ . Необходимо проинтерполировать эти значения, т. е. построить двумерную функцию  $\tilde{f}(x, y)$  вместо  $z = f(x, y)$ . При этом интерполирующая функция  $\tilde{f}$  должна «приближать» функцию  $f$ .

Рассмотрим различные варианты.

#### 3.5.1. Прямоугольная сетка

В случае прямоугольной сетки

$$\Omega_h = \Omega_h^x \times \Omega_h^y,$$

где

$$\Omega_h^x = \{x_i : a \leq x_i \leq b\}; \quad \Omega_h^y = \{y_j : c \leq y_j \leq d\}.$$

Наиболее простой вариант построения интерполяционного многочлена заключается в отдельной интерполяции по  $x$  и  $y$  и в выборе результирующего интерполянта в виде произведения одномерных. Такая **интерполяция** также называется **последовательной**. Считаем, что  $b = x_n$ ,  $d = y_m$ .

Пусть для произвольных  $y = y_j$  справедливо равенство

$$\tilde{f}(x, y_j) = \sum_{i=0}^n z_{ij} \varphi_i^x(x),$$

а для произвольных  $x = x_i$  —

$$\tilde{f}(x_i, y) = \sum_{j=0}^m z_{ij} \varphi_j^y(y).$$

Тогда

$$\tilde{f}(x, y) = \sum_{i=0}^n \sum_{j=0}^m z_{ij} \varphi_i^x(x) \varphi_j^y(y).$$

Нами получена последовательная интерполяция.

В качестве  $\varphi_i^x$ ,  $\varphi_j^y$  могут быть взяты функции типа *базисных функций конечных элементов, полинома Лагранжа* вида

$$\frac{\omega(x)}{(x - x_i)\omega'(x_i)}$$

(для  $x$ ), сплайна. Базисные сплайны соответствуют единичному значению функции в  $i$ -й точке и нулю в остальных.

Однако при последовательной интерполяции происходит завышение степени интерполяционного полинома. Так, если  $\varphi_i^x$ ,  $\varphi_j^y$  — линейные одномерные базисные функции конечных элементов, то на сеточных прямоугольниках интерполянт имеет вторую степень:

$$\varphi_i^x \varphi_j^y = a + bx + cy + dxy.$$

Но при этом никакого повышения порядка точности интерполяции по сравнению с точностью линейной одномерной (по  $x$  или  $y$ ) интерполяции не происходит.

Рассмотрим многочлен  $N$ -й степени двух переменных:

$$z = f(x, y) = \sum_{i+j=0, 0 \leq i, j \leq N}^N a_{ij} x^i y^j.$$

В этом выражении использовано следующее число коэффициентов:

$$\frac{1}{2}((N+1)^2 - (N+1)) + (N+1) = \frac{1}{2}(N+1)^2 + \frac{1}{2}(N+1) = \frac{1}{2}(N+1)(N+2).$$

Соответственно, для их определения необходимо задать столько же уравнений. Если у нас есть  $n+1$  точек по  $x$  и  $m+1$  точек по  $y$ , то должно быть выполнено равенство

$$(n+1)(m+1) = \frac{1}{2}(N+1)(N+2).$$

Удовлетворить этому условию довольно сложно. Например, если  $n = m = 1$ , то

$$(1+1)(1+1) = 4 = \frac{1}{2}(N+1)(N+2),$$

откуда  $N^2 + 3N + 2 = 8$ , и

$$N = -\frac{3}{2} \pm \sqrt{\frac{9}{4} + 6},$$

или

$$N = \frac{\sqrt{33} - 3}{2}, \quad —$$

нецелое число.

При  $N = 1$  имеем

$$\frac{1}{2}(N+1)(N+2) = 3,$$

при  $N = 2$  —

$$\frac{1}{2}(N+1)(N+2) = 6.$$

В результате для полинома первой степени с  $N = 1$  четырех точек много, а для для полинома второй степени с  $N = 2$  мало. Если выбрать  $N = 2$ , то необходимо добавить еще два уравнения. Их выбор неоднозначен, что порождает излишнюю неопределенность.

### 3.5.2. Треугольная сетка

Для построения полинома первой степени с  $N = 1$  требуется лишь три точки сетки (рис. 3.3). Таким образом, полином минимальной степени получается на треугольной сетке. Легко записать функцию  $\tilde{f}(x, y) = a + bx + cy$ , которая принимает в вершинах треугольника заданные значения:

$$\begin{aligned} a + bx_1 + cy_1 &= f_1, \\ a + bx_2 + cy_2 &= f_2, \\ a + bx_3 + cy_3 &= f_3. \end{aligned}$$

Эта система линейных алгебраических уравнений (СЛАУ) относительно коэффициентов  $a, b, c$  имеет однозначное решение, если точки 1, 2, 3 не лежат на одной прямой.

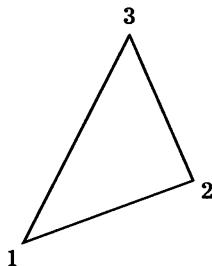


Рис. 3.3

Если рассматривать интерполяцию на всей сетке, то получим

$$z = \tilde{f}(x, y) = \sum_{k=1}^K z_k \varphi_k^{xy},$$

где  $K$  — число узлов сетки;  $\varphi_k^{xy}$  — кусочно-линейная двумерная конечно-элементная базисная функция. График линейной базисной функции двумерного конечного элемента представляет собой пирамиду (рис. 3.4).

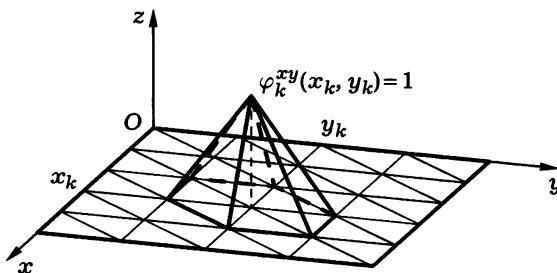


Рис. 3.4

Функция  $\varphi_k^{xy}$  принимает значения 1 в точке с номером  $k$  и 0 — во всех остальных вершинах треугольников, имеющих точку  $k$  своей вершиной. На каждом треугольнике функция  $\varphi_k^{xy}$  представляет собой плоскость. Носитель функции  $\varphi_k^{xy}$  — множество, на котором данная функция отлична от нуля — называется **двумерным конечным элементом**.

Для построения полинома второго порядка ( $N = 2$ ) требуется шесть точек. Чаще всего это три угловые точки (см. рис. 3.3) и три точки в центрах сторон треугольника. Для построения полинома третьего порядка требуется десять точек. Обычно это три угловых точки, по две точки на сторонах треугольника и одна точка в его центре.

Существуют определенные технологии работы с подобными конечными элементами и очень развитая теория.

Отметим очевидное отличие двумерной экстраполяции от одномерной (рис. 3.5): вычисление  $\tilde{f}$  в точке  $B$  есть экстраполяция, а в точке  $A$  — интерполяция. Точки  $B$  и  $A$  находятся вне и внутри выпуклого тела соответственно.

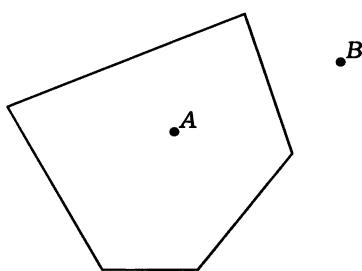


Рис. 3.5

### 3.6. Библиографические комментарии

Теория аппроксимации и ее частного случая — интерполяции — относится к наиболее разработанным областям математики, используемым в приложениях. Эта теория наиболее близка таким областям математики, как, например, функциональный анализ, теория функций, численный анализ. В качестве примера можно привести работы [8, 105], в которых содержится большой материал по аппроксимации классов функций конечномерными компактами. Там же имеется большое количество ссылок на соответствующую литературу.

Одной из самых интересных задач теории аппроксимации является задача построения равномерных приближений. Многие результаты решения этой задачи можно найти в [17] и [18]. Вопросы практического применения отражены в [80].

Теории сплайнов и их применением в вычислительной математике посвящены монографии [4, 164]. Сплайн-интерполяция, теория и вопросы практического применения (включая  $B$ -сплайны) достаточно подробно описаны в [25], а также в [80, 92, 137] и других руководствах.

Отметим, что приближение функций с помощью конечных сумм является широко распространенным приемом получения чисто математических результатов типа существования и единственности решения краевых задач, дифференциальных неравенств и т. п. [23, 93, 102–104, 133, 166].

Материал по специальным функциям математической физики можно найти во многих руководствах, например в [27, 100, 170]. Очень много информации по ним содержит справочник [2]. Особо отметим монографию [120].

Алгоритмы без насыщения представляют собой один из самых важных и красивых классов численных алгоритмов. Их построение достаточно сложно. Значительная информация о таких алгоритмах содержится в [8] и [105].

## 4. МЕТОДЫ ЧИСЛЕННОГО ИНТЕГРИРОВАНИЯ И ДИФФЕРЕНЦИРОВАНИЯ

Представлены *квадратурные формулы* для численного нахождения одномерных и многомерных интегралов. Рассмотрены квадратурные формулы *интерполяционного типа* (включая *формулы прямоугольников, трапеций, Симпсона* и др.) и *квадратурные формулы Гаусса*. Отдельно описаны способы вычисления несобственных интегралов I и II рода, интегралов от быстроосциллирующих функций. Приведены способы численного дифференцирования функций.

### 4.1. Простейшие квадратурные формулы

Согласно определению, например, интеграла Римана,

$$I = \int_a^b f(x) dx = \lim_{\lambda_R \rightarrow 0} \sum_{i=1}^N f(\xi_i) h_i, \quad \lambda_R = \max_{1 \leq i \leq N} h_i, \quad h_i = x_i - x_{i-1} > 0.$$

При этом набор точек  $x_0, x_1, \dots, x_N$ , лежащих на  $[a, b]$ , называется разбиением  $R$  этого отрезка,  $\lambda_R$  — диаметром разбиения, точки  $\xi_i \in [x_{i-1}, x_i]$  выбираются произвольно, как и точки разбиения, но  $x_0 = a$ ,  $x_N = b$ .

Отсюда возникает идея приближенного вычисления интеграла  $I$  путем его замены на

$$I_h = \sum_{k=1}^K c_k f(\tilde{x}_k).$$

**Определение 4.1.** Приближенное равенство  $I \approx I_h$  называется *квадратурной формулой*,  $I_h$  — *квадратурной суммой*, точки  $\tilde{x}_k$  — *узлами квадратурной суммы*,  $c_k$  — ее коэффициентами, разность  $\psi_h = I - I_h$  — погрешностью квадратурной формулы.

Довольно часто выражение для  $I_h$  также называется *квадратурной формулой*.

При этом мы будем считать, что функция  $f$  известна в узлах сетки

$$\tilde{\Omega}_h = \{\tilde{x}_i : a \leq \tilde{x}_1 < \tilde{x}_2 < \dots < \tilde{x}_n \leq b\}.$$

Иногда узлы сетки будут задаваться специальным образом, иногда — произвольным.

Ограничимся пока равномерной сеткой

$$h_i = h = \frac{b-a}{n}.$$

При этом сетка

$$\Omega_h = \{x_i: x_i = a + ih, i = 0, 1, \dots, n\}.$$

Для построения квадратурной формулы часто достаточно рассмотреть частичный отрезок  $[x_{i-1}, x_i]$ , так как

$$I = \int_a^b f(x) dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx.$$

#### 4.1.1. Формула прямоугольников

В случае квадратурной *формулы центральных прямоугольников*

$$I_{h,i} = f(x_{i-1/2})h,$$

где  $x_{i-1/2} = x_{i-1} + h/2 = 1/2(x_i + x_{i-1})$ . При этом истинная площадь криволинейной трапеции отличается от рассчитываемой площади прямоугольника (рис. 4.1).

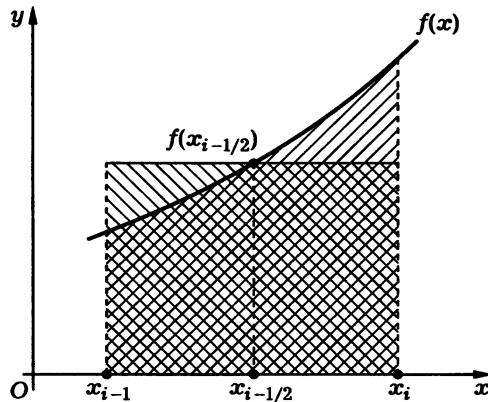


Рис. 4.1

Погрешность квадратурной формулы составляет

$$\psi_{h,i} = \int_{x_{i-1}}^{x_i} f(x) dx - I_{h,i}.$$

Если функция  $f(x)$  дважды непрерывно дифференцируема на отрезке  $[a, b]$ , так что  $\|f''\|_C \leq M_2$ , то

$$\begin{aligned} |\psi_{h,i}| &= \left| \int_{x_{i-1}}^{x_i} (f(x) - f(x_{i-1/2})) dx \right| = \\ &= \left| \int_{x_{i-1}}^{x_i} \left( f(x_{i-1/2}) + f'(x_{i-1/2})(x - x_{i-1/2}) + \right. \right. \\ &\quad \left. \left. + \frac{1}{2} f''(x_{i-1/2}^*)(x - x_{i-1/2})^2 - f(x_{i-1/2}) \right) dx \right| \leqslant \\ &\leqslant \frac{M_2}{2} \int_{x_{i-1}}^{x_i} (x - x_{i-1/2})^2 dx = M_2 \frac{h^3}{24}. \end{aligned}$$

Оценка является неулучшаемой: функция  $f = (x - x_{i-1/2})^2$  реализует равенство в данной оценке ( $M_2 = 2$ ).

Оценка погрешности квадратурной формулы для всего отрезка

$$I_h = \sum_{i=1}^n f(x_{i-1/2})h$$

имеет вид

$$|\psi_h| \leq M_2 \frac{h^3}{24} n = \frac{M_2}{24} \frac{(b-a)^3}{n^2} = O(h^2).$$

**Определение 4.2.** Если  $\psi_h = O(h^l)$ , то говорят, что **квадратурная формула** имеет  **$l$ -й порядок точности**.

Для формулы центральных прямоугольников  $l = 2$ .

Если взять **формулы левых** ( $I_{h,i} = f(x_{i-1})h$ ) или **правых** ( $I_{h,i} = f(x_i)h$ ) **прямоугольников**, то в результате вычисления  $f$  в нецентralьной точке в обоих случаях получим  $\psi_h = O(h)$ .

#### 4.1.2. Формула трапеций

Выберем линейный интерполиант на отрезке  $[x_{i-1}, x_i]$ :

$$\tilde{f}(x) = \frac{1}{h} \left( (x - x_{i-1})f(x_i) + (x_i - x)f(x_{i-1}) \right).$$

Из 3 известно, что

$$|f - \tilde{f}| \leq \frac{1}{2} M_2 (x - x_{i-1})(x_i - x)$$

для дважды непрерывно дифференцируемых функций. Тогда интеграл от  $\tilde{f}$  дает **квадратурную формулу трапеций**

$$I_{h,i} = \frac{1}{2}(f(x_{i-1}) + f(x_i))h,$$

в результате имеем оценку

$$\begin{aligned} |\psi_{h,i}| &\leq \frac{M_2}{2} \int_{x_{i-1/2}}^{x_i} (x - x_{i-1})(x_i - x) dx = \\ &= \frac{M_2}{2} \left( \frac{1}{2}(x - x_{i-1})^2(x_i - x) + \frac{1}{6}(x - x_{i-1})^3 \right) \Big|_{x_{i-1}}^{x_i} = \frac{1}{12} M_2 h^3. \end{aligned}$$

Отсюда получаем

$$|\psi_h| \leq \frac{M_2}{2} h^2(b-a) = O(h^2) \quad —$$

погрешность **формул** так называемого **интерполяционного типа**, в которой используется интерполяция.

#### 4.1.3. Формула Симпсона

Пусть  $n = 2m$  — четное число. Рассмотрим отрезок  $[x_{2i}, x_{2i+2}]$  длиной  $2h$ . На нем имеются три точки:  $x_{2i}, x_{2i+1}, x_{2i+2}$ . Требуется вычислить интеграл  $I_i = \int_{x_{2i}}^{x_{2i+2}} f(x) dx$ .

Пусть по формуле трапеций величина

$$I_h^{(1)} = \frac{1}{2}(f(x_{2i}) + f(x_{2i+2}))2h = h(f(x_{2i}) + f(x_{2i+2}))$$

построена по двум точкам  $x_{2i}, x_{2i+2}$ . При этом из оценки погрешности квадратурной формулы трапеции (для  $f \in C^{(4)}[a, b]$ ) имеем

$$I_h^{(1)} = \int_{x_{2i}}^{x_{2i+2}} f(x) dx + C(2h)^3 + O(h^5) = I_i + C(2h)^3 + O(h^5).$$

Теперь учтем наличие трех точек. Применение формулы трапеций дает:

$$\begin{aligned} I_h^{(2)} &= \frac{1}{2}(f(x_{2i}) + f(x_{2i+1}))h + \frac{1}{2}(f(x_{2i+1}) + f(x_{2i+2}))h = \\ &= I_i + 2Ch^3 + O(h^5). \end{aligned}$$

Из оценок погрешностей двух формул получаем

$$I_i = \frac{1}{3} (4I_h^{(2)} - I_h^{(1)}) + O(h^5) = \frac{h}{3} (f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})) + O(h^5).$$

Описанная процедура позволяет повысить локальный порядок точности до пяти.

В проведенных выкладках использовалась формула Тейлора для значений функции  $f(x)$  в рассматриваемых точках с центром в точке  $x_{2i+1}$ :

$$\begin{aligned} f(x) &= f_{2i+1} + f'_{2i+1}(x - x_{2i+1}) + \frac{1}{2} f''_{2i+1}(x - x_{2i+1})^2 + \\ &\quad + \frac{1}{3!} f'''_{2i+1}(x - x_{2i+1})^3 + \frac{1}{4!} \tilde{f}_{2i+1}^{(4)}(x - x_{2i+1})^4. \end{aligned}$$

Здесь  $f_{2i+1}^{(k)} = \frac{d^k f}{dx^k} \Big|_{x=x_{2i+1}}$ . В последнем слагаемом производная вычисляется в некоторой вспомогательной точке.

После интегрирования по участку  $[x_{2i}, x_{2i+2}]$  слагаемые с нечетными степенями (нечетные функции относительно середины отрезка интегрирования) дадут нулевой вклад в интеграл. Поэтому в результате останутся слагаемые  $O(h^3)$  и  $O(h^5)$ . При этом требуется, чтобы функция имела четвертую непрерывную производную.

Формула

$$I_h = \frac{h}{3} (f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2}))$$

называется **квадратурной формулой Симпсона**. Процесс ее вывода из формулы трапеций на двух сетках ( $2h$  и  $h$  — их шаги) называется **правилом Рунге**, иногда **методом экстраполяции Ричардсона**.

Окончательно погрешность формулы Симпсона по всему отрезку  $[a, b]$  равна  $\psi_h = O(h^4)$  для  $f \in C^{(4)}[a, b]$ .

## 4.2. Квадратурные формулы интерполяционного типа

Рассмотрим квадратурные формулы для вычисления интегралов вида

$$I = \int_a^b \rho(x)f(x)dx,$$

где  $\rho = \rho(x) > 0$  — весовая функция;  $\rho, f$  считаются достаточно гладкими. Как и ранее, квадратурные формулы имеют вид

$$I \approx \sum_{k=0}^n c_k f(x_k), \quad x_k \in [a, b],$$

где  $c_k, k = 0, 1, \dots, n$ , — коэффициенты.

В отличие от вывода формул в 4.1 не будем по отдельности рассматривать приближенное значение интеграла по различным сеточным интервалам, а рассмотрим интеграл сразу на всем отрезке  $[a, b]$ . Для этого проинтерполируем функцию  $f(x)$  по  $n + 1$  точкам, заменив ее на  $[a, b]$  *интерполяционным полиномом Лагранжа* вида

$$L_n(x) = \sum_{k=0}^n f(x_k) \frac{\omega(x)}{(x - x_k)\omega'(x_k)},$$

где

$$\omega(x) = \prod_{i=0}^n (x - x_i); \quad \omega'(x_k) = \prod_{\substack{i=0 \\ i \neq k}} (x_k - x_i).$$

**Определение 4.3.** Полученные с помощью процедуры интерполяции функции по всей сетке квадратурные формулы называются *квадратурными формулами интерполяционного типа*.

Имеем

$$\int_a^b \rho(x)f(x)dx \approx \sum_{k=0}^n c_k f(x_k), \quad c_k = \int_a^b \rho(x) \frac{\omega(x)}{(x - x_k)\omega'(x_k)} dx.$$

Так как интерполяционный многочлен определен единственным образом, то и коэффициенты квадратурной формулы интерполяционного типа однозначно выражаются через  $x_k, \rho, h, a, b$ .

**Пример 4.1.** Пусть  $[a, b] = [-1, 1]$ ,  $x_0 = -1$ ,  $x_1 = 0$ ,  $x_2 = 1$ . Интерполяция по трем точкам дает значения

$$c_0 = \int_{-1}^1 \frac{x(x-1)}{(-1)(-2)} dx = \frac{1}{2} \int_{-1}^1 x(x-1) dx = \frac{1}{2} \left( \frac{1}{3}x^3 - \frac{1}{2}x^2 \right) \Big|_{-1}^1 = \frac{1}{3},$$

$$c_1 = \int_{-1}^1 \frac{(x+1)(x-1)}{(+1)(-1)} dx = - \left( \frac{1}{3}x^3 - x \right) \Big|_{-1}^1 = \frac{4}{3},$$

$$c_2 = \int_{-1}^1 \frac{(x+1)x}{(2)(1)} dx = \frac{1}{3}.$$

Отсюда

$$\int_{-1}^1 f(x) dx \approx \frac{1}{3}(f(-1) + 4f(0) + f(1)),$$

т. е. мы получаем формулу Симпсона при  $h = 1$ .

**Пример 4.2.** В условиях предыдущего примера будем искать такую линейную функцию  $\varphi(x) = ax + b$ , что

$$\sum_{i=0}^2 (f(x_i) - \varphi(x_i))^2 \rightarrow \min.$$

Опуская выкладки, ранее уже проводившиеся при описании *метода наименьших квадратов*, получаем линейный полином наилучшего среднеквадратичного приближения

$$\varphi(x) = \frac{1}{3}(f(x_0) + f(x_1) + f(x_2)) + \frac{1}{2}(f(x_2) - f(x_0))x.$$

Отсюда следует приближенная неинтерполяционная квадратурная формула

$$I = \int_{-1}^{+1} f(x) dx \approx \frac{2}{3}(f(x_0) + f(x_1) + f(x_2)).$$

**Теорема 4.1.** Квадратурная формула интерполяционного типа, построенная по  $n + 1$  узлам  $x_0, x_1, \dots, x_n$ , является точной для любого полинома степени  $n$ . Для  $n + 1$  раз непрерывно дифференцируемой на  $[a, b]$  функции погрешность квадратурной формулы составляет

$$|\psi_h| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b \rho(x) |\omega(x)| dx, \quad M_{n+1} = \|f^{(n+1)}\|_C.$$

◀ Как известно, в условиях теоремы

$$f(x) = L_n(x) + r_n(x), \quad r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega(x).$$

Отсюда

$$|\psi_h| = \left| \int_a^b \rho(x) (f(x) - L_n(x)) dx \right| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b \rho(x) |\omega(x)| dx.$$

Если  $f(x)$  — полином степени меньше или равной  $n$ , то  $M_{n+1} = 0$  и потому  $|\psi_h| = 0$ , т. е. формула точна. ►

**Теорема 4.2 (обратная).** Если квадратурная формула

$$I_k = \sum_{k=0}^n d_k f(x_k)$$

точна для любого полинома степени  $n$ , то она является квадратурной формулой интерполяционного типа.

◀ Нужно доказать, что для произвольного  $k$  коэффициент  $d_k = c_k$ , где  $c_k$  — определенные выше коэффициенты квадратурной формулы интерполяционного типа. Рассмотрим многочлен степени  $n$

$$f(x) = \varphi_k(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)},$$

удовлетворяющий равенствам  $\varphi_k(x_i) = \delta_{ki}$ . Так как квадратурная формула точна для таких функций  $f(x)$ , то

$$\int_a^b \rho(x) f(x) dx = \int_a^b \rho(x) \varphi_k(x) dx = c_k = \sum_{i=0}^n \varphi_k(x_i) d_i = d_k.$$

Следовательно,  $d_k = c_k$  для всех  $k$ . ►

**Замечание 4.1.** Формулы Симпсона, трапеций, прямоугольников — квадратурные формулы интерполяционного типа. При этом значения функции интерполируются полиномами второго, первого и нулевого порядков соответственно.

**Замечание 4.2.** Формулы центральных прямоугольников имеют повышенный порядок точности по сравнению с формулами левых или правых прямоугольников за счет симметричного относительно центра сеточного интервала выбора узла квадратурной формулы. В общем случае условия симметричного выбора также позволяют повысить порядок точности.

**Замечание 4.3.** Оценка погрешности  $\psi_h$  не является неулучшаемой: для формулы Симпсона в действительности  $\psi_h$  определяется четвертой производной, а не третьей  $M_{n+1} = M_{2+1}$ . Это происходит из-за учета симметрии при определении погрешности в выводе оценки.

**Определение 4.4.** Квадратурные формулы интерполяционного типа на равномерной сетке  $a = x_0 < x_1 < \dots < x_n = b$ , где  $x_k - x_{k-1} = h$ ,  $k = 1, 2, \dots, n$ , называются **формулами Ньютона — Котеса**.

В этом случае вид выражений для  $c_k$  может быть упрощен. Мы этого делать не будем.

**Лемма 4.1.** Квадратурная формула интерполяционного типа устойчива относительно возмущений функции  $f$  при условии знакопостоянства коэффициентов квадратурной формулы.

◀ Пусть

$$I_h = \sum_{k=0}^n c_k f(x_k).$$

Так как квадратурная формула точна при  $f \equiv 1$ , то

$$\sum_{k=0}^n c_k = \int_a^b \rho(x) dx = M.$$

Следовательно, в случае знакопостоянных  $c_k$

$$\sum_{k=0}^n |c_k| = |M|$$

не зависит от  $n$ . Пусть

$$I_h + \delta I_h = \sum_{k=0}^n c_k (f(x_k) + \delta f(x_k)) = I_h + \sum_{k=0}^n c_k \delta f(x_k).$$

Тогда

$$|\delta I_h| \leq \sum_{k=0}^n |c_k| |\delta f(x_k)| \leq \|\delta f\|_C |M|.$$

Следовательно, квадратурная формула интерполяционного типа устойчива относительно возмущений функции  $f$ . ►

Ранее при рассмотрении интерполяции уже отмечалась неустойчивость глобальной интерполяции полиномами при больших  $n$ . Аналогично и при ее использовании в квадратурах с  $n \geq 5$  должны проявиться эффекты неустойчивости. Например, известно, что при  $n \geq 10$

и  $\rho = 1$  появляются как положительные, так и отрицательные  $c_k$ . В результате  $\sum_{k=0}^n |c_k|$  может превысить  $|M|$  и квадратура будет неустойчивой.

Чтобы этого не случилось, следует использовать, например, кусочно-полиномиальную аппроксимацию функций.

### 4.3. Квадратурные формулы Гаусса

Ранее в 4.1, 4.2 узлы  $x_0, x_1, \dots, x_n$  квадратурной формулы (всего  $n + 1$  узлов) считались заданными. При этом была построена квадратурная формула, точная для полиномов степени, меньшей либо равной  $n$ , при выбранных коэффициентах  $c_k$ ,  $k = 0, 1, \dots, n$ .

Построим квадратурную формулу на  $n + 1$  узлах так, чтобы она была точной для полиномов максимальной степени  $m$ . При этом будем выбирать коэффициенты  $c_k$  и узлы  $x_k$ .

Следовательно, условие точности имеет вид

$$\sum_{k=0}^n c_k x_k^\alpha = \int_a^b \rho x^\alpha dx, \quad \alpha = 0, 1, \dots, m.$$

У нас есть  $2(n + 1)$  неизвестных, следовательно, для их определения нужно  $2n + 2$  уравнений. Таким образом, максимальное значение степени  $\alpha = m = 2n + 1$ . Для нахождения  $c_k$  и  $x_k$  следует решить полученную систему.

**Пример 4.3.** Пусть  $\rho = 1$ ,  $[a, b] = [-1, 1]$ .

1. При  $n = 0$  степень  $m = 1$ . Тогда система уравнений примет вид

$$c_0 x_0^0 = \int_{-1}^{+1} x^0 dx = 2,$$

$$c_0 x_0^1 = \int_{-1}^{+1} x^1 dx = 0.$$

Следовательно,  $c_0 = 2$ ,  $x_0 = 0$  и квадратурная формула

$$\int_{-1}^{+1} f dx \approx 2f(0)$$

точна для многочленов первой степени (формула центральных прямоугольников).

2. При  $n = 1$  степень  $m = 3$ . Тогда система уравнений примет вид

$$c_0x_0^0 + c_1x_1^0 = \int_{-1}^1 x^0 dx = 2,$$

$$c_0x_0^1 + c_1x_1^1 = \int_{-1}^1 x^1 dx = 0,$$

$$c_0x_0^2 + c_1x_1^2 = \int_{-1}^1 x^2 dx = \frac{2}{3},$$

$$c_0x_0^3 + c_1x_1^3 = \int_{-1}^1 x^3 dx = 0,$$

откуда

$$c_0 = c_1 = 1, \quad x_0 = -x_1 = -\frac{1}{\sqrt{3}},$$

и квадратурная формула

$$\int_{-1}^{+1} f dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

точна на полиномах степени, меньшей или равной трем.

**Теорема 4.3.** Квадратурная формула вида

$$\sum_{k=0}^n c_k f(x_k)$$

точна для любых многочленов степени  $m \leq 2n + 1$  тогда и только тогда, когда выполнены два условия:

1)  $\int_a^b \rho(x)\omega(x)q(x) dx = 0$  для любого многочлена  $q(x)$  степени, меньшей  $n + 1$ , т. е. полином  $\omega(x) = \prod_{k=0}^n (x - x_k)$  ортогонален  $q(x)$  с весом  $\rho(x)$ ;

2) квадратурная формула является квадратурной формулой интерполяционного типа, т. е.

$$c_k = \int_a^b \rho(x) \frac{\omega(x)}{(x - x_k)\omega'(x_k)} dx, \quad k = 0, 1, \dots, n.$$

◀ Докажем необходимость приведенных условий. Формула точна для любого многочлена степени не выше  $m = 2n + 1$ . Значит, она точна и для многочлена  $\omega(x)q(x)$ , имеющего степень не выше  $m$ . Следовательно,

$$\int_a^b \rho(x)\omega(x)q(x) dx = \sum_{k=0}^n c_k \omega(x_k)q(x_k) = 0,$$

так как  $\omega(x_k) = 0$  для любого  $k$ . Условие 2 выполняется в силу теоремы 4.2 (так как степень  $m = 2n + 1 \geq n$ , для которой теорема и доказана).

Докажем достаточность. Пусть  $f(x)$  — многочлен степени  $m \leq 2n + 1$ . Тогда, поделив его на  $\omega(x)$ , получим  $f(x) = \omega(x)q(x) + r(x)$ , где  $q(x)$  и  $r(x)$  — многочлены степени не выше  $n$ , так как  $\omega(x)$  имеет степень  $n + 1$ . Вследствие того, что полученное выражение является квадратурной формулой интерполяционного типа, она точна для  $r(x)$ :

$$\begin{aligned} \int_a^b \rho(x)r(x) dx &= \sum_{k=0}^n c_k (f(x_k) - \omega(x_k)q(x_k)) = \\ &= \sum_{k=0}^n c_k f(x_k) = \int_a^b \rho(x)(f(x) - \omega(x)q(x)) dx = \int_a^b \rho(x)f(x) dx. \end{aligned}$$

Таким образом, формула точна и для многочлена степени  $m$ . ►

**Следствие 4.1.** Уравнения для определения  $n + 1$  узлов квадратурной формулы можно записать в виде

$$\int_a^b \rho(x)\omega(x)x^\alpha dx = 0, \quad \alpha = 0, 1, \dots, n.$$

Это утверждение вытекает из условия 1 теоремы 4.3. Оно значительно упрощает процедуру нахождения узлов квадратурной формулы.

Полученные уравнения для определения узлов квадратурной формулы фигурируют не только в теории методов численного интегрирования. В классической математике они появились значительно раньше при определении классических ортогональных полиномов (Чебышева, Лагерра, Эрмита и др.). В зависимости от веса и области изменения аргумента появляются те или иные полиномы.

Теорема 4.3 не дает ответа на вопросы, сколько различных решений — наборов узлов  $x_0, x_1, \dots, x_n$  — будет найдено в результате решения полученной системы; будут ли они лежать на  $[a, b]$ ; будут

ли они все различны и т. д. Мы эти вопросы рассматривать не будем. Укажем лишь, что при  $\rho(x) > 0$  такой многочлен  $\omega(x)$  существует, единственен, все его корни различны и расположены на  $(a, b)$ .

**Определение 4.5.** Построенные квадратурные формулы называются *квадратурными формулами наивысшей алгебраической степени точности* или *формулами Гаусса*.

Легко видеть, что для многочлена степени  $2n + 2$  формула Гаусса, вообще говоря, не является точной. Пусть  $f(x) = \omega^2(x)$  — многочлен степени  $2n + 2$ . Точное значение интеграла

$$I = \int_a^b \rho(x)f(x)dx > 0.$$

В то же время

$$\sum_{k=0}^n c_k f(x_k) = \sum_{k=0}^n c_k \omega^2(x_k) = 0.$$

Поскольку квадратурная формула Гаусса точна вплоть до степени  $m = 2n + 1$ , она точна и для многочлена степени  $2n$

$$f(x) = \varphi_k^2(x) = \left( \frac{\omega(k)}{(x - x_k)\omega'(x_k)} \right)^2.$$

Так как

$$\int_a^b \rho \varphi_i^2 dx = \sum_{k=0}^n c_k \varphi_i^2(x_k) = \sum_{k=0}^n c_k \delta_{ik}^2 = c_i = \int_a^b \rho \varphi_i^2 dx > 0,$$

то все коэффициенты  $c_i > 0$ . Следовательно, квадратурная формула Гаусса является устойчивой относительно возмущений функции  $f(x)$  (см. лемму 4.1). Поэтому на практике формулы Гаусса применяют до  $n \leq 100$ .

В чем же причина устойчивости? Очевидно, что устойчивость есть результат неравномерности сетки, специально построенной для исследуемой функции (см. 3.3.3).

Погрешность квадратурной формулы Гаусса может быть представлена в виде

$$\psi_h = \frac{1}{(2n+2)!} \int_a^b \rho(x) \omega^2(x) f^{(2n+2)}(\xi) dx,$$

где  $\xi = \xi(x)$ .

Для частных случаев весов  $\rho(x)$  узлы и коэффициенты Гаусса вычислены и приведены в справочниках.

## 4.4. Интегрирование быстроосциллирующих функций

Пусть необходимо вычислить интеграл

$$I = \int_a^b f(x) e^{i\nu x} dx$$

с большим значением  $\nu$  (здесь  $i$  — мнимая единица:  $i^2 = -1$ ). При использовании стандартного подхода каждый полупериод  $\pi/\nu$  необходимо разбить хотя бы на 10 отрезков. Тогда на  $(a, b)$  нужно иметь следующее число узлов *квадратурной формулы*:

$$N \approx 10 \frac{(b-a)}{\pi} \nu.$$

Нас интересует случай больших значений  $N$ .

Решить поставленную задачу можно, считая экспоненту под знаком интеграла весом, т. е. положив  $e^{i\nu x} = \rho(x)$ , и используя *квадратурные формулы* типа *Ньютона — Котеса*. Так, если  $L_n(x)$  — *интерполяционный полином Лагранжа*, то

$$I \approx \sum_{k=0}^n c_k f(x_k), \quad c_k = \int_a^b e^{i\nu x} \frac{\nu(x)}{(x - x_k)\nu(x_k)} dx, \quad k = 0, 1, \dots, n.$$

При этом погрешность

$$\psi_h = \int_a^b e^{i\nu x} (f - L_n) dx.$$

Получаемые квадратурные формулы носят названия *квадратурных формул Филона*.

При интегрировании быстроосциллирующих функций можно использовать не только глобальную, но и локальную *интерполяцию*, например *сплайны* различного порядка. Такие формулы также называются формулами Филона.

Рассмотрим вариант замены подынтегральной функции  $f(x)$  на ее среднее значение  $f(x_{i-1/2})$ , где  $x_{i-1/2} = x_{i-1} + h/2 = 1/2(x_i + x_{i-1})$  на равномерной (для простоты) сетке. Тогда получим квадратурную формулу Филона типа формул центральных прямоугольников:

$$I_{h,i} = f(x_{i-1/2}) \int_{x_{i-1}}^{x_i} e^{i\nu x} dx = \frac{2}{\nu} f(x_{i-1/2}) e^{i\nu x_{i-1/2}} \sin \frac{\nu h}{2}.$$

Погрешность построенной квадратурной формулы может быть оценена аналогично 4.1.

Ясно, что при необходимости можно построить формулы на основе линейных или более высокого порядка локальных сплайнов.

#### 4.5. Вычисление несобственных интегралов I и II рода

Пусть требуется вычислить несобственный интеграл, т. е. интеграл по неограниченному участку или интеграл от неограниченной функции.

**Интеграл I рода.** Пусть требуется вычислить интеграл

$$I = \int_a^{\infty} f(x) dx,$$

где  $f$  — гладкая функция. Существует ряд приемов расчета.

1. Замена переменной с целью получения интеграла по конечному участку.

**Пример 4.4.** Выполним замену переменной

$$x = \frac{a}{1-t} : \quad (a, \infty) \longrightarrow (0, 1),$$

$$dx = \frac{a}{(1-t)^2} dt, \quad I = \int_0^1 \frac{a}{(1-t)^2} f\left(\frac{a}{1-t}\right) dt. \quad \#$$

Если будет получена ограниченная результирующая подынтегральная функция, то далее для расчета следует применять обычные квадратурные формулы.

2. Обрезание верхнего предела. По определению несобственного интеграла I рода,

$$I = \int_a^{\infty} f(x) dx = \lim_{A \rightarrow \infty} \int_a^A f(x) dx = \lim_{A \rightarrow \infty} I_A.$$

Вычисление  $I_A$  можно проводить обычным способом. Иногда, заменив точное значение  $I$  интеграла I рода приближенным значением  $I_A$ , удается оценить ошибку

$$\delta I_A = \int_A^{\infty} f dx,$$

что позволяет выбрать  $A$  исходя из наперед заданной точности. Чаще же всего вычисляют значения  $I_{A_1}, I_{A_2}$ . Если  $A_2 > A_1$  (существенно), а  $I_{A_1}, I_{A_2}$  отличаются менее, чем на заданную малую величину, то процесс прекращают. Правда, это весьма ненадежный алгоритм при плохой сходимости интеграла.

3. Использование квадратурной формулы Гаусса. Для интеграла

$$I = \int_a^{\infty} \rho(x) f(x) dx$$

и для соответствующего веса  $\rho(x)$  ищется набор узлов  $\{x_k\}$ . А далее применяется квадратурная формула Гаусса.

**Пример 4.5.** Пусть необходимо вычислить значения функции Эйри

$$E_i(x) = \int_x^{\infty} \frac{e^{-t}}{t} dt = \int_0^{\infty} \frac{e^{-t-x}}{t+x} dt = e^{-x} \int_0^{\infty} \frac{e^{-t}}{t+x} dx.$$

В этом случае положим  $\rho(t) = \exp(-t)$ . Соответствующие узлы квадратурной формулы являются нулями полинома Лагерра  $\{x_k\}$ . Следовательно, получим некоторую аппроксимирующую формулу

$$E_i(x) = e^{-x} \sum_{k=0}^n \frac{\gamma_k}{x_k + x}.$$

Здесь  $\gamma_k = c_k$  — вес квадратурной формулы. #

4. Использование нестандартных формул. Необходимо приблизить  $f(x)$  какой-то функцией, интеграл от которой легко вычисляется. Например, если приблизить  $f(x)$  экспонентой на участке  $[x_n - h/2, +\infty)$ , потребовав точной передачи производной и функции в точке  $x_n$ , то получим

$$f(x) \approx \alpha e^{-\beta x}, \quad \alpha = f(x_n) e^{\beta x_n}, \quad f'(x_n) = -f(x_n) e^{\beta x_n - \beta x_n} \beta,$$

откуда

$$\beta = -\frac{f'(x_n)}{f(x_n)}, \quad \alpha = f(x_n) e^{\beta x_n}$$

и

$$\int_{x_n - h/2}^{+\infty} f(x) dx \approx \frac{\alpha}{\beta} e^{-\beta(x_n - h/2)},$$

где  $\alpha$  и  $\beta$  известны. Для вычисления интеграла по участку  $[a, x_n - h/2]$  используются обычные квадратурные формулы.

**Интеграл II рода.** Пусть требуется вычислить интеграл

$$I = \int_a^b f(x) dx,$$

где функция  $f(x)$  не ограничена в точке  $a$ . Существуют разнообразные приемы вычислений.

**1. Аддитивное выделение особенности.** Пусть  $f = \varphi + \psi$ , где функция  $\varphi$  содержит особенность, т. е. является неограниченной, и интегрируется аналитически, а функция  $\psi$  не содержит особенности и интегрируется численно обычным образом.

**Пример 4.6.** Используем аддитивное выделение особенности для вычисления интеграла

$$\int_0^1 \frac{f(t)}{\sqrt{t}} dt = \int_0^1 \frac{1}{\sqrt{t}} (f(t) - f(0) - tf'(0)) dt + \int_0^1 \frac{f(0)}{\sqrt{t}} dt + \int_0^1 f'(0) \sqrt{t} dt.$$

Здесь два последних интеграла вычисляются аналитически. В первом выделено столько слагаемых, чтобы можно было использовать подходящую квадратурную формулу.

**2. Мультипликативное выделение особенности.** Предположим, что  $f(x) = \varphi(x)\rho(x)$ , где  $\varphi(x)$  — гладкая и ограниченная функция, а функция  $\rho(x) > 0$  и интегрируема, но содержит особенность. В этом случае можно использовать квадратурную формулу Гаусса для вычисления интеграла, считая функцию  $\rho(x)$  весом.

**Пример 4.7.** Используем мультипликативное выделение особенности для вычисления интеграла

$$\int_{-1}^{+1} \frac{e^x}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n} \sum_{i=1}^n e^{x_i}.$$

Здесь

$$x_i = \cos \frac{\pi}{n} \left( i - \frac{1}{2} \right), \quad i = 1, 2, \dots, n, \quad —$$

нули многочленов Чебышева. #

3. Использование нестандартных аппроксимаций. Пусть  $f(x) = \rho(x)\varphi(x)$ , где функция  $\varphi(x)$  меняется слабо по сравнению с функцией  $\rho(x)$ . Тогда на участке  $[x_{i-1}, x_i]$  функция  $f(x) \approx \rho(x)\varphi(x_{i-1/2})$  и интеграл

$$\int_{x_i}^{x_i} f dx \approx \varphi(x_{i-1/2}) \int_{x_{i-1}}^{x_i} \rho(x) dx.$$

Часто последний интеграл можно вычислить аналитически (см. пример 4.7).

Напомним, что одним из методов вычисления интегралов от быстроосцилирующих функций является использование указанных аппроксимаций.

4. Обрезание нижнего предела. По определению несобственного интеграла II рода,

$$I = \int_a^b f(x) dx = \lim_{c \rightarrow a+0} \int_c^b f(x) dx = \lim_{c \rightarrow a+0} I_c.$$

Далее алгоритм вычислений аналогичен алгоритму обрезания верхнего предела для вычисления интегралов I рода.

## 4.6. Вычисление кратных интегралов

Пусть необходимо вычислить интеграл

$$I = \iint_G f(x, y) dx dy.$$

**Метод ячеек.** Пусть область  $G$  представляет собой прямоугольник:  $G = [a, b] \times [c, d]$ . Разобьем отрезки  $[a, b]$ ,  $[c, d]$  на  $M$  и  $N$  частей соответственно. В каждом элементарном прямоугольнике выберем центральную точку и запишем

$$I_h = \sum_{i=1}^M \sum_{j=1}^N f(x_{i-1/2}, y_{j-1/2}) h_{x,i} h_{y,j}, \quad h_{x,i} = x_i - x_{i-1}, \quad h_{y,j} = y_j - y_{j-1}.$$

Как и в одномерном случае, получим погрешность  $\psi_h = O(h_x^2 + h_y^2)$ , поскольку  $f(x, y)$  вычисляется в центре прямоугольника. Оценка погрешности определяется вторыми производными  $f''_{x^2}$ ,  $f''_{y^2}$ ,  $f''_{xy}$ . Данная

оценка может быть получена практически так же, как и соответствующая оценка для квадратурной формулы центральных прямоугольников.

Если элементарная ячейка не является прямоугольником, то тот же порядок точности будет установлен при вычислении  $f$  в точке, являющейся центром тяжести данной ячейки. Координаты  $\bar{x}$ ,  $\bar{y}$  центра тяжести определяются выражениями

$$\bar{x} = \frac{1}{S_h} \iint_{G_h} x \, dx \, dy, \quad \bar{y} = \frac{1}{S_h} \iint_{G_h} y \, dx \, dy, \quad S_h = \iint_{G_h} dx \, dy.$$

При получении оценки погрешности интегралы от линейных слагаемых в формуле Тейлора оказываются равными нулю. В результате погрешность остается той же.

**Последовательное интегрирование.** Пусть область интегрирования  $G$  можно представить в виде

$$G = \{x \in [a, b]: \varphi_1(x) \leq y \leq \varphi_2(x)\} = \{y \in [c, d]: \psi_1(y) \leq x \leq \psi_2(y)\},$$

тогда двойной интеграл можно вычислить с помощью последовательного интегрирования:

$$I = \iint_G f(x, y) \, dx \, dy = \int_a^b F(x) \, dx = \int_c^d R(y) \, dy,$$

$$F(x) = \int_{\varphi_1(x)}^{\varphi_2(x)} f(x, y) \, dy, \quad R(y) = \int_{\psi_1(y)}^{\psi_2(y)} f(x, y) \, dx.$$

В этом случае можно выполнить численное интегрирование по  $y$  для фиксированного набора  $x$ , а далее аналогично по  $x$ . Последовательное применение квадратурных формул по обеим переменным приводит к кубатурным формулам, являющимся прямым произведением одномерных квадратурных формул. При этом можно использовать все те квадратурные формулы, которые получены выше.

**Конечно-элементный подход.** Если

$$f(x, y) \approx \sum_{k=1}^N f(x_i, y_i) \varphi_i(x, y),$$

где  $\varphi_i(x, y)$  — двумерные базисные функции конечных элементов, то

$$I_n \approx \sum_{k=1}^N f(x_k, y_k) \int_G \varphi_i(x, y) dx dy.$$

При этом используется двумерная интерполяция на треугольных сетках.

#### 4.7. Численное дифференцирование

Пусть известно  $n + 1$  значений функции  $f(x)$  в  $n + 1$  точках  $x_0, x_1, \dots, x_n$ . Необходимо вычислить производную  $f'(x)$ . Рассмотрим возможные методы решения.

**Использование сплайн-интерполяции.** Мы знаем, что для функции  $f \in C^{(4)}(a, b)$  ее кубическая сплайн-интерполяция  $\varphi(x)$  такова, что

$$\|\varphi' - f'\| \leq C_1 M_4 h^3, \quad \|\varphi'' - f''\| \leq C_2 M_4 h^2.$$

Таким образом, сплайн позволяет получить  $f'$ ,  $f''$  путем прямого дифференцирования и с хорошей точностью. Здесь  $M_4 = \|f^{(4)}\|_C$ .

**Глобальная полиномиальная интерполяция.** Рассмотрим интерполяционный полином в форме полинома Ньютона:

$$\begin{aligned} L_n(x) &= f(x_0) + (x - x_0)f(x_0, x_1) + (x - x_0)(x - x_1)f(x_0, x_1, x_2) + \dots \\ &\quad \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1})f(x_0, x_1, \dots, x_n). \end{aligned}$$

**Лемма 4.2.** Пусть  $f(x) \in C^{(n)}[a, b]$ . Тогда

$$\exists \xi \in [a, b] : \quad f(x_0, x_1, \dots, x_n) = \frac{f^{(n)}(\xi)}{n!}.$$

◀ Рассмотрим остаток  $r_n(x) = f(x) - L_n(x)$ . Эта функция имеет  $n + 1$  нулей на  $[a, b]$ . Поэтому ее производная  $r'_n$  имеет  $n$  нулей, расположенных между нулями  $r_n$ , так как нули  $r_n$  являются некратными, и т. д. Следовательно,  $n$ -я производная имеет хотя бы один нуль  $\xi \in [a, b]$ :

$$r_n^{(n)}(\xi) = 0 = f^{(n)}(\xi) - n!f(x_0, x_1, \dots, x_n),$$

откуда  $f(x_0, x_1, \dots, x_n) = \frac{f^{(n)}(\xi)}{n!}$ . ►

Результат леммы показывает, что первая производная функции может быть приближенно вычислена с помощью формулы

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

а вторая производная — с помощью формулы

$$2!f(x_0, x_1, x_2) = \frac{2!}{x_2 - x_0} \left( \frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0} \right)$$

и т. д.

Однако при этом не понятно, к какой точке относится значение производной:  $\xi \in [a, b]$  ( $[a, b]$  — границы участка интерполяции).

Для вычисления  $y' = f'$  также можно использовать интерполянт  $L_n$ . Очевидно, что для вычисления  $f^{(n)}$  нужно использовать не менее  $n + 1$  точек интерполяции.

**Лемма 4.3.** Пусть  $f(x) \in C^{(n+1)}[a, b]$ ,  $L_n(x)$  — интерполяционный полином, построенный по значениям функции в точках сетки  $\Omega_h = \{a = x_0 < x_1 < x_2 \dots < x_n = b\}$ . Тогда погрешность вычисления производной  $q$ -го порядка ( $q = 1, 2, \dots, n$ ) удовлетворяет оценке

$$\|f^{(q)} - L_n^{(q)}\|_C \leq \frac{1}{(n+1-q)!} (b-a)^{n+1-q} \|f^{(n+1)}\|_C.$$

◀ Рассмотрим остаток  $r_n(x) = f(x) - L_n(x)$ . Эта функция имеет  $n + 1$  нулей на  $[a, b]$ . Таким образом, ее производная  $r'_n$  имеет  $n$  нулей, расположенных между нулями  $r_n$ , так как указанные нули являются некратными, и т. д. Следовательно,  $q$ -я производная функции  $r_n$  имеет не менее  $n + 1 - q$  нулей на отрезке  $[a, b]$ .

Это означает, что значения  $f^{(q)}(x)$  и  $L_n^{(q)}(x)$  совпадают по крайней мере в  $n + 1 - q$  точках отрезка  $[a, b]$ , т. е. полином  $L_n^{(q)}(x)$  является интерполяционным полиномом для функции  $f^{(q)}(x)$ . Последняя имеет производную  $(n + 1 - q)$ -го порядка, что позволяет стандартным образом оценить остаточный член интерполяции. При этом используется функция  $\tilde{\omega}(x) = \prod_{i=0}^{n-q} (x - \tilde{x}_i)$ , где точки  $a \leq \tilde{x}_0 < \tilde{x}_1 < \dots < \tilde{x}_{n-q} \leq b$  являются узлами интерполяции функции  $f^{(q)}(x)$ . Неопределенность положения точек интерполяции заставляет оценивать полученный остаток с помощью максимума  $|\tilde{\omega}|$  по всем возможным расположениям точек. ►

**Следствие 4.2.** Рассмотрим случай равномерной сетки, на которой  $b - a = hn$ . Тогда при выполнении условий леммы оценка погрешности

численного дифференцирования принимает вид

$$\|f^{(q)} - L_n^{(q)}\|_C \leq \frac{h^{n+1-q}}{(n+1-q)!} n^{n+1-q} \|f^{(n+1)}\|_C.$$

Полученная оценка бесполезна в крайнем случае  $q = n$ , так как при этом происходит сравнение численной производной с точной на всем участке интерполирования, в то время как точное совпадение этих величин в общем случае имеет место лишь в одной точке.

Рассмотрим другой крайний случай, когда число  $n + 1 - q$  велико. Здесь применение формулы Стирлинга дает оценку

$$\|f^{(q)} - L_n^{(q)}\|_C \leq \frac{h^{n+1-q}}{\sqrt{2\pi(n+1-q)}} e^n \|f^{(n+1)}\|_C.$$

**Пример 4.8.** Рассмотрим в качестве примера квадратичный интерполянт, построенный по заданным в трех точках  $x_{i-1}$ ,  $x_i$ ,  $x_{i+1}$  значениям функции  $f_{i-1}$ ,  $f_i$ ,  $f_{i+1}$ .

Интерполяция по трем точкам дает полином

$$\begin{aligned} L_2(x) = f_{i-1} \frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} + \\ + f_i \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} + f_{i+1} \frac{(x - x_{i-1})(x - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)}. \end{aligned}$$

Его производная представляет собой линейную функцию

$$\begin{aligned} L'_2(x) = f_{i-1} \frac{(2x - x_i - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} + \\ + f_i \frac{(2x - x_{i-1} - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} + f_{i+1} \frac{(2x - x_{i-1} - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)}. \end{aligned}$$

Из доказательства леммы следует, что значения  $L'_2(x)$  на отрезке  $[x_{i-1}, x_{i+1}]$  хотя бы дважды совпадут с точными значениями производной  $f'(x)$ . Точки этих совпадений заранее неизвестны. Тем не менее они лежат на отрезках  $[x_{i-1}, x_i]$  и  $[x_i, x_{i+1}]$ . Отсюда понятно, что значения  $L'_2(x_{i-1})$ ,  $L'_2(x_i)$ ,  $L'_2(x_{i+1})$  должны с некоторой точностью аппроксимировать точные значения производной исходной функции в этих точках. #

Далее мы увидим, что широко распространенные разностные соотношения представляют собой подобные выражения для производных соответствующего порядка.

**Использование разностных соотношений.** Пусть  $f(x_i)$  — значения функции  $y = f(x)$  в точках сетки. Для простоты рассмотрим равномерную сетку

$$x_i = a + ih, \quad i = 0, 1, \dots, n, \quad h = \frac{b - a}{n}.$$

Необходимо вычислить приближенные значения производных в точках  $x = x_i$ .

Возможны следующие варианты:

$y_{\bar{x},i} = (y_i - y_{i-1})/h$  — **разностная производная назад**, или **левая** разностная производная;

$y_{x,i} = (y_{i+1} - y_i)/h$  — **разностная производная вперед**, или **правая** разностная производная;

$y_{\circ,x,i} = (y_{i+1} - y_{i-1})/(2h)$  — **центральная разностная производная**.

Найдем ошибку приведенных разностных соотношений. Пусть  $y_i = f(x_i)$  — значение трижды непрерывно дифференцируемой функции  $f(x)$  в точке  $x_i$ . Тогда

$$\begin{aligned} f(x_{i-1}) &= y_{i-1} = f(x_i) + \frac{1}{1!}(x_{i-1} - x_i)f'(x_i) + \\ &\quad + \frac{1}{2!}(x_{i-1} - x_i)^2f''(x_i) + \frac{1}{3!}(x_{i-1} - x_i)^3f'''(\xi_i), \end{aligned}$$

$$f(x_{i+1}) = y_{i+1} = f(x_i) + \frac{1}{1!}hf'(x_i) + \frac{1}{2}h^2f''(x_i) + \frac{1}{6}h^3f'''(\tilde{\xi}_i).$$

Следовательно,

$$\begin{aligned} y_{\bar{x},i} &= f'(x_i) - \frac{1}{2}hf''(\xi_i^1), \quad y_{x,i} = f'(x_i) + \frac{1}{2}hf''(\xi_i^2), \\ y_{\circ,x,i} &= f'(x_i) + \frac{1}{6}h^2f'''(\xi_i^3). \end{aligned}$$

Заметим, что  $y_{\bar{x},i} = y_{x,i-1}$ ,  $y_{\bar{x},i+1} = y_{x,i}$ , а центральная разностная производная  $y_{\circ,x,i} = (y_{x,i} + y_{\bar{x},i})/2$  имеет повышенный порядок аппроксимации. Это соответствует и разным знакам в погрешностях  $y_{\bar{x},i}, y_{x,i}$ . При этом левая и правая разностные производные являются производными линейного сплайна в соответствующих точках, а центральная разностная производная есть производная квадратичного интерполянта в центре участка интерполирования на равномерной сетке.

**Пример 4.9.** Рассмотрим **метод Рунге — Ромберга** повышения точности формул численного дифференцирования на примере вычисления левой разностной производной на равномерной сетке. Для этого выберем точки  $x_{i-1}$ ,  $x_i$ ,  $x_{i+1}$  с заданными на них значениями

функции. Вычислим левую разностную производную в точке  $x_{i+1}$  двумя способами:

- 1) считая, что она принадлежит сетке с шагом  $h$ ;
- 2) считая, что она принадлежит сетке с шагом  $2h$ .

Воспользуемся полученным выше выражением для погрешности такой разностной производной и в результате запишем

$$\begin{aligned} y_{\bar{x}, i+1}^{(1)} &= \frac{1}{h}(y_{i+1} - y_i) = f'(x_{i+1}) + Ch + O(h^2), \\ y_{\bar{x}, i+1}^{(2)} &= \frac{1}{2h}(y_{i+1} - y_{i-1}) = f'(x_{i+1}) + C2h + O(h^2). \end{aligned}$$

Легко видеть, что главный член погрешности численного дифференцирования может быть исключен путем элементарных преобразований:

$$\tilde{y}_{\bar{x}, i+1} = \frac{1}{2h}(3y_{i+1} - 4y_i + y_{i-1}) = f'(x_{i+1}) + O(h^2).$$

Это выражение представляет собой одностороннюю («левую») разностную производную второго порядка аппроксимации на трех точках.

Точно такая же формула может быть получена при вычислении производной с использованием квадратичного интерполянта на равномерной сетке.

Заметим, что полученная формула дает приближенное значение производной на краю рассматриваемого участка, а не в его центре. Отсюда и следует отличие данной производной от центральной разностной производной.

Такой же прием использован выше при получении квадратурной формулы Симпсона. Очевидно, что этот прием имеет общий характер и применим всегда при наличии двух или более выражений (соответствующих разным шагам сетки) с известной асимптотикой ошибки по некоторому малому параметру (в данном случае — шагу сетки). Не составляет труда записать его формулы при наличии главной части остаточного члена ошибки вида  $O(h^p)$ .

**Пример 4.10.** В случае метода Рунге — Ромберга повышение точности происходит за счет удаления главного слагаемого (в рассмотренном примере  $Ch$  — ошибки некоторого численного выражения). Очевидно, что удаление главного слагаемого может быть выполнено в результате нахождения данного слагаемого из системы двух уравнений. Другим неизвестным является сама уточненная величина. Если бы главная часть ошибки равнялась  $Ch^p$ , то алгоритм был бы практически тем же самым.

Рассмотрим случай, в котором асимптотика главного члена погрешности некоторой численной формулы неизвестна, т. е. неизвестны

параметры  $C$  и  $p$ . Формулы, связывающие точные и приближенные значения, содержат три неизвестных, поэтому для их определения нужны три уравнения. Пусть в нашем распоряжении есть численные значения  $f_1, f_2, f_3$  — приближения некоторой величины  $f$  — на сетках с шагами  $h, qh, q^2h$  соответственно. Тогда имеем систему трех уравнений

$$\begin{aligned}f_1 &= f + Ch^p + O(h^{p+1}), \\f_2 &= f + Cq^ph^p + O(h^{p+1}), \\f_3 &= f + Cq^{2p}h^p + O(h^{p+1}).\end{aligned}$$

Решая ее с точностью  $O(h^{p+1})$ , получаем в итоге

$$f = f_1 + \frac{(f_1 - f_2)^2}{2f_2 - f_1 - f_3} + O(h^{p+1}), \quad p \approx \frac{1}{\ln q} \ln \frac{f_3 - f_2}{f_2 - f_1},$$

т. е. не только уточненное значение вычисляемой величины, но и эффективный порядок точности.

Описанный алгоритм называется *процессом Эйткена*. #

Введем разностную аппроксимацию *второй производной*:

$$\begin{aligned}y_{\bar{x}x} &= \frac{1}{h}(y_{\bar{x},i+1} - y_{\bar{x},i}) = \frac{1}{h}(y_{x,i} - y_{x,i-1}) = \\&= \frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}) = f''(x_i) + \frac{h^2}{12}f^{(4)}(\xi)\end{aligned}$$

в случае наличия непрерывной четвертой производной.

Можно ввести аппроксимацию третьей производной и производных более высокого порядка.

Отметим некорректность операции дифференцирования как в случае вычисления обычной производной функции непрерывного аргумента, так и при *численном дифференцировании*: операция вычитания и деления на малое число  $h$  ведет к существенному возрастанию ошибок. Так, если значения функции  $f$  известны с точностью  $\Delta f$ , то

$$\Delta f' = \frac{2\Delta f}{h}, \quad \Delta f'' = \frac{4\Delta f}{h^2}$$

и т. д. Полученная ошибка тем существеннее, чем меньше погрешность аппроксимации приближенной формулы.

Приведенные соотношения показывают, что формально правильное с точки зрения математического анализа устремление шага сетки к нулю с целью получения более точного результата численного дифференцирования может в случае приближенно заданной функции дать крайне неудовлетворительный результат.

Отсюда следует, что при численном дифференцировании необходимо предпринимать специальные меры для того, чтобы получать надежные результаты. Обычно подобные меры называют регуляризацией дифференцирования. Чаще всего они сводятся к поиску решения исходной задачи на некотором подпространстве исходного пространства, в котором содержатся более гладкие функции. Ранее термин «регуляризация» встречался при решении плохо обусловленных систем линейных алгебраических уравнений (СЛАУ). Более подробно этот алгоритм будет исследован при решении интегральных уравнений первого рода. Эти уравнения имеют самое прямое отношение к вычислению производной, так как процедура дифференцирования может быть сведена к решению интегрального уравнения Вольтерра первого рода.

Простейший пример регуляризации состоит в выборе шага сетки, согласованного с точностью задания функции. Если, например,  $|f''| \leq M_2$ , то ошибка приближенного вычисления производной скажется слабо на результате при

$$\frac{2\Delta f}{h} \approx \frac{1}{2}hM_2,$$

т. е.

$$\Delta f \approx \frac{1}{4}h^2M_2.$$

В этом случае ошибка численного дифференцирования есть величина порядка ошибки аппроксимации. То же выражение можно использовать для определения числа сетки:

$$h \approx h_0 = 2\frac{\sqrt{\Delta f}}{\sqrt{M_2}}.$$

Легко видеть, что приведенное значение шага сетки является оптимальным, обеспечивающим минимум суммы  $\frac{2\Delta f}{h} + h\frac{M_2}{2}$ .

Для старших производных условие выбора оптимального шага еще жестче.

Приведенные выражения показывают, что параметры сетки должны быть согласованы с решением задачи. В частности, на участках быстрых изменений решения (и, соответственно, больших значений производных) сетка должна быть мелкой, и наоборот. Отсюда вытекает необходимость использования неравномерных сеток, которые позволяют получить более точное решение задачи с меньшими затратами.

Некорректность операции дифференцирования имеет место и в случае функции непрерывного аргумента. Рассмотрим пример:

$$f(x) = \frac{1}{n} \sin(n^2 x), \quad \|f\|_C = \frac{1}{n} \ll 1$$

при  $n \gg 1$ . Тогда

$$f'(x) = n \cos(n^2 x), \quad \|f'\|_C = n \gg 1.$$

Следовательно, значения производной малой (в смысле исходного пространства  $C$ ) функции являются огромными.

#### 4.8. Библиографические комментарии

Интегрирование и вычисление различных интегралов являются неизменной оставляющей любого курса математического анализа. Практически любой курс методов вычислений содержит материал, посвященный численному интегрированию. То же самое можно сказать и о дифференцировании.

Поэтому задачи численного интегрирования и дифференцирования рассматриваются во всех книгах учебного характера, указанных в библиографии.

Отметим монографии [121] и [161], в которых рассмотрены вопросы построения квадратурных формул для различных классов функций, обладающих какими-либо особыми свойствами типа оптимальности. Данные монографии указывают на очертную связь численных методов и теории функций. Справочный материал по численному интегрированию и дифференцированию содержится также в [2] и [75].

В [80] приведен обширный материал по регуляризации численного дифференцирования, включая чисто практические приемы. В [81] описано применение так называемых квазиравномерных сеток, в том числе для расчета интегралов и производных.

Метод регуляризации подробно представлен в работах [118, 145, 168, 169].

Материал по классическим ортогональным полиномам и их применению в вычислительной математике можно найти в справочнике [2] и монографии [120].

В [32] дана интересная оценка ошибки численного дифференцирования в случае функции, значительно менее гладкой, чем рассмотренная нами.

Мы совсем не касаемся такого важного метода решения задач вычислительной математики, в том числе и задачи вычисления интегралов, как метод Монте-Карло [162]. Он в особенности эффективен при расчете многомерных интегралов.

В монографии [91] содержатся и многие другие методы численного интегрирования, в особенности полезные при решении больших задач.

# 5. ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Содержится описание и анализ методов типа *метода Рунге — Кутты* и *многошаговых разностных методов* для решения одного обыкновенного дифференциального уравнения (ОДУ) и систем таких уравнений. Даны основные понятия, характеризующие численные методы, предназначенные для численного решения. Для систем введено понятие *жесткой системы ОДУ*. Представлены методы решения таких систем и связанные с ними основные понятия.

## 5.1. Постановка задачи и простейшие методы

Рассмотрим *задачу Коши* для обыкновенного дифференциального уравнения (ОДУ)  $n$ -го порядка в форме, разрешенной относительно старшей производной:

$$u^{(n)} = f(x, u, u', \dots, u^{(n-1)}), \quad x > x_0,$$
$$x = x_0 : \quad u = u_0, \quad u' = u'_0, \quad \dots, \quad u^{(n-1)} = u_0^{(n-1)}.$$

Известно, что с помощью замены  $u_k = u^{(k-1)}$ ,  $k = 1, 2, \dots, n$ , задачу для ОДУ  $n$ -го порядка можно свести к системе ОДУ первого порядка:

$$\begin{cases} u'_k = u_{k+1}, \quad k = 1, 2, \dots, n-1, \quad x > x_0, \\ u'_n = f(x, u_1, u_2, \dots, u_n), \\ x = x_0 : \quad u_k(x_0) = u_0^{(k-1)}, \quad k = 1, 2, \dots, n. \end{cases}$$

Это частный случай системы ОДУ первого порядка в так называемой нормальной форме:  $u' = f(x, u)$ , где  $u$  и  $f$  — вектор-функции из  $n$  компонент. Тогда исходную задачу можно переписать в виде

$$u' = f(x, u), \quad x > x_0,$$
$$u(x_0) = u_0.$$

Параметр  $x_0$  и вектор  $u_0$  считаются заданными.

Будем изучать численные методы решения такой задачи, полагая, что условия существования и единственности ее решения выполнены, т. е. функция  $f(x, u)$  определена и непрерывна в прямоугольнике:

$$D = \{(x, u) : |x - x_0| \leq a, |u_i - u_0^i| \leq b, i = 1, 2, \dots, n\}.$$

При этих условиях в прямоугольнике  $D$  все компоненты  $|f_i| \leq M$ . Пусть функция  $f(x, u)$  является липшиц-непрерывной с постоянной  $L$  по переменным  $u_1, u_2, \dots, u_n$ :

$$|f(x, u^{(1)}) - f(x, u^{(2)})| \leq L \sum_{i=1}^n |u_i^{(1)} - u_i^{(2)}|.$$

Тогда существует единственное решение задачи Коши на участке

$$|x - x_0| \leq \tilde{x}_0 = \min \left\{ a, \frac{b}{M}, \frac{1}{L} \right\}.$$

Далее большинство методов численного решения будет излагаться на примере задачи Коши для одного уравнения. Как правило, методы для одного уравнения обобщаются на случай систем. Однако мы рассмотрим и методы, специально ориентированные на решение систем.

Будем считать  $x_0 = 0$  и положим  $x = t$ , так как обычно ОДУ связаны с эволюцией решения во времени. Введем на участке  $t \in (0, T)$  равномерную разностную сетку

$$\omega_\tau = \{t_k = k\tau, k = 0, 1, \dots\}.$$

Рассмотрим для начала простейшие методы.

**Метод Эйлера.** Запишем вместо исходного ОДУ систему алгебраических уравнений

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n),$$

где  $y_n = y(t_n)$  — приближенное решение, известное только на сетке  $\omega_\tau$ . Следовательно,  $y_n$  — сеточная функция. Имеем

$$\begin{aligned} y_{n+1} &= y_n + \tau f(t_n, y_n), \quad n = 0, 1, \dots; \\ y_0 &= u_0. \end{aligned}$$

Здесь и далее индекс, в том числе и у знака производной, указывает на номер точки, в которой вычисляется функция.

**Определение 5.1.** Говорят, что **численный метод** обладает **сходимостью** в точке  $t = t_n = n\tau$ , если

$$|y_n - u(t_n)| \rightarrow 0 \quad \text{при } n \rightarrow \infty,$$

где  $y_n$  — приближенное решение в данной точке;  $u_n = u(t_n)$  — точное решение. При этом  $\{y_n\}$  — последовательность приближенных решений на последовательности сеток  $\{\omega_\tau\}$ , таких, что  $t_n = n\tau$ .

**Определение 5.2.** Метод сходится на участке  $(0, T)$ , если он сходится в каждой точке  $(0, T)$ .

**Определение 5.3.** Метод имеет *p-й порядок точности*, если  $|y_n - u(t_n)| = O(\tau^p)$  при  $\tau \rightarrow 0$ .

**Определение 5.4.** *Погрешностью метода* называется сеточная функция

$$z_n = y_n - u(t_n).$$

Из выражения для  $y_{n+1}$  имеем

$$\frac{z_{n+1} - z_n}{\tau} = f(t_n, z_n + u(t_n)) - \frac{u_{n+1} - u_n}{\tau} = \psi_h^{(1)} + \psi_h^{(2)},$$

где

$$\psi_h^{(1)} = f(t_n, u_n) - \frac{u_{n+1} - u_n}{\tau}; \quad \psi_h^{(2)} = f(t_n, z_n + u_n) - f(t_n, u_n).$$

**Определение 5.5.** Функция  $\psi_h^{(1)}$  называется *невязкой* или *погрешностью аппроксимации* разностного уравнения на решении исходного уравнения.

Невязка может быть вычислена в результате подстановки точного решения в разностное уравнение. Если  $y_n = u_n$ , то невязка обращается в нуль.

**Определение 5.6.** Говорят, что имеет место *аппроксимация разностного уравнения* на решении исходного дифференциального уравнения, если

$$|\psi_h^{(1)}| \rightarrow 0$$

при  $\tau \rightarrow 0$ . Если же

$$|\psi_h^{(1)}| = O(\tau^p),$$

то говорят, что имеет место *аппроксимация p-го порядка*.

По формуле конечных приращений Лагранжа

$$\psi_h^{(2)} = \frac{\partial f}{\partial u}(t_n, u_n + \vartheta z_n) z_n,$$

где  $\vartheta \in (0, 1)$ . Будем считать здесь и далее, что все производные нужного порядка существуют.

Тогда

$$\psi_h^{(1)} = f(t_n, u_n) - u'(t_n) - \frac{1}{2} u''(t_n + \tilde{\vartheta}\tau)\tau = O(\tau)$$

в случае ограниченной производной  $u''$ , так как на точном решении справедливо равенство  $f(t_n, u_n) = u'(t_n)$ .

**Замечание 5.1.** Метод Эйлера встречается в качественной теории ОДУ при доказательстве теоремы существования с помощью так называемых ломаных Эйлера.

### 5.1.1. Симметричная схема

Пусть выбрана следующая схема:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{f(t_n, y_n) + f(t_{n+1}, y_{n+1})}{2}.$$

Она является нелинейной, так как для нахождения решения  $y_{n+1}$  в новой временной точке требуется решить нелинейное уравнение.

Для этой схемы

$$\begin{aligned} \psi_h^{(1)} &= \frac{f(t_n, u_n) + f(t_{n+1}, u_{n+1})}{2} - \frac{u_{n+1} - u_n}{\tau} = \frac{u'_n + u'_{n+1}}{2} - \frac{u_{n+1} - u_n}{\tau} = \\ &= \frac{1}{2} \left( u'_{n+1/2} - \frac{\tau}{2} u''_{n+1/2} + \frac{\tau^2}{8} \tilde{u}'''_{n+1/2} + u'_{n+1/2} + \frac{\tau}{2} u''_{n+1/2} + \frac{\tau^2}{8} \tilde{u}'''_{n+1/2} \right) - \\ &\quad - \frac{1}{\tau} \left( u_{n+1/2} + \frac{\tau}{2} u'_{n+1/2} + \frac{\tau^2}{8} u''_{n+1/2} + \frac{\tau^3}{48} \tilde{u}'''_{n+1/2} - u_{n+1/2} + \right. \\ &\quad \left. + \frac{\tau}{2} u'_{n+1/2} - \frac{\tau^2}{8} u''_{n+1/2} + \frac{\tau^3}{48} \tilde{u}'''_{n+1/2} \right) = O(\tau^2) \end{aligned}$$

на трижды непрерывно дифференцируемых решениях.

Здесь и далее использованы обозначения  $u'_n$ ,  $u'_{n+1/2}$ ,  $u'''_{n+1/2}$  и им подобные для значений соответствующих производных в указанных точках сетки. Индекс  $n + 1/2$  указывает на середину отрезка между  $n$ -й и  $(n + 1)$ -й точками:  $t_{n+1/2} = t_n + 0,5\tau$ .

**Метод Эйлера и симметричная схема решения ОДУ** — простейшие примеры так называемых разностных схем. Методы Рунге — Кутты в отличие от них допускают вычисление правой части не только в точках сетки, но и в промежуточных точках.

### 5.1.2. Метод Рунге — Кутты второго порядка

Для нахождения  $y_{n+1}$  выполним сначала одно действие по методу Эйлера с шагом  $\tau/2$ :

$$y_{n+1/2} = y_n + \frac{\tau}{2} f(t_n, y_n),$$

а далее вычислим

$$y_{n+1} = y_n + \tau f(t_{n+1/2}, y_{n+1/2}).$$

Тогда

$$\begin{aligned} \psi_h^{(1)} &= f(t_n + 0,5\tau, u_n + 0,5\tau f(t_n, u_n)) - \frac{1}{\tau}(u_{n+1} - u_n) = \\ &= f(t_n, u_n) + 0,5\tau(f'_t(t_n, u_n) + f(t_n, u_n)f'_u(t_n, u_n)) + O(\tau^2) - u'_n - \frac{\tau}{2}u''_n = \\ &= u'_n + 0,5\tau u''_n + O(\tau^2) - u'_n - \frac{\tau}{2}u''_n = O(\tau^2), \end{aligned}$$

так как из уравнения  $u' = f(t, u)$  следует равенство  $u'' = f'_t + f'_u u' = f'_t + f'_u f$ .

Таким образом, данный метод имеет второй порядок аппроксимации и в отличие от симметричной схемы является явным, т. е. для нахождения решения в новой временной точке не требуется решать нелинейное уравнение.

**Определение 5.7.** Реализация данного метода в виде двух шагов называется **методом предиктор-корректор** (предсказание-исправление).

**Определение 5.8.** Реализация того же метода в виде

$$k_1 = f(t_n, y_n), \quad k_2 = f(t_n + 0,5\tau, y_n + 0,5\tau k_1), \quad y_{n+1} = y_n + \tau k_2$$

называется **двухшаговым методом Рунге — Кутты** (или двухэтапным).

Существуют две большие группы методов численного решения задач Коши для ОДУ: многошаговые разностные методы и методы Рунге — Кутты. Не все методы естественным образом вкладываются в это разделение, но большинство все же принадлежит указанным группам.

Перейдем к детальному исследованию методов. В частности, далее будет показано, что порядок точности рассмотренных методов совпадает с порядком их аппроксимации.

## 5.2. Методы Рунге — Кутты

Рассмотрим задачу Коши для одного дифференциального уравнения:

$$\begin{aligned} u' &= f(t, u), \quad t > 0, \\ u(0) &= u_0. \end{aligned}$$

**Явный  $m$ -шаговый метод Рунге — Кутты** состоит в задании коэффициентов  $a_i$ ,  $b_{ij}$  и  $\sigma_i$ ,  $j = 1, 2, \dots, i-1$ ,  $i = 1, 2, \dots, m$ , и последовательном вычислении функций:

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + a_2\tau, y_n + b_{21}\tau k_1), \\ k_3 &= f(t_n + a_3\tau, y_n + b_{31}\tau k_1 + b_{32}\tau k_2), \\ &\dots \dots \dots \dots \dots \dots \dots \\ k_m &= f(t_n + a_m\tau, y_n + \tau \sum_{j=1}^{m-1} b_{mj} k_j), \end{aligned}$$

при этом приближенное решение  $y_{n+1}$  находят из уравнения

$$\frac{1}{\tau}(y_{n+1} - y_n) = \sum_{j=1}^m \sigma_j k_j.$$

Параметры  $a_i$ ,  $b_{ij}$ ,  $\sigma_i$  выбирают из соображений аппроксимации и точности, равно как и параметр  $m$ . Формулы Рунге — Кутты используются обычно при  $m \leq 5$ .

Исходя из аппроксимации обыкновенного дифференциального уравнения (ОДУ) на постоянной правой части получаем необходимое условие:

$$\sum_{j=1}^m \sigma_j = 1.$$

**Варианты метода.** При  $m = 1$  получаем схему Эйлера в своем обычном виде.

При  $m = 2$  получаем семейство методов

$$\begin{aligned} k_1 &= f(t_n, y_n), \quad k_2 = f(t_n + a_2\tau, y_n + b_{21}\tau k_1), \\ y_{n+1} &= y_n + \tau(\sigma_1 k_1 + \sigma_2 k_2). \end{aligned}$$

Для исследования погрешности аппроксимации запишем приближенное уравнение в виде

$$\frac{y_{n+1} - y_n}{\tau} = \sigma_1 f(t_n, y_n) + \sigma_2 f\left(t_n + a_2\tau, y_n + b_{21}\tau f(t_n, y_n)\right).$$

Отсюда

$$\begin{aligned}\psi_h^{(1)} &= \sigma_1 f(t_n, u_n) + \sigma_2 f\left(t_n + a_2 \tau, u_n + b_{21} \tau f(t_n, u_n)\right) - \frac{1}{\tau}(u_{n+1} - u_n) = \\ &= \sigma_1 u'_n + \sigma_2 \left( u'_n + a_2 \tau f'_t(t_n, u_n) + \right. \\ &\quad \left. + b_{21} \tau f(t_n, u_n) f'_u(t_n, u_n) + O(\tau^2) \right) - u'_n - \frac{1}{2} \tau u''_n = \\ &= \tau f'_t(t_n, u_n) \left( \sigma_2 a_2 - \frac{1}{2} \right) + \tau f(t_n, u_n) f'_u(t_n, u_n) \left( \sigma_2 b_{21} - \frac{1}{2} \right) + O(\tau^2),\end{aligned}$$

так как  $u'' = f'_t + f f'_u$ .

Выше уже использовано условие  $\sigma_1 + \sigma_2 = 1$ . Если все остальные параметры произвольны, то порядок аппроксимации равен единице.

Методы второго порядка могут быть получены при  $\sigma_2 a_2 = \sigma_2 b_{21} = 0,5$ . При  $a = \frac{1}{2\sigma}$  их можно записать в виде

$$\frac{y_{n+1} - y_n}{\tau} = (1 - \sigma) f(t_n, y_n) + \sigma f(t_n + a\tau, y_n + a\tau f(t_n, y_n)).$$

Если  $\sigma = 1$ , то получим метод, рассмотренный в 5.1. На практике применяется и метод с  $\sigma = 1/2$ .

Покажем, что методов Рунге — Кутты третьего порядка аппроксимации при  $m = 2$ , вообще говоря, не существует.

Пусть  $f(t, u) = u$ , тогда полученный метод (по крайней мере второго порядка аппроксимации) имеет вид

$$\frac{y_{n+1} - y_n}{\tau} = (1 - \sigma) y_n + \sigma (y_n + a\tau y_n) = y_n + \frac{\tau y_n}{2}.$$

При этом

$$\begin{aligned}\psi_h^{(1)} &= \left( 1 + \frac{1}{2} \tau \right) u_n - \frac{1}{\tau} (u_{n+1} - u_n) = \\ &= u'_n + \frac{1}{2} \tau u''_n - u'_n - \frac{1}{2} \tau u''_n - \frac{1}{6} \tau^2 \tilde{u}'''_n = -\frac{1}{6} \tau^2 u(t_n + \vartheta \tau),\end{aligned}$$

так как  $u' = u'' = u''' = u$ , т. е. наивысший порядок аппроксимации равен двум.

На практике в различных пакетах прикладных программ в основном используются методы Рунге — Кутты третьего и четвертого порядков.

Приведем без вывода наиболее употребительный в пакетах прикладных программ метод Рунге — Кутты четвертого порядка точности. Он состоит в последовательном вычислении значений:

$$\begin{aligned}k_1 &= f(t_n, y_n), \\k_2 &= f(t_n + 0,5\tau, y_n + 0,5\tau k_1), \\k_3 &= f(t_n + 0,5\tau, y_n + 0,5\tau k_2), \\k_4 &= f(t_n + \tau, y_n + \tau k_3),\end{aligned}$$

при этом приближенное решение  $y_{n+1}$  находят из выражения

$$y_{n+1} = y_n + \frac{\tau}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

**Доказательство сходимости.** В соответствии с методом Рунге — Кутты

$$\begin{aligned}\frac{y_{n+1} - y_n}{\tau} &= \sum_{j=1}^m \sigma_j k_j, \\k_j &= f\left(t_n + a_j \tau, y_n + \sum_{i=1}^{j-1} b_{ji} \tau k_i\right), \quad j = 1, 2, \dots, m, \quad a_1 = 0.\end{aligned}$$

Запишем приближенное решение метода в виде  $y_n = u_n + z_n$ , где  $u_n$  — точное решение;  $z_n$  — погрешность. Тогда

$$\frac{z_{n+1} - z_n}{\tau} = \psi_h^{(1)} + \psi_h^{(2)},$$

где погрешность аппроксимации ОДУ на точном решении (невязка) равна

$$\begin{aligned}\psi_h^{(1)} &= \sum_{i=1}^m \sigma_i k_i(t_n, u_n, \tau) - \frac{u_{n+1} - u_n}{\tau}, \\\psi_h^{(2)} &= \sum_{i=1}^m \sigma_i (k_i(t_n, y_n, \tau) - k_i(t_n, u_n, \tau)).\end{aligned}$$

Считаем, что  $y_0 = u_0$ , т. е. начальные данные задаются точно,  $t \in (0, T)$ ,  $t_n = n\tau \leq T$  для произвольного  $n$ .

**Теорема 5.1.** Пусть правая часть ОДУ удовлетворяет условию Липшица по второму аргументу с постоянной  $L$ ,  $\psi_h^{(1)}$  — погрешность

аппроксимации ОДУ на точном решении (невязка). Тогда для погрешности метода Рунге — Кутты при  $n\tau \leq T$  справедлива оценка

$$|z_n| = |y_n - u(t_n)| \leq Te^{\alpha T} \max_{0 \leq j \leq n-1} |\psi_j^{(1)}|,$$

где

$$\alpha = \sigma Lm(1 + Lb\tau)^{m-1}; \quad \sigma = \max_{1 \leq i \leq m} |\sigma_i|; \quad b = \max_{\substack{1 \leq i \leq m \\ 1 \leq j \leq i-1}} |b_{ij}|.$$

◀ Из соотношений для  $k_i$  имеем

$$\begin{aligned} & |k_i(t_n, y_n, \tau) - k_i(t_n, u_n, \tau)| \leq \\ & \leq L \left( |y_n - u_n| + \sum_{j=1}^{i-1} \tau b_{ij} |k_j(t_n, y_n, \tau) - k_j(t_n, u_n, \tau)| \right), \quad i = 1, 2, \dots, m. \end{aligned}$$

В частности,  $|k_1(t_n, y_n, \tau) - k_1(t_n, u_n, \tau)| \leq L|y_n - u_n|$ . Обозначим

$$g = L|y_n - u_n| = L|z_n|, \quad r_i = |k_i(t_n, y_n, \tau) - k_i(t_n, u_n, \tau)|.$$

Тогда, используя  $b$ , запишем

$$r_i \leq g + L\tau b \sum_{j=1}^{i-1} r_j.$$

В результате получаем

$$\begin{aligned} r_1 &\leq g, \\ r_2 &\leq g(1 + L\tau b), \\ r_3 &\leq g + L\tau b n + L\tau b g(1 + L\tau b) = g(1 + L\tau b)^2. \end{aligned}$$

Допустим, что для некоторого  $i$  справедлива оценка  $r_i \leq g\rho^{i-1}$ ,  $\rho = 1 + L\tau b$ . Тогда из основного неравенства верна оценка

$$r_{i+1} \leq g + (\rho - 1) \sum_{j=1}^i g\rho^{j-1} = g + (\rho - 1)g \frac{1 - \rho^i}{1 - \rho} = g(1 - 1 + \rho^i) = g\rho^i.$$

Отсюда для любого  $i$  справедлива оценка  $r_i \leq g\rho^{i-1}$ ,  $\rho = 1 + L\tau b$ . Следовательно,

$$\begin{aligned} |\psi_h^{(1)}| &\leq \sum_{j=1}^m |\sigma_j| |r_j| \leq \sigma g \sum_{j=1}^m \rho^{j-1} \leq \sigma g m \rho^{m-1} \leq \\ &\leq \sigma L |z_n| m (1 + \tau L b)^{m-1} = \alpha |z_n|. \end{aligned}$$

Здесь использованы введенные в формулировке теоремы параметры  $\alpha$  и  $\sigma$ . Тогда получим

$$|z_{n+1}| \leq |z_n|(1 + \alpha\tau) + \tau|\psi_{h,n}^{(1)}| \leq (1 + \alpha\tau)^{n+1}|z_0| + \sum_{j=0}^n \tau(1 + \alpha\tau)^{n-j}|\psi_{h,j}^{(1)}|.$$

Учитывая, что  $z_0 = 0$ , имеем

$$\begin{aligned} |z_{n+1}| &\leq (n+1)\tau(1 + \alpha\tau)^n \max_j |\psi_{h,j}^{(1)}| \leq \\ &\leq t_{n+1} e^{\alpha t_n} \max_{0 \leq j \leq n} |\psi_h^{(1)}| \leq T e^{\alpha T} \max_{0 \leq j \leq n} |\psi_{h,j}^{(1)}|. \end{aligned} \quad \blacktriangleright$$

**Следствие 5.1.** В условиях теоремы порядок точности метода Рунге — Кутты совпадает с порядком аппроксимации.

◀ Это следует из полученной оценки и равномерной по  $\tau$  ограниченности  $\alpha$ :

$$\alpha = \sigma L m (1 + L b \tau)^{m-1} \leq \sigma L m e^{(m-1)Lb\tau} \leq \sigma L m e^{(m-1)LbT}. \quad \blacktriangleright$$

**Замечание 5.2.** Оценка погрешности численного решения, представленная в доказательстве сходимости, является весьма громоздкой и вряд ли применима для практической оценки точности полученных численных результатов. Для практических целей значительно более употребимы *правило Рунге* или *процесс Эйткена* уточнения решения и определения реальной погрешности вычислений с эффективным порядком метода. Отметим, что и для вполне известного метода порядок может отличаться от теоретического, если решение задачи содержит какие-либо особенности.

Рассмотрим, например, правило Рунге применительно к вопросу оценки ошибки проведенных вычислений. Пусть в нашем распоряжении есть численные значения  $y^{(1)}$ ,  $y^{(2)}$  — приближения искомого решения  $u$  в данной точке — на сетках с шагами  $h$  и  $qh$  соответственно. Считаем известным порядок  $p$  главного члена ошибки численного решения. В результате имеем систему двух уравнений:

$$\begin{aligned} y^{(1)} &= u + Ch^p + O(h^{p+1}), \\ y^{(2)} &= u + Cq^p h^p + O(h^{p+1}). \end{aligned}$$

Решая ее с точностью  $O(h^{p+1})$ , запишем

$$u = y^{(1)} - \frac{y^{(2)} - y^{(1)}}{q^p - 1} + O(h^{p+1}), \quad Ch^p \approx \frac{y^{(2)} - y^{(1)}}{q^p - 1}.$$

В итоге мы получаем возможность найти более точное значение решения. Кроме того, зная  $C$ , легко оценить шаг сетки, такой, чтобы

вычисления проводились с наперед заданной точностью, т. е. выполнялось условие  $|Ch^p| \leq \varepsilon$  с заданным  $\varepsilon$ .

Отметим, что полученные соотношения не носят характер гарантированной точности. Они имеют асимптотический смысл. Тем не менее правило Рунге широко применяется для расчетов, в частности для автоматического выбора шага в программных комплексах.

### 5.3. Многошаговые разностные методы

Рассмотрим снова задачу Коши для одного дифференциального уравнения (ОДУ):

$$\begin{aligned} u' &= f(t, u), \quad t > 0, \\ u(0) &= u_0 \end{aligned}$$

и сетку  $\omega_\tau$  с постоянным шагом  $\tau > 0$ .

**Линейный  $m$ -шаговый разностный метод** решения ОДУ характеризуется системой разностных уравнений

$$\frac{a_0 y_n + a_1 y_{n-1} + \dots + a_m y_{n-m}}{\tau} = b_0 f_n + b_1 f_{n-1} + \dots + b_m f_{n-m},$$

$$n = m, m+1, \dots$$

Решение системы начинается со значения  $n = m$ . Следовательно, для обеспечения возможности расчета необходимо задать  $y_0, y_1, \dots, y_{m-1}$ . Обычно эти значения при постановке задачи неизвестны, но их можно вычислить с помощью каких-либо иных методов, например метода Рунге — Кутты.

В рассматриваемом методе  $a_0 \neq 0$ , все  $a_i, b_i$  — заранее заданные числовые коэффициенты. Они определены с точностью до постоянного множителя (все одновременно). Поэтому будем полагать, что выполнено условие нормировки

$$\sum_{i=0}^m b_i = 1.$$

При этом условии правая часть разностного уравнения может аппроксимировать правую часть исходного ОДУ. Обозначим  $f_{n-k} = f(t_{n-k}, y_{n-k})$ .

**Определение 5.9.** Если  $b_0 = 0$ , то **метод** называется **явным**, если  $b_0 \neq 0$ , то **неявным**.

Если  $b_k = 0$ ,  $k = 1, 2, \dots, m$ ,  $b_0 = 1$ , то такой **метод** часто называют **полностью неявным**.

В случае неявного метода для нахождения  $y_n$  необходимо решать, вообще говоря, нелинейное уравнение.

**Определение 5.10.** Линейный  $m$ -шаговый метод с  $a_0 = -a_1 = 1$ ,  $a_k = 0$ ,  $k = 2, 3, \dots, m$ , называется **методом Адамса**:

$$\frac{y_{n+1} - y_n}{\tau} = \sum_{k=0}^m b_k f_{n-k}.$$

### 5.3.1. Погрешность аппроксимации многошаговых методов

Справедливо выражение

$$\psi_h^{(1)} = \sum_{k=0}^m b_k f(t_{n-k}, u_{n-k}) - \frac{1}{\tau} \sum_{k=0}^m a_k u_{n-k}.$$

Предположив наличие у решения производных нужного порядка, получим

$$u_{n-k} = u(t_n - k\tau) = \sum_{l=0}^p \frac{(-k\tau)^l u^{(l)}(t_n)}{l!} + O(\tau^{p+1}),$$

$$f(t_{n-k}, u_{n-k}) = u'_{n-k} = u'(t_n - k\tau) = \sum_{l=0}^{p-1} \frac{(-k\tau)^l u^{(l+1)}(t_n)}{l!} + O(\tau^p).$$

В результате имеем

$$\begin{aligned} \psi_h^{(1)} &= \sum_{k=0}^m b_k \sum_{l=0}^{p-1} \frac{(-k\tau)^l u^{(l+1)}(t_n)}{l!} - \sum_{k=0}^m \frac{1}{\tau} a_k \sum_{l=0}^p \frac{(-k\tau)^l u^{(l)}(t_n)}{l!} + O(\tau^p) = \\ &= - \left( \sum_{k=0}^m \frac{a_k}{\tau} \right) u(t_n) + \sum_{l=1}^p \sum_{k=0}^m \left( b_k \frac{(-k\tau)^{l-1}}{(l-1)!} - \frac{1}{\tau} a_k \frac{(-k\tau)^l}{l!} \right) u^{(l)}(t_n) + O(\tau^p). \end{aligned}$$

Следовательно, для обеспечения погрешности аппроксимации порядка  $p$  должны выполняться  $p+1$  уравнений

$$\sum_{k=0}^m \frac{1}{\tau} a_k = 0, \quad \sum_{k=0}^m k^{l-1} \left( b_k + a_k \frac{k}{l} \right) = 0, \quad l = 1, 2, \dots, p.$$

Условие нормировки

$$\sum_{k=0}^m b_k = 1$$

и полученные  $p+1$  уравнений дают систему из  $p+2$  уравнений для определения  $2(m+1)$  неизвестных  $a_i$ ,  $b_i$ ,  $i = 0, 1, \dots, m$ .

Эта система уравнений не является переопределенной только при  $2(m+1) \geq p+2$ , т. е. при  $p \leq 2m$ . Следовательно, порядок аппроксимации  $m$ -шаговых методов не может быть больше  $2m$ . Для явных методов он, очевидно, не может быть выше  $2m-1$ .

Для методов Адамса получаем уравнения

$$l \sum_{k=0}^m k^{l-1} b_k = 1, \quad l = 1, 2, \dots, p.$$

Таким образом, имеется  $p$  уравнений (вместе с условием нормировки) для определения  $m+1$  параметров. Следовательно, в общем случае  $p \leq m+1$ .

**Замечание 5.3.** Количество уравнений равно  $p$ , так как уравнение

$$l \sum_{k=0}^m k^{l-1} b_k = 1$$

при  $l=1$  совпадает с условием нормировки. Явные методы Адамса ( $b_0 = 0$ ) имеют порядок аппроксимации  $p \leq m$ .

### 5.3.2. Устойчивость и сходимость разностных методов

Методы высокого порядка аппроксимации практически не используются, так как они неустойчивы. Остановимся на основных понятиях, относящихся к устойчивости и сходимости, но без подробностей и доказательств.

Рассмотрим наряду с исходным  $m$ -шаговым линейным разностным методом однородное разностное уравнение с постоянными коэффициентами

$$\sum_{k=0}^m a_k y_{n-k} = 0, \quad n = m, m+1, \dots,$$

и будем искать его решение вида  $y_k = q^k$ . Тогда для любой точки  $n$  получим

$$\sum_{k=0}^m a_k q^{m-k} = 0$$

**Определение 5.11.** Уравнение

$$\sum_{k=0}^m a_k q^{m-k} = 0$$

называется *характеристическим уравнением линейного  $m$ -шагового разностного метода*.

**Определение 5.12.** Говорят, что линейный  $m$ -шаговый разностный метод удовлетворяет **условию корней**, если все корни  $q_1, q_2, \dots, q_m$  характеристического уравнения лежат внутри или на границе единичного круга комплексной плоскости, причем на границе нет кратных корней. Тогда разностный метод называют **устойчивым**.

**Замечание 5.4.** Теория линейных разностных уравнений с постоянными коэффициентами весьма близка к теории линейных ОДУ. В частности, общее решение неоднородного уравнения в обоих случаях можно представить в виде суммы общего решения однородного уравнения и частного решения неоднородного уравнения. Общее решение однородного уравнения может быть найдено с помощью элементарного решения экспоненциального вида в случае дифференциального уравнения или приведенной выше степенной зависимости для алгебраического уравнения. Далее необходимо решить полученное характеристическое уравнение. Каждому простому корню соответствует свое линейно независимое решение. Если же корень кратный, то для получения нового элемента фундаментальной системы решений элементарное решение необходимо домножить на степень  $t$  в дифференциальном случае или на степень  $k$  в алгебраическом случае.

Отсюда понятны причины появления условия корней: для устойчивости общего решения необходима устойчивость решения однородного уравнения. Для этого никакой элемент фундаментальной системы решений, составляющих общее решение, не должен расти с увеличением  $k$ .

**Теорема 5.2.** Пусть разностный  $m$ -шаговый метод удовлетворяет условию корней и имеет порядок аппроксимации  $p$ . Тогда  $p \leq m + 1$  при нечетном  $m$  и  $p \leq m + 2$  при четном  $m$ . Для явных устойчивых  $m$ -шаговых устойчивых методов порядок аппроксимации не превосходит  $m$ .

Без доказательства.

**Теорема 5.3.** Пусть разностный  $m$ -шаговый метод удовлетворяет условию корней и  $|f'_u| \leq L$ . Тогда для любого  $m\tau \leq t_n = n\tau \leq T$  при малом  $\tau$  выполнена оценка

$$|y_n - u(t_n)| \leq M \left( \max_{0 \leq j \leq m-1} |y_j - u(t_j)| + \max_{m \leq j \leq n} |\psi_{h,j}^{(1)}| \right),$$

где  $M$  — постоянная, не зависящая от  $m$ ;  $|y_j - u(t_j)|$ ,  $j = 0, 1, \dots, m-1$ , — погрешность в задании начальных условий;  $\psi_{h,j}^{(1)}$ ,  $j = m, m+1, \dots, n$ , — погрешность аппроксимации (невязка).

Без доказательства.

Из оценки теоремы 5.3 следует сходимость метода, если начальные погрешности сходятся к нулю при  $\tau \rightarrow 0$  и имеет место аппроксимация.

Методы Адамса всегда удовлетворяют условию корней, так как  $a_0 = 1$ ,  $a_1 = -1$ , откуда характеристическое уравнение имеет вид  $a_0 q + a_1 = 0$ , и, следовательно,  $q = 1$ .

### 5.3.3. Примеры методов Адамса

Пусть рассматривается чисто явный метод, т. е.  $b_0 = 0$ . Имеем условия  $p$ -го порядка аппроксимации

$$\sum_{k=0}^m k^{l-1} b_k = l^{-1}, \quad l = 1, 2, \dots, p = m.$$

При  $m = 1$  получим  $b_0 = 0$ ,  $b_1 = 1$ , т. е. метод Эйлера.

При  $m = 2$  имеем систему

$$\begin{aligned} b_0 &= 0, \\ b_1 + b_2 &= 1, \\ b_1 + 2b_2 &= \frac{1}{2}, \end{aligned}$$

откуда

$$b_0 = 0, \quad b_1 = \frac{3}{2}, \quad b_2 = -\frac{1}{2}.$$

При  $m = 3$  имеем систему

$$\begin{aligned} b_0 &= 0, \\ b_1 + b_2 + b_3 &= 1, \\ b_1 + 2b_2 + 3b_3 &= \frac{1}{2}, \\ b_1 + 4b_2 + 9b_3 &= \frac{1}{3}, \end{aligned}$$

откуда

$$b_0 = 0, \quad b_1 = \frac{23}{12}, \quad b_2 = -\frac{16}{12}, \quad b_3 = \frac{5}{12}.$$

и т. д.

Если же рассматривать и неявные методы, то  $p = m + 1$ . Тогда

$$\sum_{k=0}^m k^{l-1} b_k = l^{-1}, \quad l = 1, 2, \dots, p = m + 1.$$

При  $m = 1$  получим систему

$$\begin{aligned} b_0 + b_1 &= 1, \\ b_1 &= \frac{1}{2}, \end{aligned}$$

откуда  $b_0 = 1/2$ , и таким образом приходим к симметричной схеме:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{f_n + f_{n+1}}{2}.$$

Процедура может быть продолжена.

## 5.4. Понятие о методах решения жестких систем

### 5.4.1. Условно устойчивые и безусловно устойчивые разностные методы

*Условие корней* очень общее и никак не учитывает структуру правой части обыкновенного дифференциального уравнения (ОДУ) и, следовательно, характерных особенностей решения. Рассмотрим, например, следующую задачу Коши:

$$\begin{aligned} u' &= -\alpha^2 u, \quad t > 0, \\ u(0) &= u_0. \end{aligned}$$

Тогда  $u(t) = u_0 \exp(-\alpha^2 t)$ , откуда  $|u(t_{n+1})| < |u(t_n)|$ , т. е. решение монотонно (с сохранением знака) убывает.

Рассмотрим метод Эйлера

$$\frac{y_{n+1} - y_n}{\tau} = -\alpha^2 y_n,$$

или  $y_{n+1} = y_n(1 - \alpha^2 \tau)$ , откуда  $|y_{n+1}| \leq |y_n|$  при  $0 \leq \tau \leq 2/\alpha^2$ .

Таким образом, метод Эйлера (*явный*) устойчив в смысле удовлетворения оценки  $|y_{n+1}| \leq |y_n|$  лишь при выполнении условия  $0 < \tau \leq 2/\alpha^2$ .

**Определение 5.13.** Разностный метод называется *условно устойчивым*, если он устойчив при некоторых ограничениях на шаг  $\tau$ , и *безусловно устойчивым*, если он устойчив при произвольных  $\tau$ .

Рассмотрим пример безусловно устойчивого метода — *неявный метод Эйлера*:

$$\frac{u_{n+1} - u_n}{\tau} = -\alpha^2 u_{n+1},$$

откуда

$$|u_{n+1}| = |u_n(1 + \alpha^2 \tau)^{-1}| \leq |u_n|.$$

Чаще всего явные схемы являются условно устойчивыми, а среди неявных схем есть безусловно устойчивые. В последнем случае приходится решать нелинейное уравнение для  $y_{n+1}$ , если правая часть  $f$  нелинейным образом зависит от решения. При этом ограничения на шаг не накладываются. Однако явные методы намного проще в реализации.

### 5.4.2. Понятие жесткой системы ОДУ

Многие из рассмотренных выше методов решения одного ОДУ можно без проблем перенести на случай систем ОДУ. Однако при этом могут появиться трудности, связанные именно с наличием системы.

**Пример 5.1.** Рассмотрим задачу Коши:

$$\begin{aligned} u'_1 &= -u_1, \quad t > 0, \\ u'_2 &= -\varepsilon^2 u_2. \end{aligned}$$

Тогда  $u_1 = u_{1,0} \exp(-t)$ ,  $u_2 = u_{2,0} \exp(-\varepsilon^2 t)$ .

Пусть задача решается на участке  $(0, T)$ , причем  $\varepsilon^2 T \gg 1$ , например,  $\varepsilon^2 T = 5$ , т. е.  $T = 5\varepsilon^{-2}$ . Если система решается явным методом Эйлера, то должно быть выполнено условие

$$\tau \leq \min \left\{ 2, \frac{2}{\varepsilon^2} \right\}.$$

Если  $\varepsilon^2 \ll 1$ , то эти два ограничивающих шаг  $\tau$  значения отличаются в  $\varepsilon^{-2} \gg 1$  раз. В результате при расчете на равномерной сетке потребуется  $T/\tau = 5/2\varepsilon^{-2} = 2,5\varepsilon^{-2}$  временных слоев. Количество слоев определяется условием устойчивости самой быстроменяющейся составляющей части решения. Это становится бессмысленным с некоторого момента времени, так как функция  $u_1$  к моменту времени  $t = 5$  станет пренебрежимо малой, а именно связанный с ней шаг  $\tau$  определяет количество слоев. #

Искусственность данного примера является только видимой. Рассмотрим более общую ситуацию. Пусть необходимо решить систему  $u' = Au$  с постоянной матрицей, которую можно привести к диагональной матрице преобразованием вида  $Q^{-1}AQ$ . Тогда замена  $u = Qu$  приводит исходное уравнение к системе  $v_t = Q^{-1}AQv$  с диагональной матрицей, которая имеет те же самые собственные числа, что и матрица  $A$ .

**Определение 5.14.** Система ОДУ  $u' = Au$  с постоянной матрицей  $A = A_{m \times m}$  называется **жесткой**, если:

1) все собственные числа матрицы  $A$  имеют отрицательную действительную часть, т. е.  $\operatorname{Re} \lambda_i < 0$ ,  $i = 1, 2, \dots, m$ ;

2) число

$$S = \frac{\max_{1 \leq k \leq m} |\operatorname{Re} \lambda_k|}{\min_{1 \leq k \leq m} |\operatorname{Re} \lambda_k|}$$

велико. Число  $S$  называется **числом жесткости**.

Если  $\lambda_k = \lambda_k(t)$ , то вводят понятие **жесткости на временном интервале**. Тогда число

$$\sup_{t \in (0, T)} S(t)$$

должно быть велико.

Нечто подобное можно ввести и для нелинейных систем, рассмотрев их локальную линеаризацию.

Отметим, что существуют и другие определения жестких систем и жесткости.

#### 5.4.3. Решение жестких систем

При разработке методов решения жестких систем ОДУ, как правило, каждый новый метод прежде всего тестируется на модельном уравнении  $u' = \lambda u$ , где  $\lambda$  — параметр. Для того чтобы проявленные на нем свойства алгоритма как-то соответствовали системе  $u' = Au$ , необходимо просмотреть всю область изменения  $\lambda$ , соответствующую диапазону собственных чисел матрицы  $A$ .

Чаще всего рассматриваются методы вида

$$\frac{1}{\tau} \sum_{k=0}^m a_k y_{n-k} = \sum_{k=0}^m b_k f_{n-k}.$$

Применительно к уравнению  $u' = \lambda u$  метод сводится к решению следующей системы алгебраических уравнений:

$$\sum_{k=0}^m (a_k - \lambda \tau b_k) y_{n-k} = 0.$$

Обозначим  $\lambda \tau = \mu$ .

Тогда характеристическое уравнение для данного метода, которое есть следствие поиска его решения в виде  $y_n = q^n$ , записывается следующим образом:

$$\sum_{k=0}^m (a_k - \mu b_k) q^{m-k} = 0.$$

Это уравнение отличается от записанного при обсуждении условия корней в 5.3.2.

**Определение 5.15.** *Областью устойчивости  $m$ -шагового линейного разностного метода будем называть множество точек  $\mu = \lambda \tau$  комплексной плоскости, для которых данный метод применительно к уравнению  $u' = \lambda u$  будет устойчив (т. е.  $|y_{n+1}| \leq |y_n|$ , решение харак-*

теристического уравнения удовлетворяет условию  $|q| \leq 1$ , на границе области устойчивости  $|q| = 1$  нет кратных корней).

**Пример 5.2.** Запишем явный метод Эйлера в используемых обозначениях:  $y_{n+1} = y_n(1 + \mu)$ . Отсюда получим уравнение для области устойчивости  $|1 + \mu| \leq 1$ , или, полагая  $\mu = \mu_x + i\mu_y$  ( $i$  — мнимая единица),

$$(1 + \mu_x)^2 + \mu_y^2 \leq 1.$$

Область устойчивости представляет собой круг единичного радиуса с центром в точке  $(-1, 0)$  (рис. 5.1).

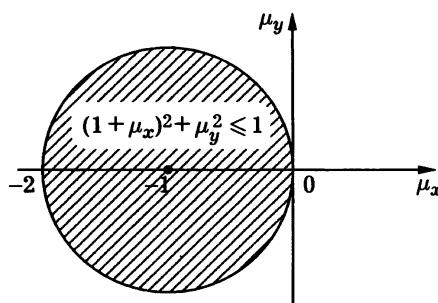


Рис. 5.1

**Пример 5.3.** Неявный метод Эйлера имеет вид  $y_{n+1} = y_n(1 - \mu)^{-1}$ . Отсюда получим неравенство для области устойчивости  $|(1 - \mu)^{-1}| \leq 1$ , или, полагая  $\mu = \mu_x + i\mu_y$ ,

$$(1 - \mu_x)^2 + \mu_y^2 \geq 1,$$

т. е. внешность единичного круга с центром в точке  $(1, 0)$  (рис. 5.2).

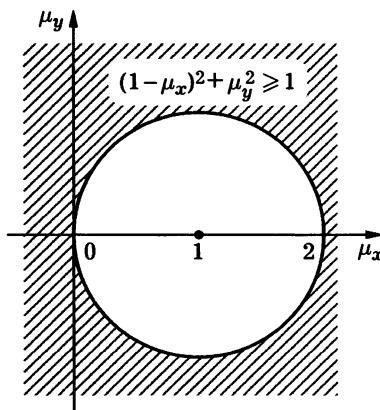


Рис. 5.2

**Определение 5.16.** *Разностный метод называется  $A$ -устойчивым*, если область его устойчивости содержит левую полуплоскость  $\operatorname{Re} \mu < 0$ .

**Определение 5.17.** *Разностный метод называется  $A(\alpha)$ -устойчивым*, если существует угол  $\alpha \in (0, \pi/2]$ , такой, что область его устойчивости содержит сектор комплексной плоскости переменной  $\mu$ , определяемый неравенством  $|\arg(-\mu)| < \alpha$ .

В рамках данного определения  $A$ -устойчивость есть  $A(\pi/2)$ -устойчивость.

В случае асимптотически устойчивых систем  $\operatorname{Re} \lambda < 0$  и, следовательно, для таких систем  $A$ -устойчивые методы устойчивы при любых  $\tau > 0$ , что означает безусловную устойчивость.

**Пример 5.4.** Рассмотрим *полунеявный метод*

$$\frac{y_{n+1} - y_n}{\tau} = \lambda \frac{y_n + y_{n+1}}{2},$$

или

$$y_{n+1} = y_n \frac{1 + \mu/2}{1 - \mu/2}.$$

Отсюда получаем неравенство для определения области устойчивости:

$$\left(1 + \frac{\mu_x}{2}\right)^2 + \left(\frac{\mu_y}{2}\right)^2 \leq \left(1 - \frac{\mu_x}{2}\right)^2 + \left(\frac{\mu_y}{2}\right)^2,$$

откуда следует, что  $\mu_x \leq 0$ , т. е. данный метод является  $A$ -устойчивым. #

Отметим, что среди линейных явных  $m$ -шаговых разностных методов нет  $A$ -устойчивых.

Покажем это. Из характеристического уравнения в силу равенства  $b_0 = 0$  получаем, что для любого  $q$  справедливо равенство

$$\mu = \frac{\sum_{k=0}^m a_k q^{m-k}}{\sum_{k=1}^m b_k q^{m-k}}.$$

Таким образом, при больших (здесь и далее по модулю) значениях  $q$  параметр  $\mu$  растет линейно как  $a_0/b_1 q$ , если  $b_1 \neq 0$ , либо как более высокая степень  $q$  в случае равенства  $b_1 = 0$ . Следовательно, для любого достаточно большого  $\mu$  найдется  $q$  из левой полуплоскости, в том числе

с  $|q| > 1$ , для которого справедливо характеристическое уравнение. В результате  $A$ -устойчивость не имеет места.

Точно так же доказано, что ни для какого  $\alpha$  не существует явного  $A(\alpha)$ -устойчивого линейного многошагового метода.

Поэтому для решения жестких систем в последнее время используют так называемый **метод Гира** — полностью неявный многошаговый разностный метод высокого порядка аппроксимации, т. е.

$$\frac{1}{\tau} \sum_{k=0}^m a_k y_{n-k} = f(t_n, y_n).$$

Так как  $b_0 = 1$ ,  $b_k = 0$ ,  $k = 1, 2, \dots, m$ , то этот метод является полностью неявным.

Условия  $p$ -го порядка аппроксимации имеют вид

$$\sum_{k=0}^m a_k = 0,$$

$$\sum_{k=0}^m k^{l-1} \left( b_k + a_k \frac{k}{l} \right) = 0, \quad l = 1, 2, \dots, p;$$

$$b_0 = 1, \quad b_k = 0, \quad k = 1, 2, \dots, m.$$

Отсюда получаем  $p+1$  уравнений для  $m+1$  неизвестных:

$$\sum_{k=0}^m a_k = 0,$$

$$\sum_{k=0}^m a_k k = -1,$$

$$\sum_{k=0}^m k^l a_k = 0, \quad l = 2, 3, \dots, p.$$

Следовательно,  $p \leq m$ .

**Пример 5.5.** Рассмотрим варианты метода Гира:

при  $m = 1$  получаем неявный метод Эйлера;

при  $m = 2$  получаем метод второго порядка точности

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = \tau f_n;$$

при  $m = 3$  имеем метод третьего порядка точности:

$$\frac{11}{6}y_n - 3y_{n-1} + \frac{3}{2}y_{n-2} - \frac{1}{3}y_{n-3} = \tau f_n.$$

**Пример 5.6.** Рассмотрим последний вариант метода Гира при  $m = 3$  и найдем область его устойчивости. Для модельного уравнения, на котором тестируются все методы, получаем

$$\frac{11}{6}y_n - 3y_{n-1} + \frac{3}{2}y_{n-2} - \frac{1}{3}y_{n-3} = \lambda\tau y_n = \mu y_n.$$

Далее ищем его решение вида  $y_n = q^n$ . После подстановки и сокращения на  $q^{n-3}$  получим характеристическое уравнение

$$\frac{11}{6}q^3 - 3q^2 + \frac{3}{2}q - \frac{1}{3} = \mu q^3.$$

Записать формально аналитическое решение кубического уравнения еще можно. Однако найти из его решения область устойчивости метода крайне сложно. Существует простой выход: нужно найти границу области устойчивости, после чего определить области устойчивости и неустойчивости обычным методом пробных точек, а именно: граница области устойчивости соответствует таким  $q$ , что  $|q| = 1$ . Для них  $q = e^{i\varphi}$ , где  $i$  — мнимая единица, а  $\varphi$  — аргумент комплексного числа (действительное число).

Таким методом получаем параметрическое задание границы области устойчивости в виде

$$\frac{11}{6} - 3e^{-i\varphi} + \frac{3}{2}e^{-2i\varphi} - \frac{1}{3}e^{-3i\varphi} = \mu.$$

Здесь  $\varphi \in [0, 2\pi]$ . В частности,

$$\mu(0) = 0, \quad \mu\left(\frac{\pi}{2}\right) = \frac{5}{6} + \frac{10}{3}i, \quad \mu(\pi) = \frac{20}{3}.$$

Граница области устойчивости симметрична относительно действительной оси. Нетрудно видеть, что область устойчивости находится вне области, граница которой описывается данной кривой.

## 5.5. Библиографические комментарии

Приведенный в данной главе материал является традиционным [149]. Основы теории ОДУ можно найти, например, в [3].

Многие подробности рассматриваемых алгоритмов и другие способы решения задачи Коши можно найти в [6, 12, 14, 17, 18, 25, 53, 61, 62, 80, 83, 84, 90, 103] и других руководствах по численным методам.

Апробированные и новые подходы к решению жестких задач представлены в [80], [103], [176].

Особо отметим двухтомник [184, 185], являющийся на сегодня, пожалуй, самым полным изложением методов численного решения ОДУ.

В последнее время в связи с появлением хороших пакетов программ для решения больших жестких систем ОДУ новую жизнь обрел метод прямых, предназначенный для решения эволюционных уравнений и систем таких уравнений с частными производными. В случае метода прямых проводится дискретизация по пространственным переменным, в результате чего получается задача Коши для системы ОДУ большой размерности. Полученная система, как правило, оказывается жесткой. Подробности данного метода можно найти в [25, 62, 95, 96] и другой литературе, представленной в библиографии.

# 6. РЕШЕНИЕ КРАЕВЫХ ЗАДАЧ ДЛЯ СИСТЕМ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Изложены простейшие методы решения краевых задач для обыкновенных дифференциальных уравнений (ОДУ). Представлен *метод стрельбы* для одного и нескольких уравнений. На примере простейшей краевой задачи для уравнений второго порядка (линейного и нелинейного) описан *разностный метод* с доказательством сходимости, обоснована применимость *метода Ньютона* в нелинейном случае. Приведен пример точной разностной схемы — *схемы с экспоненциальной подгонкой*. Представлены *методы Ритца* и *Галеркина*.

## 6.1. Постановка задачи. Метод стрельбы

Пусть на участке  $t \in (a, b)$  задана система  $n$  обыкновенных дифференциальных уравнений (ОДУ) первого порядка в нормальной форме:

$$u' = f(t, u), \quad a < t < b.$$

Здесь  $u$  и  $f(t, u)$  — векторные функции.

Для определения единственного решения этой задачи необходимо задать  $n$  дополнительных условий. Если они заданы более чем в одной точке, то такая задача называется краевой. При этом возможно наличие двух и более точек, в том числе связанных между собой. Ограничимся случаем двух точек, т. е. дополнительными условиями вида

$$\varphi_k(u_1(a), u_2(a), \dots, u_n(a)) = 0, \quad k = 1, 2, \dots, m;$$

$$\varphi_k(u_1(b), u_2(b), \dots, u_n(b)) = 0, \quad k = m + 1, m + 2, \dots, n.$$

При этом уравнения в точках  $a$  и  $b$  напрямую не связаны между собой,  $\varphi_k$  — заданные функции.

В данной главе рассмотрим метод приближенного решения такой задачи, основанный на ее сведении к задаче Коши для той же системы, так называемый *метод стрельбы*.

**Случай  $n = 2$ .** Интерес представляет лишь ситуация с  $n = 2, m = 1$ . Только в этом случае данная задача является краевой, т. е. на решение накладываются дополнительные условия в точках  $a$  и  $b$ .

Выбрав одно из граничных значений равным  $\eta$ , зададим  $y_1(a), y_2(a)$  так, чтобы  $\varphi_1(y_1(a), y_2(a)) = 0$ . Пусть, например,  $y_2(a) = \eta$ , тогда

$$\varphi_1(y_1(\eta, a), \eta) = 0.$$

Очевидно, что описываемый алгоритм имеет смысл только в случае, когда последнее уравнение разрешимо и дает в результате  $y_1(\eta, a) = \xi(\eta)$ .

Таким образом, получаем задачу Коши для той же системы. В результате ее решения имеем  $y_1(\eta, b)$  и  $y_2(\eta, b)$ , причем очевидно, что

$$\varphi_2(y_1(\eta, b), y_2(\eta, b)) = \psi(\eta) \neq 0.$$

Здесь  $\eta$  — параметр. Необходимо подобрать его так, чтобы получить  $\psi(\eta) = 0$ . Наиболее простой способ решения уравнения  $\psi(\eta) = 0$  — метод деления пополам некоторого отрезка  $[\eta_1, \eta_2]$ , такого, что  $\psi(\eta_1) < 0$ , а  $\psi(\eta_2) > 0$  (знаки  $\psi$  могут быть обратными), т. е. «перелет»—«недолет» (рис. 6.1).

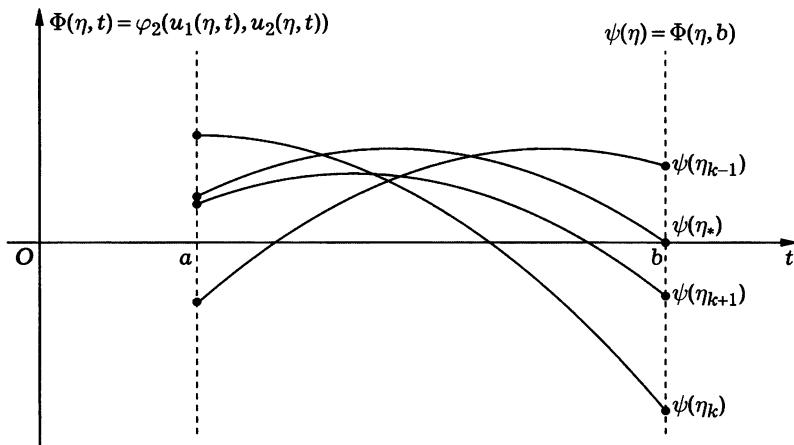


Рис. 6.1

Можно сконструировать и реализовать варианты метода Ньютона или метода секущих, или метода парабол, которые бы позволили решать эту задачу быстрее.

Особенно просто метод стрельбы можно реализовать для решения линейных задач, так как в этом случае очевидно, что решение линейным образом зависит от  $\eta$ . Рассмотрим следующую задачу:

$$\begin{cases} u'_1 = a_{11}(x)u_1 + a_{12}(x)u_2 + f_1(x), & a < t < b, \\ u'_2 = a_{21}(x)u_1 + a_{22}(x)u_2 + f_2(x), \\ \alpha_1 u_1(a) + \alpha_2 u_2(a) = \gamma_1, \\ \beta_1 u_1(b) + \beta_2 u_2(b) = \gamma_2. \end{cases}$$

Если  $y_2(a) = \eta$ ,  $\alpha_1 \neq 0$ , то функции

$$y_1(x) = y_1^0(x) + (y_1^1(x) - y_1^0(x))\eta, \quad y_2(x) = y_2^0(x) + (y_2^1(x) - y_2^0(x))\eta$$

являются линейными относительно  $\eta$ .

Очевидно, что в такой записи  $y_1^0$ ,  $y_2^0$  — решения исходной задачи при  $\eta = 0$ , а  $y_1^1$ ,  $y_2^1$  — при  $\eta = 1$ . Тогда получаем, что на правой границе должно быть выполнено условие

$$\beta_1(y_1^0(b) + (y_1^1(b) - y_1^0(b))\eta) + \beta_2(y_2^0(b) + (y_2^1(b) - y_2^0(b))\eta) = \gamma_2,$$

т. е.  $\eta$  является решением данного линейного уравнения.

Следовательно, нужно решить всего две задачи Коши для  $\eta = 0$  и  $\eta = 1$ , найти  $\eta$  из граничного условия и получить  $y_1$ ,  $y_2$ . Такова общая схема.

При конкретной реализации возникает вопрос, с какой границы ( $a$  или  $b$ ) нужно стартовать при стрельбе. Бывает, что краевая задача для исходной системы устойчива, а задача Коши — нет. Тогда решение задачи Коши будет иметь низкую точность. Необходимо выбрать и подходящий метод решения начальной задачи.

**Пример 6.1.** Рассмотрим в качестве примера следующую краевую задачу:

$$\begin{aligned} u'' - u &= 0, \quad 0 < x < 1; \\ u(0) &= 1, \quad u(1) = 2. \end{aligned}$$

Ее точное решение имеет вид

$$u(x) = \frac{2 \operatorname{sh} x - \operatorname{sh}(x-1)}{\operatorname{sh} 1}.$$

Таким образом, решение содержит две экспоненты: растущую и убывающую. Следовательно, задача Коши для данного уравнения будет являться неустойчивой вне зависимости от того, на каком конце отрезка  $[0, 1]$  ставятся так называемые начальные условия. В то же время численное решение данной задачи именно методами решения краевых задач, как мы увидим ниже, не представляет никакой сложности.

**Случай  $n \geq 3$ .** В такой постановке задачи присутствуют  $m$  условий в точке  $a$ , где  $1 \leq m \leq n-1$ , и  $n-m$  условий в точке  $b$ . При решении задачи методом стрельбы необходимо задать  $n-m$  параметров  $\eta_1, \eta_2, \dots, \eta_{n-m}$ , присвоив их, например, величинам  $y_{m+1}(a), y_{m+2}(a), \dots, y_n(a)$ . Тогда величины  $y_1(a), y_2(a), \dots, y_m(a)$  выбираются так, чтобы удовлетворить  $m$  уравнений

$$\varphi_k(y_1(a), y_2(a), \dots, y_n(a)) = 0, \quad k = 1, 2, \dots, m.$$

В результате получаем обычную задачу Коши. Решив ее, вычисляем  $n - m$  значений

$$\varphi_k(y_1(b), y_2(b), \dots, y_n(b)) = \psi_k(\eta_1, \eta_2, \dots, \eta_{n-m}), \quad k = m+1, m+2, \dots, n,$$

отличных, вообще говоря, от нуля. Нужно подобрать параметры  $\eta_1, \eta_2, \dots, \eta_{n-m}$  так, чтобы все  $\psi_k = 0, k = m+1, m+2, \dots, n$ .

В нелинейном случае, когда вид функций  $\psi_k$  неизвестен, подбор параметров  $\eta_1, \eta_2, \dots, \eta_{n-m}$  очень сложен. Поэтому в нелинейном случае метод стрельбы при  $n - m > 1$  практически не используется.

В случае линейной задачи изложенный выше метод для случая системы двух уравнений ( $m = 2$ ) легко обобщается на системы большего числа уравнений ( $n \geq 3$ ). При этом необходимо решить  $n - m + 1$  раз задачу Коши и еще одну задачу для системы линейных алгебраических уравнений (СЛАУ) для определения  $\eta_1, \eta_2, \dots, \eta_{n-m}$ .

**Замечание 6.1.** В настоящее время методы стрельбы сравнительно редко применяются для нахождения решений собственно краевых задач. Однако они широко встречаются при решении задач на собственные значения, в которых требуется найти параметр, при котором решение однородной задачи является нетривиальным. Здесь мы их не рассматриваем.

## 6.2. Разностные методы

Рассмотрим конечно-разностные методы решения краевой задачи. При этом на  $(a, b)$  введем сетку с шагом (пока для простоты постоянным)  $h$ :

$$\omega_h = \{x_i = a + ih, i = 0, 1, \dots, N\}, \quad h = \frac{b-a}{N}.$$

Вместо непрерывной функции (или вектор-функции) будем искать приближенное решение лишь в точках сетки, т. е. сеточную функцию. Производные в исходной задаче заменим разностными соотношениями.

Если внутри  $(a, b)$  для определения решения в новой точке записать уравнения методов Рунге — Кутты или линейных многошаговых методов вплоть до правой граничной точки, то получится система линейных уравнений или нелинейных уравнений для определения сеточного решения.

Матрица полученной системы уже не является треугольной.

Подробно конечно-разностные методы будут изучаться в последующих главах. Сейчас мы ограничимся рассмотрением простейшего примера.

### 6.2.1. Линейная краевая задача второго порядка

Рассмотрим следующую задачу:

$$\begin{aligned} u'' - p(x)u &= f(x), \quad a < x < b; \\ u(a) &= \alpha, \quad u(b) = \beta. \end{aligned}$$

Введем сетку и заменим вторую производную разностным соотношением. Тогда получим систему линейных алгебраических уравнений (СЛАУ) с трехдиагональной матрицей:

$$\begin{aligned} \frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}) - p(x_i)y_i &= f(x_i), \quad i = 1, 2, \dots, N-1; \\ y_0 = y(x_0) &= \alpha, \quad y_N = y(x_N) = \beta. \end{aligned}$$

Ее можно решать методом прогонки. Если  $p(x) > 0$ , то алгоритм прогонки устойчив и легко реализуется. А именно, если  $p(x) > 0$ , то в системе имеет место строгое диагональное преобладание:

$$\frac{2}{h^2} + p(x_i) > \frac{1}{h^2} + \frac{1}{h^2}.$$

**Теорема 6.1.** Если  $p, f \in C^2(a, b)$ ,  $p(x) > 0$ ,  $x \in [a, b]$ , то разностное решение равномерно сходится к точному со скоростью  $O(h^2)$ .

◀ Будем пользоваться теми же терминами, что и при исследовании методов для задачи Коши. Тогда для  $z_i = y_i - u_i$  имеем задачу

$$\begin{aligned} \frac{1}{h^2}(z_{i+1} - 2z_i + z_{i-1}) - p_i z_i &= \\ &= f(x_i) - \frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1}) + p_i u_i = \psi_h^{(1)} + \psi_h^{(2)}, \end{aligned}$$

где невязка на точном решении задачи имеет вид

$$\begin{aligned} \psi_h^{(1)} &= f(x_i) - \frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1}) + p_i u_i = \\ &= u_i'' - \frac{1}{h^2} \left( u_i'' h^2 + \frac{1}{12} u_i^{(4)} h^4 \right) = -\frac{1}{12} u_i^{(4)} h^2, \end{aligned}$$

а  $\psi_h^{(2)} = 0$ .

Существование и ограниченность  $u^{(4)}$  гарантируются условиями теоремы ( $f, p \in C^{(2)}$ ). Имеем  $|u^{(4)}| \leq M_4$ ,  $z_0 = z_N = 0$ . Тогда  $z_i$  в какой-то внутренней точке достигает своего максимума (по модулю). Пусть это точка  $i_0$ . Для этой точки

$$(2 + p_{i_0} h^2) z_{i_0} = z_{i_0+1} + z_{i_0-1} - h^2 \psi_h^{(1)},$$

откуда

$$|z_{i_0}| \leq \frac{h^2}{12} \frac{u^{(4)}}{p_{i_0}},$$

и

$$\|z\|_C \leq \frac{h^2}{12} \max_{x, x^* \in (a, b)} \left| \frac{u^{(4)}(x)}{p(x^*)} \right|,$$

т. е.  $\|z\|_C = O(h^2)$ . ▶

**Специализированная схема.** Вернемся к только что рассмотренной задаче и разберем на ее примере алгоритм построения специализированных схем, позволяющих точно передать типичное для данной задачи точное решение. А именно, потребуем от конструируемой разностной схемы, чтобы она была точна на решениях исходной задачи в случае  $p = p(x_i) = \text{const}$ ,  $f = f(x_i) = \text{const}$ . Ограничимся, как и ранее, равномерной сеткой.

В рассматриваемом варианте точное решение записывается в виде

$$u(x) = -\frac{f(x_i)}{p(x_i)} + A \exp((p(x_i))^{1/2} x) + B \exp(-(p(x_i))^{1/2} x).$$

Зададим разностную схему

$$\frac{1}{h^2} (y_{i+1} - 2y_i + y_{i-1}) - a_i y_i = \frac{a_i}{p(x_i)} f(x_i), \quad i = 1, 2, \dots, N-1,$$

и подберем параметр  $a_i$  так, чтобы записанное выше точное решение исходной дифференциальной задачи являлось и точным решением выписанной системы алгебраических уравнений. Легко видеть, что указанному требованию удовлетворяет параметр

$$a_i = \frac{4}{h^2} (\operatorname{sh}(0,5(p(x_i))^{1/2} h))^2.$$

По построению схема является точной на решениях данного вида. Такие схемы часто называют точными разностными схемами.

Конкретная схема, построенная для данной задачи, иногда называется *схемой с экспоненциальной подгонкой*. Особенно успешно по сравнению с традиционными такие схемы применяются в случае больших значений параметра  $p$ , т. е. для решения задач с малым параметром при старшей производной. Такие задачи обладают неприятным для численного решения свойством: размеры областей резкого изменения решения в них намного меньше размера рассматриваемого пространственного (или временного) участка. По этой причине их также иногда называют жесткими краевыми задачами.

Отметим, что при малом  $p$ , таком, что  $|0,5(p(x_i))^{1/2} h| \ll 1$ , получаем  $a_i \approx p(x_i)$ , т. е. традиционную схему.

**6.2.2. Нелинейные задачи**

Опять ограничимся рассмотрением простейшей задачи

$$\begin{aligned} u'' &= f(x, u), \quad x \in (a, b); \\ u(a) &= \alpha, \quad u(b) = \beta. \end{aligned}$$

Пусть  $f'_u \geq m > 0$ .

Для решения задачи выберем разностную схему того же вида, что и в линейном случае:

$$\begin{aligned} \frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}) &= f(x_i, y_i), \quad i = 1, 2, \dots, N-1; \\ y_0 &= \alpha, \quad y_N = \beta. \end{aligned}$$

Докажем сходимость приближенного решения к точному в предположении их существования и выполнения условий

$$|u^{(4)}| \leq M_4, \quad f'_u \geq m > 0.$$

Аналогично линейному случаю получим

$$\begin{aligned} \psi_h^{(1)} &= f(x_i, u_i) - \frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1}) = -\frac{1}{12}h^2u_i^{(4)}, \\ \psi_h^{(2)} &= f(x_i, u_i + z_i) - f(x_i, u_i) = f'_u(x_i, u_i + \vartheta z_i)z_i. \end{aligned}$$

Уравнения для погрешности  $z_i$  имеют вид

$$\begin{aligned} \frac{1}{h^2}(z_{i+1} - 2z_i + z_{i-1}) - f'_u(x_i, u_i + \vartheta z_i)z_i &= \psi_h^{(1)}, \\ z_0 &= z_N = 0. \end{aligned}$$

В результате точно так же, как и в линейном случае, получаем оценку

$$\|z\|_C \leq \frac{h^2}{12} \frac{M_4}{m},$$

т. е.  $\|z\|_C = O(h^2)$ .

Для нахождения решения нелинейной системы алгебраических уравнений необходимо использовать какой-либо итерационный метод. Наиболее употребительным является метод Ньютона.

Рассмотрим его подробнее. Пусть  $y_i^s$  — приближенное решение на  $s$ -й итерации. Запишем результат линеаризации правой части:

$$f(x_i, y_i^{s+1}) \cong f(x_i, y_i^s) + f'_u(x_i, y_i^s)(y_i^{s+1} - y_i^s).$$

Пусть  $\delta y_i^s = y_i - y_i^s$ . Тогда из уравнений итерационного метода

$$\frac{1}{h^2} (y_{i+1}^{s+1} - 2y_i^{s+1} + y_{i-1}^{s+1}) = f(x_i, y_i^s) + f'_u(x_i, y_i^s)(y_i^{s+1} - y_i^s)$$

и исходных нелинейных уравнений после вычитания получим

$$\begin{aligned} \frac{1}{h^2} (\delta y_{i+1}^{s+1} - 2\delta y_i^{s+1} + \delta y_{i-1}^{s+1}) - f'_u(x_i, y_i^s)\delta y_i^{s+1} = \\ = f(x_i, y_i^s + \delta y_i^s) - f(x_i, y_i^s) - f'_u(x_i, y_i^s)\delta y_i^s = \frac{1}{2} f''_{uu}(x_i, y_i^s + \vartheta \delta y_i^s)(\delta y_i^s)^2. \end{aligned}$$

Следовательно, точно таким же образом получим

$$\|\delta y^{s+1}\|_C \leq \frac{1}{2} \left\| \frac{f''_{uu}}{f'_u} \right\|_C \|\delta y^s\|_C^2.$$

В результате при выборе начального приближения недалеко от корня итерации метода Ньютона будут сходиться к решению, и притом с квадратичной скоростью. Если такие итерации сходятся, то в силу непрерывности функции  $f(x, u)$  они сходятся к точному решению исходной системы нелинейных алгебраических уравнений. Критерий прекращения итераций может быть выбран в форме, подходящей для данного случая.

Отметим, что наличие известного порядка сходимости делает возможным применение правила Рунге уточнения решения. Иногда этот прием может существенно сократить трудозатраты для получения решения нужного качества.

### 6.3. Методы Ритца и Галеркина

С данными методами мы лишь познакомимся и притом на простейшем примере. Пусть требуется найти решение  $u \in U$  следующей задачи:

$$Au = f, \quad a < x < b;$$

$$u(a) = \alpha, \quad u(b) = \beta.$$

Оператор  $A$  задан,  $f \in F$ ,  $D(A) \subset U$ ,  $\text{im } A \subset F$ .

Будем искать приближенное решение  $u$  в виде

$$u \approx y_h(x) = \varphi_0(x) + \sum_{i=1}^n c_i \varphi_i(x),$$

где  $\varphi_0(x)$  — некоторая гладкая функция, удовлетворяющая граничным условиям, т. е.  $\varphi_0(a) = \alpha$ ,  $\varphi_0(b) = \beta$ ;  $\{\varphi_i\}$  — выбранная система линейно-независимых функций, полная в пространстве  $U$ , причем  $\varphi_i \in D(A)$ ,  $i = 1, 2, \dots$ , и все функции  $\varphi_i$ ,  $i = 1, 2, \dots$ , обращаются в нуль в точках  $a$  и  $b$ . Найдем условия, определяющие коэффициенты  $c_i$ .

### 6.3.1. Метод Ритца

Рассмотрим функционал

$$\Phi[u] = \int_a^b (Au - f)^2 \rho dx,$$

где  $\rho > 0$  — вес. Очевидно, что решение нашего уравнения  $Au = f$  обеспечивает абсолютный минимум этому функционалу. Вместе с тем абсолютный минимум функционала  $\Phi$ , равный нулю, заведомо дает решение исходного уравнения, так как этот минимум соответствует функциям  $u$ , таким, что  $Au = f$ . При этом мы предполагаем существование точки абсолютного минимума (на функциях, удовлетворяющих заданным граничным условиям), так как функционал ограничен снизу и непрерывно зависит от  $Au$ .

Метод сведения задачи  $Au = f$  к задаче минимизации  $\Phi[u]$  обычно называют **методом наименьших квадратов**.

В случае линейного самосопряженного положительного оператора  $A$ , т. е. при при  $A = A^*$ ,  $A > 0$ , можно указать и другой функционал, минимизация которого дает решение исходной задачи:

$$\Phi[u] = (u, Au) - 2(u, f).$$

Пусть  $u = \bar{u} + \lambda \delta u$ . Тогда

$$\Phi[u] = \Phi[\bar{u}] + \lambda^2(\delta u, A\delta u) + 2\lambda(A\bar{u} - f, \delta u).$$

Если  $\bar{u}$  таково, что  $A\bar{u} = f$ , то  $\Phi[u] \geq \Phi[\bar{u}]$  для любых  $\lambda$ ,  $\delta u$  в силу положительности оператора  $A$ , т. е.  $\bar{u}$  реализует минимум  $\Phi[u]$ . В то же время из условия достижения минимума при  $u = \bar{u}$  имеем необходимое условие

$$\left. \frac{\partial \Phi}{\partial \lambda} \right|_{\lambda=0} = 0,$$

откуда следует, что для произвольного  $\delta u$  выполняется равенство  $(A\bar{u} - f, \delta u) = 0$ , в том числе и для  $\delta u = A\bar{u} - f$ . Следовательно,  $A\bar{u} = f$ .

Таким образом, задача отыскания минимума этого функционала также эквивалентна поиску решения задачи  $Au = f$ .

Пусть

$$\Phi[u] = (u, Au) - 2(u, f).$$

Будем искать решение  $u$  в виде

$$u \approx y_h(x) = \varphi_0(x) + \sum_{i=1}^n c_i \varphi_i(x).$$

Тогда .

$$\begin{aligned}\Phi[y_h] = & \sum_{i=1}^n \sum_{j=1}^n c_i c_j (\varphi_i, A\varphi_j) + \\ & + 2 \sum_{k=1}^n c_k ((\varphi_k, A\varphi_0) - (\varphi_k, f)) + (\varphi_0, A\varphi_0 - 2f) \rightarrow \min.\end{aligned}$$

Ищем коэффициенты  $c_i$ , при которых  $\Phi$  на функциях данного класса достигает минимума. Условия экстремума  $\Phi$  имеют вид

$$\sum_{i=1}^n c_i (\varphi_i, A\varphi_j) = -(\varphi_j, A\varphi_0 - f), \quad j = 1, 2, \dots, n.$$

Решив полученную систему линейных алгебраических уравнений (СЛАУ), находим приближенное решение  $y_h$ .

Описанный выше метод приближенного решения называют **методом Ритца**.

### 6.3.2. Метод Галеркина

Ищем решение в виде

$$u \approx y_h(x) = \varphi_0(x) + \sum_{i=1}^n c_i \varphi_i(x),$$

где  $\{\varphi_i(x)\}$ ,  $i = 1, 2, \dots$ , — некоторая полная система функций. Тогда, если

$$(F, \varphi_i) = 0, \quad i = 1, 2, \dots,$$

то  $F \equiv 0$  в соответствии с определением полноты системы функций.

Тем самым, если найти функцию  $u$ , такую, что

$$(Au - f, \varphi_i) = 0, \quad i = 1, 2, \dots,$$

то это означало бы, что  $u$  — решение нашей исходной задачи. Если же такая ортогональность есть лишь при  $i \leq n$ , то, очевидно, полученная функция  $u$  есть приближенное решение исходной задачи с точностью до функций  $\varphi_{n+1}, \varphi_{n+2}, \dots$ . Точность определяется выбором системы функций и размерностью  $n$ .

Возьмем приближенное решение в указанной форме и потребуем, чтобы

$$(Ay_h - f, \varphi_i) = 0, \quad i = 1, 2, \dots, n.$$

Это есть система  $n$  уравнений для  $n$  коэффициентов:

$$\sum_{i=1}^n c_i(A\varphi_i, \varphi_j) = -(\varphi_j, A\varphi_0 - f), \quad j = 1, 2, \dots, n.$$

Таким образом, получаем те же уравнения, что и в случае метода Ритца. Описанный здесь метод называется **методом Галеркина**.

Вопрос о сходимости метода мы не рассматриваем.

### 6.3.3. Выбор системы функций

Качество получаемого приближенного решения в методах Ритца и Галеркина в значительной степени зависит от выбранной системы функций (при заданном  $n$ ). Так, обычно

$$\varphi_0 = \alpha + \frac{\beta - \alpha}{b - a}(x - a),$$

а  $\varphi_i$  могут быть самыми разнообразными.

Если  $\varphi_i(x)$  — базисные функции метода конечных элементов, построенных в соответствии с некоторой сеткой, то будет получен вариант *метода конечных элементов*. По форме полученная СЛАУ для коэффициентов очень будет напоминать обычную разностную схему. Если на сетке задана система тригонометрических функций, то будут получены СЛАУ так называемого спектрального метода.

Вообще методы, основанные на проектировании исходного уравнения на некоторые последовательности подпространств  $U_n \subset D(A)$  и  $F_n \subset F$  и поиске решений  $y_n \in U_n$ , называются проекционными. Часто при этом решение ищется в виде

$$y_h = \varphi_0 + \sum_{i=1}^n c_i \varphi_i.$$

Если базисные функции заданы с помощью какой-либо сетки, то такие методы называют проекционно-сеточными. Отметим, что при этом находится функция  $y_h$ , определенная при произвольном  $x$ , а не только в узлах сетки, так как функции  $\varphi_i$  заданы на всем участке  $(a, b)$ .

## 6.4. Библиографические комментарии

Материал данной главы является классическим.

Многие подробности рассматриваемых алгоритмов и другие способы решения краевых задач можно найти в [6, 12, 14, 17, 18, 25, 53, 61, 62, 80, 83, 84, 90, 103] и других руководствах по численным методам.

Апробированные и новые подходы к решению жестких задач представлены в [80, 103, 176]. Описание и анализ схем с экспоненциальной подгонкой можно найти в [12, 14, 176].

Особо отметим двухтомник [184, 185], являющийся на сегодня, пожалуй, самым полным изложением методов численного решения ОДУ.

Укажем, что существуют методы решения уравнений в частных производных, сводящие процедуру решения исходной задачи к поиску решения некоторой краевой задачи. В [95, 96], они называются поперечными методами прямых. При этом в случае, например, эволюционных задач происходит дискретизация по времени. В результате остается некоторая краевая задача для нахождения решения на данном временном слое. При наличии хороших пакетов программ для решения больших жестких систем ОДУ такой алгоритм может оказаться весьма эффективным.

Описание вариационных и проекционных методов, только упомянутых выше, можно найти в [61, 62, 110, 112, 117] и многих других руководствах.

Алгоритмы численного решения экстремальных задач, которые могут быть использованы при реализации вариационных методов, приведены в [24].

## 7. ЭЛЕМЕНТЫ ТЕОРИИ РАЗНОСТНЫХ СХЕМ

Введены основные понятия теории разностных схем: *сетка, сеточная функция, разностная схема*. Получены полезные разностные соотношения, необходимые для проведения теоретического исследования схем. Продемонстрированы методы и приемы построения разностных схем. Введены и подробно обсуждены фундаментальные понятия — *аппроксимация, устойчивость и сходимость*. Установлена связь между ними. Решена задача Штурма — Лиувилля в дискретном случае для простейшего разностного оператора второго порядка. Установлен *принцип максимума* для разностных схем. Приведены различные признаки устойчивости разностных схем.

### 7.1. Постановка задачи и основные понятия

#### 7.1.1. Постановка задачи

Ранее было рассмотрено решение систем линейных алгебраических уравнений (СЛАУ) и обыкновенных дифференциальных уравнений (ОДУ). Главным содержанием данной и последующих глав является решение уравнений в частных производных и частично интегральных уравнений. Основным же методом их решения будет конечно-разностный, в котором производные по некоторым правилам заменяются на свои конечно-разностные аппроксимации.

Напомним основные типы уравнений в частных производных для случая двух (для определенности) переменных:

$$A \frac{\partial^2 u}{\partial x^2} + 2B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D \frac{\partial u}{\partial x} + E \frac{\partial u}{\partial y} + Fu + P = 0.$$

В дальнейшем для сокращения записи мы будем использовать обозначения производных без знака дифференцирования там, где это не вызывает неверного истолкования. В частности, рассматриваемые типы уравнений в такой форме имеют вид

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + P = 0.$$

Действительные коэффициенты  $A, B, C, D, E, F, P$  могут зависеть, вообще говоря, от  $x, y, u$  и производных  $u$  по  $x, y$ . Если они зависят от  $u$  и производных  $u$ , то уравнение является нелинейным. Если

не зависят, то уравнение линейно по  $u$ . Если  $A, B, C, D, E, F$  не зависят и от  $x, y$ , то записанное уравнение является линейным уравнением с постоянными коэффициентами. Если  $A, B, C, D, E, F, P$  зависят от  $u$  и не зависят от  $u_x, u_y, \dots$ , то обычно такое уравнение называется квазилинейным. Зависимость  $A, B, C, D, E, F, P$  от  $u_x, u_y, \dots$  повышает «степень» нелинейности. В основном мы будем рассматривать линейные уравнения с постоянными коэффициентами.

Пусть  $\Delta = B^2 - AC$ .

**Определение 7.1.** Если  $A^2 + B^2 + C^2 = 0$ , но  $E \neq 0, F \neq 0$ , то уравнение имеет первый порядок и называется уравнением переноса. Если  $\Delta > 0$ , то рассматриваемое уравнение имеет гиперболический тип, если  $\Delta = 0$  — то параболический (при этом  $A^2 + B^2 + C^2 \neq 0$ ), если  $\Delta < 0$  — то эллиптический.

К перечисленным типам относятся следующие уравнения:

1) к гиперболическому типу — **уравнение колебаний**

$$u_{xx} - a^2 u_{yy} = f$$

и **уравнение переноса**

$$u_x + cu_y = f;$$

2) к параболическому типу — **уравнение теплопроводности**

$$u_x - a^2 u_{yy} = f;$$

3) к эллиптическому типу — **уравнение Лапласа**

$$\Delta u = u_{xx} + u_{yy} = 0$$

и **уравнение Пуассона**

$$\Delta u = -f.$$

Разделение уравнений на типы проведено в связи с существенным различием свойств решений задач для уравнений разных типов. Поэтому далее мы будем изучать способы их численного решения по отдельности и по мере необходимости напоминать свойства решений рассматриваемых уравнений.

Уравнения необходимо решать в некоторой области  $G$  изменения независимых переменных. Для нахождения единственного решения требуется специальным образом сформулировать задачи, в постановку которых входят некоторые заданные дополнительные условия.

Отметим, что уравнение переноса является простейшим уравнением в частных производных вообще. Оно является частным случаем не только указанного выше уравнения 2-го порядка, но и системы  $n$  уравнений 1-го порядка вида  $u_x + Au_y = f$ , где  $A$  — квадратная матрица порядка  $n$ . Такая система является гиперболической, если у матрицы  $A$

все собственные значения действительны и существует базис из  $n$  левых собственных векторов. В указанной форме может быть записано уравнение колебаний.

В одном из самых распространенных случаев  $G = G(\vec{r}, t) = \Omega(\vec{r}) \times [0, T]$ , т. е.  $G$  является цилиндрической областью в области изменения пространственных переменных  $\vec{r} = (x, y)$  (в двумерном случае) и временной переменной  $t$ . Дополнительные условия, заданные на границе  $t = 0$ , называют начальными, а заданные на границе  $\Omega$ , или на боковой стороне цилиндра  $G$  при произвольных  $t$  — краевыми, или граничными. В данном случае мы рассматриваем задачи, в которых одна из переменных —  $t$  — является выделенной и называется временем. Если область  $\Omega$  не ограничена, то поставленная задача называется **задачей Коши**. Если в задаче есть время  $t$  и ограниченная область  $\Omega$ , то поставленную задачу чаще всего называют **начально-краевой**, или **смешанной**. Если же времени  $t$  среди независимых переменных нет, то решают **краевую задачу**.

Далее при исследовании общих конструкций мы будем рассматривать задачу о нахождении решения уравнения  $Au = f$  в области  $G$  с дополнительными условиями  $Ru = \mu$  на границе  $\Gamma = \partial G$  области  $G$ :

$$Au = f \text{ в } G, \quad Ru = \mu \text{ на } \Gamma.$$

Здесь  $A, R$  — заданные операторы;  $f, \mu$  — заданные функции.

Считаем задачу корректно поставленной по Адамару.

**Определение 7.2.** Задача называется **корректно поставленной**, если ее решение существует, единственно и непрерывно зависит от входных данных. Если же не выполнено хотя бы одно из этих условий, то задача называется **некорректно поставленной**.

Примеры различных конструкций или построений в данной главе даются либо для уравнения теплопроводности, либо для обыкновенного дифференциального уравнения второго порядка.

### 7.1.2. Сетка и сеточные функции

Рассмотрим задачу нахождения приближенных численных значений решения хотя бы в некоторых точках. Она не всегда может быть решена до конца лишь аналитическими методами даже для линейных уравнений с постоянными коэффициентами, поэтому решение ищем численно. А так как решение исходной задачи — элемент бесконечномерного пространства, и его численно найти невозможно (за исключением тривиальных случаев), мы заменяем бесконечномерное пространство конечномерным.

**Определение 7.3.** Дискретное множество точек, «заменяющих» область  $G$  изменения независимых переменных с указанием связей между ними, называется *сеткой*. При этом точки сетки обычно называют *узлами*, а расстояния между ними вдоль связей — *шагами*.

Далее, если не указано противное, *сетку* будем считать *равномерной*, т. е. такой, в которой шаги по каждой из переменных являются постоянными (рис. 7.1).

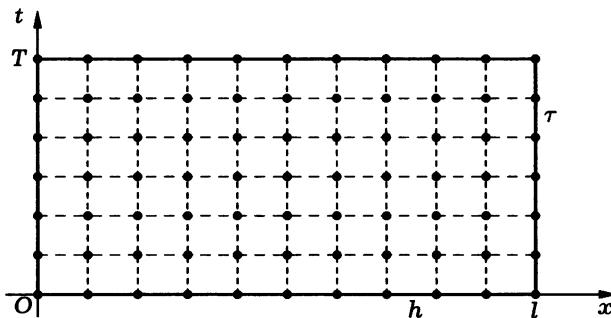


Рис. 7.1

**Определение 7.4.** *Функция*, определенная только в узлах сетки, называется *сеточной*.

Далее, если не будет противоречий с обозначениями независимых переменных,  $y$  будет обозначать сеточную функцию, даже без указания индексов. При этом будем обозначать знаком  $y_i^j$  значение сеточной функции  $y$  в узле сетки с координатами  $(x_i, t_j)$ , т. е.  $y_i^j = y(x_i, t_j)$  в случае области  $G = [0, l] \times [0, T]$ .

**Определение 7.5.** Алгебраические уравнения для сеточных функций, «заменяющие» уравнения и дополнительные условия исходной задачи и полученные путем замены производных в исходных уравнениях конечными разностями, интегралов — квадратурными формулами, прочих членов — алгебраическими соотношениями, называются *разностной схемой*.

В дальнейшем через  $h$  будем обозначать шаг (может быть, с индексом) сетки по пространственным переменным, через  $\tau$  — шаг сетки по времени, а индексом  $h$  помечать разностные аппроксимации рассматриваемых операторов либо сеточные функции:

$$A_h y = \varphi \text{ в } G_h, \quad R_h y = \nu \text{ на } \Gamma_h,$$

где  $y, \varphi, \nu$  — сеточные функции;  $G_h, \Gamma_h$  — сетки.

**Пример 7.1.** Рассмотрим начально-краевую задачу для уравнения теплопроводности:

$$\begin{aligned} u_t - ku_{xx} &= 0, \quad 0 < x < l, \quad t > 0; \\ u(0, t) &= u(l, t) = 0, \quad u(x, 0) = u_0(x). \end{aligned}$$

Для численного решения введем сетку

$$\bar{\omega}_h = \{x_i = ih, i = 0, 1, \dots, n\}, \quad hn = l,$$

по переменной  $x$  и сетку

$$\bar{\omega}_\tau = \{t_j = j\tau, j = 0, 1, \dots, M\}, \quad M\tau = T,$$

по переменной  $t$  (см. рис. 7.1).

Рассмотрим узел с координатами  $(x_i, t_j)$  и заменим в дифференциальном уравнении производные разностными соотношениями на выбранной совокупности узлов (рис. 7.2). Такая совокупность называется 6-точечным шаблоном.

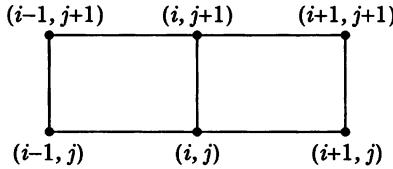


Рис. 7.2

Существуют разные варианты. Пусть, например, для аппроксимации обычной временной производной  $u_t$  выбрана разностная производная

$$\frac{y_i^{j+1} - y_i^j}{\tau},$$

а для аппроксимации второй пространственной производной  $u_{xx}$  принята разностная аппроксимация

$$\frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2}.$$

Обозначения для разностных производных будут приведены ниже. Запишем с их помощью разностную схему

$$y_{t,i}^j - ky_{xx,i}^j = 0,$$

в которой используются узлы сетки  $(x_i, t_j), (x_{i-1}, t_j), (x_{i+1}, t_j), (x_i, t_{j+1})$  (рис. 7.3). Такая схема в дальнейшем будет называться явной схемой на 4-точечном шаблоне.

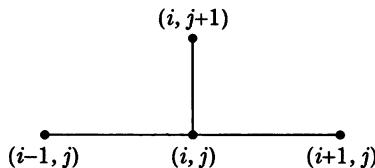


Рис. 7.3

В этом случае значение  $y_i^{j+1}$  может быть вычислено непосредственно по данным о сеточном решении с предыдущего временного слоя  $t_j$ , а в конечном итоге — по начальным и граничным данным, последовательно от слоя к слою.

Если же для аппроксимации второй производной  $u_{xx}$  выбрать аппроксимацию

$$\frac{y_{i+1}^{j+1} - 2y_i^{j+1} + y_{i-1}^{j+1}}{h^2},$$

то в записи сеточного уравнения будут использоваться точки  $(x_i, t_{j+1})$ ,  $(x_{i-1}, t_{j+1})$ ,  $(x_{i+1}, t_{j+1})$ ,  $(x_i, t_j)$  (рис. 7.4). Такая схема в дальнейшем будет называться неявной схемой на 4-точечном шаблоне.

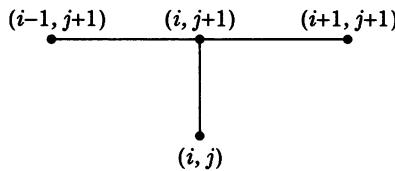


Рис. 7.4

В этом случае для нахождения решения на новом слое  $t_{j+1}$  уже необходимо решить систему алгебраических уравнений с трехдиагональной матрицей.

**Определение 7.6.** Совокупность узлов сетки, используемых для записи разностной схемы, называют **шаблоном** разностной схемы (см. рис. 7.3, 7.4).

Шаблоны, используемые в примере 7.1, называют 4-точечными.

Очевидно, что шаблон разностной схемы зависит от рассматриваемого узла сетки. Тем не менее довольно часто слово «шаблон» применяют только к большей части внутренних узлов сетки или ко всем узлам. Поэтому используется также следующее определение.

**Определение 7.7.** Узлы, в которых разностная схема записана на шаблоне, называются **регулярными**, остальные — **нерегулярными**.

В частности, нерегулярными являются граничные узлы, попадающие на границу  $\Gamma$  рассматриваемой области. Иногда такими оказываются узлы сетки, прилегающие к границе области, но не попадающие на нее. Такая ситуация легко возникает при использовании прямоугольной сетки в непрямоугольной области (рис. 7.5). Очевидно, что если область  $G$  не является прямоугольником (параллелепипедом), то попасть приграничными узлами на границу в случае использования прямоугольной сетки практически невозможно. Тогда возникают проблемы с аппроксимацией граничных условий даже первого рода и приходится использовать, например, треугольную сетку либо аппроксимировать граничные условия каким-то специальным образом.

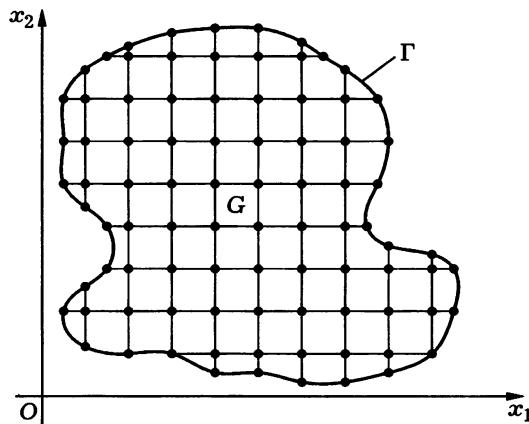


Рис. 7.5

Пусть решается эволюционная задача в области  $G_h = \Omega_h \times \omega_\tau$  (индекс  $\tau$  у  $G$  для краткости писать не будем). Рассмотрим временной слой  $j+1$ , на котором сеточное решение еще неизвестно, в то время как на слоях с номерами  $j, j-1, \dots, 0$  оно уже получено.

**Определение 7.8.** Если разностная схема относительно любого узла сетки содержит лишь одно значение неизвестной сеточной функции на новом слое, то схема называется **явной**, если более одного, то **неявной**. Иногда схему, в которой используется лишь одно значение со старого слоя, называют **полностью неявной**.

Для явной схемы можно найти решение на  $(j+1)$ -м слое непосредственно без решения системы алгебраических уравнений (или, точнее, решением системы с диагональной матрицей). Для неявной схемы систему алгебраических уравнений с недиагональной матрицей приходится решать.

Пусть точное решение исходной (дифференциальной) задачи — элемент пространства  $U$ . Очевидно, что сеточная функция ему не принадлежит и является элементом пространства  $Y$  сеточных функций. Пусть на пространстве  $U$  введена норма  $\|\cdot\|_U$ , позволяющая оценивать близость функций, а на пространстве  $Y$  — норма  $\|\cdot\|_Y$ . Главная цель численного решения — найти приближение точного решения исходной задачи. Но решение исходной задачи и решение сеточной задачи являются элементами разных пространств. Чтобы получить возможность сравнить такие решения, необходимо рассматриваемые функции свести в одно пространство. Так как пространство  $Y$  намного «беднее», то проще всего решению и исходной задачи поставить в соответствие элемент пространства  $Y$ , а не наоборот.

**Определение 7.9.** Линейный оператор  $p_h$ , ставящий в соответствие каждой функции  $u \in U$  некоторую сеточную функцию  $u_h \in Y$ , называется **оператором проектирования** на сетку:  $u_h = p_h u$ .

**Пример 7.2.** 1. Рассмотрим пространство непрерывных функций  $U = C[a, b]$ . Оператор проектирования  $p_h$  на нем можно определить следующим образом:  $p_h u(x_i) = u(x_i) = u_i$ .

2. Рассмотрим пространство интегрируемых функций  $U = L_1[a, b]$ . Оператор проектирования можно ввести так, что во внутренних узлах отрезка значения сеточной функции определяются равенством

$$p_h u(x_i) = u_i = \frac{1}{h} \int_{x_i-h/2}^{x_i+h/2} u(\xi) d\xi,$$

а в граничных узлах — равенствами

$$p_h u(x_0) = u_0 = \frac{2}{h} \int_a^{a+h/2} u(\xi) d\xi, \quad p_h u(x_n) = u_n = \frac{2}{h} \int_{b-h/2}^b u(\xi) d\xi.$$

3. Рассмотрим пространство непрерывных функций  $U = C[a, b]$  еще раз. Введем оператор проектирования, который во внутренних узлах отрезка  $[a, b]$  принимает значения

$$p_h u(x_i) = \frac{1}{2} \left( u\left(x_i + \frac{h}{2}\right) + u\left(x_i - \frac{h}{2}\right) \right).$$

Очевидно, что список вариантов задания оператора проектирования может быть продолжен. #

Далее будем считать, что мы имеем дело с проектором  $p_h$ , таким, что  $p_h u(x_i) = u(x_i)$ .

**Определение 7.10.** Скажем, что последовательность  $y$  сеточных функций сходится к  $u$  при  $h \rightarrow 0$ , если  $\|y - p_h u\|_Y \rightarrow 0$  при  $h \rightarrow 0$ .

Сформулированное определение корректно, если указанный предел единственный.

**Определение 7.11.** Будем требовать выполнения следующего условия: **нормы** в  $Y$  и в  $U$  должны быть **согласованными**, т. е. для любой функции  $u \in U$

$$\lim_{h \rightarrow 0} \|p_h u\|_Y = \|u\|_U.$$

**Предложение 7.1.** Если нормы в  $Y$  и  $U$  согласованы, то предел последовательности сеточных функций при  $h \rightarrow 0$  является единственным.

◀ Пусть  $y$  — последовательность сеточных функций,  $u$  и  $v$  — две функции, являющиеся пределами этой последовательности при  $h \rightarrow 0$ , т. е. выполняются условия

$$\lim_{h \rightarrow 0} \|p_h u - y\|_Y = \lim_{h \rightarrow 0} \|p_h v - y\|_Y = 0.$$

Из соотношений

$$\|p_h u - p_h v\|_Y = \|(p_h u - y) + (y - p_h v)\|_Y \leq \|p_h u - y\|_Y + \|y - p_h v\|_Y$$

заключаем, что

$$\|u - v\|_U = \lim_{h \rightarrow 0} \|p_h(u - v)\|_Y = \lim_{h \rightarrow 0} \|p_h u - p_h v\|_Y = 0.$$

Следовательно,  $u = v$ . ▶

**Замечание 7.1.** Условие согласованности для единственности предела последовательности сеточных функций существенно. Можно построить вариации сеточных норм с помощью умножения на  $h^p$ ,  $p > 0$ , в которых, например, последовательность  $y$ , которая на каждой сетке равна тождественно единице, будет сходитьсь в смысле определения 7.10 к произвольной функции  $u$ .

Далее мы, как правило, будем писать  $u_h$  или  $u$ , а не  $p_h u$ , в случаях, когда очевидно, что рассматривается не «непрерывная» функция, т. е. функция непрерывного аргумента, а ее проекция на сетку.

**Пример 7.3.** Рассмотрим некоторые варианты сеточных норм.

1. Норма  $\|y\|_C = \max_i |y_i|$  согласована с равномерной нормой (доказательство может быть проведено через сравнение множества ступенчатых функций  $\tilde{y}$ , равных  $y$  на соответствующем участке, с функцией  $u$  на основе использования равномерной непрерывности функций, непрерывных на отрезке).

2. Нормы  $\|[y]\|_2 = \sqrt{\sum_{i=0}^n hy_i^2}$ ,  $\|y\|_2 = \sqrt{\sum_{i=1}^n hy_i^2}$ ,  $\|[y]\|_2 = \sqrt{\sum_{i=0}^{n-1} hy_i^2}$  согласованы с  $\|u\|_{L_2}$  (доказательство основано на сходимости квадратурных формул для функций указанного класса, что совпадает с исходной конструкцией интеграла Лебега как предела интегралов от последовательности ступенчатых функций).

3. Норма  $\|y\|_2^* = \sqrt{\sum_{i=0}^n y_i^2}$  не согласована с  $\|u\|_{L_2}$ , так как при  $h \rightarrow 0$  ( $n \rightarrow \infty$ ) эта сумма может расходиться, хотя  $u \in L_2[a, b]$ . #

Вообще же выбор нормы определяется решаемой задачей и используемой разностной схемой.

Переход к пределу при  $h \rightarrow 0$  означает рассмотрение последовательности конечномерных линейных пространств, в некотором смысле аппроксимирующих бесконечномерное, и переход к этому бесконечномерному пространству. Как следствие, различные сеточные нормы становятся неэквивалентными в отличие от сеточных норм в задачах линейной алгебры, в которых рассматривается одно линейное пространство и размерность является фиксированной.

В конечномерном случае все нормы эквивалентны: из сходимости последовательности элементов в одной норме следует сходимость в другой. Но множители в неравенствах

$$m\|y\|_1 \leq \|y\|_2 \leq M\|y\|_1,$$

определяющих эквивалентность норм, зависят, вообще говоря, от выбора норм и размерности пространства  $n$ . Поэтому как только появляется последовательность конечномерных пространств, эквивалентность чаще всего исчезает. Выбор нормы оказывает очень существенное влияние на те характеристики схем, которые получаются с ее помощью.

**Пример 7.4. 1.** Нетрудно видеть, что для введенных выше норм справедливы неравенства

$$\sqrt{h}\|y\|_C \leq \|[y]\|_2 \leq \sqrt{l+h}\|y\|_C.$$

При этом каждое из неравенств является точным, т. е. существуют функции, на которых достигается равенство. Приведенная оценка представляет собой условие эквивалентности сеточных норм при фиксированном шаге сетки, но эквивалентность уже не имеет места в случае бесконечномерного пространства, элементами которого являются функции непрерывного аргумента (примеры можно найти в курсе математического анализа).

2. Пусть

$$y_{h,i} = \begin{cases} 1 + h^{-1/4}, & i = 0; \\ 1, & i \geq 1. \end{cases}$$

Легко видеть, что для непрерывной функции  $u \equiv 1$  имеем  $\|y - u\|_2 \rightarrow 0$  при  $h \rightarrow 0$ . В то же время  $\|y - u\|_C = \max |y_i - u_i| = h^{-1/4} \rightarrow +\infty$  при  $h \rightarrow 0$ .

## 7.2. Обозначения и некоторые разностные соотношения

Договоримся индексами внизу у сеточной функции  $y$  указывать ее значение, вычисленное в узле пространственной сетки с данными индексами, т. е.  $y_{ij}$  есть значение сеточной функции  $y$  в узле пространственной сетки с номером  $(i, j)$ . Индекс вверху будет относиться к временной сетке: обозначение  $y_i^k$  указывает на значение сеточной функции в  $i$ -м узле по пространству и  $k$ -м по времени.

**Определение 7.12.** Совокупность узлов сетки с временной координатой  $t = t_k$  называется **временным слоем**  $t_k$  или  $k$ -м временными слоями.

Далее мы, как правило, будем опускать индексы  $i, j$  и использовать так называемые безындексные обозначения. При этом мы ограничимся случаем задач с двумя переменными. Если не оговорено противное, при рассмотрении разностного уравнения будем считать, что написано оно с «центром» в сеточном узле с номером  $(i, j)$ .

Будем рассматривать сетку с постоянным шагом  $h$  на участке  $[0, l]$ :

$$\bar{\omega}_h = \{ih, i = 0, 1, \dots, n\}, \quad nh = l.$$

Тогда

$$y_i^j = y, \quad y_{i+1}^j = y_+, \quad y_{i-1}^j = y_-, \quad y_i^{j+1} = \hat{y}, \quad y_{i-1}^{j+1} = \hat{y}_-.$$

и т. д. Введем также обозначение  $\check{y} = y_i^{j-1}$ .

Напомним стандартные аппроксимации пространственных производных на равномерной сетке:

*правая разностная производная*, или производная «вперед»,

$$y_x = \frac{y_{+1} - y}{h} = y_{\bar{x},+1};$$

*левая разностная производная*, или производная «назад»,

$$y_{\bar{x}} = \frac{y - y_{-1}}{h} = y_{x,-1};$$

*центральная разностная производная*

$$y_x^{\circ} = \frac{y_{+1} - y_{-1}}{2h} = \frac{y_x + y_{\bar{x}}}{2};$$

аппроксимации временных производных на *равномерной сетке*

$$y_t = \frac{\hat{y} - y}{\tau}, \quad y_{\bar{t}} = \frac{y - \check{y}}{\tau}, \quad \hat{y}_{\bar{t}} = y_t;$$

*аппроксимации вторых производных*

$$\begin{aligned} y_{\bar{x}x} &= \frac{y_{+1} - 2y + y_{-1}}{h^2} = (y_{\bar{x}})_x = (y_x)_{\bar{x}}, \\ y_{\bar{t}t} &= \frac{\hat{y} - 2y + \check{y}}{\tau^2} = (y_{\bar{t}})_t = (y_t)_{\bar{t}}. \end{aligned}$$

Введем аппроксимации *скалярного произведения* сеточных функций  $y, z$  в различных видах:

$$\begin{aligned} (y, z) &= \sum_{i=1}^{n-1} hy_i z_i, \quad [y, z] = \sum_{i=0}^n hy_i z_i, \\ [y, z] &= \sum_{i=0}^{n-1} hy_i z_i, \quad (y, z) = \sum_{i=1}^n hy_i z_i. \end{aligned}$$

Все эти выражения являются квадратурными формулами для вычисления скалярного произведения

$$\int_0^l u(x)v(x) dx = (u, v)$$

в случае обычных квадратично-интегрируемых функций.

Запишем еще одно скалярное произведение:

$$\widetilde{[y, z]} = \frac{h}{2}(y_0 z_0 + y_n z_n) + (y, z),$$

в котором взята квадратурная формула трапеций для вычисления интеграла. Выбор той или иной формулы определяется решаемой задачей.

**Определение 7.13.** Величину  $z = y - u_h$ , т. е. сеточную функцию, равную разности приближенного решения и проекции точного решения на сетку, будем называть *погрешностью разностной схемы*.

Наша схема  $A_h y = \varphi$  «заменяет» (или аппроксимирует) исходную задачу  $Au = f$ . Слову «заменяет» можно придать количественный смысл, если доказать близость  $y$  к  $u$ , т. е. малость  $z$ . При этом должны быть близки как операторы  $A_h$  и  $A$ , так и правые части  $\varphi$  и  $f$ .

Запишем задачу для  $z$ , воспользовавшись соотношением  $y = z + u_h$ . Тогда

$$A_h z = \varphi + A_h z - A_h(z + u_h) = (\varphi - f_h) + (f_h + A_h z - A_h(z + u_h)) = \psi_h,$$

где

$$\psi_h = \psi_h^{(1)} + \psi_h^{(2)}, \quad \psi_h^{(1)} = (Au)_h + A_h z - A_h(z + u_h), \quad \psi_h^{(2)} = \varphi - f_h.$$

В случае линейного оператора  $A_h$  получаем, что для  $\psi_h^{(1)}$  справедливо следующее равенство:

$$\psi_h^{(1)} = (Au)_h - A_h u_h.$$

**Определение 7.14.** Определенная выше сеточная функция  $\psi_h$  называется *погрешностью аппроксимации разностной задачи*  $A_h y = \varphi$  на решении исходной задачи  $Au = f$ .

**Определение 7.15.** Сеточные функции  $\psi_h^{(1)}$  и  $\psi_h^{(2)}$  называются *погрешностью аппроксимации оператора*  $A$  разностным оператором  $A_h$  и *погрешностью аппроксимации правой части* соответственно.

Далеко не всегда есть возможность проверять малость погрешности аппроксимации на точном решении задачи. Во избежание этой трудности введем следующее понятие.

**Определение 7.16.** Назовем погрешностью аппроксимации разностной задачи  $A_h y = \varphi$  сеточную функцию  $\psi_h$ , определяемую следующим образом:

$$\psi_h = \psi_h^{(1)} + \psi_h^{(2)}, \quad \psi_h^{(1)} = (Av)_h - A_h v_h, \quad \psi_h^{(2)} = \varphi - f_h.$$

Определяемая таким общим способом погрешность аппроксимации является просто разностью невязок разностной и дифференциальной задач на функции  $v$ :  $\psi_h = (\varphi - A_h v_h) - (f_h - (Av)_h)$ .

**Замечание 7.2.** Данное определение не предполагает линейность разностного оператора, используемая в нем функция  $v$  — произвольная функция из области определения оператора  $A$ .

**Замечание 7.3.** Погрешность аппроксимации линейного разностного оператора  $A_h$  на точном решении исходной задачи задает правую часть разностного уравнения для погрешности разностной схемы.

Выше речь шла об уравнениях задачи внутри области. Аналогичные соотношения получаются и на ее границе.

**Формула интегрирования по частям.** Пусть  $y, z$  — сеточные функции. Тогда

$$\begin{aligned} (y, z_{\bar{x}}] &= \sum_{i=1}^n y_i z_{\bar{x}, i} h = \sum_{i=1}^n y_i (z_i - z_{i-1}) = \sum_{i=1}^n y_i z_i - \sum_{i=0}^{n-1} y_{i+1} z_i = \\ &= y_n z_n - y_0 z_0 - [y_x, z] = y_n z_n - y_1 z_0 - (y_x, z). \end{aligned}$$

Данные соотношения представляют собой разностные аналоги формулы интегрирования по частям.

Отсюда, в частности, следует

$$\sum_{i=1}^n y_i z_{\bar{x}, i} h + y_0 z_0 = y_n z_n - \sum_{i=0}^{n-1} y_{x, i} z_i h,$$

т. е. оператор левой разностной производной, определенный соотношением

$$A_h^l y = \begin{cases} \frac{y_0}{h}, & i = 0; \\ \frac{y_i - y_{i-1}}{h}, & 1 \leq i \leq n, \end{cases}$$

является сопряженным оператору правой разностной производной

$$A_h^r y = \begin{cases} \frac{y_{i+1} - y_i}{h}, & 0 \leq i \leq n-1; \\ -\frac{y_n}{h}, & i = n, \end{cases}$$

взятыму со знаком минус, относительно скалярного произведения  $[\cdot, \cdot]$ .

**Разностная формула Грина.** Возьмем тождество

$$(y_x, z) = y_n z_n - y_1 z_0 - (y, z_{\bar{x}}]$$

и подставим вместо  $u$  комбинацию  $az_{\bar{x}}$ . В результате получим первую разностную формулу Грина:

$$((az_{\bar{x}})_x, z) = (az_{\bar{x}})_n z_n - (az_{\bar{x}})_1 z_0 - [a, (z_{\bar{x}})^2].$$

**Неравенство типа вложения.** Пусть  $z_n = 0$ . Тогда

$$z_i = - \sum_{j=i+1}^n h z_{\bar{x},j}.$$

Отсюда с использованием неравенства Коши — Буняковского для сумм получим

$$|z_i|^2 \leq \left( \sum_{j=i+1}^n h \sum_{j=i+1}^n h z_{\bar{x},j}^2 \right) \leq (l - x_i) \sum_{j=i+1}^n h z_{\bar{x}}^2 \leq l \|z_{\bar{x}}\|_2^2,$$

где  $\|z_{\bar{x}}\|_2^2 = \sum_{j=1}^n h z_{\bar{x}}^2 = (z_{\bar{x}}, z_{\bar{x}})$ . Следовательно, для любой сеточной функции  $z$ , такой, что  $z_n = 0$ , справедливо неравенство

$$\|z\|_C \leq \sqrt{l} \|z_{\bar{x}}\|_2.$$

### 7.3. Методы и приемы конструирования разностных схем

Рассмотрим некоторые способы, позволяющие поставить в соответствие исходной дифференциальной задаче конечно-разностную задачу, т. е. способы конструирования *разностных схем* для рассматриваемых задач.

#### 7.3.1. Метод разностной аппроксимации

Данный метод является самым прямым и непосредственным. Он состоит в буквальной замене производных разностными соотношениями, т. е. замене предела отношений бесконечно малых отношениями конечных разностей (см. пример 7.1). Этим способом легко выводятся схемы первого или второго порядка *аппроксимации* для случая гладких решений и непрерывных коэффициентов. Однако повышение порядка аппроксимации потребует анализа погрешности аппроксимации и каких-то дополнительных действий. То же самое необходимо выполнять и на неравномерной сетке, и в непрямоугольной области и т. д.

### 7.3.2. Интегро-интерполяционный метод

Иногда этот метод называют **методом баланса**. В настоящее время, следуя западной литературе, этот метод называют также методом конечных объемов (см. также метод Галеркина — Петрова). Он заключается в выборе шаблона для схемы, в построении разностной ячейки, связанной с этим шаблоном, в интегрировании дифференциальной задачи по ячейке и последующей аппроксимации полученного интегрального соотношения. Поэтому в определении 7.5 разностной схемы использовалось упоминание о квадратурных формулах.

Причиной использования данного метода является тот факт, что многие уравнения математической физики имеют вид законов сохранения некоторой величины, например

$$\rho_t + \operatorname{div} \vec{j} = f.$$

Формальным интегрированием по выделенному объему  $V$  это уравнение можно преобразовать к интегральной форме

$$\left( \int_V \rho dV \right)_t + \oint_{\partial V} \vec{j} d\vec{S} = \int_V f dV,$$

которая в действительности является первичной. Здесь  $\rho$  — объемная плотность рассматриваемой величины;  $\vec{j}$  — ее поток через границу;  $f$  — скорость производства или убывания этой величины за счет источников или стоков. Непосредственная аппроксимация интегрального тождества позволяет обеспечить для разностной схемы выполнение тех же балансовых соотношений (точнее, их конечномерных аналогов), которые справедливы и в дифференциальном случае.

Рассмотрим в качестве примера *уравнение теплопроводности* с переменным коэффициентом температуропроводности  $k = k(x, t)$  на неравномерной сетке:

$$u_t = (ku_x)_x + f, \quad x \in (0, l), \quad t > 0.$$

Пусть

$$h_i = x_i - x_{i-1}, \quad h_{i+1} = x_{i+1} - x_i,$$

$$\hbar_i = \frac{x_{i+1/2} - x_{i-1/2}}{2}, \quad x_{i+1/2} = \frac{x_i + x_{i+1}}{2}.$$

Выберем разностную ячейку (рис. 7.6) и проинтегрируем уравнение теплопроводности по данной ячейке. Получим точное соотношение

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \int_{t_j}^{t_{j+1}} dx dt (u_t - (ku_x)_x - f) = \int_{x_{i-1/2}}^{x_{i+1/2}} (u(x, t_{j+1}) - u(x, t_j)) dx + \\ + \int_{t_j}^{t_{j+1}} (W(x_{i+1/2}, t) - W(x_{i-1/2}, t)) dt - \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{t_j}^{t_{j+1}} f dx dt = 0,$$

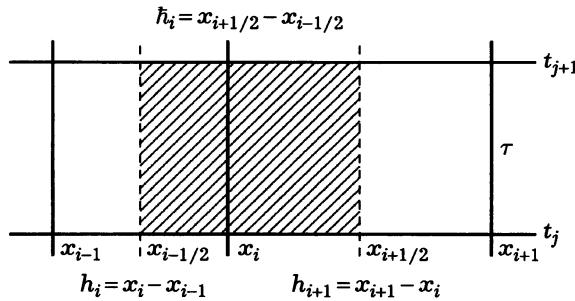


Рис. 7.6

справедливое для решения исходной задачи. Здесь  $W(x, t) = -ku_x$  — **тепловой поток**. Для него справедливо следующее соотношение:

$$-\int_{x_{i-1}}^{x_i} W(x, t)/k(x, t) dx = u(x_i, t) - u(x_{i-1}, t).$$

Пока все соотношения точны. Теперь проведем их аппроксимацию: используем  $y$ ,  $\widetilde{W}$  — соответствующие сеточные функции. Пусть

$$\varphi_i = \frac{1}{\tau \hbar_{x,i}} \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{t_j}^{t_{j+1}} f dx dt, \quad a_{i-1/2} = \left( \frac{1}{h_i} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x, t_{j+1})} \right)^{-1}.$$

Тогда из соотношений для потока и результата интегрирования по ячейке имеем

$$-\hat{\tilde{W}}_{i-1/2} h_i = a_{i-1/2} (\hat{y}_i - \hat{y}_{i-1}), \quad -\hat{\tilde{W}}_{i+1/2} h_{i+1} = a_{i+1/2} (\hat{y}_{i+1} - \hat{y}_i), \\ \hbar_{x,i} (\hat{y}_i - y_i) + (\hat{\tilde{W}}_{i+1/2} - \hat{\tilde{W}}_{i-1/2}) \tau - \hbar_{x,i} \tau \varphi_i = 0.$$

Отсюда получаем следующее разностное уравнение:

$$\frac{1}{\tau}(\hat{y}_i - y_i) = \frac{1}{h_{x,i}} \left( a_{i+1/2} \frac{\hat{y}_{i+1} - \hat{y}_i}{h_{i+1}} - a_{i-1/2} \frac{\hat{y}_i - \hat{y}_{i-1}}{h_i} \right) + \varphi_i.$$

Для случая сетки с постоянными шагами и задачи с постоянным коэффициентом температуропроводности  $k$  и нулевой правой частью  $f = 0$  получим уже рассмотренную выше неявную схему.

Отметим, что границы разностной ячейки проведены через середины отрезков между точками  $x_{i-1}$ ,  $x_i$  для того, чтобы при получении разностной схемы использовать формулу центральных прямоугольников, позволяющую повысить порядок аппроксимации.

Еще раз подчеркнем, что данный метод особенно полезен в случае наличия каких-то неоднородностей в задаче.

**Замечание 7.4.** Представленные выше выкладки проделаны в предположении, что коэффициент температуропроводности  $k(x, t) \neq 0$ . Из этого следует, что буквальное применение формул численного интегрирования к полученному для  $a_{i-1/2}$  выражению возможно лишь при выполнении указанного ограничения. Например, если коэффициент  $k(x, t)$  является кусочно-постоянным с разрывом в середине отрезка  $[x_{i-1}, x_i]$ , то применение полученных соотношений дает

$$a_{i-1/2} = \frac{2k_{i-1}k_i}{k_{i-1} + k_i}.$$

Если же обращение в нуль коэффициента  $k(x, t)$  возможно, то применимы формулы, в которых  $a_{i-1/2}$  вычисляется с помощью квадратурной формулы для среднеквадратичного значения коэффициента температуропроводности.

### 7.3.3. Метод неопределенных коэффициентов

Данный метод состоит в задании шаблона схемы и ее записи с пока неизвестными коэффициентами, которые должны быть определены из условия минимума погрешности аппроксимации. В качестве примера рассмотрим уравнение теплопроводности с постоянным коэффициентом температуропроводности  $u_t - ku_{xx} = 0$ . Зададимся шаблоном (см. рис. 7.4) и запишем на нем схему в виде

$$a\hat{y} + b\hat{y}_{-1} + c\hat{y}_{+1} + dy = 0.$$

Пусть сетка равномерная. Подсчитаем погрешность аппроксимации данной схемы, для чего подставим в разностную схему точное решение

задачи, считая его дважды непрерывно дифференцируемым по  $t$  и четырежды по  $x$ . Тогда применим формулу Тейлора:

$$u = \hat{u} - \frac{1}{1!} \hat{u}_t \tau + O(\tau^2) = \hat{u} - k \hat{u}_{xx} \tau + O(\tau^2),$$

$$\hat{u}_{\pm 1} = \hat{u} \pm \frac{1}{1!} \hat{u}_x h + \frac{1}{2!} \hat{u}_{xx} h^2 \pm \frac{1}{3!} \hat{u}_{xxx} h^3 + O(h^4).$$

Далее получим

$$\begin{aligned} \psi_h &= \hat{u}(a + b + c + d) + \hat{u}_x h(c - b) + \hat{u}_{xx} \left( \frac{1}{2} h^2(b + c) - kd\tau \right) + \\ &\quad + \hat{u}_{xxx} \frac{1}{6} h^3(c - b) + (b + c)O(h^4) + dO(\tau^2). \end{aligned}$$

Здесь в целях сокращения записи у функций неперерывного аргумента применен знак «крышки», указывающий на значения с верхнего временного слоя, нижний индекс указывает на пространственный сдвиг. В правых частях формул Тейлора стоят обычные, а не разностные производные.

Приравняв нулю коэффициенты у старших (по порядку параметров дискретизации) коэффициентов полученного выражения, получим следующую систему:

$$a + b + c + d = 0, \quad c - b = 0, \quad \frac{1}{2} h^2(b + c) - d\tau k = 0,$$

откуда  $b = c = dk\tau/h^2$ ,  $a = -d - 2dk\tau/h^2$ . Очевидно, что параметры такой схемы определены с точностью до произвольного коэффициента (сомножителя). Поэтому можно положить  $d = -1/\tau$ , откуда получаем  $a = 1/\tau + 2k/h^2$ ,  $b = c = -k/h^2$ , т. е. уже известную схему

$$\frac{\hat{y} - y}{\tau} - \frac{k(\hat{y}_{+1} - 2\hat{y} + \hat{y}_{-1})}{h^2} = 0.$$

Погрешность аппроксимации полученной схемы  $\psi_h = O(\tau + h^2)$ .

Этот метод можно применять и в более сложных ситуациях. При этом, очевидно, возрастет и громоздкость выкладок.

### 7.3.4. Другие методы получения алгебраических уравнений

Существует большой класс так называемых **проекционно-сеточных методов**. В них решение разыскивается в виде суммы

$$u \approx y = \varphi_0(x) + \sum_{i=1}^n c_i \varphi_i(x),$$

где  $\{\varphi_i(x)\}$  — некоторая выбранная последовательность функций, заданная с помощью сетки и удовлетворяющая ряду условий гладкости и полноты. Мы рассматривали эти методы при решении краевых задач для обыкновенных дифференциальных уравнений (ОДУ). Чаще всего оказывается, что  $c_i$  — значения искомой функции  $y_i$  в узлах сетки. Определив  $c_i$ , можно найти всю функцию  $y$ , заданную с помощью данной суммы. Для построения уравнений, описывающих  $y_i$ , используются различные методы.

Если уравнения для определения  $c_i$  строятся из условия минимума некоторого функционала путем его варьирования, то полученная система уравнений называется вариационно-разностной схемой.

Если для построения системы используется *метод Галеркина* (возможны различные вариации: методы Галеркина — Бубнова, Галеркина — Петрова и т. д.), то полученная система уравнений обычно называется схемой Галеркина.

Если  $\varphi_i$  — сплайновые функции, то схема называется сплайновой, если  $\varphi_i$  — конечные элементы, то схема называется конечноэлементной.

Возможны комбинации описанных схем, т. е. выбор схемы одного типа по временной переменной и другого типа — по пространственной переменной.

Несмотря на различие методов получения схем, уравнения для определения значений  $y_i$  могут оказаться одинаковыми, так что по окончательной схеме не всегда можно восстановить, каким методом она построена. Отличием проекционно-сеточных методов является изначальное задание способа интерполирования сеточных значений на весь промежуток изменения независимого переменного.

Мы лишь коснемся этих методов.

### 7.3.5. Аппроксимации в нерегулярных точках

Наибольшие сложности возникают при аппроксимации граничных условий второго и третьего рода или вообще граничных условий (в том числе и первого рода) в случае непрямоугольной пространственной области и прямоугольной сетки.

Мы ограничимся рассмотрением прямоугольной области, на границе которой заданы граничные условия второго или третьего рода. Очевидно, что в этом случае условия первого рода (как и начальные в случае уравнения с производными первого порядка по  $t$ ) легко аппроксимировать точно.

Рассмотрим уравнение

$$u_t = (ku_x)_x, \quad 0 < x < l, \quad t > 0.$$

Пусть на левой границе  $x = 0$  задано условие  $-ku_x + \alpha u = p(t)$ . Ограничимся аппроксимацией одного такого условия на трехточечном шаблоне (рис. 7.7). Пусть  $k = \text{const}$ . В качестве схемы на таком шаблоне возьмем, например, неявную схему (см. 7.3.3), имеющую погрешность аппроксимации  $\psi_h = O(\tau + h^2)$ .

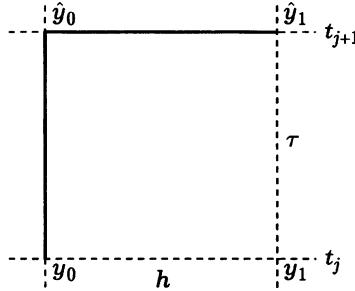


Рис. 7.7

Рассмотрим различные варианты.

1. Аппроксимируем граничное условие непосредственно, т. е. заменим дифференциальные производные разностными. Тогда получим

$$-\frac{k(\hat{y}_1 - \hat{y}_0)}{h} + \alpha \hat{y}_0 = \hat{p}.$$

Однако подстановка в это выражение точного решения задачи, т. е. вычисление погрешности аппроксимации, дает

$$\psi_{h,0} = -k \hat{u}_{xx} \frac{h}{2} + O(h^2),$$

из чего следует, что данное граничное условие аппроксимируется с точностью, меньшей, чем уравнение внутри области, что вносит дополнительную ошибку в получаемое разностное решение.

2. Проинтегрируем исходное уравнение по ячейке, примыкающей к левой границе:

$$\begin{aligned} \int_{x_0}^{x_{1/2}} (u(x, t_{j+1}) - u(x, t_j)) dx &= \int_{t_j}^{t_{j+1}} (ku_x(x_{1/2}, t) - ku_x(x_0, t)) dt = \\ &= \int_{t_j}^{t_{j+1}} (ku_x(x_{1/2}, t) - \alpha u(x_0, t) + p(t)) dt, \end{aligned}$$

откуда, как и ранее, получим разностную аппроксимацию

$$\frac{h}{2} \frac{\hat{y}_0 - y_0}{\tau} = k \frac{\hat{y}_1 - \hat{y}_0}{h} - \alpha \hat{y}_0 + \hat{p},$$

т. е.

$$-k\hat{y}_{x,0} + \alpha\hat{y}_0 + \frac{h}{2}\hat{y}_{t,0} = \hat{p}.$$

Тогда вычисление погрешности аппроксимации на точном решении исходной задачи дает

$$\psi_{h,0} = -k\hat{u}_{xx}\frac{h}{2} + \frac{h}{2}\hat{u}_t + O(\tau h + h^2) = O(\tau^2 + h^2),$$

так как на точном решении выполнено равенство  $u_t = ku_{xx}$ .

Таким образом, порядок аппроксимации удается повысить. Тот же результат можно получить путем уменьшения погрешности аппроксимации за счет введения дополнительных слагаемых, уничтожающих главный член погрешности аппроксимации  $-1/2k\hat{u}_{xx}h$ , т. е. добавки  $+1/2k\hat{u}_{xx}h$ , что мы и имеем во втором варианте в результате применения интегро-интерполяционного метода.

## 7.4. Основные качественно-количественные характеристики разностных схем и их виды

### 7.4.1. Аппроксимация

Рассмотрим точную задачу

$$Au = f \text{ в } G, \quad Ru = \mu \text{ на } \Gamma$$

и «заменяющую» ее разностную схему

$$A_h y = \varphi \text{ в } G_h, \quad R_h y = \nu \text{ на } \Gamma_h.$$

В этом случае

$$\psi_h = (\varphi - f_h) + ((Av)_h - A_h v_h), \quad \chi_h = (\nu - \mu_h) + ((Rv)_h - R_h v_h) \quad —$$

погрешности аппроксимации разностной задачи (для общего случая произвольной функции  $v$ ) в  $G_h$  и на  $\Gamma_h$  соответственно.

**Определение 7.17.** Разностная схема аппроксимирует исходную задачу, если  $\|\psi_h\|_\psi \rightarrow 0$ ,  $\|\chi_h\|_\chi \rightarrow 0$  при  $h \rightarrow 0$ ; аппроксимация имеет  $p$ -й порядок ( $p > 0$ ), если  $\|\psi_h\|_\psi = O(h^p)$ ,  $\|\chi_h\|_\chi = O(h^p)$  при  $h \rightarrow 0$ .

Очевидно, что порядок аппроксимации зависит от выбора норм. Выбор норм зависит, в свою очередь, от того, элементом какого пространства является соответствующая функция непрерывного аргумента. Например,  $f$  и  $\mu$  являются функциями даже разного числа переменных, откуда очевидно следует и различие норм, связанных с ними.

При возможности желательно выбирать как можно более слабые нормы, связанные с входными данными  $f, \mu$  (норма  $\|\cdot\|_1$  сильнее нормы  $\|\cdot\|_2$ , если из сходимости последовательности в норме  $\|\cdot\|_1$  следует сходимость в норме  $\|\cdot\|_2$ , но не наоборот).

В дальнейшем, как правило, будем оценивать невязку в равномерной норме  $C$ . При этом *аппроксимацию* будем называть *локальной*.

Выкладки будут проводиться применительно к  $\psi_h$ , величину  $\chi_h$  рассматривать не будем. Она может быть вычислена аналогичным образом.

В случае нескольких переменных  $\psi_h, \chi_h$  зависят не от одного шага. Например,  $\psi_h = O(\tau + h^2)$  в случае приведенной выше явной схемы для уравнения теплопроводности. Тогда за счет выбора соотношения шагов погрешность  $\psi_h$  может быть изменена.

**Пример 7.5.** Рассмотрим уравнение  $u_t - ku_{xx} = 0$ . Выберем явную схему. Вычислим погрешность аппроксимации на точном решении, считая, что оно (функция  $u$ ) имеет необходимое количество производных,  $k = \text{const}$ . Тогда

$$\begin{aligned}\psi_h &= \frac{1}{\tau}(\hat{u} - u) - \frac{k}{h^2}(u_{+1} - 2u + u_{-1}) = u_t + \frac{1}{2}u_{tt}\tau + \frac{1}{6}u_{ttt}\tau^2 + O(\tau^3) - \\ &\quad - k\left(u_{xx} + \frac{1}{12}u_{xxxx}h^2 + \frac{1}{360}u_{xxxxx}h^4 + O(h^6)\right) = \\ &= \frac{1}{2}k^2u_{xxxx}\tau + O(\tau^2) - \frac{1}{12}ku_{xxxx}h^2 + O(h^4) = O(h^4),\end{aligned}$$

если  $\tau = h^2/(6k)$ . Для других  $\tau$  и  $h$  погрешность  $\psi_h = O(\tau + h^2)$ . При выводе использовано уравнение  $u_t = ku_{xx}$ , откуда  $u_{tt} = ku_{txx} = k^2u_{xxxx}$ .

**Определение 7.18.** *Аппроксимацию* в случае нескольких переменных называют *безусловной* или *абсолютной*, если ее погрешность стремится к нулю при любом законе стремления к нулю шагов по разным переменным. Если же для стремления погрешности к нулю необходим некоторый определенный закон поведения шагов дискретизации, то аппроксимацию называют *условной*.

**Замечание 7.5.** На примере 7.5 хорошо видно различие аппроксимации на точном решении и на произвольной функции: даже для приведенного ограничения на шаги порядок аппроксимации на произвольной функции остается равным  $O(\tau + h^2)$ .

**Пример 7.6.** Запишем схему для уравнения теплопроводности на зигзагообразном шаблоне, состоящем из узлов  $(i, j), (i-1, j), (i, j+1), (i+1, j+1)$  (см. рис. 7.2):

$$y_t - \frac{k(\hat{y}_x - \bar{y}_{\bar{x}})}{h} = 0,$$

т. е.

$$y_t - k y_{\bar{x}x} - \frac{\tau k}{h} y_{x,t} = 0.$$

Отличие погрешности аппроксимации от случая обычной явной схемы состоит в появлении следующего лишнего слагаемого:

$$\tilde{\psi}_h = \frac{k\tau}{h} u_{x,t} = \frac{k}{h^2} (\hat{u}_{+1} - \hat{u} - u_{+1} + u) = O\left(\frac{\tau}{h}\right).$$

Следовательно, определена полная погрешность  $\psi_h = O(\tau + h^2 + \tau/h)$ . Она стремится к нулю лишь при  $\tau/h \rightarrow 0$ . Таких схем по возможности нужно избегать.

**Замечание 7.6.** Полученное выражение для погрешности аппроксимации справедливо на функциях, имеющих четвертые непрерывные смешанные производные. Довольно часто вычисление аппроксимации разностных операторов второго и более высокого порядков, действующих по разным переменным, вызывает затруднения. Источником затруднений является формула Тейлора для функций многих переменных, буквальное применение которой ведет к появлению проблем с остаточным членом. Покажем на примере способ вычисления, позволяющий избежать осложнений. Это можно сделать путем введения параметра следующим образом:

$$\begin{aligned} u_{x,t} &= \frac{1}{h\tau} (\hat{u}_{+1} - \hat{u} - u_{+1} + u) = \\ &= \frac{1}{h\tau} (u(x + \alpha h, t + \alpha\tau) - u(x, t + \alpha\tau) - u(x + \alpha h, t) + u(x, t)) \end{aligned}$$

при  $\alpha = 1$ . Далее к полученному выражению можно применить формулу Тейлора по переменной  $\alpha$  с центром  $\alpha = 0$ . Проводя вычисления до слагаемых третьего порядка, получаем

$$\begin{aligned} u_{x,t} &= u_{x,t}(x, t) + \frac{1}{3!h\tau} \left( u_{x^3}(x + \alpha^* h, t + \alpha^*\tau) h^3 + \right. \\ &\quad + 3u_{x^2t}(x + \alpha^* h, t + \alpha^*\tau) h^2\tau + 3u_{xt^2}(x + \alpha^* h, t + \alpha^*\tau) h\tau^2 + \\ &\quad \left. + u_{t^3}(x + \alpha^* h, t + \alpha^*\tau) \tau^3 - u_{t^3}(x, t + \alpha^*\tau) \tau^3 - u_{x^3}(x + \alpha^* h, t) h^3 \right). \end{aligned}$$

В правой части данного выражения использованы обычные производные. В остаточном члене используется некоторое промежуточное значение  $\alpha^* \in (0, 1)$ . Дальнейшие выкладки опустим. Очевидно, что

применение формулы Лагранжа конечных приращений в остаточном члене даст

$$u_{x,t} = u_{x,t}(x, t) + O(h^2 + h + \tau + \tau^2) = u_{x,t}(x, t) + O(h + \tau).$$

Нужная оценка получена.

Отметим также, что в случае неравномерной сетки условие  $h \rightarrow 0$  следует понимать как стремление к нулю некоторой характеристики всех шагов сетки типа их нормы ( $\max_i h_i \rightarrow 0$  и т. п.).

#### 7.4.2. Устойчивость

Ранее это понятие уже встречалось при решении *задачи Коши* для ОДУ.

Пусть  $y^I$ ,  $y^{II}$  — решения двух разностных задач с одинаковым оператором, соответствующие правым частям  $\varphi^I$ ,  $\varphi^{II}$  и граничным условиям  $\nu^I$ ,  $\nu^{II}$ .

**Определение 7.19.** Будем называть *разностную схему устойчивой*, если решение уравнений разностной схемы непрерывно зависит от входных данных и эта зависимость равномерна по  $h$ , т. е.

$$\forall \varepsilon > 0 \exists \delta(\varepsilon) > 0 : \|\varphi^I - \varphi^{II}\|_\varphi < \delta, \quad \|\nu^I - \nu^{II}\|_\nu < \delta \Rightarrow \|y^I - y^{II}\|_Y < \varepsilon.$$

Если схема линейна, то это определение переходит в требование существования таких постоянных  $M_1$ ,  $M_2$ , не зависящих от шага сетки, что справедливо неравенство

$$\|y^I - y^{II}\|_Y \leq M_1 \|\varphi^I - \varphi^{II}\|_\varphi + M_2 \|\nu^I - \nu^{II}\|_\nu,$$

которое дает равномерную по  $h$  ограниченность линейных операторов, определяющих  $y$ , по  $\varphi$ ,  $\nu$ .

В случае линейных разностных схем устойчивость в смысле общего определения очевидным образом следует из определения устойчивости линейных схем и, наоборот (существование необходимых постоянных в оценке разности решений двух линейных разностных задач), эквивалентность следует из известной теоремы функционального анализа об ограниченности линейного непрерывного оператора.

**Определение 7.20.** В случае нескольких независимых переменных *устойчивость* называют *безусловной* (или *абсолютной*), если устойчивость имеет место для любого соотношения шагов, и *условной* в противном случае.

**Определение 7.21.** Непрерывную зависимость решения разностной задачи от  $\varphi$  называют *устойчивостью по правой части*, от  $\nu$  на границе пространственной области — *устойчивостью по граничным условиям*, от  $\nu$  на гиперплоскости  $t = 0$  — *устойчивостью по начальным данным*.

Отметим, что существует еще устойчивость относительно коэффициентов уравнений, но мы ее касаться не будем.

Устойчивость разностной схемы наиболее критична для эволюционных задач, в которых решение определяется начальными и граничными условиями и правой частью. Рассмотрим более подробно одно из таких понятий для так называемых двухслойных разностных схем.

**Определение 7.22.** *Разностные схемы*, содержащие неизвестные значения сеточной функции на одном временном слое, а известные — также на одном, называются *двуслойными*.

Очевидно, что для двухслойной разностной схемы решение на любом временном слое можно рассматривать как начальное условие для последующих слоев.

**Определение 7.23.** Двухслойная *разностная схема* называется *равномерно устойчивой* по начальным данным, если при постановке начальных данных на любом слое  $t_j$ ,  $(0 < t_j < T)$  схема равномерно устойчива по ним, т. е. в случае линейных схем для любых  $t_j$ ,  $t_k$ ,  $0 \leq t_j \leq t_k \leq T$ , существует постоянная  $K$ , не зависящая от  $h$ ,  $\tau$ ,  $j$  и  $k$ , такая, что выполнено неравенство

$$\|y^I(t_k) - y^{II}(t_k)\|_Y \leq K \|y^I(t_j) - y^{II}(t_j)\|_Y.$$

При этом правые части и граничные условия для  $y^I$ ,  $y^{II}$  должны совпадать.

При исследовании свойств разностных схем для эволюционных задач часто выделяют *асимптотическую устойчивость*, означающую аккуратную передачу разностным решением асимптотики точного решения при больших значениях параметра  $t$ .

### 7.4.3. Сходимость

Наличие сходимости является главным требованием к схеме.

**Определение 7.24.** Разностное решение  $y$  сходится к решению  $u$  точной задачи, если  $\|y - p_h u\|_Y \rightarrow 0$  при  $h \rightarrow 0$ . Говорят, что имеет место *сходимость* с  $p$ -м ( $p > 0$ ) порядком, если  $\|y - p_h u\|_Y = O(h^p)$  при  $h \rightarrow 0$ .

**Определение 7.25.** Разностная схема  $A_h y = \varphi$  в  $G_h$ ,  $R_h y = \nu$  на  $\Gamma_h$  называется **корректной**, если ее решение существует, единственно и непрерывно зависит от входных данных, т. е. если схема устойчива.

**Теорема 7.1.** Если решение  $u$  задачи  $Au = f$  в  $G$ ,  $Ru = \mu$  на  $\Gamma$  существует, разностная схема  $A_h y = \varphi$  в  $G_h$ ,  $R_h y = \nu$  на  $\Gamma_h$  корректна и аппроксимирует задачу для  $u$ , то разностное решение сходится к точному.

◀ Запишем следующие соотношения:

$$\begin{aligned} A_h u_h &= A_h u_h - \varphi + \varphi + f_h - (Au)_h = \varphi + (f_h - \varphi + A_h u_h - (Au)_h) = \varphi - \psi_h, \\ R_h u_h &= R_h u_h + \nu - \nu + \mu_h - (Ru)_h = \nu + (\mu_h - \nu + R_h u_h - (Ru)_h) = \nu - \chi_h, \end{aligned}$$

где  $\psi_h, \chi_h$  — погрешности аппроксимации исходной задачи на точном решении. Оказывается, что  $u_h$  есть решение данной разностной схемы с входными данными (в правых частях), возмущенными на величину невязок  $\psi_h, \chi_h$ . Поскольку схема корректна, разность решений задач для  $u_h$  и  $y$  стремится к нулю при стремлении к нулю невязок  $\psi_h, \chi_h$ , а именно: для любого  $\varepsilon > 0$  существует такое  $\delta(\varepsilon) > 0$ , что  $\|y - u_h\|_Y < \varepsilon$  при  $\|\psi_h\|_\psi < \delta$  и  $\|\chi_h\|_\chi < \delta$ . А так как схема аппроксимирует исходную задачу, то для любого  $\delta > 0$  существует такое  $h_0$ , что  $\|\psi_h\|_\psi < \delta$  и  $\|\chi_h\|_\chi < \delta$  при  $h < h_0$ . Таким образом, для произвольного  $\varepsilon > 0$  мы указали  $h_0 > 0$ , такое, что  $\|y - u_h\|_Y < \varepsilon$ ,  $h < h_0$ . ►

**Замечание 7.7.** Теорема сформулирована и доказана В.С. Рябеньким и А.А. Филипповым и на языке вычислителей-прикладников звучит обычно так: «из аппроксимации и устойчивости разностной схемы следует ее сходимость».

Данная теорема свидетельствует, что наибольший интерес при исследовании схем необходимо уделять изучению устойчивости, так как аппроксимация проверяется сравнительно просто путем применения формулы Тейлора, а сходимость есть следствие аппроксимации и устойчивости.

**Замечание 7.8.** Утверждения о простоте относятся только к случаю линейных разностных схем. Для нелинейного случая все существенно усложняется.

**Теорема 7.2.** Если, как и в теореме 7.1, решение  $u$  задачи  $Au = f$  в  $G$ ,  $Ru = \mu$  на  $\Gamma$  существует, разностная схема  $A_h y = \varphi$  в  $G_h$ ,  $R_h y = \nu$  на  $\Gamma_h$  корректна и аппроксимирует исходную задачу, и, кроме того, операторы  $A_h$  и  $R_h$  являются линейными, а порядок аппроксимации

равен  $p$ , то разностное решение сходится к точному со скоростью не ниже  $\tilde{O}(h^p)$  при  $h \rightarrow 0$ .

◀ Для нахождения погрешности  $z = y - u_h$  (здесь  $y$  — решение разностной задачи,  $u_h$  — проекция на сетку точного решения исходной задачи) разностной схемы имеем следующую разностную задачу:

$$A_h z = \psi_h, \quad R_h z = \chi_h.$$

Поскольку схема линейна, существуют такие постоянные  $M_1, M_2$ , не зависящие от шага сетки, что справедливо неравенство

$$\|z\|_Y \leq M_1 \|\psi_h\|_\varphi + M_2 \|\chi_h\|_\nu.$$

Из этой оценки следует утверждение теоремы. ►

**Замечание 7.9.** Коротко утверждение теоремы обычно формулируют следующим образом: для линейных схем порядок точности не ниже порядка аппроксимации.

#### 7.4.4. Качественно-количественные виды схем

**Консервативные (дивергентные) схемы.**

**Определение 7.26.** *Разностные схемы* называют **консервативными**, если их решение удовлетворяет дискретному аналогу закона сохранения (баланса), присущему исходной задаче. В противном случае схему называют **неконсервативной**, или дисбалансной.

**Пример 7.7.** Рассмотрим задачу

$$(k(x)u_x)_x = 0, \quad 0 < x < 1; \\ u(0) = 1, \quad u(1) = 0.$$

Пусть

$$k(x) = \begin{cases} 2, & 0 \leq x < 1/2; \\ 1, & 1/2 \leq x < 1. \end{cases}$$

Ее точное решение имеет вид

$$u(x) = \begin{cases} 1 - \frac{2x}{3}, & 0 \leq x < \frac{1}{2}; \\ \frac{4(1-x)}{3}, & \frac{1}{2} \leq x < 1. \end{cases}$$

В точке разрыва коэффициента справедливы соотношения  $u|_{0,5-0} = u|_{0,5+0}$ ,  $(ku_x)|_{0,5-0} = (ku_x)|_{0,5+0}$ .

Для решения исходной задачи всюду, кроме точки  $x = 0,5$ , справедливо уравнение  $ku_{xx} = 0$ . Для такого уравнения можно записать разностную схему  $ky_{\bar{x}\bar{x}} = 0$ . Если узел сетки не попадает на точку разрыва коэффициента, то такая процедура выглядит на первый взгляд вполне приемлемой. Коэффициент  $k = 1$  или  $k = 2$  в зависимости от точки. Уравнение  $ky_{\bar{x}\bar{x}} = 0$  можно на него разделить и в результате получить совершенно точное решение такой разностной схемы:  $y = 1 - x$ ,  $y_i = 1 - ih$ ,  $i = 0, 1, \dots, n$ ,  $nh = 1$ . При этом норма разности решений точной и приближенной задач равна

$$\|y - u_h\|_C = \frac{2}{3} - \frac{1}{2} = \frac{1}{6} \not\rightarrow 0$$

при  $h \rightarrow 0$ . #

Итак, при написании схем необходимо пользоваться (быть может, неявно) интегро-интерполяционным методом и не следует дифференцировать выражения вида  $(ku_x)_x$  без необходимости. Отметим, что формальный учет в разностной схеме производной коэффициента  $k_x$  не дает сходимости решения разностной схемы.

### Однородные схемы.

**Определение 7.27.** *Разностные схемы*, в которых расчет ведется по одним формулам и на одном шаблоне во всех узлах сетки без какого-то специального выделения имеющихся особенностей, называются *однородными*.

**Пример 7.8.** Для уравнения  $(ku_x)_x = 0$  из примера 7.7 схема  $(ku_x)_{\bar{x}} = 0$  является однородной, а схема с выделенной точкой  $x_0 = 0,5$  и условием  $(ku_x)|_{x_0-0} = (ku_x)|_{x_0+0}$  — неоднородной. #

Однородные схемы еще называют схемами сквозного счета.

**Монотонные схемы.** При определенных условиях решение точной дифференциальной задачи удовлетворяет принципу максимума в той или иной форме. Например, решение уравнения Лапласа достигает своих минимального и максимального значений на границе. Возможно также, что решение сохраняет монотонность начальных данных по пространству при увеличении времени. Отметим, что свойство монотонности функции в строгом и однозначном смысле имеет место только в одномерном случае, поэтому в многомерном приходится рассматривать *принцип максимума*.

**Определение 7.28.** Схемы, решение которых удовлетворяет принципу максимума или сохраняет пространственную монотонность (в одномерном случае) при условии, что соответствующие свойства справедливы для исходных задач, называются **монотонными**.

Примеры будут приведены далее при рассмотрении конкретных схем.

**Другие виды схем.** Часто выделяют отдельно так называемые «экономичные» схемы, которые позволяют решать многомерные задачи как последовательность одномерных.

При решении систем уравнений, описывающих физически содержащиеся задачи, выделяют «полностью консервативные» разностные схемы, которые передают не только дискретные аналоги основных законов сохранения (балансов), присущих решению исходной задачи, но и некоторые дополнительные соотношения, необходимость выполнения которых диктуется физическими соображениями.

## 7.5. Разделение переменных в дискретном случае

Напомним факт из теории уравнений в частных производных. Пусть решаемую задачу можно записать в виде

$$Au = A_t u + A_{\vec{r}} u = 0,$$

где линейные операторы  $A_t$ ,  $A_{\vec{r}}$  действуют по временной и пространственной переменной соответственно и зависят только от  $t$  и  $\vec{r}$ . Пусть область  $G = \Omega \times (0, T)$  — цилиндр, в основании которого лежит пространственная область  $\Omega$ , на границе которой заданы нулевые граничные условия, а на гиперплоскости  $t = 0$  — какие-то начальные данные. Тогда задача имеет частное решение, которое можно представить в виде произведения двух функций  $u = T(t)R(\vec{r})$ , одна из которых зависит только от  $t$ , а вторая — от  $\vec{r}$ . Эти функции удовлетворяют следующим уравнениям:

$$A_t T + \lambda^2 T = A_{\vec{r}} R - \lambda^2 R = 0.$$

Функция  $R$  удовлетворяет нулевым граничным условиям и, значит, является собственной функцией оператора  $A_{\vec{r}}$ , соответствующей *собственному значению*  $\lambda$ . При определенных условиях система таких собственных функций является *полной ортонормированной*, что позволяет искать решение исходной задачи в виде разложения по системе таких функций с коэффициентами вида  $T(t)$ . Начальные данные для  $T(t)$  определяются разложением начальных данных исходной задачи по

системе собственных функций. Описанный метод построения решения называют **методом разделения переменных**. Задачу нахождения собственных функций  $R(\vec{r})$  часто называют **задачей Штурма — Лиувилля**. Очевидно, что в случае **разностных** задач должен существовать какой-то ее аналог.

Рассмотрим разностный аналог простейшей задачи Штурма — Лиувилля — разностную задачу

$$y_{\bar{x}x} + \lambda^2 y = 0$$

на сетке

$$\bar{\omega}_h = \{x_i = ih, i = 0, 1, \dots, n\}, \quad nh = l,$$

с граничными условиями  $y_0 = y_n = 0$ . По аналогии с решением соответствующей дифференциальной задачи

$$u_k(x) = \sin \frac{\pi kx}{l}, \quad \lambda_k^2 = \left( \frac{\pi k}{l} \right)^2, \quad k = 1, 2, \dots,$$

будем искать решение разностной задачи в виде

$$y_{k,i} = \sin \frac{\pi kih}{l}.$$

Тогда после подстановки решения в приведенной форме в разностную задачу получим

$$\lambda_k^2 = \frac{4}{h^2} \sin^2 \frac{\pi kh}{2l}.$$

В этих выражениях  $i = 0, 1, \dots, n$ ,  $k$  — целое число.

Если  $k = 0$ , то  $y_k = 0$ , если  $k = n$ , то  $y_{k,i} = \sin \pi i = 0$ , т. е. это не собственные функции (которые должны быть нетривиальными). Пусть  $k = n + 1$ . Тогда

$$\frac{\pi kih}{l} = 2\pi i - \frac{\pi(n-1)ih}{l}.$$

Следовательно,  $y_{n+1,i} = -y_{n-1,i}$ ,

$$\lambda_{n+1}^2 = \frac{4}{h^2} \sin^2 \frac{\pi(2n-(n-1))h}{2l} = \frac{4}{h^2} \sin^2 \frac{\pi(n-1)}{2l} h = \lambda_{n-1}^2.$$

Таким образом, при  $k = n + 1$  мы снова получаем ту же (с точностью до знака) собственную функцию и то же собственное значение. Очевидно, что то же мы получим и при  $k = n + 2, n + 3, \dots$  Следовательно,

$$y_{k,i} = \sin \frac{\pi kih}{l}, \quad \lambda_k^2 = \frac{4}{h^2} \sin^2 \frac{\pi kh}{2l}, \quad k = 1, 2, \dots, n-1,$$

что в точности соответствует размерности рассматриваемого пространства  $(n+1)-2=n-1$ , так как в точках  $i=0, i=n$  заданы граничные условия  $y_0=y_n=0$ .

Очевидно, что

$$\begin{aligned} 0 < \lambda_1^2 < \lambda_2^2 < \dots < \lambda_{n-1}^2 = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l} < \frac{4}{h^2}, \\ \lambda_1^2 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} = \frac{\pi^2}{l^2} \left( \frac{\sin \alpha}{\alpha} \right)^2, \quad f(\alpha) = \frac{\sin \alpha}{\alpha}, \quad \alpha = \frac{\pi h}{2l}. \end{aligned}$$

Найдем оценку снизу для  $\lambda_1^2$ , для чего вычислим производную функции  $f(\alpha)$ :

$$f'_\alpha = -\frac{\sin \alpha - \alpha \cos \alpha}{\alpha^2} = \frac{\alpha \cos \alpha - \sin \alpha}{\alpha^2}.$$

Так как  $\sin \alpha < \alpha < \tan \alpha$  при  $\alpha \in (0, \pi/2)$ , то  $f'_\alpha < 0$ . Если  $h < l/3$ , то  $\alpha \in (0, \pi/6)$ . Отсюда

$$f(\alpha) \geq \frac{1}{2} \cdot \frac{6}{\pi} = \frac{3}{\pi}.$$

Таким образом, получим оценки

$$\lambda_1^2 > \frac{\pi^2}{l^2} \frac{9}{\pi^2},$$

и

$$0 < \frac{9}{l^2} < \lambda_1^2 < \lambda_2^2 < \dots < \lambda_{n-1}^2 < \frac{4}{h^2}.$$

Используем формулу интегрирования по частям

$$(y_x, z) = y_n z_n - y_1 z_0 - (y, z_{\bar{x}}).$$

Пусть  $y = y_{k,\bar{x}}, z = y_m, k \neq m$ . Тогда

$$(y_{k,\bar{x}x}, y_m) = 0 - 0 - (y_{k,\bar{x}}, y_{m,\bar{x}}).$$

Совершенно аналогичным образом выведем равенство

$$(y_{m,\bar{x}x}, y_k) = 0 - 0 - (y_{m,\bar{x}}, y_{k,x}).$$

Вычитая полученные равенства друг из друга, имеем

$$(\lambda_k^2 - \lambda_m^2)(y_k, y_m) = 0,$$

так как  $y_{k,\bar{x}x} = -\lambda_k^2 y_k$  (то же самое верно и для  $y_m$ ). Отсюда в силу различия собственных значений  $\lambda_k \neq \lambda_m$  при  $k \neq m$  следует  $(y_k, y_m) = 0$ , т. е. ортогональность собственных функций относительно данного скалярного произведения.

Подсчитаем норму  $\|y_k\|_2^2$ :

$$\begin{aligned} \|y_k\|_2^2 &= \sum_{i=1}^{n-1} h \left( \sin \frac{\pi k i h}{l} \right)^2 = \frac{1}{2} \sum_{i=1}^{n-1} h \left( 1 - \cos \frac{2\pi k i h}{l} \right) = \\ &= \frac{h}{2}(n-1) - \frac{1}{4} h \sum_{i=1}^{n-1} \left( \exp \left( \tilde{i} \frac{2\pi k i h}{l} \right) + \exp \left( -\tilde{i} \frac{2\pi k i h}{l} \right) \right) = \frac{h}{2}(n-1) - \\ &- \frac{h}{4} \left( \frac{\exp \left( \tilde{i} \frac{2\pi k h}{l} \right) - \exp \left( \tilde{i} 2\pi k \right)}{1 - \exp \left( \tilde{i} \frac{2\pi k h}{l} \right)} + \frac{\exp \left( -\tilde{i} \frac{2\pi k h}{l} \right) - \exp \left( -\tilde{i} 2\pi k \right)}{1 - \exp \left( -\tilde{i} \frac{2\pi k h}{l} \right)} \right) = \\ &= \frac{h}{2}(n-1) + \frac{h}{2} = \frac{h}{2} n = \frac{l}{2}. \end{aligned}$$

В этих выражениях  $\tilde{i}$  — мнимая единица:  $\tilde{i}^2 = -1$ . Следовательно, нормированные собственные функции имеют вид

$$y_k(x_i) = \sqrt{\frac{2}{l}} \sin \frac{\pi k x_i}{l} = \sqrt{\frac{2}{l}} \sin \frac{\pi k i h}{l}, \quad k = 1, 2, \dots, n-1.$$

Отсюда следует, что любая сеточная функция, заданная на  $\bar{\omega}_h$  и обращающаяся в нуль на концах, может быть разложена по системе функций  $y_k(x_i)$ .

Отметим, что построенная система функций применима и при решении задач с ненулевыми граничными условиями. Рассмотрим в качестве примера задачу:

$$\begin{aligned} y_{\bar{x}x,i} &= -f_i, \quad i = 1, 2, \dots, n-1; \\ y_0 &= \mu_0, \quad y_n = \mu_n. \end{aligned}$$

Введем новую исковую функцию  $\tilde{y}_i = y_i - \mu_0 \delta_{i0} - \mu_n \delta_{in}$ , где  $\delta_{ij}$  — сеточная функция, определяемая символом Кронекера. Тогда задача переформулируется следующим образом:

$$\begin{aligned} \tilde{y}_{\bar{x}x,i} &= -\tilde{f}_i, \quad i = 1, 2, \dots, n-1; \\ \tilde{y}_0 &= \tilde{y}_n = 0, \end{aligned}$$

где

$$\tilde{f}_i = f_i - \frac{\delta_{i1}\mu_0}{h^2} - \frac{\delta_{i,n-1}\mu_n}{h^2}, \quad i = 1, 2, \dots, n-1.$$

Правая часть  $\tilde{f}$  задана при  $1 \leq i \leq n-1$ , поэтому можно считать, что  $\tilde{f}_0 = \tilde{f}_n = 0$ . Отсюда получаем

$$\tilde{y}_i = \sum_{k=1}^{n-1} c_k y_k(x_i), \quad \tilde{f}_i = \sum_{k=1}^{n-1} b_k y_k(x_i), \quad b_k = (\tilde{f}, y_k).$$

Подстановка таких  $\tilde{y}$  и  $\tilde{f}$  в решаемое уравнение дает

$$\tilde{y}_{\bar{x}x} = \sum_{k=1}^{n-1} c_k y_{k,\bar{x}x} = - \sum_{k=1}^{n-1} c_k \lambda_k^2 y_k = - \sum_{k=1}^{n-1} b_k y_k.$$

В итоге получаем следующий результат:  $c_k = b_k / \lambda_k^2$  — точное решение разностной задачи.

Очевидно, что найти решение разностной задачи Штурма — Лиувилля можно далеко не всегда. Даже в случае рассмотренной задачи на неравномерной сетке написать точное решение в аналитическом виде уже не представляется возможным. В то же время полученные соотношения иногда позволяют получать полезные оценки.

Рассмотрим, например, разностный оператор на той же равномерной сетке, действующий по правилу

$$\begin{aligned} Ay &= -(ay_{\bar{x}})_x; \\ y_0 &= y_n = 0. \end{aligned}$$

Тогда

$$y = \sum_{k=1}^{n-1} c_k y_k,$$

где  $y_k$  — собственные функции. Пусть  $0 < a_{\min} \leq a \leq a_{\max}$ , коэффициент  $a$  зависит от  $x$ . По первой формуле Грина имеем

$$(Ay, y) = (a, (y_{\bar{x}})^2).$$

Если  $a = 1$ , то

$$A_0 y = -y_{\bar{x}x}, \quad (A_0 y, y) = \sum_{k=1}^{n-1} \lambda_k^2 c_k^2 = (1, (y_{\bar{x}})^2).$$

Следовательно, учитывая равенство  $\|y\|_2^2 = \sum_{k=1}^{n-1} c_k^2$  и оценки собственных значений, получаем

$$\frac{9}{l^2} \|y\|_2^2 \leq (A_0 y, y) \leq \frac{4}{h^2} \|y\|_2^2.$$

Запишем то же самое неравенство в виде неравенства вложения:

$$\frac{9}{l^2} \|y\|_2^2 \leq \|y_{\bar{x}}\|_2^2 \leq \frac{4}{h^2} \|y\|_2^2.$$

Используя это неравенство, для оператора  $A$  имеем оценки

$$\frac{9}{l^2}a_{\min}\|y\|_2^2 \leq (Ay, y) \leq \frac{4}{h^2}a_{\max}\|y\|_2^2,$$

а также

$$a_{\min}(A_0y, y) \leq (Ay, y) \leq a_{\max}(A_0y, y).$$

Таким образом, операторы  $A_0$  и  $A$  энергетически эквивалентны и оба строго положительно определены.

Мы также доказали энергетическую эквивалентность оператора  $A$  и единичного оператора  $E$  с постоянными энергетической эквивалентности

$$\gamma_1 = \frac{9}{l^2}a_{\min}, \quad \gamma_2 = \frac{4}{h^2}a_{\max}.$$

## 7.6. Принцип максимума для разностных схем

Любую линейную разностную схему можно записать в виде

$$A(x)y(x) = \sum_{\xi \in S'(x)} B(x, \xi)y(\xi) + F(x), \quad x \in G_h.$$

Такая запись называется канонической формой записи разностной схемы. В ней  $x$  — одна или несколько независимых переменных,  $S'(x) = S(x) \setminus x$  — множество, которое может быть и пустым,  $S(x)$  — шаблон разностной схемы, которому принадлежит и узел  $x$ . Будем считать, что  $S(x)$  может быть множеством узлов сетки, связанных какими-то разностными уравнениями и в нерегулярных узлах сетки. Следовательно, приведенная запись представляет собой всю разностную схему, в том числе и уравнения в начальных и граничных узлах. Каждый узел  $x$  сетки и соответствующее значение сеточной функции  $y(x)$  в левой части канонической записи фигурируют лишь один раз. Именно это условие обеспечивает единственность представления разностной схемы в данной канонической форме. Здесь величины  $A(x) \neq 0$ ,  $B(x, \xi) \neq 0$  — коэффициенты.

Определим сеточный оператор

$$Ly = D(x)y(x) - \sum_{\xi \in S'(x)} B(x, \xi)(y(\xi) - y(x))$$

и обозначим  $D(x) = A(x) - \sum_{\xi \in S'(x)} B(x, \xi)$ . Тогда  $Ly = F$ .

**Определение 7.29.** Назовем *сетку*  $G_h$  *связной*, если для любых двух узлов  $\tilde{x}, \tilde{\tilde{x}} \in G_h$ , имеющих непустую окрестность, существует набор узлов  $x_i \in G_h$ ,  $i = 1, 2, \dots, m$ , таких, что

$$x_1 \in S'(\tilde{x}), x_2 \in S'(x_1), \dots, \tilde{\tilde{x}} \in S'(x_m),$$

т. е. от одного узла к другому можно перейти, используя заданные шаблоны, причем  $B(x_i, x_{i+1}) \neq 0$ ,  $i = 1, 2, \dots, m-1$ ;  $B(\tilde{x}, x_1) \neq 0$ ,  $B(x_m, \tilde{\tilde{x}}) \neq 0$ .

Отметим, что, очевидно, понятие связности определяется и шаблоном разностной схемы, а не только сеткой. Пример несвязной области для шаблона типа «крест» приведен на рис. 7.8: от узла  $A$  к узлу  $B$  нельзя перейти, используя указанный шаблон.

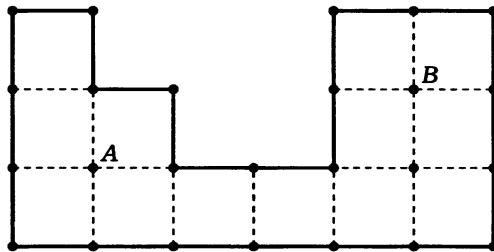


Рис. 7.8

**Определение 7.30.** Будем говорить, что в узле  $x$  выполнено *условие положительности коэффициентов*, если

$$A(x) > 0, \quad D(x) \geq 0; \quad B(x, \xi) > 0, \quad \xi \in S'(x).$$

Введем обозначение для объединения всех шаблонов с центрами в узлах  $x \in \omega$ :

$$\bar{\omega} = \bigcup_{x \in \omega} S(x).$$

Например, если  $\omega$  — множество внутренних узлов сетки  $G_h$ , то  $\bar{\omega} = G_h$ . Множество  $\bar{\omega}$  представляет собой замыкание  $\omega$  относительно указанной системы шаблонов.

**Теорема 7.3 (принцип максимума).** Пусть  $G_h$  и  $\omega$  связны,  $\bar{\omega} \in G_h$ , в  $\omega$  выполнены условия положительности коэффициентов. Тогда, если функция  $y$  не является постоянной на  $\bar{\omega}$  и  $Ly \leq 0$  при  $x \in \omega$  ( $Ly \geq 0$  при  $x \in \bar{\omega} \setminus \omega$ ), функция  $y$  не может принимать наибольшего положительного (наименьшего отрицательного) значения на  $\omega$  среди всех ее значений на  $\bar{\omega}$  (т. е. они принимаются на  $\bar{\omega} \setminus \omega$ ).

◀ Рассмотрим случай  $Ly \leq 0$  и допустим, что теорема неверна: в точке  $x_0 \in \omega$  достигается положительный максимум:  $y(x_0) > 0$ . Тогда имеем

$$Ly(x_0) = D(x_0)y(x_0) - \sum_{\xi \in S'(x_0)} B(x_0, \xi)(y(\xi) - y(x_0)) \geq 0.$$

Но  $Ly(x_0) \leq 0$ . Следовательно,  $Ly(x_0) = 0$ . Из положительности коэффициентов  $B$  следует равенство  $y(\xi) = y(x_0)$  в  $S'(x_0)$ . Функция  $y \neq \text{const}$  в  $\bar{\omega}$ , поэтому найдется узел  $x'_0 \in \bar{\omega}$ , такой, что  $y(x'_0) < y(x_0)$ . Сетка связна, поэтому от узла  $x_0$  до узла  $x'_0$  можно перейти по системе шаблонов через узлы  $x_1, x_2, \dots, x_m \in \omega$ . Причем, так как  $x_1 \in S'(x_0)$ , то  $y(x_1) = y(x_0)$ . Но для узла  $x_1$  можно провести те же рассуждения, что и для остальных  $x_2, x_3, \dots, x_m$ , и в результате получить

$$y(x_0) = y(x_1) = \dots = y(x_m).$$

Отсюда

$$\begin{aligned} Ly(x_m) &= D(x_m)y(x_m) + \sum_{\xi \in S'(x_m)} B(x_m, \xi)(y(x_m) - y(\xi)) \geq \\ &\geq B(x_m, x'_0)(y(x_m) - y(x'_0)) > 0, \end{aligned}$$

но  $Ly \leq 0$  всюду. Следовательно, допущение неверно. Случай  $Ly \geq 0$  сводится к предыдущему случаю заменой  $y$  на  $-y$ . ►

**Замечание 7.10.** При доказательстве теоремы использована связность только  $\omega$ . Связность  $\bar{\omega}$  и  $G_h \setminus \bar{\omega}$  является, вообще говоря, необязательной.

**Следствие 7.1.** Теорема 7.3 верна и при  $\omega = G_h$ . Если сетка  $\omega = G_h$  связна, для любого  $x \in G_h$  выполнены условия положительности коэффициентов,  $Ly \leq 0$  ( $Ly \geq 0$ ),  $D(x_0) > 0$  для некоторого  $x_0 \in G_h$ , то  $y(x) \leq 0$  ( $y(x) \geq 0$ ) для всех  $x \in G_h$ .

◀ Если  $y(x) \neq \text{const}$  в  $G_h$ , то результат следствия является просто иной формулировкой результата теоремы 7.3, так как если предположить  $y > 0$  в  $G_h$ , то в некотором узле в  $G_h$  будет находиться положительный максимум, что невозможно. Если же  $y = \text{const}$ , то в узле  $x_0$

$$Ly(x_0) = D(x_0)y(x_0) \leq 0,$$

откуда  $y(x_0) \leq 0$ . Случай  $Ly \geq 0$  аналогичен рассмотренному. ►

**Следствие 7.2.** Пусть в исходной разностной схеме  $Ly = F$  выполнены условия положительности коэффициентов и  $D(x_0) > 0$  для некоторого  $x_0$ . Тогда разностная схема имеет единственное решение.

◀ Достаточно показать, что однородная задача  $Ly = 0$  имеет только нулевое решение. Но в силу следствия 7.1 ее решение одновременно  $y \leq 0$  и  $y \geq 0$ , что возможно лишь при  $y = 0$ . ►

**Замечание 7.11.** Рассмотрим подробнее узлы, в которых задано решение задачи, т. е. узлы с граничным условием первого рода или начальными данными. Если на каком-то участке границы  $G_h$  заданы такие условия, то для соответствующего узла  $x$  справедливо  $S'(x) = \emptyset$ . Формально эти узлы не участвуют в определении связности сетки. В рамках доказательства принципа максимума такие узлы не могут оказаться узлами вида  $x_0$ , в которых предполагалось наличие экстремума, так как в силу условия о знаке  $Ly(x_0) \leq 0$  получаем  $y(x_0) \leq 0$ . Однако они могут оказаться среди узлов  $\bar{\omega} \setminus \omega$  как элементы шаблонов узлов из  $\omega$ . В результате очевидно, что никаких ограничений на применимость принципа максимума наличие таких узлов не накладывает.

Таким образом, при выполнении указанных условий для обеспечения единственности решения достаточно одного узла с заданным решением.

**Теорема 7.4 (теорема сравнения).** Пусть для всех  $x \in G_h$  выполнены условия положительности коэффициентов и  $D(x_0) > 0$  для некоторого  $x_0 \in G_h$ . Рассмотрим две задачи:  $Ly = F$  и  $LY = \bar{F}$ . Тогда, если  $|F| \leq \bar{F}$  для всех  $x \in G_h$ , то  $|y| \leq Y$ .

◀ В силу линейности задачи имеем оценку  $L(y + Y) = F + \bar{F} \geq 0$ ,  $L(Y - y) = \bar{F} - F \geq 0$ , откуда  $y + Y \geq 0$ ,  $Y - y \geq 0$ , т. е.  $-Y \leq y \leq Y$ . Следовательно,  $|y| \leq Y$ . ►

Данная теорема далее будет использоваться для сравнения решений разностных схем, сеточная функция  $Y = Y(x)$  при этом будет называться *мажорантой*.

## 7.7. Устойчивость разностных схем

Наиболее сложным для исследования является свойство *устойчивости* разностной схемы, в особенности важное для эволюционных задач, которые мы и будем рассматривать. Ограничимся *двухслойными разностными схемами*.

### 7.7.1. Применение принципа максимума к исследованию устойчивости по граничным условиям первого рода и начальным данным

Рассмотрим схему

$$A_h y = \varphi \text{ в } G_h, \quad R_h y = \nu \text{ на } \Gamma_h.$$

Выделим на  $\Gamma_h$  начальную гиперплоскость  $t = 0$  и границу пространственной области с граничными условиями первого рода, которые аппроксимируются точно, обозначив эти подмножества через  $\Gamma_h^*$ . Рассмотрим две разностные задачи I, II, отличающиеся данными на  $\Gamma_h^*$ :

$$Ly^I = F \text{ на } (G_h \cup \Gamma_h) \setminus \Gamma_h^*, \quad y^I = \nu^I \text{ на } \Gamma_h^*;$$

$$Ly^{II} = F \text{ на } (G_h \cup \Gamma_h) \setminus \Gamma_h^*, \quad y^{II} = \nu^{II} \text{ на } \Gamma_h^*;$$

при этом  $Ly^I = Ly^{II}$  на  $(G_h \cup \Gamma_h) \setminus \Gamma_h^*$ . Отметим, что оператор  $L$  равен операторам  $A_h$  и  $R_h$  на соответствующих подмножествах.

**Теорема 7.5.** Пусть для любого узла связной сетки на  $(G_h \cup \Gamma_h) \setminus \Gamma_h^*$  коэффициенты разностной схемы удовлетворяют *условию положительности*. Тогда для решений задач I и II справедлива оценка

$$\|y^I - y^{II}\|_{C(G_h \cup \Gamma_h)} \leq \|\nu^I - \nu^{II}\|_{C(\Gamma_h^*)}.$$

◀ Для разности решений  $y = y^I - y^{II}$  получаем задачу  $Ly = 0$  на  $(G_h \cup \Gamma_h) \setminus \Gamma_h^*$ ,  $y = \nu^I - \nu^{II}$  на  $\Gamma_h^*$ . Наряду с ней рассмотрим задачу  $LY = 0$  на  $(G_h \cup \Gamma_h) \setminus \Gamma_h^*$  и  $Y = \alpha$  на  $\Gamma_h^*$ ,  $\alpha = \|\nu^I - \nu^{II}\|_{C(\Gamma_h^*)}$ . Тогда выполнены все условия теоремы сравнения, откуда получаем оценку  $|y| \leq Y$ . Для  $v = \alpha - Y$  имеем  $Lv = D(x)\alpha \geq 0$ ,  $v = 0$  на  $\Gamma_h^*$ . Отсюда  $v(x) \geq 0$ . Следовательно,  $Y(x) \leq \alpha$  и  $|y(x)| \leq Y(x) \leq \alpha = \|\nu^I - \nu^{II}\|_{C(\Gamma_h^*)}$ . ►

**Пример 7.9.** Полностью неявная разностная схема для уравнения теплопроводности удовлетворяет всем приведенным условиям. Следовательно, ее решение устойчиво по начальным данным и граничным условиям первого рода.

### 7.7.2. Признаки равномерной устойчивости

Рассмотрим влияние данных на начальной гиперплоскости  $t = 0$ .

**Теорема 7.6.** Если  $A_h y^I = A_h y^{II}$ , то для равномерной устойчивости по начальным данным достаточно, чтобы на любом временном слое выполнялось неравенство

$$\|\hat{y}^I - \hat{y}^{II}\| \leq (1 + C\tau) \|y^I - y^{II}\|, \quad C \geq 0, \quad \tau = \hat{t} - t.$$

◀ Ошибка  $\delta y = y^I - y^{II}$  возрастает при переходе на слой  $\hat{t}$  в  $(1 + C\tau) \leq \exp\{C\tau\}$  раз. Для перехода со слоя  $t_j$  на слой  $t_k$  нужно сделать  $m = (t_j - t_k)/\tau$  шагов. Отсюда

$$\|\delta y(t_k)\| \leq \exp\{C\tau m\} \|\delta y(t_j)\| \leq \exp\{CT\} \|\delta y(t_j)\|,$$

т. е.  $K = \exp\{CT\}$ . ▶

**Замечание 7.12.** Если  $T \rightarrow \infty$ , то, очевидно, устойчивость имеет место лишь при  $C = 0$ . Вообще же допустимое значение  $C$  должно соотноситься с характером самого решения, так как чаще всего важна не абсолютная ошибка  $\|\delta y\|$ , а относительная  $\|\delta y\|/\|y\|$ , которая не должна возрастать с течением времени. Часто при  $y \sim \exp(C_0 t)$  выполняется соотношение

$$\frac{\|\delta y\|}{\|y\|} \sim \exp\{(C - C_0)T\}.$$

Если  $\exp\{(C - C_0)T\}$  много больше единицы, то схема слабо устойчива, если много меньше единицы — хорошо устойчива, при  $T \rightarrow \infty$  и  $C \leq C_0$  схема называется асимптотически устойчивой.

**Теорема 7.7.** Пусть схема  $A_h y = \varphi$  равномерно устойчива по начальным данным,  $A_h y^l = \varphi^l$ ,  $l = I, II$ . При этом, если  $y^I = y^{II}$  на некотором слое, то на следующем выполняется неравенство  $\|y^I - y^{II}\| \leq \alpha\tau \|\varphi^I - \varphi^{II}\|$ ,  $\alpha = \text{const}$ . Тогда схема устойчива по правой части.

◀ Введем последовательность сеточных функций  $w_m(t)$ , определенных при  $t \geq t_{m-1}$  (здесь  $t_0 = 0$ ), следующим образом:  $w_1(0) = y^I(0)$ ,  $w_{m+1}(t_m) = w_m(t_m)$ ,  $m = 1, 2, \dots$ , причем

$$A_h w_0 = \varphi^I, \quad t \geq 0, \quad A_h w_m = \begin{cases} \varphi^{II}, & t_{m-1} \leq t \leq t_m, \\ \varphi^I, & t > t_m, \end{cases} \quad m = 1, 2, \dots, M.$$

Поясним подробнее (рис. 7.9): при  $m = 0$   $w_0(0) = y^I(0)$ , при  $m = 1$   $w_1(0) = y^I(0) = y^{II}(0)$ , т. е. считаем, что начальные данные совпадают, так как предметом рассмотрения в теореме является устойчивость по правой части. Следовательно,

$$A_h w_1 = \begin{cases} \varphi^{II}, & 0 = t_0 \leq t \leq t_1; \\ \varphi^I, & t > t_1, \end{cases}$$

откуда  $w_1 \equiv y^{II}$  при  $t \leq t_1$ . При  $m = 2$  получаем

$$w_2(t_1) = w_1(t_1), \quad A_h w_2 = \begin{cases} \varphi^{II}, & t_1 \leq t \leq t_2; \\ \varphi^I, & t > t_2, \end{cases}$$

откуда  $w_2 \equiv y^{II}$  при  $t \leq t_2$  и т. д.

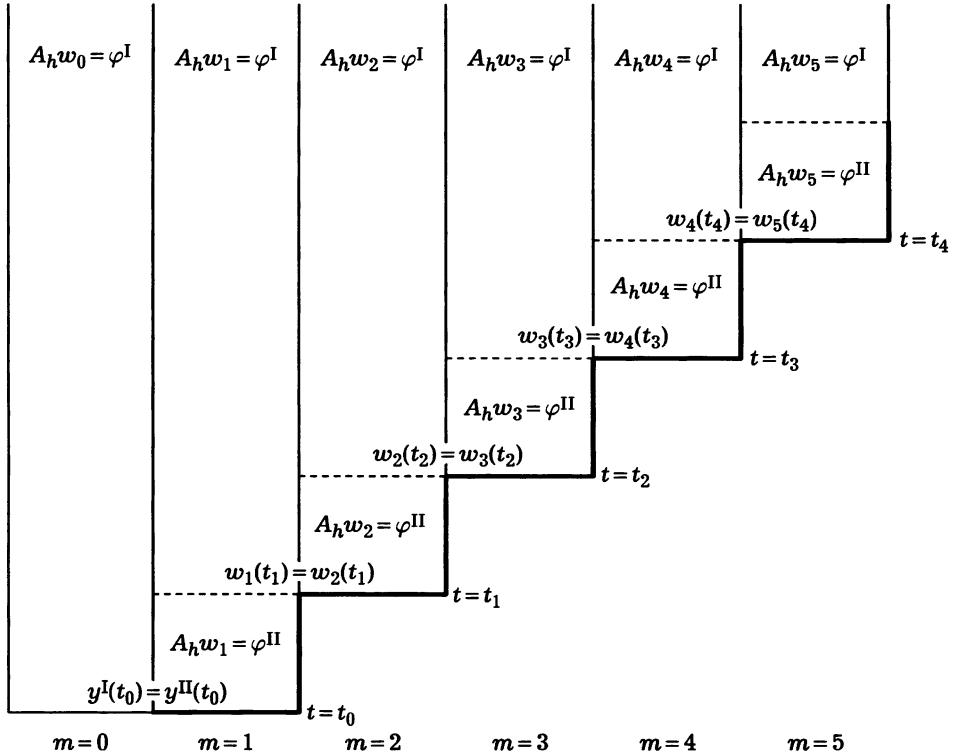


Рис. 7.9

При  $t \leq t_m$  справедливо равенство  $w_m \equiv y^{\text{II}}$ . Функции  $w_m$  и  $w_{m+1}$  совпадают на слое  $t_m$  и отличаются лишь правой частью в описываемых их уравнениях при  $t_m \leq t < t_{m+1}$ . Следовательно, по условиям теоремы справедливо неравенство

$$\|w_m(t_{m+1}) - w_{m+1}(t_{m+1})\| \leq \alpha \tau \|\varphi^{\text{I}} - \varphi^{\text{II}}\|.$$

При  $t \geq t_{m+1}$  эти функции удовлетворяют разностным схемам с одной правой частью, но разными начальными данными. Поэтому из условия равномерной устойчивости по начальным данным верна оценка

$$\|w_m(t_M) - w_{m+1}(t_M)\| \leq K \alpha \tau \|\varphi^{\text{I}} - \varphi^{\text{II}}\|.$$

Далее запишем  $w_0 = y^{\text{I}}$ . Отсюда

$$\begin{aligned} \|y^{\text{II}}(t_M) - y^{\text{I}}(t_M)\| &= \left\| \sum_{m=0}^{M-1} (w_{m+1} - w_m) \right\| \leq \sum_{m=0}^{M-1} \|w_{m+1} - w_m\| \leq \\ &\leq K \alpha \tau M \|\varphi^{\text{I}} - \varphi^{\text{II}}\| = K T \alpha \|\varphi^{\text{I}} - \varphi^{\text{II}}\|, \end{aligned}$$

что и требовалось доказать. ►

**Следствие 7.3.** Если для любых временных слоев справедливо

$$\|\hat{y}^I - \hat{y}^{II}\| \leq (1 + C\tau) \|y^I - y^{II}\| \text{ при } A_h y^I = A_h y^{II},$$

$$\|\hat{y}^I - \hat{y}^{II}\| \leq \alpha \tau \|\varphi^I - \varphi^{II}\| \text{ при } y^I = y^{II},$$

то разностная схема равномерно устойчива по начальным данным и правой части.

Данное следствие просто объединяет формулировки двух последних теорем.

Рассмотрим применение теорем 7.6 и 7.7 к случаю линейных двухслойных разностных схем. Запишем уравнение для  $i$ -го узла в виде

$$\sum_k \alpha_k \hat{y}_{+k} = \sum_l \beta_l y_{+l} + \varphi,$$

где суммирование проводится по узлам шаблона схемы на верхнем ( $k$ ) и нижнем ( $l$ ) слоях. Обозначим через  $\alpha_{k_0}$  величину, такую, что

$$|\alpha_{k_0}| = \max_k |\alpha_k|.$$

Будем для простоты рассматривать случай независимых от  $i$  коэффициентов  $\alpha_k$ ,  $\beta_k$ .

**Теорема 7.8.** Двухслойная разностная схема равномерно устойчива по начальным данным, если

$$(1 + C\tau) \left( |\alpha_{k_0}| - \sum_{k \neq k_0} |\alpha_k| \right) \geq \sum_l |\beta_l|, \quad C = \text{const}, \quad C \geq 0.$$

Схема устойчива по правой части, если при этом выполнено условие

$$|\alpha_{k_0}| - \sum_{k \neq k_0} |\alpha_k| \geq \frac{\chi}{\tau}, \quad \chi = \text{const}.$$

◀ Начнем с доказательства устойчивости по начальным данным. Считаем правую часть  $\varphi = \text{const}$ , а в решение  $y$  внесем ошибку  $\delta y$ . Тогда ошибка решения на следующем временном слое есть решение уравнения

$$\sum_k \alpha_k \delta \hat{y}_{+k} = \sum_l \beta_l \delta y_{+l}.$$

Рассмотрим точку  $i$ , такую, что  $|\delta \hat{y}_{+k_0}| = \|\delta \hat{y}\|_C$ . Тогда

$$|\alpha_{k_0}| |\delta \hat{y}_{+k_0}| \leq \left| \sum_l \beta_l \delta y_{+l} - \sum_{k \neq k_0} \alpha_k \delta \hat{y}_{+k} \right|,$$

откуда

$$|\alpha_{k_0}| \|\delta\hat{y}\|_C \leq \|\delta\hat{y}\|_C \sum_{k \neq k_0} |\alpha_k| + \|\delta y\|_C \sum_l |\beta_l|$$

и, следовательно,

$$\|\delta\hat{y}\|_C \left( |\alpha_{k_0}| - \sum_{k \neq k_0} |\alpha_k| \right) \leq \|\delta y\|_C \sum_l \beta_l \leq (1 + C\tau) \left( |\alpha_{k_0}| - \sum_{k \neq k_0} |\alpha_k| \right) \|\delta y\|_C.$$

В результате получим

$$\|\delta\hat{y}\|_C \leq (1 + C\tau) \|\delta y\|_C,$$

т.е. выполнение условия теоремы 7.6. В итоге имеем равномерную устойчивость по начальным данным.

Рассмотрим доказательство устойчивости по правой части. Пусть внесено возмущение в правую часть  $\varphi$ . Тогда для возмущения решения на новом слое имеем уравнение

$$\sum_k \alpha_k \delta\hat{y}_{+k} = \delta\varphi,$$

из которого точно такими же рассуждениями получаем

$$\|\delta\hat{y}\|_C \left( |\alpha_{k_0}| - \sum_{k \neq k_0} |\alpha_k| \right) \leq \|\delta\varphi\|_C,$$

откуда

$$\|\delta\hat{y}\|_C \leq \frac{\tau}{\chi} \|\delta\varphi\|_C,$$

т.е. выполнение условий теоремы 7.7. Таким образом, требуемый результат получен. ►

**Замечание 7.13.** Условия теоремы 7.8 являются достаточными. Их невыполнение не означает неустойчивости схемы.

**Замечание 7.14.** По существу теорема 7.8 представляет собой некоторый вариант принципа максимума, но здесь не требуется положительности коэффициентов схемы.

**Замечание 7.15.** Вариант теоремы 7.8 справедлив и в случае задач с непостоянными коэффициентами.

**Замечание 7.16.** Рассмотренная двухслойная схема имеет тот же самый вид и в граничных узлах, где  $\varphi$  — краевые условия. Следовательно, речь в теореме идет и об устойчивости по граничным данным.

### 7.7.3. Использование метода разделения переменных

Рассмотрим применение этого метода к *двуухслойным линейным разностным схемам*, записанным в канонической форме вида

$$By_t + Ay = \varphi,$$

где  $A, B$  — линейные разностные операторы, действующие по пространственным переменным. Очевидно, что любую линейную двухслойную (в том числе неявную) разностную схему можно записать в таком виде с помощью замены  $\hat{y} = y + \tau y_t$ .

**Теорема 7.9.** Пусть операторы  $A, B$  не зависят от  $t$ , причем  $A = A^* > 0, B = B^* > 0$ . Тогда двухслойная разностная схема устойчива по начальным данным в норме  $\|\cdot\|_A = \sqrt{(A \cdot, \cdot)}$  при

$$B \geq \frac{\tau}{2} A.$$

◀ Обозначим через  $\mu_k$  обобщенные собственные функции, являющиеся решениями следующей задачи:

$$A\mu_k = \lambda_k B\mu_k.$$

Они ортогональны, т. е.  $(B\mu_k, \mu_l) = \delta_{kl}$ , и образуют базис. Будем искать решение в виде

$$\hat{y} = \sum_k \hat{c}_k \mu_k, \quad y = \sum_k c_k \mu_k.$$

Отсюда

$$\sum_k \frac{\hat{c}_k B\mu_k - c_k B\mu_k}{\tau} + \sum_k c_k A\mu_k = 0.$$

В результате получаем

$$\sum_k \left( \frac{\hat{c}_k - c_k}{\tau} + \lambda_k c_k \right) B\mu_k = 0.$$

После домножения последнего выражения на  $\mu_l$  из условия ортогональности собственных функций имеем

$$\frac{\hat{c}_k - c_k}{\tau} + \lambda_k c_k = 0,$$

или  $\hat{c}_k = (1 - \tau\lambda_k)c_k$ , т. е.

$$\hat{y} = \sum_k (1 - \tau\lambda_k)c_k\mu_k,$$

откуда

$$\|\hat{y}\|_A^2 = (A\hat{y}, \hat{y}) = \sum_k (1 - \tau\lambda_k)^2 \lambda_k c_k^2 \leq \max_k (1 - \tau\lambda_k)^2 \|y\|_A^2.$$

В результате имеем нужную оценку  $\|\hat{y}\|_A \leq \|y\|_A$  (неравенство из теоремы 7.6 с коэффициентом  $C = 0$ ), если  $|1 - \tau\lambda_k| \leq 1$  для всех  $k$ , т. е.  $-1 \leq 1 - \tau\lambda_k \leq 1$  или  $0 \leq \lambda_k \leq 2/\tau$ . Покажем, что это условие эквивалентно неравенству  $B \geq \tau A/2$ , для чего вычислим скалярное произведение

$$\left( \left( B - \frac{\tau}{2}A \right) y, y \right) = \left( \sum_k c_k \left( 1 - \frac{\tau}{2}\lambda_k \right) B\mu_k, \sum_k c_k\mu_k \right) = \sum_k c_k^2 \left( 1 - \frac{\tau}{2}\lambda_k \right) \geq 0.$$

Из него получаем, что из неравенства  $0 \leq \lambda_k \leq 2/\tau$  следует неравенство  $B - \tau/2A \geq 0$ . И наоборот, из  $B - \tau/2A \geq 0$  следует, что для всех  $k$  справедливо  $1 - \tau/2\lambda_k \geq 0$ , т. е.  $0 \leq \lambda_k \leq 2/\tau$  ( $\lambda_k > 0$  в силу положительности операторов  $A$  и  $B$ ). ►

#### 7.7.4. Необходимый «спектральный» признак устойчивости схемы по начальным данным

По сути, данный признак есть вариант метода разделения переменных в случае пренебрежения граничными условиями, а именно: условие устойчивости схемы по начальным данным  $\|y\|_Y \leq M\|\mu\|_\mu$ , где  $\mu = y^0$ , должно быть выполнено для произвольного  $\mu$ , в том числе и для начальных данных специального вида  $\mu_i = \exp(i\varphi)$  — так называемой **гармоники**. Будем искать частное решение задачи в виде  $y_i^j = \rho^j \exp(i\varphi)$ , где  $\varphi$  — некоторая постоянная. В этих выражениях  $i$  — мнимая единица,  $i^2 = -1$ . Подставим такое решение в схему  $By_t + Ay = 0$ , откуда после сокращения на  $\rho^j \exp(i\varphi)$  имеем уравнение для  $\rho$ . Очевидно, что схема может быть устойчивой лишь при  $|\rho| \leq 1 + C\tau$ ,  $C$  не должно зависеть от  $\varphi$ .

Этот признак устойчивости называют **признаком Неймана** или **спектральным признаком**. Последнее название определяется тем, что  $\rho$  есть собственное значение (с оговоркой о граничных условиях) оператора перехода  $\hat{y} = y - \tau B^{-1}Ay = (E - \tau B^{-1}A)y$ , а  $\exp(i\varphi)$  — его собственная функция.

Рассматриваемый признак является лишь необходимым признаком устойчивости, так как данная гармоника не учитывает граничных условий. К тому же ограниченность частного решения специального вида совсем не гарантирует ограниченности решения задачи. В то же время невыполнение спектрального признака устойчивости является достаточным условием неустойчивости, так как любые начальные условия содержат подобные гармоники. Если  $|\rho| \geq 1 + C\tau$ , то амплитуды гармоник растут и очень быстро ведут к переполнению арифметического устройства ЭВМ.

Описанный метод исследования устойчивости называют **методом гармоник**.

**Пример 7.10.** В случае явной разностной схемы для уравнения теплопроводности  $y_t = ky_{xx}$  получим

$$\rho = 1 - \frac{4k\tau}{h^2} \sin^2 \frac{\varphi}{2}.$$

Таким образом, устойчивость с коэффициентом  $C = 0$  (см. теорему 7.6) может иметь место лишь при  $4k\tau/h^2 \leq 2$ , т. е.  $\tau \leq h^2/(2k)$ .

### 7.7.5. Метод энергетических неравенств

На наш взгляд, это наиболее общий и мощный метод исследования устойчивости. Для знакомства с ним рассмотрим простейший признак устойчивости двухслойных разностных схем.

**Теорема 7.10.** Пусть  $A = A^* > 0$ ,  $B = B^* > 0$  не зависят от временного слоя и  $B \geq \frac{\tau}{2}A$ . Тогда схема  $By_t + Ay = 0$  устойчива по начальным данным в энергетической норме, при этом для решения однородного уравнения справедлива оценка  $\|\hat{y}\|_A \leq \|y\|_A$ .

◀ По определению,  $\|y\|_A^2 = (Ay, y)$ . Запишем схему в виде  $y_t + B^{-1}Ay = 0$ . Сделаем замену  $\eta = A^{1/2}y$ . Тогда

$$\dot{\eta} - \eta + \tau A^{1/2}B^{-1}A^{1/2}\eta = 0,$$

или

$$\dot{\eta} = S\eta, \quad S = E - \tau A^{1/2}B^{-1}A^{1/2}.$$

Нужно доказать, что  $\|S\| \leq 1$ . Из условия теоремы  $B \geq \frac{\tau}{2}A$  и положительности  $A, B$  следует, что  $A^{-1} \geq \frac{\tau}{2}B^{-1}$ , откуда  $A^{-1/2} \geq \frac{\tau}{2}B^{-1}A^{1/2}$ , или  $E \geq \frac{\tau}{2}A^{1/2}B^{-1}A^{1/2}$ . Следовательно,  $S = E - \tau A^{1/2}B^{-1}A^{1/2}\eta \geq -E$  и  $S \leq E$ , так как  $A > 0$ ,  $B > 0$ . Таким образом,  $\|S\| \leq 1$  и  $\|\hat{\eta}\| \leq \|\eta\|$ . Но

$$\|\eta\|^2 = (\eta, \eta) = (A^{1/2}y, A^{1/2}y) = (Ay, y) = \|y\|_A^2.$$

Приведем другое доказательство. Умножим скалярно однородное уравнение исследуемой разностной схемы на  $y_t$  и проделаем элементарные преобразования. В результате получим

$$(By_t, y_t) + (Ay, y_t) = ((B - 0,5\tau A)y_t, y_t) + (0,5\tau Ay_t + Ay, y_t) = 0.$$

Воспользуемся соотношением

$$0,5\tau Ay_t + Ay = 0,5A(\hat{y} + y).$$

Его использование совместно с условиями самосопряженности и независимости оператора  $A$  от номера временного слоя дает

$$\begin{aligned} (0,5\tau Ay_t + Ay, y_t) &= (0,5A(\hat{y} + y), y_t) = \\ &= 0,5\tau^{-1}((A\hat{y}, \hat{y}) - (Ay, y)) = 0,5(\|y\|_A^2)_t. \end{aligned}$$

В результате получаем следующее тождество:

$$((B - 0,5\tau A)y_t, y_t) + 0,5(\|y\|_A^2)_t = 0.$$

Поскольку по условиям теоремы первое слагаемое неотрицательно, то второе неположительно. Следовательно, операторная норма решения  $\|y\|_A$  не растет от слоя к слою, что и означает выполнение требуемого нами условия устойчивости. ►

**Замечание 7.17.** Приведенное второе доказательство ясно показывает, что условия теоремы, наложенные на оператор  $B$ , являются избыточными.

**Замечание 7.18.** Из приведенного доказательства с очевидностью следует, что неравенство  $B \geq \frac{\tau}{2}A$  является необходимым и достаточным условием устойчивости рассмотренного разностного уравнения в операторной норме.

**Замечание 7.19.** Примеры применения этой теоремы будут в последующих главах, хотя самый простой и естественный уже приведен (см. 7.7.3).

## 7.8. Библиографические комментарии

Математические модели, в которых используются уравнения в частных производных и для нахождения численного решения которых разрабатываются численные алгоритмы, описаны во множестве книг. Укажем в качестве основных [27, 50, 70, 99, 100, 102, 108, 170]. В них же описаны свойства решений основных задач математической физики.

Разностные методы решения уравнений математической физики также рассматриваются во многих руководствах, первоочередными из которых являются [6, 12, 14, 22, 25, 28, 80, 84, 110, 113, 128, 134, 172, 176, 183, 187].

Выделим отдельно книги [139–141, 149, 151, 154, 155, 171], в которых наиболее полно изложена теория разностных схем. Отметим также монографию [150], полностью посвященную устойчивости разностных схем, включая трехслойные, и книги [144, 146, 152, 155], посвященные математическому и численному моделированию различных физико-механических явлений. Например, в [139] подробно изучен вопрос о необходимости обеспечения условия консервативности разностной схемы на примере уравнения второго порядка с разрывным коэффициентом. При этом в неверной схеме учтена и производная коэффициента.

Помимо многих теоретических сведений и результатов в [137] содержится исторический обзор работ по теории устойчивости.

Мы лишь касаемся столь важного метода решения уравнений математической физики, как метод конечных элементов. Его описание, теорию и практику применения можно найти в следующих книгах: [11, 23, 45, 60, 66, 72, 73, 78, 93, 114, 117, 122, 157, 158, 165, 166, 189].

## 8. ЧИСЛЕННОЕ РЕШЕНИЕ ПАРАБОЛИЧЕСКИХ УРАВНЕНИЙ

В главе содержатся описание и исследование разностных схем для решения одномерного *уравнения теплопроводности* — уравнения параболического типа. Подробно изучена *схема с весами*. Описана схема для решения нелинейного уравнения. Рассмотрены примеры нестандартных *схем*, в том числе *трехслойные*.

### 8.1. Линейное одномерное уравнение теплопроводности с постоянными коэффициентами. Схема с весами

Рассмотрим уравнение *теплопроводности*

$$u_t = ku_{xx} + f, \quad 0 < x < l, \quad 0 < t < T,$$

с начальными и граничными условиями

$$\begin{aligned} u(x, 0) &= u_0(x), \\ u(0, t) &= \mu_1(t), \quad u(l, t) = \mu_2(t). \end{aligned}$$

Считаем, что  $k = \text{const} > 0$ .

Для решения этого уравнения построим двухслойную *схему с весами* на обычном шеститочечном (см. рис. 7.2) шаблоне и равномерной сетке  $\bar{\omega}_h$ :

$$y^{(\sigma)} = \sigma \hat{y} + (1 - \sigma)y,$$

где  $\sigma$  — вес. Например,  $y^{(1)} = \hat{y}$  при  $\sigma = 1$ ,  $y = y^{(0)}$  при  $\sigma = 0$ . Считаем, что начальные и граничные условия аппроксимируются точно. Тогда схема имеет вид

$$y_t = ky_{\bar{x}\bar{x}}^{(\sigma)} + \varphi,$$

т. е.

$$\frac{\hat{y} - y}{\tau} = \frac{k}{h^2} (\sigma(\hat{y}_{+1} - 2\hat{y} + \hat{y}_{-1}) + (1 - \sigma)(y_{+1} - 2y + y_{-1})) + \varphi.$$

Та же схема в виде, каноническом для проверки применимости принципа максимума, записывается следующим образом:

$$\left( \frac{1}{\tau} + \frac{2\sigma k}{h^2} \right) \hat{y} = \frac{\sigma k}{h^2} (\hat{y}_{+1} + \hat{y}_{-1}) + \left( \frac{1}{\tau} - \frac{2(1-\sigma)k}{h^2} \right) y + \frac{(1-\sigma)k}{h^2} (y_{+1} + y_{-1}).$$

При  $\sigma \neq 0$  для нахождения решения на новом слое необходимо решить систему линейных алгебраических уравнений (СЛАУ) с трехдиагональной матрицей *методом прогонки*. Видно, что эта СЛАУ имеет строгое диагональное преобладание:

$$\frac{1}{\tau} + \frac{2k\sigma}{h^2} > \frac{2k\sigma}{h^2},$$

поэтому ее решение методом прогонки является корректным.

*Условие положительности коэффициентов* разностной схемы выполнено при  $0 \leq \sigma \leq 1$  и при

$$\frac{1}{\tau} - \frac{2(1-\sigma)k}{h^2} \geq 0,$$

т. е.

$$\sigma \geq 1 - \frac{h^2}{2k\tau}.$$

Внутри сетки коэффициент  $D = 0$ , а на границе области, т. е. в точках с номерами  $i = 0$ ,  $i = n$  или  $j = 0$  выполнено равенство  $D = 1$ .

### 8.1.1. Аппроксимация схемы с весами

Для вычисления ошибки *аппроксимации* подставим в схему проекцию точного решения  $u_h$  в узлах сетки. Используя разностные производные и опустив индекс  $h$ , запишем

$$\psi_h = \varphi + ku_{\bar{x}\bar{x}}^{(\sigma)} - u_t = \varphi + k\tau\sigma u_{t\bar{x}\bar{x}} + ku_{\bar{x}\bar{x}} - u_t,$$

так как  $\hat{u} = u + \tau u_t$ .

Выполняя разложение  $u_h$  в узлах сетки по формуле Тейлора с центром в точке  $(i, j)$ , получим невязку в следующем виде:

$$\psi_h = \varphi + k\tau\sigma u_{txx} + ku_{xx} + kh^2 \frac{1}{12} u_{xxxx} - u_t - \frac{1}{2} \tau u_{tt} + O(\tau^2 + h^4)$$

(здесь используются обычные производные).

Воспользуемся исходным уравнением. При  $k = \text{const}$  получим

$$u_{tt} = ku_{txx} + f_t, \quad u_{txx} = ku_{xxxx} + f_{xx},$$

откуда вытекает равенство

$$u_{tt} = k^2 u_{xxxx} + kf_{xx} + f_t.$$

Следовательно,

$$\begin{aligned} \psi_h &= (\varphi + ku_{xx} - u_t) + k\tau\sigma(ku_{xxxx} + f_{xx}) + \frac{1}{12}kh^2u_{xxxx} - \\ &- \frac{1}{2}\tau(k^2u_{xxxx} + kf_{xx} + f_t) + O(\tau^2 + h^4) = \varphi - f + k\tau\left(\sigma - \frac{1}{2}\right)f_{xx} - \\ &- \frac{1}{2}\tau f_t + k\left(k\tau\left(\sigma - \frac{1}{2}\right) + \frac{1}{12}h^2\right)u_{xxxx} + O(\tau^2 + h^4). \end{aligned}$$

Отсюда следует, что при  $\sigma = 1/2$  погрешность равна

$$\psi_h = \varphi - f - \frac{1}{2}\tau f_t + O(\tau^2 + h^2) = O(\tau^2 + h^2),$$

если  $\varphi_i^j = f(x_i, t_j + 1/2\tau)$ , или  $\varphi = f + 1/2\tau f_t$ . При

$$\sigma = \frac{1}{2} - \frac{h^2}{12k\tau} = \sigma^*$$

погрешность равна

$$\psi_h = \varphi - f - \frac{1}{2}\tau f_t - \frac{1}{12}h^2f_{xx} + O(\tau^2 + h^4) = O(\tau^2 + h^4),$$

если

$$\varphi_i^j = f\left(x_i, t_j + \frac{1}{2}\tau\right) + \frac{1}{12}h^2f_{xx}(x_i, t_j).$$

Если же  $\sigma \neq 1/2$ ,  $\sigma \neq \sigma^*$ ,  $\varphi = f$ , то схема имеет порядок аппроксимации  $O(\tau + h^2)$ .

При вычислении погрешности считаем, что все необходимые производные существуют и ограничены.

Схему с весом  $\sigma = \sigma^*$  обычно называют схемой повышенного порядка аппроксимации (точности).

### 8.1.2. Устойчивость схемы с весами

Рассмотрим условия *устойчивости* данной схемы в различных нормированных пространствах.

**Устойчивость по начальным данным в  $L_2$ .** Запишем схему с однородными граничными условиями и нулевой правой частью в виде

$$y_t - k\sigma\tau y_{t\bar{x}x} = ky_{\bar{x}x}.$$

Для оценки решения воспользуемся *методом разделения переменных*:

$$y = \sum_{j=1}^{n-1} c_j \mu_j,$$

где  $\mu_j$  — собственные функции оператора второй разностной производной:

$$\mu_{j,\bar{x}x} + \lambda_j^2 \mu_j = 0.$$

В результате подстановки получим

$$\sum_{j=1}^{n-1} ((1 + k\sigma\tau\lambda_j^2)c_{j,t} + k\lambda_j^2 c_j) \mu_j = 0.$$

Так как функции  $\mu_j$  ортонормированы, то

$$(1 + k\sigma\tau\lambda_j^2)\hat{c}_j = (1 + k\sigma\tau\lambda_j^2)c_j - k\sigma\tau\lambda_j^2 c_j.$$

Следовательно, коэффициент у  $j$ -й гармоники меняется в

$$\rho_j = \frac{1 - k\lambda_j^2\tau(1 - \sigma)}{1 + k\sigma\tau\lambda_j^2}$$

раз. Отсюда имеем

$$\|\hat{y}\|^2 = \sum_{j=1}^{n-1} \hat{c}_j^2 = \sum_{j=1}^{n-1} \rho_j^2 c_j^2 \leq \|y\|^2$$

при  $|\rho_j| \leq 1$  для произвольного  $j$ .

Поскольку решение исходной дифференциальной задачи со временем убывает (при нулевой правой части), имеет смысл требовать того же и от разностного решения, т. е. выполнения неравенства  $|\rho_j| \leq 1$ .

Допустим, что знаменатель  $\rho_j$  положителен, т. е.  $\sigma > -\frac{1}{k\tau\lambda_j^2}$ . Тогда условие устойчивости можно записать в виде неравенств:

$$-(1 + k\sigma\tau\lambda_j^2) \leq 1 + k\sigma\lambda_j^2\tau - k\lambda_j^2\tau \leq 1 + k\sigma\tau\lambda_j^2.$$

В результате сделанного предположения относительно  $\sigma$  получаем, что правое неравенство справедливо для любого положительного  $\tau$ , а левое неравенство дает условие

$$2k\sigma\lambda_j^2\tau \geq k\lambda_j^2\tau - 2,$$

откуда

$$\sigma \geq \frac{1}{2} - \frac{1}{k\lambda_j^2\tau}.$$

Это неравенство гарантирует выполнение сделанного относительно веса  $\sigma$  предположения. Далее, для любого  $j$  имеем  $\lambda_j^2 \leq \lambda_{n-1}^2 < 4/h^2$  и поэтому

$$\sigma \geq \frac{1}{2} - \frac{h^2}{4k\tau} = \sigma_0.$$

**Пример 8.1.** 1. Рассмотрим явную схему с  $\sigma = 0$ . Тогда условие устойчивости дает  $0 \geq \frac{1}{2} - \frac{h^2}{4k\tau}$ , т. е.  $\tau \leq \frac{h^2}{2k}$ .

2. Пусть  $\sigma \geq 1/2$ . Следовательно, для схем с такими весами устойчивость в  $L_2$  имеет место для любого  $\tau$ .

3. Пусть  $\sigma = \sigma^*$ . В этом случае также имеет место безусловная устойчивость, так как  $\sigma^* > \sigma_0$ . #

Доказанное неравенство для норм  $\hat{y}$  и  $y$  (см. теорему 7.6) гарантирует *равномерную устойчивость* построенной схемы по начальным данным в  $L_2$ .

**Замечание 8.1.** Отметим естественный характер условия устойчивости  $\tau = O(h^2)$  для решения уравнения теплопроводности. Эта «естественность» объясняется следующими причинами.

1. Характером точного решения (например, функцией Грина) задачи, показывающим что теплота распространяется по закону  $x \sim t^{1/2}$ . Следовательно, для обеспечения верного характера распространения теплоты требуется выполнение условия  $h \sim \tau^{1/2}$ .

2. Бесконечностью области зависимости точного решения уравнения теплопроводности в случае задачи Коши на всей числовой прямой. Рассмотрим это подробнее на примере явной схемы. Пусть ее шаблон включает  $m+1$  точек с нижнего слоя. Тогда решение на новом временном слое в каждой своей точке определяется данными с предыдущего

слоя на отрезке длиной  $mh$ . Увеличение числа временных слоев влечет пропорциональный рост размера области зависимости, данные с которой определяют решение в точке. Так, при переходе с нулевого временного слоя на слой, соответствующий моменту времени  $t$  (кратному  $\tau$ ), решение в каждой точке определяется начальными данными с отрезка размером  $mht/\tau$ . Если мы требуем сходимости приближенного решения к точному, то необходимо требовать и роста размера области зависимости при измельчении сетки. Тогда получаем условие  $\tau = o(h)$  при стремлении параметров дискретизации к нулю.

**Устойчивость по правой части.** Пусть теперь в разностной схеме присутствует ненулевая правая часть

$$\varphi = \sum_{j=1}^{n-1} b_j \mu_j.$$

Тогда точно так же, как и в предыдущем случае, получим

$$\hat{c}_j = \rho_j c_j + \frac{\tau}{1 + k\sigma\tau\lambda_j^2} b_j.$$

Отсюда имеем

$$\|\hat{y}\| \leq \max_j |\rho_j| \|y\| + \max_j \left| \frac{\tau}{1 + k\sigma\tau\lambda_j^2} \right| \|\varphi\|.$$

Следовательно, при  $\sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4k\tau}$  и  $\sigma > 0$  получим  $\|\hat{y}\| \leq \|y\| + \tau \|\varphi\|$ , т. е. условие, гарантирующее устойчивость решения по правой части и начальным данным (см. теоремы 7.6 и 7.7).

Если же  $\sigma < 0$ , но

$$\sigma \geq \frac{1}{2} - \frac{(1-\varepsilon)h^2}{4k\tau} = \sigma_\varepsilon, \quad \varepsilon \in (0, 1),$$

то

$$\begin{aligned} 1 + k\sigma\tau\lambda_j^2 &= 1 + k\tau\lambda_j^2(\sigma - \sigma_\varepsilon) + k\tau\lambda_j^2\sigma_\varepsilon \geq \\ &\geq 1 + \frac{1}{2}k\tau\lambda_j^2 - \frac{1-\varepsilon}{4}h^2\lambda_j^2 > 1 - \frac{1-\varepsilon}{4}h^2\frac{4}{h^2} = \varepsilon. \end{aligned}$$

Следовательно,

$$\|\hat{y}\| \leq \|y\| + \frac{\tau}{\varepsilon} \|\varphi\|.$$

Таким образом, для устойчивости по правой части достаточно выполнения условия  $\sigma \geq \sigma_\varepsilon$ , при котором обеспечена и равномерная устойчивость по начальным данным.

**Устойчивость в равномерной норме  $C$ .** Вернемся к исходной форме записи схемы. Из нее следует, что принцип максимума к ней применим и имеет место при условии

$$\sigma \geqslant 1 - \frac{h^2}{2k\tau} = 2\sigma_0.$$

При выполнении этого неравенства, согласно теореме 7.5, имеем **устойчивость решения в норме  $C$**  по начальным данным и граничным условиям.

Устойчивость по правой части  $\varphi$  также имеет место, так как при  $\sigma > 0$

$$|\alpha_{j_0}| - \sum_{j \neq j_0} |\alpha_j| = \frac{1}{\tau} + \frac{2k\sigma}{h^2} - \frac{2\sigma k}{h^2} = \frac{1}{\tau},$$

т. е. выполнены условия теоремы 7.8, в которых участвуют лишь коэффициенты при  $\hat{y}$ .

**Замечание 8.2.** Разностная схема с весом  $\sigma = 1/2$ , которая является безусловно устойчивой в норме  $L_2$ , согласно полученным условиям является устойчивой в норме  $C$  лишь при выполнении ограничения  $\tau \leq h^2/k$ . Однако эта оценка получена из принципа максимума. Известно, что принцип максимума дает достаточные условия устойчивости. В действительности рассматриваемая схема также является безусловно устойчивой и в равномерной норме. В общем случае равномерная норма оператора перехода больше единицы. При этом схема является монотонной только при условии справедливости принципа максимума.

**Асимптотическая устойчивость.** При больших значениях времени  $T \rightarrow \infty$  для получения качественно верного поведения численного решения необходимо, чтобы возможные возмущения росли медленнее растущего решения или убывали быстрее убывающего решения. Данное свойство выражается в такой характеристике схемы, как **асимптотическая устойчивость**. Это формально может быть описано следующими соотношениями: при изменении ошибки по закону  $\|\delta u\| \sim \exp(Ct)$  и самого решения по закону  $u \sim \exp(C_0 t)$  должно выполняться условие  $C \leq C_0$ . То же условие может быть переписано другим способом: с использованием множителей роста (уменьшения) решения от слоя к слою.

Рассмотрим схему с весами.

Известно, что решение рассматриваемого уравнения теплопроводности при больших значениях времени определяется первой гармоникой

$u \sim \exp\left(-k \frac{\pi^2}{l^2} t\right)$ . Таким образом, за один переход со слоя на слой решение меняется в

$$p_1 = \exp\left(-k \frac{\pi^2}{l^2} \tau\right) = 1 - k \frac{\pi^2}{l^2} \tau + O(\tau^2)$$

раз.

Множитель перехода  $j$ -й гармоники разностного решения равен

$$\rho_j = \frac{1 - k\lambda_j^2 \tau(1 - \sigma)}{1 + k\sigma\tau\lambda_j^2} = 1 - \frac{k\lambda_j^2 \tau}{1 + k\sigma\tau\lambda_j^2} = 1 - \frac{\sin^2 \frac{\pi j h}{2l}}{\frac{h^2}{4k\tau} + \sigma \sin^2 \frac{\pi j h}{2l}}.$$

Нетрудно видеть, что максимальное значение  $\rho_j$  достигается при  $j = 1$ , а минимальное — при  $j = n - 1$ .

Для установления свойства асимптотической устойчивости достаточно ограничиться выражениями с точностью  $O(\tau^2)$ . Такая величина при возведении в степень, обратно пропорциональную  $\tau$ , равна единице с точностью  $O(\tau)$ .

Исходя из изложенного для обеспечения асимптотической устойчивости требуется выполнить для любого номера  $j$  следующую систему неравенств:

$$-1 + k \frac{\pi^2}{l^2} \tau \leq 1 - \frac{\sin^2 \frac{\pi j h}{2l}}{\frac{h^2}{4k\tau} + \sigma \sin^2 \frac{\pi j h}{2l}} \leq 1 - k \frac{\pi^2}{l^2} \tau.$$

Для малых значений  $h$  имеем

$$\rho_1 \approx 1 - \frac{\left(\frac{\pi h}{2l}\right)^2}{\frac{h^2}{4k\tau} + \sigma \left(\frac{\pi h}{2l}\right)^2} \approx 1 - k \frac{\pi^2}{l^2} \tau.$$

Таким образом, правое неравенство выполнено.

При рассмотрении левого неравенства используем множитель  $\rho_{n-1}$ , считая, что  $\sin^2 \frac{\pi(n-1)h}{2l} \approx 1$ .

Решив полученное неравенство, запишем следующее условие асимптотической устойчивости:

$$\sigma \geq \frac{1}{2} + k \frac{\pi^2}{4l^2} \tau - \frac{h^2}{4k\tau}.$$

Данное условие является более жестким, чем обычное условие устойчивости в  $L_2$ , в котором вес должен быть не меньше  $\sigma_0$ . Выражение

в правой части, являющееся функцией  $\tau$ , линейно растет до плюс бесконечности при больших значениях шага и убывает до минус бесконечности при стремлении шага к нулю. Следовательно, безусловно асимптотически устойчивых схем в данном классе схем не существует.

Условие для веса  $\sigma$  может быть переписано в виде условия для шага  $\tau$ .

**Пример 8.2.** Рассмотрим симметричную схему с  $\sigma = 1/2$ . Из полученного результата имеем следующее условие асимптотической устойчивости:  $\tau \leq \frac{lh}{\pi k}$ . Напомним, что в норме  $L_2$  данная схема является безусловно устойчивой (по начальным данным и по правой части).

**Метод энергетических неравенств.** Запишем схему в канонической форме  $By_t + Ay = 0$ . Тогда, согласно теореме 7.10, для устойчивости решения разностной схемы по начальным данным в энергетической норме достаточно выполнения неравенства

$$B \geq \frac{\tau}{2} A.$$

При этом операторы  $A$  и  $B$  постоянны.

В данном случае  $Ay = -ky_{\bar{x}x}$ ,  $B = E + \tau\sigma A$ . Следовательно, условие устойчивости имеет вид

$$E \geq \tau \left( \frac{1}{2} - \sigma \right) A.$$

Поэтому устойчивость имеет место для всех  $\tau$ , если  $\sigma \geq 1/2$ .

Запишем неравенство (см. 7.5)

$$k \frac{9}{l^2} \|y\|^2 \leq (Ay, y) \leq k \frac{4}{h^2} \|y\|^2.$$

Отсюда при  $\sigma < 1/2$  получаем условие

$$\tau \left( \frac{1}{2} - \sigma \right) \frac{4k}{h^2} \leq 1$$

и, следовательно,

$$\sigma \geq \frac{1}{2} - \frac{h^2}{4k\tau}.$$

В итоге имеем то же самое условие, что и условие устойчивости по начальным данным в  $L_2$ . Оно объединяет оба рассмотренных случая:  $\sigma \geq 1/2$  и  $\sigma < 1/2$ .

Используя **метод энергетических неравенств**, можно показать и устойчивость по правой части при

$$\sigma \geq \sigma_\varepsilon = \frac{1}{2} - \frac{(1-\varepsilon)h^2}{4\tau k}.$$

### 8.1.3. Сходимость и точность схемы с весами

Как показано в 8.1, в разных пространствах условия устойчивости различны.

1. Рассмотрим оценки в пространстве  $L_2$ . Если  $\sigma \geq \sigma_0 = \frac{1}{2} - \frac{h^2}{4k\tau}$ , то при  $\sigma > 0$  (или при  $\sigma \geq \sigma_\epsilon$ ) решение схемы с весами устойчиво как по начальным данным, так и по правой части. Следовательно, имеет место оценка  $\|y\| \leq M_1\|u_0\| + M_2\|\varphi\|$ , где фигурируют интегральные нормы.

Для определения ошибки разностного решения  $z = y - u_h$  имеем ту же разностную задачу, правая часть которой равна погрешности аппроксимации  $\psi_h$ . Таким образом, из устойчивости по правой части имеем оценку  $\|z\| \leq M_2\|\varphi\|$ .

Отсюда получаем следующие характеристики сходимости разностной схемы:

- а) при  $\sigma = 1/2$ :  $\|y - u_h\| = O(\tau^2 + h^2)$ ;
- б) при  $\sigma = \sigma^* = \frac{1}{2} - \frac{h^2}{12k\tau}$ :  $\|y - u_h\| = O(\tau^2 + h^4)$ ;
- в) при  $\sigma \neq 1/2$ ,  $\sigma \neq \sigma^*$ :  $\|y - u_h\| = O(\tau + h^2)$ .

4. Рассмотрим оценки в пространстве  $C$ .

Точно такие же скорости сходимости имеют место в норме  $C$  при  $\sigma \geq 1 - \frac{h^2}{2k\tau}$ , но при этом даже в случае  $\sigma = 1/2$  данное условие накладывает на  $\tau$  ограничение  $\tau \leq \frac{h^2}{k}$ .

Как уже отмечалось в 8.1.2, условие устойчивости в  $C$  получено из условия выполнимости принципа максимума и является достаточным условием устойчивости. В этом случае схема с весами при  $\sigma \geq 1 - \frac{h^2}{2k\tau}$  является и монотонной. В частности, если начальные и граничные данные и правая часть неотрицательны, то таким же будет и решение на всех временных слоях. Это свойство очень важно для расчета температуры (или концентрации) в сложных задачах.

## 8.2. Некоторые другие задачи и схемы

Ниже представлена двухслойная разностная схема для уравнения теплопроводности с переменным коэффициентом и описана задача для квазилинейного уравнения. Рассмотрена нестандартная двухшаговая разностная схема с условной аппроксимацией. Описаны трехслойные разностные схемы и проведено их краткое исследование.

В данной главе не рассмотрены многомерные задачи для уравнения теплопроводности. Они будут изучены в 10 в качестве одного из средств нахождения решения эллиптических краевых задач. Там же

описаны так называемые *экономичные разностные схемы* для решения уравнения теплопроводности.

В 10 построены аппроксимации оператора Лапласа в криволинейных координатах, необходимые для численного решения задач в таких координатах и для решения уравнения теплопроводности.

### 8.2.1. Задача с переменными коэффициентами

Рассмотрим уравнение теплопроводности с переменным коэффициентом температуропроводности  $k = k(x, t)$ :

$$\begin{aligned} u_t &= (ku_x)_x + f, \quad 0 < x < l, \quad 0 < t < T; \\ u(x, 0) &= u_0(x), \\ u(0, t) &= \mu_1(t), \quad u(l, t) = \mu_2(t). \end{aligned}$$

Для такой задачи нами в 7.3.2 *интегро-интерполяционным методом* построена разностная схема на неравномерной сетке. Если представить ее в виде схемы с весами, то получим

$$y_t = (ay_{\bar{x}}^{(\sigma)})_x + \varphi.$$

Пусть сетка равномерная. Тогда  $A_h y = -(ay_{\bar{x}})_x$  и в соответствии с 7.5 запишем неравенство

$$(A_h y, y) \leq a_{\max} \frac{4}{h^2} \|y\|^2.$$

Применяя энергетический метод, получаем условие *устойчивости*:

$$\sigma \geq \frac{1}{2} - \frac{h^2}{4a_{\max}\tau}.$$

Из этой оценки вытекает следующий принцип: схема должна удовлетворять условию устойчивости для всего интервала изменения  $a$ , реализуемого в данной задаче. Этот принцип часто называют *принципом замороженных коэффициентов*: берется фиксированное значение коэффициента, схема исследуется на устойчивость при этом значении параметра, после чего выбирается такой вес  $\sigma$  (или  $h, \tau$ ), чтобы удовлетворить всему интервалу изменения  $a$ . Данный прием является нестрогим, но часто дает нужные результаты. Используя его, например, на неравномерной сетке можно получить условие

$$\sigma \geq \frac{1}{2} - \frac{h_{\min}^2}{4a_{\max}\tau}.$$

Однако существуют разностные схемы, применение к которым принципа замороженных коэффициентов дает ошибочный результат.

**Пример 8.3.** Рассмотрим в качестве примера задачу для квазилинейного уравнения теплопроводности со степенным коэффициентом  $k = \kappa_0 u^\sigma$  при  $\sigma > 0$ :

$$u_t = (ku_x)_x + f, \quad 0 < x < +\infty, \quad 0 < t < T;$$

$$u(x, 0) = u_0(x), \quad u(0, t) = \mu_0(t), \quad |u(x, t)| < +\infty \text{ при } x \rightarrow +\infty.$$

Такая задача является хорошим приближением для описания многих физико-химических процессов, таких как горение, развитие микроорганизмов и т. п. Одновременно она является хорошим тестом для проверки работоспособности разнообразных методов численного решения. Для выполнения тестовых расчетов необходимо иметь «эталонное» решение, с которым можно сравнивать вычисленное с помощью тестируемого алгоритма. Идеальным вариантом является наличие аналитического решения задачи и выбор его в качестве «эталона».

Обычно поиск тестового решения в сложных задачах осуществляется следующим образом: ищется аналитическое решение рассматриваемого уравнения, после чего задаются те начальные и граничные условия, которые ему должны соответствовать. При нахождении аналитического решения весьма распространенным является поиск *автомодельных решений*, в которых решение — функция двух или более переменных — зависит от них через какую-то одну комбинацию. Тогда решение фактически является функцией одного аргумента, сложным образом зависящим от всех независимых переменных.

Простейшая автомодельная переменная представляет собой переменную  $\xi = Dt - x$ , где  $D$  — некоторый дополнительный параметр, имеющий смысл скорости распространения сигнала в среде. Решение, зависящее от одной переменной  $\xi$ , обычно называют решением типа **бегущей волны**.

Для рассматриваемой задачи нетрудно получить следующее решение:

$$u(x, t) = \begin{cases} u_0 \left( t - \frac{x}{D} \right)^{1/\sigma}, & x \in [0, Dt]; \\ 0, & x \geqslant Dt. \end{cases}$$

Данное решение соответствует нулевым начальным данным и правой части, а также степенному закону изменения граничных условий  $\mu_0(t) = u_0 t^{1/\sigma}$ . При этом скорость  $D$  распространения теплоты в таком решении, называемом тепловой волной, однозначно определяется из уравнения теплопроводности параметрами процесса:  $D = (\kappa_0 u_0^\sigma / \sigma)^{1/2}$ .

Как видно, данное решение представляет собой волну, бегущую по нулевому фону. На фронте волны происходит скачок производной решения при непрерывности самого решения и теплового потока.

Для численного решения данного уравнения применяют в том числе и схемы, описанные в 8.2.1, полученные интегро-интерполяционным методом. При выводе схемы стоит аппроксимировать все входящие в тепловой поток величины с одного временного слоя, что обеспечит лучшую передачу энергетических соотношений. Проще выбирать полностью неявную схему, которая заведомо обеспечит монотонность профиля численного решения для любых шагов сеток по пространству и времени. Возникшую систему нелинейных алгебраических уравнений можно решать методом Ньютона либо методом типа простой итерации, считая коэффициент  $a$  заданным на предыдущей итерации во всех точках сетки. Второй метод намного проще первого, однако потребует малого временного шага для своей реализации.

При выборе способов вычисления коэффициента  $a$  стоит отдать предпочтение выражениям  $a_{i+1/2} = \frac{k(y_i) + k(y_{i+1})}{2}$  или  $a_{i+1/2} = k\left(\frac{ky_i + y_{i+1}}{2}\right)$  для случая равномерной сетки, на неравномерной лучше брать линейную интерполяцию. Обращение решения в нуль на фронте может привести к тому, что иные аппроксимации дадут совершенно неверное решение. Например, варианты  $a_{i+1/2} = (k(y_i)k(y_{i+1}))^{1/2}$  или  $a_{i+1/2} = \frac{2k(y_i)k(y_{i+1})}{k(y_i) + k(y_{i+1})}$  обеспечивают нулевое решение во всех точках, кроме граничной, во все моменты времени.

### 8.2.2. Схема «бегущего» счета для решения уравнения теплопроводности

Рассмотрим одну необычную схему для решения уравнения теплопроводности и продемонстрируем на ней простые приемы исследования схем. Эта схема возникла из желания построить безусловно устойчивую схему второго порядка аппроксимации, которая, будучи неявной, позволяет находить решение так же просто, как и явная схема.

Рассмотрим, как и ранее, задачу с краевыми условиями первого рода. Для перехода с предыдущего слоя на новый введем средний промежуточный слой с номером  $j + 1/2$  (отстоит на  $\tau/2$  от старого) и промежуточное (вспомогательное) решение, обозначаемое через  $\bar{y}$ . Будем переходить на новый слой за два шага. На первом найдем решение  $\bar{y}$ , а на втором — решение  $\hat{y}$ . При переходе используем схемы на зигзагообразных шаблонах. Первый состоит из точек  $(i - 1, j + 1/2)$ ,  $(i, j + 1/2)$ ,  $(i, j)$ ,  $(i + 1, j)$ , второй — из точек  $(i - 1, j + 1/2)$ ,  $(i, j + 1/2)$ ,

$(i, j+1), (i+1, j+1)$ . Решаемая разностная задача имеет вид

$$\frac{2}{\tau}(\bar{y} - y) = \frac{k}{h^2}(\bar{y}_{-1} - \bar{y} - y + y_{+1}),$$

$$\frac{2}{\tau}(\hat{y} - \bar{y}) = \frac{k}{h^2}(\bar{y}_{-1} - \bar{y} - \hat{y} + \hat{y}_{+1}).$$

Из приведенных выражений видно, что на первом этапе происходит вычисление промежуточного решения слева направо непосредственным пересчетом по явным формулам. На втором этапе движение идет, наоборот, справа налево, аналогичным явным образом. Тем самым в решение на новом слое вносятся данные о граничных условиях.

Для анализа данной разностной схемы исключим из уравнений промежуточный слой. С этой целью нужно вычесть одно уравнение схемы из другого. После элементарных преобразований запишем

$$\bar{y} = y^{(1/2)} - \frac{k\tau^2}{4h}y_{xt}.$$

Теперь сложим уравнения схемы и подставим полученное для  $\bar{y}$  выражение. В итоге получим следующее разностное уравнение:

$$y_t = ky_{xx}^{(1/2)} + \frac{k^2\tau^2}{4h^2}y_{xtt}.$$

Из результирующего уравнения видно, что погрешность аппроксимации данной схемы равна  $\psi_h = O(\tau^2 + h^2 + \tau^2/h^2)$ . Она стремится к нулю лишь при  $\tau/h \rightarrow 0$ , т. е. является условной.

Заметим, что подобная схема рассмотрена в 7.4.1. Состояла она всего из одного шага. При этом была получена погрешность аппроксимации  $\psi_h = O(\tau + h^2 + \tau/h)$ . Из чего следует, что наличие двух шагов приводит к увеличению порядка аппроксимации за счет того, что на разных шагах возникающие ошибки частично гасят друг друга, имея разные знаки.

Для исследования устойчивости запишем формулу для перехода от одного временного веса к другому:

$$y^{(\sigma_1)} = y^{(\sigma_2)} + \tau(\sigma_1 - \sigma_2)y_t.$$

Сопоставив данное выражение с правой частью исследуемой схемы, заметим, что она представляет собой выражение  $ky_{xx}^{(\sigma_1)}$  при  $\sigma_2 = 1/2$  и  $\sigma_1 = \frac{1}{2} + \frac{k\tau}{4h^2}$ . Отсюда тривиально следует безусловная устойчивость рассматриваемой схемы «бегущего» счета в  $L_2$ , так как всегда

$$\sigma_1 = \frac{1}{2} + \frac{k\tau}{4h^2} > \sigma_0 = \frac{1}{2} - \frac{h^2}{4k\tau}$$

из условия устойчивости.

### 8.2.3. Трехслойные схемы

Для уравнения теплопроводности нетрудно написать и *трехслойные по времени разностные схемы*. Рассмотрим две из них.

**Схема Ричардсона («крест»).** Выберем шаблон из точек с номерами  $(i-1, j)$ ,  $(i, j)$ ,  $(i+1, j)$ ,  $(i, j-1)$ ,  $(i, j+1)$ , называемый чаще всего «крест» за свою форму, и запишем на нем схему

$$y_t^o = ky_{\bar{x}\bar{x}}$$

для уравнения  $u_t = ku_{xx}$ . Такая схема называется *схемой Ричардсона*.

Нетрудно видеть, что схема имеет погрешность аппроксимации  $O(\tau^2 + h^2)$ , соответствующую наличию центральной разностной производной по времени и второй разностной производной по пространству. Схема является явной, т. е. легко реализуемой.

Исследуем устойчивость схемы методом гармоник, подставив в схему решение вида  $u_i^j = \rho^j \exp(i\varphi)$ , где  $i$  — мнимая единица ( $i^2 = -1$ ). Тогда для определения  $\rho$  получим уравнение

$$\frac{\rho - \rho^{-1}}{\tau} = -\frac{8k}{h^2} \sin^2 \frac{\varphi}{2},$$

откуда

$$\rho^2 + \frac{8\tau k}{h^2} \sin^2 \frac{\varphi}{2} \rho - 1 = 0.$$

Видно, что его дискриминант

$$D = \left( \frac{4\tau k}{h^2} \sin^2 \frac{\varphi}{2} \right)^2 + 1 > 0,$$

следовательно, корни действительны и различны, причем  $\rho_1 \rho_2 = -1$ . Отсюда получаем, что один из корней заведомо больше единицы (по модулю). Следовательно, схема является безусловно неустойчивой для любых соотношений  $\tau$ ,  $h$  и непригодной для расчетов.

**Схема Дюфорта — Франкела («ромб»).** Для построения устойчивой схемы исключим в предыдущем шаблоне точку  $(i, j)$  и заменим в правой части  $y$  на  $\frac{1}{2}(\hat{y} + \hat{y})$ . Тогда

$$y_t^o = ky_{\bar{x}\bar{x}} - k \frac{\tau^2}{h^2} y_{\bar{t}\bar{t}}.$$

Шаблон «крест» без центральной точки превращается в «ромб». Записанная схема называется *схемой Дюфорта — Франкела*.

Нетрудно видеть, что погрешность аппроксимации данной схемы есть величина  $O(t^2 + h^2 + \tau^2/h^2)$ , т. е. является условной. Для обеспечения сходимости параметры  $\tau, h$  должны стремиться к нулю так, чтобы  $\tau/h \rightarrow 0$ .

Данная схема является безусловно устойчивой, но условная аппроксимация все портит. Доказательство мы не приводим. Проверка справедливости необходимого признака устойчивости, выполняемая методом гармоник, проводится аналогично изучению устойчивости схемы Ричардсона и сводится к нахождению значений корней квадратного уравнения.

### 8.3. Библиографические комментарии

Разностные методы решения уравнений параболического типа рассматриваются практически во всех основных руководствах, указанных в списке литературы. Так, дополнительные сведения о предмете данной главы можно найти в [6, 12, 14, 22, 25, 80, 84, 110, 113, 130, 134, 137, 139, 141, 149, 151, 154, 155, 176].

В монографии [139] подробно рассмотрены и трехслойные схемы, устойчивость которых исследована в [150].

Укажем, что в [150] приведены примеры, показывающие, что в результате применения принципа замороженных коэффициентов можно получить неверные выводы.

В монографии [139] кратко представлены основные свойства квазилинейного уравнения теплопроводности, а в [148] описаны многие необычные проявления нелинейности.

## 9. ЧИСЛЕННОЕ РЕШЕНИЕ ГИПЕРБОЛИЧЕСКИХ УРАВНЕНИЙ

Описаны и исследованы *разностные схемы* для численного решения одномерного линейного *уравнения переноса* — простейшего гиперболического уравнения. Рассмотрены схемы для случая переменной скорости. Описаны различные способы конструирования разностных схем для решения данного уравнения. Изучена *монотонность* схем. Приведен алгоритм исследования схем с помощью *первого дифференциального приближения*. Изучена схема для решения *волнового уравнения*.

### 9.1. Линейное одномерное уравнение переноса

Простейшим представителем семейства гиперболических уравнений является *уравнение переноса*

$$u_t + cu_x = 0, \quad c = \text{const} > 0.$$

Класс гиперболических уравнений (и систем) определяется наличием действительных характеристик, число которых совпадает с числом неизвестных для системы уравнений первого порядка.

*Характеристикой уравнения переноса* является множество точек  $(x, t)$ , удовлетворяющее уравнению

$$\frac{dx}{dt} = c,$$

т.е. множество  $x - ct = \xi_0$ . Введение характеристики позволяет записать уравнение переноса в виде  $\frac{du}{dt} = 0$  вдоль характеристики. Результатом является общее решение исходного уравнения  $u = \Phi(x - ct)$ , где  $\Phi$  — произвольная достаточно гладкая функция.

Перечислим наиболее естественные задачи для рассматриваемого уравнения:

1) *задача Коши* (начальная задача) — поиск решения уравнения переноса на множестве  $-\infty < x < +\infty, t > 0$ , принимающего заданное значение в начальный момент времени  $u(x, 0) = u_0(x)$ ;

2) *начально-краевая задача* — поиск решения уравнения переноса на множестве  $0 < x < +\infty, t > 0$ , принимающего заданные значения в начальный момент времени  $u(x, 0) = u_0(x)$  и на левой границе  $u(0, t) = v_0(t)$ .

Легко видеть, что решением задачи Коши является функция  $u = u_0(x - ct)$ , а решением начально-краевой задачи функция

$$u(x, t) = \begin{cases} u_0(x - ct), & x \geqslant ct; \\ v_0\left(t - \frac{x}{c}\right), & x \leqslant ct, \end{cases}$$

при этом считаем  $u_0(0) = v_0(0)$ .

Решение заключается в сносе неизменного профиля по характеристикам. На рис. 9.1 показаны характеристики уравнения переноса при  $c = \text{const} > 0$ . А именно, на данном решении  $u = \text{const}$  при  $x - ct = \xi_0$ , т. е. при  $x = ct + \xi_0$ . Следовательно, данное значение решения перемещается по характеристике с заданной скоростью  $c$ .

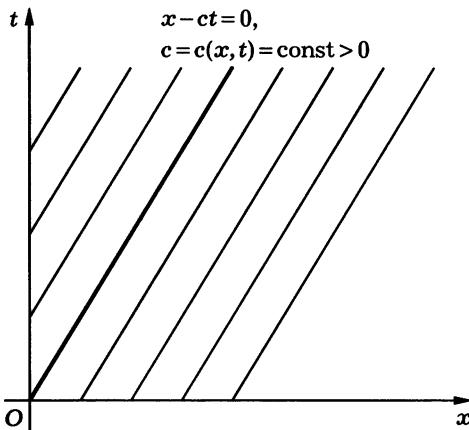


Рис. 9.1

Важнейшее свойство рассматриваемого решения — сохранение начального профиля: если начальное решение представляет собой, например, профиль буквы М, то оно будет таким всегда.

Рассмотрим разностные схемы для уравнения переноса на сетке  $\bar{\omega}_h$  с постоянным шагом  $h$ .

Целью изучения методов численного решения данного уравнения является не собственно нахождение его решения, а исследование численного алгоритма на простейшем примере. Как мы видели, точное решение данного уравнения известно и тривиально. В первом же случае мы увидим, что нет никакого труда получить точную разностную схему, которая в точках разностной сетки дает проекцию точного решения на сетку. Однако в чрезвычайно большое количество математических моделей оператор переноса входит в качестве составной части. Это модели газо- и гидродинамики, переноса частиц и излучения, электродинамики и многие другие. Разработать для их численного решения

численный метод можно только в том случае, если метод будет построен и успешно применен для данного простейшего (рафинированного) уравнения, описывающего перенос.

### 9.1.1. Явная схема с левой разностью (схема 1)

Рассмотрим на шаблоне (рис. 9.2) явную схему с левой разностью, называемую также **«левый уголок»**, для данной задачи:

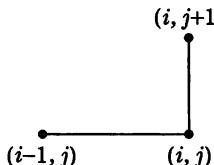


Рис. 9.2

$$y_t + cy_{\bar{x}} = 0.$$

Следовательно,

$$\hat{y} = (1 - \gamma)y + \gamma y_{-1},$$

где  $\gamma = ct/h$  — **число Куранта**.

Видно, что эта схема есть схема так называемого **«бегущего счета»**, позволяющая легко рассчитать решение.

Найдем погрешность аппроксимации на точном решении:

$$\psi_h = u_t + cu_{\bar{x}} = u_t + cu_x + \frac{1}{2}u_{tt}\tau - \frac{1}{2}cu_{xx}h + \frac{1}{6}u_{ttt}\tau^2 + \frac{1}{6}cu_{xxx}h^2 = O(\tau + h),$$

где  $\tau$  и  $h$  произвольны. Так как для точного решения справедливы равенства

$$u_{tt} = -cu_{tx} = c^2u_{xx}, \quad u_{ttt} = c^2u_{txx} = -c^3u_{xxx},$$

то

$$\psi_h = \frac{1}{2}cu_{xx}h(\gamma - 1) + \frac{1}{6}cu_{xxx}h^2(1 - \gamma^2) + O(\tau^3 + h^3).$$

Отсюда легко видеть, что при  $\gamma = 1$  справедливо  $\hat{y} = y_{-1}$ , вследствие чего выполнено тождество  $\psi_h \equiv 0$ . Это полностью соответствует точному решению исходной задачи, так как при  $c = \text{const}$  решение смещается на величину  $ct = h$  вдоль по характеристике. Следовательно, при  $\gamma = 1$  схема является точной.

Принцип максимума для данной схемы справедлив при  $1 - \gamma \geq 0$ , что верно при  $\gamma \leq 1$  (см. теорему 7.8).

**Определение 9.1.** Условие  $\gamma \leq 1$  является достаточным условием устойчивости решения в равномерной норме по начальным данным и называется **критерием Куранта**. Соответственно шаг  $\tau = h/c$  называется **курантовским временным шагом**.

Отметим, что при выполнении критерия Куранта схема является **равномерно устойчивой по начальным данным**. Из схемы видно, что ее решение устойчиво по правой части (см. теорему 7.8) при выполнении указанного условия.

Найдем достаточное условие неустойчивости, для чего воспользуемся *методом гармоник*. Подставим решение вида  $y_i^j = \rho^j \exp(i\varphi)$ , где  $i$  — мнимая единица, и получим

$$\rho = 1 - \gamma + \gamma \exp(-i\varphi) = ((1 - \gamma) + \gamma \cos \varphi) - i\gamma \sin \varphi,$$

откуда

$$|\rho|^2 = (1 - \gamma)^2 + \gamma^2 + 2\gamma(1 - \gamma) \cos \varphi = 1 - 4\gamma(1 - \gamma) \sin^2 \frac{\varphi}{2}.$$

Таким образом, условие  $\gamma > 1$  является достаточным условием неустойчивости ( $|\rho| > 1$ ) данной схемы по начальным данным.

В результате получаем, что условие  $\gamma \leq 1$  — критерий Куранта — является необходимым и достаточным для устойчивости явной схемы с левой разностью в норме  $C$ .

Исследуем ту же самую схему на устойчивость по начальным данным в норме  $L_2$ . Ограничимся **задачей Коши с финитными начальными данными**. Домножим схему на

$$y = y^{(0.5)} - \frac{\tau y_t}{2} = y_{-0.5} + \frac{hy_{\bar{x}}}{2},$$

где

$$y^{(0.5)} = \frac{\hat{y} + y}{2}, \quad y_{-0.5} = \frac{y + y_{-1}}{2}.$$

Тогда

$$y^{(0.5)} y_t = \frac{(y^2)_t}{2}, \quad y_{-0.5} y_{\bar{x}} = \frac{(y^2)_{\bar{x}}}{2},$$

откуда

$$(y^2)_t - \tau(y_t)^2 + c(y^2)_{\bar{x}} + hc(y_{\bar{x}})^2 = 0.$$

В силу уравнений схемы получаем

$$(y^2)_t + c(y^2)_{\bar{x}} + (hc - c^2\tau)(y_{\bar{x}})^2 = 0.$$

Домножим последнее равенство на  $h$  и просуммируем по точкам сетки. Воспользуемся финитностью начальных данных, в результате чего получим

$$\|\hat{y}\|_2^2 - \|y\|_2 + \tau hc(1 - \gamma) \|y_{\bar{x}}\|_2^2 = 0.$$

Следовательно,  $\|\hat{y}\|_2 \leq \|y\|_2$  при  $\gamma \leq 1$ , т. е. имеем достаточное условие устойчивости схемы в норме  $L_2$ .

Поскольку данная схема при  $\gamma \leq 1$  устойчива по правой части (см. теорему 7.7), для ошибки  $z = y - u$  имеем оценку  $\|z\|_C \leq K\|\psi_h\|_C$ . Следовательно, схема сходится со скоростью  $O(\tau + h)$ . Если же  $\gamma = 1$ , то  $z \equiv 0$ , значит, схема точна.

**Замечание 9.1.** Убедимся в естественности критерия Куранта в случае, например, решения задачи Коши. Рассмотрим точку с координатой  $x = 0$ . Исследуемая явная схема на первом временном слое даст решение, определяемое начальными данными на отрезке  $[-h, 0]$ , на втором временном слое — на отрезке  $[-2h, 0]$  и т. д. На слое, относящемся к моменту времени  $t$ , кратному  $\tau$ , решение в нулевой точке определяется данными на отрезке  $[-ht/\tau, 0]$ . Запишем точное решение исходной задачи  $u(0, t) = u_0(-ct)$ . Отсюда следует, что сходимость приближенного решения к точному возможна только тогда, когда точка с координатой  $-ct$  попадет на отрезок  $[-ht/\tau, 0]$ , т. е. будет выполнено неравенство  $-ct \geq -\frac{ht}{\tau}$  или  $\gamma = \frac{c\tau}{h} \leq 1$ .

**Замечание 9.2.** Нетрудно привести пример, демонстрирующий неустойчивость в случае невыполнения критерия Куранта.

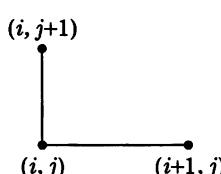
Рассмотрим задачу Коши и выберем начальные условия, такие, что на нулевом временном слое  $y_i = u_0(x_i) = (-1)^i \varepsilon$  для некоторого (возможно, малого)  $\varepsilon$ . Расчетная формула данной схемы дает решение на первом слое  $\hat{y}_i = (1 - 2\gamma)(-1)^i \varepsilon$ . Тогда на  $j$ -м слое имеем решение  $y_i^j = (1 - 2\gamma)^j (-1)^i \varepsilon$ . В случае  $\gamma > 1$  получаем  $|1 - 2\gamma| > 1$ , что приводит к экспоненциальному росту решения от слоя к слою, т. е. к неустойчивости. Она очень быстро превратит малое значение  $\varepsilon$  в машинную бесконечность.

**Замечание 9.3.** Отметим, что «естественность» критерия Куранта следует и из вида точного решения задачи. Точное решение распространяется с постоянной скоростью  $c$  по закону  $x = ct + \text{const}$ . Таким образом, для верного описания точного решения численное решение за один временной шаг также должно смещаться на один пространственный шаг, т. е. должно быть выполнено условие  $\tau = O(h)$ .

### 9.1.2. Явная схема с правой разностью (схема 2)

Рассмотрим явную схему с *правой разностью*, часто называемую **«правый уголок»** (рис. 9.3):

$$y_t + cy_x = 0.$$



Эта схема очень похожа на предыдущую и имеет тот же порядок аппроксимации, но является безусловно (абсолютно) неустойчивой. Как обычно, возьмем  $y_i^j = \rho^j \exp(i\varphi)$ . Тогда из уравнений схемы получим

Рис. 9.3

$$\rho - 1 + \gamma(\exp(i\varphi) - 1) = 0$$

и  $\rho = (1 + \gamma) - \gamma \exp(i\varphi)$ . Отсюда

$$\begin{aligned} |\rho|^2 &= ((1 + \gamma) - \gamma \cos \varphi)^2 + \gamma^2 \sin^2 \varphi = \\ &= (1 + \gamma^2) + \gamma^2 - 2\gamma(1 + \gamma) \cos \varphi = 1 + 2\gamma(1 + \gamma) - 2\gamma(1 + \gamma) \cos \varphi = \\ &= 1 + 4\gamma(1 + \gamma) \sin^2 \frac{\varphi}{2} \geq 1 \end{aligned}$$

и  $|\rho|^2 > 1$  при  $\sin^2 \frac{\varphi}{2} \neq 0$ .

Следовательно, данная схема является неустойчивой для любых  $t, h$ .

Рис. 9.4 иллюстрирует условие устойчивости для схемы с левой разностью: если  $\gamma \leq 1$ , то решение  $\hat{y}$  является результатом линейной *интерполяции* решений  $y$  и  $y_{-1}$  в точках  $x$  и  $x_{-1}$ . Если же  $\gamma > 1$ , то решение на новом слое есть уже результат линейной *экстраполяции*. Известно (см. 3), что последняя процедура всегда неустойчива, что и выражается в данном случае в неустойчивости разностной схемы. На рис. 9.4 изображены соответствующие характеристики, вдоль которых точное решение сохраняется. Если  $\gamma \neq 1$ , то решение  $\hat{y}$  всегда есть только интерполяция или экстраполяция решения с нижнего временного слоя.

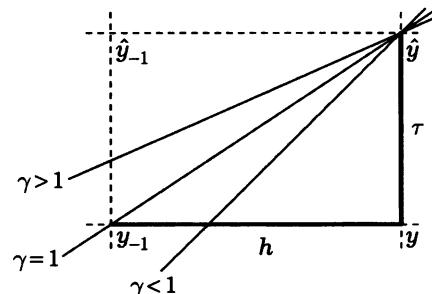


Рис. 9.4

Для схемы с правой разностью (рис. 9.5) при любом  $\gamma > 0$  решение есть результат экстраполяции с нижнего слоя. Отсюда и следует неустойчивость.

Отметим, что данная интерпретация не является доказательством, а лишь иллюстрирует полученные оценки.

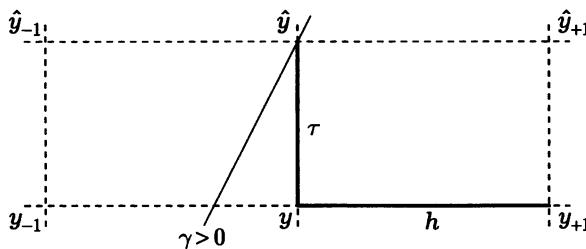


Рис. 9.5

### 9.1.3. Явная схема с центральной разностью (схема 3)

Рассмотрим схему

$$y_t + c y_{\bar{x}} = 0,$$

являющуюся полусуммой двух схем, рассмотренных выше. Шаблон схемы показан на рис. 7.3. Тот факт, что данная схема есть полусумма условно устойчивой и безусловно неустойчивой схем, вызывает сомнение в ее устойчивости. Исследуем этот вопрос. Выберем решение вида  $y_i^j = \rho^j \exp(i\varphi)$ , тогда

$$\rho - 1 + \gamma \frac{\exp(i\varphi) - \exp(-i\varphi)}{2} = 0, \quad \rho = 1 - \gamma i \sin \varphi.$$

В результате получим

$$|\rho|^2 = 1 + \gamma^2 \sin^2 \varphi \geqslant 1$$

для произвольного действительного  $\varphi$ . При этом модуль перехода заведомо больше единицы при  $\sin \varphi \neq 0$ . Следовательно, имеем достаточное условие неустойчивости по начальным данным.

Заметим, что геометрическая иллюстрация (см. рис. 9.4, 9.5) при  $\gamma \leqslant 1$  здесь вполне удовлетворительна.

Тем не менее данная схема имеет погрешность аппроксимации  $O(\tau + h^2)$ , что делает ее привлекательной для использования в качестве составной части более сложных схем.

### 9.1.4. Неявная схема с левой разностью (схема 4)

Рассмотрим теперь полностью **неявный** аналог **схемы «левый уголок»**:

$$y_t + \hat{y}_{\bar{x}} = 0,$$

т. е.

$$(1 + \gamma)\hat{y} = \gamma \hat{y}_{-1} + y.$$

Шаблон этой схемы («левый уголок вверх») показан на рис. 9.6. Несмотря на свою неявность, счет по данной схеме столь же прост,

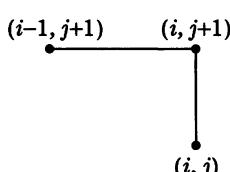


Рис. 9.6

как и по схеме 1, и проводится по явным формулам с использованием известного решения с нижнего слоя и с левой границы. Очевидно, что данная схема также имеет порядок аппроксимации  $O(\tau + h)$ . Она удовлетворяет условиям применимости принципа максимума для любого  $\gamma$ , являясь безусловно устойчивой по правой

части и начальным данным. Следовательно, скорость сходимости приближенного решения есть величина  $O(\tau + h)$ , если точное решение имеет ограниченные вторые производные. Безусловная устойчивость схемы также следует из геометрической интерпретации: для любого  $\gamma$  характеристика, вдоль которой сносится решение, проходит между точками с известным решением  $y$  и  $\hat{y}_{-1}$ , т. е. для любого  $\gamma$  происходит интерполяция решения.

Продемонстрируем безусловную устойчивость данной схемы в норме  $L_2$  для случая задачи Коши с финитными начальными данными. Для этого умножим схему на

$$\hat{y} = y^{(0,5)} + \frac{\tau y_t}{2} = \hat{y}_{-0,5} + \frac{h\hat{y}_{\bar{x}}}{2},$$

откуда получим

$$(y^2)_t + \tau(y_t)^2 + c(y^2)_{\bar{x}} + ch(\hat{y}_{\bar{x}})^2 = 0.$$

В результате использования уравнений самой схемы имеем

$$\|\hat{y}\|_2^2 - \|y\|_2^2 + c\tau(c\tau + h)\|\hat{y}_{\bar{x}}\|_2^2 = 0,$$

т. е.  $\|\hat{y}\|_2^2 \leq \|y\|_2^2$  для любого временного слоя.

### 9.1.5. Неявная схема с правой разностью (схема 5)

Рассмотрим *неявную схему*

$$y_t + c\hat{y}_x = 0$$

на шаблоне «*правый уголок*» (рис. 9.7). Очевидно, что в ней идет расчет значения  $\hat{y}_{+1}$ :

$$\hat{y}_{+1} = \left(1 - \frac{1}{\gamma}\right)\hat{y} + \frac{1}{\gamma}y.$$

Условия справедливости принципа максимума, являющегося достаточным условием устойчивости данной схемы по начальным данным, выполнены при  $1 - 1/\gamma \geq 0$ , т. е.  $\gamma \geq 1$ . Это же условие дает устойчивость и по правой части. Следовательно, скорость сходимости данного разностного решения равна  $O(\tau + h)$ .

Данное условие соответствует и геометрической интерпретации: получаемое решение есть результат интерполяции известных данных

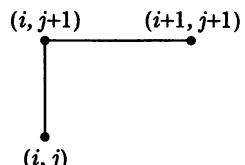


Рис. 9.7

лишь при  $\gamma \geq 1$ . Отметим, что при  $\gamma = 1$  данная схема является точной, как и схема 1.

Продемонстрируем устойчивость схемы в норме  $L_2$ , домножив уравнение схемы на

$$\hat{y} = y^{(0,5)} + \frac{\tau y_t}{2} = \hat{y}_{+0,5} - \frac{h\hat{y}_x}{2}.$$

В результате запишем

$$(y^2)_t + \tau(y_t)^2 + c(\hat{y}^2)_x - ch(\hat{y}_x)^2 = (y^2)_t + c(\hat{y}^2)_x + (c^2\tau - ch)(\hat{y}_x)^2 = 0.$$

Отсюда получаем неравенство  $\|\hat{y}\|_2 \leq \|y\|_2$ , т. е. условие устойчивости, при  $c\tau \geq h$  или  $\gamma \geq 1$ . Как и ранее, мы рассматриваем задачу Коши с финитными начальными данными. Отметим не очень естественный характер данного условия, так как, казалось бы, для хорошей передачи нужного решения необходимо малое  $\tau$ , а здесь оно должно быть больше некоторого минимального. Однако эта кажущаяся неестественность легко устраняется, если обнаружить симметрию переменных  $x$  и  $t$  в данной задаче и аналогию данной схемы явной схеме с левой разностью, в которой переменные переставлены местами. Тогда становится ясным, что условие устойчивости последней схемы является тем же критерием Куранта, но для задачи с переставленными независимыми переменными.

#### 9.1.6. Неявная схема с центральной разностью (схема 6)

Рассмотрим схему

$$y_t + c\hat{y}_x^\circ = 0$$

на шаблоне рис. 7.4. Она имеет погрешность аппроксимации  $O(\tau + h^2)$ , что делает ее в некоторых случаях привлекательной.

Счет по данной схеме возможен лишь при задании решения на верхнем слое в двух левых точках: граничной и приграничной. В приграничной точке его нужно задать из дополнительных уравнений, вычислив, например, по схеме 4. Использовать правую границу и условия на ней нельзя, так как граничные условия на правой границе неизвестны.

Из геометрической интерпретации условия устойчивости получим, что данная схема может быть устойчива лишь при  $\gamma \geq 1$ , как и схема 5. Однако для аналогичной явной схемы та же геометрическая интерпретация указывает на устойчивость при  $\gamma \leq 1$ , что оказалось неверным. В данном случае ситуация аналогична, но неравенство является обратным.

Метод гармоник дает для данной схемы множитель перехода

$$\rho = \frac{1}{1 + \gamma \tilde{\iota} \sin \varphi},$$

причем для любого  $\varphi$

$$|\rho|^2 = \frac{1}{1 + \gamma^2 \sin^2 \varphi} \leq 1.$$

Исследуем устойчивость схемы энергетическим методом, домножив уравнение схемы на величину

$$\hat{y} = y^{(0,5)} + \frac{\tau y_t}{2} = \hat{y}_{+0,5} - \frac{h \hat{y}_x}{2} = \hat{y}_{-0,5} + \frac{h \hat{y}_{\bar{x}}}{2}.$$

При этом учтем, что

$$\hat{y}_x = \frac{\hat{y}_{\bar{x}} + \hat{y}_x}{2}.$$

Тогда в результате, домножив на 4, получим

$$2(y^2)_t + 2\tau(y_t)^2 + c(\hat{y}^2)_{\bar{x}} + ch(\hat{y}_{\bar{x}})^2 + c(\hat{y}^2)_x - ch(\hat{y}_x)^2 = 0.$$

Если рассматривается задача Коши с финитными начальными данными, то суммирование даст  $\|\hat{y}\|_2^2 \leq \|y\|_2^2$ , так как слагаемые с пространственными производными уничтожаются. Таким образом, устойчивость по начальным данным в такой задаче в норме  $L_2$  имеет место.

Однако данная схема не годится для вычислений, поскольку она неустойчива по граничным данным, а эти данные всегда используются, даже если это задача Коши, так как для обеспечения возможности проведения расчета необходимо задать значения решения в двух крайних левых точках.

Рассмотрим эту неустойчивость. Для определения решения на верхнем временном слое имеем уравнение

$$\hat{y} - y + \frac{1}{2} \gamma (\hat{y}_{+1} - \hat{y}_{-1}) = 0.$$

Чтобы проверить устойчивость по граничным данным, оставим лишь часть, связанную с верхним слоем, и будем искать ее решение в виде  $\hat{y}_i = \rho^i$  (см. 5.3.2). Тогда для  $\rho$  имеем уравнение

$$\rho^2 + \frac{2\rho}{\gamma} - 1 = 0,$$

т. е.

$$\rho_{1,2} = -\frac{1}{\gamma} \pm \sqrt{\frac{1}{\gamma^2} + 1}.$$

Это значит, что оба корня действительны и различны, их произведение  $\rho_1 \rho_2 = -1$ . Следовательно, существует корень, по модулю больший единицы.

### 9.1.7. Уравнение переноса с отрицательной или переменной скоростью

**Случай  $c = \text{const} < 0$ .** При этом знаке скорости  $c$  характеристики меняют свое направление с роста на убывание при увеличении  $t$  (рис. 9.8). Следовательно, перенос решения по характеристикам происходит справа налево, в том числе и с границы. Таким образом, граничные условия в соответствующей задаче должны ставиться на границе  $x = l$  (при решении задачи в области  $x < l, t > 0$ ). Изменение наклона характеристик приводит к изменению свойств соответствующих разностных схем: явная схема с левой разностью становится абсолютно (безусловно) неустойчивой, а с правой разностью — условно устойчивой. Аналогичные изменения происходят и с неявными схемами. Это связано с тем, что изменение знака скорости эквивалентно изменению знака координаты  $x$ .

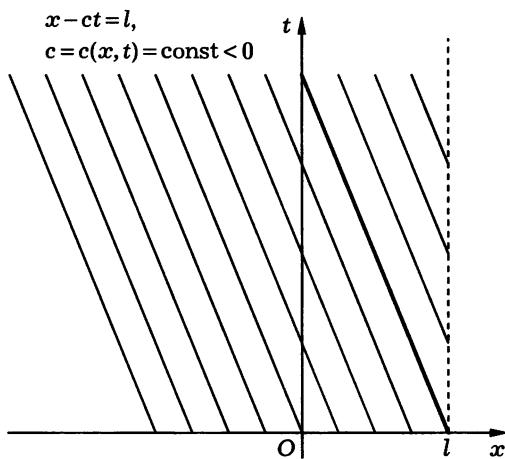


Рис. 9.8

**Знакопеременная скорость  $c = c(x, t)$ .** Пусть в задаче

$$u_t + cu_x = 0$$

скорость  $c$  зависит от  $(x, t)$  и может менять свой знак. В теории уравнений математической физики доказывается, что начально-краевая задача для такого уравнения поставлена корректно, если заданы граничные условия на всех участках границы, с которых характеристики идут внутрь области. Считаем, что на левой границе характеристика идет внутрь области, если она имеет положительный наклон, на правой границе характеристика идет внутрь области, если она имеет отрицательный наклон. При решении систем уравнений число условий должно

совпадать с числом характеристик, входящих в область, например в случае, изображенном на рис. 9.9, следует задать начальные данные и граничные условия при  $x = 0$  и  $x = l$ . Составить разностную схему нужного качества проще всего с помощью следующего приема:

$$c = c(x, t) = c_+ + c_-, \quad c_+ = \frac{c + |c|}{2} \geq 0, \quad c_- = \frac{c - |c|}{2} \leq 0,$$

а затем получить схему, например, следующего вида:

$$y_t + c_+ y_{\bar{x}} + c_- y_x = 0,$$

являющуюся условно устойчивой схемой первого порядка аппроксимации.

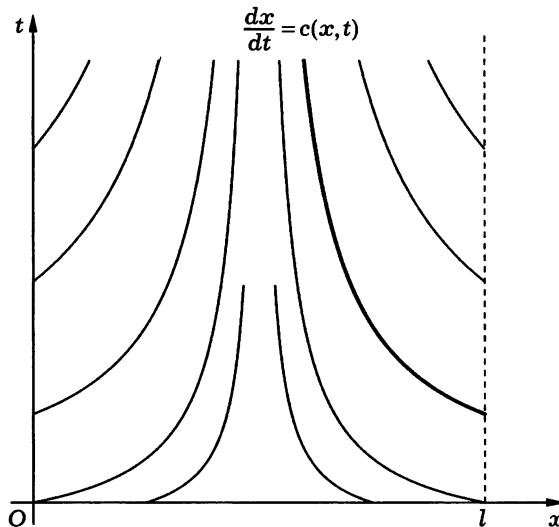


Рис. 9.9

### 9.1.8. Интерполяционный метод построения некоторых других схем для уравнения переноса

Рассмотрим следующий алгоритм. Пусть у решения рассматриваемого уравнения переноса имеются все необходимые для дальнейших выкладок производные. Воспользуемся тем, что в соответствии с формулой Тейлора для точного решения

$$\hat{u} = \sum_{l=0}^p \frac{1}{l!} u_t^{(l)} \tau^l + O(\tau^{p+1}).$$

Кроме того, для точного решения данного уравнения справедливы равенства  $u_t = -cu_x$ ,  $u_{tt} = -cu_{tx} = +c^2u_{xx}$  и т. д. Подобные выкладки позволяют производную решения по времени любого порядка заменить на производную по пространству того же порядка. В результате точное решение на новом временном слое выражается через пространственные производные на старом слое.

Изложенное выше показывает, что схему для уравнения переноса можно построить следующим образом (рассмотрение проведем на примере двухслойных схем): нужно задать шаблон схемы, по точкам со старого слоя построить интерполяционный полином  $\tilde{u} = L_k(x)$ , где  $k$  — степень полинома, на единицу меньшая числа используемых точек. Далее запишем схему с помощью формулы Тейлора, формально положив

$$\hat{y} = \sum_{l=0}^p \frac{(-c)^l}{l!} \tilde{u}_{x^l}^{(l)} \tau^l.$$

Таким образом будет построена явная разностная схема  $p$ -го порядка аппроксимации по времени и  $k$ -го порядка по пространству. Очевидно, что в описанном алгоритме для порядков аппроксимации верно неравенство  $p \leq k$ . При этом величины  $\tilde{u}_{x^l}^{(l)}$ , определяющие производные  $\tilde{u}_{t^l}^{(l)}$ , должны вычисляться в точках  $x = x_i$ ,  $t = t_j$ . Заметим, что использованная формула Тейлора имеет локальный ( $p+1$ )-й порядок по времени при переходе на один временной слой. При интегрировании по всей сетке получаем  $p$ -й порядок.

**Пример 9.1. 1.** Рассмотрим шаблон «левый уголок». Соответствующий интерполянт имеет вид

$$\tilde{u} = \frac{y(x - x_i + h) - y_{-1}(x - x_i)}{h},$$

откуда  $\tilde{u}_x = \frac{y - y_{-1}}{h}$ . Тогда в точке  $x_i$  имеем

$$\hat{y} = y - \frac{c\tau(y - y_{-1})}{h},$$

т. е. схему с *левой разностью и аппроксимацией  $O(\tau + h)$* .

**2.** Рассмотрим шаблон явной схемы с *центральной разностью*. Соответствующий интерполянт имеет вид

$$\begin{aligned} \tilde{u} = y_{+1} & \frac{(x - x_i)(x - x_i + h)}{2h^2} - y \frac{(x - x_i - h)(x - x_i + h)}{h^2} + \\ & + y_{-1} \frac{(x - x_i)(x - x_i - h)}{2h^2}, \end{aligned}$$

откуда

$$\begin{aligned}\tilde{u}_x &= y_{+1} \frac{2x - 2x_i + h}{2h^2} - y \frac{2x - 2x_i}{h^2} + y_{-1} \frac{2x - 2x_i - h}{2h^2}, \\ \tilde{u}_{xx} &= y_{+1} \frac{1}{h^2} - y \frac{2}{h^2} + y_{-1} \frac{1}{h^2} = y_{\bar{x}x}.\end{aligned}$$

Тогда в точке  $\bar{x} = x_i$  получаем  $\tilde{u}_x = y_{\bar{x}}$  и имеем **схему Лакса — Вендроффа**:

$$\hat{y} = y - c\tau y_{\bar{x}} + \frac{1}{2}(c\tau)^2 y_{\bar{x}x}$$

с локальной аппроксимацией  $O(\tau^2 + h^2)$ . #

Заметим, что данные схемы имеют максимальный порядок аппроксимации по времени и пространству, соответствующий степени полинома. В то же время интерполяционный метод может дать порядок аппроксимации по времени, меньший, чем по пространству, если параметры связаны неравенством  $p < k$ .

Попробуем также пересчитать решение с нижнего временного слоя простым его сносом по характеристике в соответствии со структурой точного решения. Для этого за решение на следующем временном слое  $\hat{t}$  примем значение интерполянта в точке  $\bar{x} = x_i - c\tau$ , соответствующей характеристике *уравнения переноса*. При использовании данного метода построения происходит перенос решения с предыдущего временного слоя в соответствии с характером точного решения. Поэтому полученная схема будет иметь максимально возможный порядок временной аппроксимации, соответствующий погрешности пространственной интерполяции.

**Пример 9.2.** 1. Рассмотрим шаблон из примера 9.1, п. 1:

$$\hat{y} = \frac{y(h - c\tau) + y_{-1}c\tau}{h} = y - c\tau y_{\bar{x}}.$$

В результате получим ту же схему с левой разностью.

2. Рассмотрим шаблон из примера 9.1, п. 2:

$$\hat{y} = -y_{+1} \frac{(h - c\tau)c\tau}{2h^2} + y \frac{h^2 - c^2\tau^2}{h^2} + y_{-1} \frac{(h + c\tau)c\tau}{2h^2},$$

откуда

$$\hat{y} = y - c\tau y_{\bar{x}} + \frac{1}{2}(c\tau)^2 y_{\bar{x}x},$$

т. е. снова получена схема **Лакса — Вендроффа**.

## 9.2. Монотонность схем для уравнения переноса

Рассмотрим более подробно вопрос о сохранении монотонности профиля численного решения. Это свойство (сохранение или несохранение) в особенности важно при решении гиперболических уравнений, точное решение которых воспроизводит (с необходимыми оговорками) исходные профили.

**Теорема 9.1.** Явная двухслойная линейная однородная схема

$$\hat{y} = \sum_k \beta_k y_{+k}$$

с постоянными коэффициентами, не зависящими от точки сетки  $i$ , монотонна в смысле сохранения монотонности профиля  $y$  тогда и только тогда, когда все коэффициенты  $\beta_k \geq 0$ .

◀ Найдем разность

$$\hat{y} - \hat{y}_{-1} = \sum_k \beta_k (y_{+k} - y_{+k-1}).$$

Если все  $\beta_k \geq 0$ , то знак левой части совпадает со знаком правой части и тогда данное условие влечет монотонность. Докажем теперь необходимость условия  $\beta_k \geq 0$ . Для этого выберем невозрастающий профиль

$$y_{i+k} = \begin{cases} 1, & i+k \leq i_0 - 1; \\ 0, & i+k \geq i_0, \end{cases}$$

предположив наличие  $\beta_{k_0} < 0$ . Тогда в точке  $i = i_0 - k_0$  выполнено равенство  $\hat{y} - \hat{y}_{-1} = -\beta_{k_0} > 0$ , а это значит, что в данной точке будет происходить возрастание профиля решения, что противоречит условию. Следовательно, предположение  $\beta_{k_0} < 0$  является неверным. ►

**Замечание 9.4.** Данное свойство является качественно более сильным, чем просто удовлетворение *принципу максимума*. К сожалению, на многомерный случай понятие монотонности однозначно не обобщается. Поэтому для многомерных задач используется принцип максимума. В рассматриваемом случае он справедлив при некоторых дополнительных ограничениях на параметры.

**Замечание 9.5.** Теорема 9.1 верна для любых двухслойных схем, а не только соответствующих *уравнению переноса*.

**Замечание 9.6.** Двухслойную неявную схему можно преобразовать к явному виду. При этом пределы суммирования по  $k$  станут бесконечными.

**Теорема 9.2 (С.К. Годунова).** Двухслойная линейная монотонная разностная схема для уравнения переноса

$$u_t + cu_x = 0$$

не может иметь второй или более высокий порядок точности.

◀ Допустим, что у нас есть схема второго или более высокого порядка точности (а значит, и аппроксимации) вида

$$\hat{y} = \sum_k \beta_k y_{+k}, \quad \beta_k \geq 0, \quad k = 1, 2, \dots$$

Рассмотрим равномерную сетку  $x_i = ih$ . Выберем в качестве начальных данных функцию

$$u_0(x) = \left( \frac{x}{h} - \frac{1}{2} \right)^2 - \frac{1}{4}.$$

Точное решение задачи также является квадратичной функцией. Погрешность аппроксимации (на точном решении) схемы второго или более высокого порядка выражается через производные третьего и более высокого порядка. В данном случае невязка  $\psi_h = 0$ , так как  $u_{xxx} = 0$ . Следовательно, решения точной и приближенной задач должны совпадать:

$$y_i^1 = \left( i - \frac{c\tau}{h} - \frac{1}{2} \right)^2 - \frac{1}{4}, \quad y_i^0 = \left( i - \frac{1}{2} \right)^2 - \frac{1}{4} \geq 0.$$

Тогда на первом слое должно выполняться равенство

$$\left( i - \frac{c\tau}{h} - \frac{1}{2} \right)^2 - \frac{1}{4} = \sum_k \beta_k \left( \left( i - \frac{1}{2} \right)^2 - \frac{1}{4} \right) \geq 0,$$

в котором оценка правой части следует из условия и вида решения. Но при нецелом числе Куранта  $\gamma = c\tau/h$  левая часть хотя бы в одной точке  $x_i$  будет отрицательна, а именно при

$$\left| i - \gamma - \frac{1}{2} \right| < \frac{1}{2},$$

т. е.  $\gamma < i < \gamma + 1$ . Данное противоречие доказывает теорему. ►

**Следствие 9.1.** Линейные монотонные схемы для уравнения переноса могут иметь только первый порядок точности, а схемы второго и более высокого порядка немонотонны (в смысле сохранения монотонности исходного профиля решения).

**Пример 9.3. 1.** Явная схема с левой разностью имеет аппроксимацию  $O(\tau + h)$  и монотонна при выполнении критерия Куранта  $\gamma \leq 1$ .

2. Неявная схема с левой разностью на шаблоне «левый уголок» может быть записана в виде

$$\begin{aligned}\hat{y} &= \frac{1}{\gamma+1}(\gamma\hat{y}_{-1} + y) = \frac{1}{\gamma+1}y + \frac{\gamma}{\gamma+1}\left(\frac{1}{\gamma+1}y_{-1} + \frac{\gamma}{\gamma+1}\hat{y}_{-2}\right) = \dots \\ &\dots = \frac{1}{\gamma+1} \sum_{k=0}^{\infty} \left(\frac{\gamma}{\gamma+1}\right)^k y_{-k}.\end{aligned}$$

Здесь все коэффициенты являются неотрицательными. Следовательно, схема является монотонной.

### 9.3. Дифференциальное приближение

На примере *уравнения переноса* рассмотрим один из методов исследования разностных схем — **метод дифференциального приближения**.

Пусть численно решается задача  $Au = f$ . Исходная задача поставлена корректно, решение ее имеет нужное количество производных. Разностная схема  $A_h u = \varphi$  аппроксимирует исходную задачу с  $p$ -м порядком. Следовательно, на решении исходной задачи невязка равна

$$\psi_h = \varphi - A_h u_h = (\varphi - f_h) + ((Au)_h - A_h u_h) = h^p (Bu)_h + o(h^p),$$

где  $B$  — некоторый оператор. Отсюда получаем, что

$$A_h u_h - \varphi = (Au)_h - f_h - h^p (Bu)_h + o(h^p).$$

Это выражение означает, что схема  $A_h u = \varphi$  аппроксимирует уравнение

$$Au - f - h^p Bu = 0$$

с порядком выше  $p$ .

**Определение 9.2.** *Первым дифференциальным приближением разностной схемы*  $A_h u = \varphi$  называют уравнение  $Au - h^p Bu = f$ , где  $Bu = \lim_{h \rightarrow 0} h^{-p} \psi_h$ .

Так как разностная схема с большей точностью аппроксимирует решение первого дифференциального приближения, то его решение будет больше соответствовать решению разностной схемы, чем решение исходной задачи. К тому же в указанном приближении содержится информация о конкретной разностной схеме, чего нет в исходной задаче. Отсюда следует и алгоритм исследования схем без их непосредственного численного решения.

Таким образом, данный метод позволяет исследовать свойства разностных схем путем качественного исследования решений дифференциальных уравнений.

Нахождение уравнения первого дифференциального приближения можно трактовать как поиск такого уравнения, которое обеспечивает на своих решениях минимум погрешности аппроксимации рассматриваемой разностной схемы с учетом первых слагаемых, зависящих от параметра дискретизации.

**Пример 9.4.** 1. Рассмотрим явную разностную схему с левой разностью:  $y_t + cy_{\bar{x}} = 0$ . Для нее невязка

$$\psi_h = -\frac{u_{tt}\tau - chu_{xx}}{2} + O(\tau^2 + h^2).$$

Отсюда получаем уравнение первого дифференциального приближения

$$u_t + cu_x + \frac{\tau u_{tt} - chu_{xx}}{2} = 0.$$

Решение данного уравнения с точностью  $O(\tau + h)$  таково, что

$$u_{tt} = c^2 u_{xx}.$$

Тогда  $u_{tt}$  можно заменить на  $u_{xx}$ , возмутив уравнение в целом на величину  $O(h^2 + \tau^2)$ , что соответствует отброшенным слагаемым. В результате получаем

$$u_t + cu_x + \frac{ch}{2}(\gamma - 1)u_{xx} = 0, \quad \gamma = \frac{c\tau}{h},$$

или

$$u_t + cu_x = \frac{ch}{2}(1 - \gamma)u_{xx}.$$

Данное уравнение обычно называют *уравнением конвекции-диффузии* (или теплопроводности) с коэффициентом диффузии

$$d^2 = \frac{ch}{2}(1 - \gamma).$$

Известно, что его решение при  $d^2 < 0$ , т. е. при  $1 - \gamma < 0$ , является неустойчивым. Это соответствует и неустойчивости схемы. Видно, что  $d$  растет с уменьшением  $\gamma$ . Следовательно, профили решения расплываются со временем тем больше, чем меньше  $\gamma$ . Это следует, например, из того, что начальный  $\delta$ -образный импульс будет эволюционировать

по закону, описываемому фундаментальным решением уравнения теплопроводности

$$t^{-0,5} \exp\left(-\frac{(x-ct)^2}{4d^2t}\right),$$

расплываюсь со временем. Подобным образом эволюционируют и другие решения.

2. Рассмотрим схему второго порядка (*Лакса — Вендроффа*) аппроксимации:

$$y_t + cy_{\dot{x}} = \frac{1}{2}c^2\tau y_{\ddot{x}x}.$$

Найдем невязку  $\psi_h$  данной схемы:

$$\begin{aligned} \psi_h = & \frac{1}{2}c^2\tau u_{xx} + O(\tau h^2) - u_t - \frac{1}{2}\tau u_{tt} - \frac{1}{6}\tau^2 u_{ttt} - cu_x - \frac{c}{6}h^2 u_{xxx} + \\ & + O(\tau^3 + h^3) = -\frac{1}{6}\tau^2 u_{ttt} - \frac{c}{6}h^2 u_{xxx} + O(\tau^3 + h^3 + \tau h^2). \end{aligned}$$

С учетом соотношений

$$u_t = -cu_x, \quad u_{tt} = -cu_{tx} = c^2 u_{xx}, \quad u_{ttt} = c^2 u_{txx} = -c^3 u_{xxx}$$

получим

$$\psi_h = \frac{1}{6}ch^2 u_{xxx}(\gamma^2 - 1) + O(\tau^3 + h^3 + \tau h^2).$$

В результате имеем первое дифференциальное приближение данной схемы

$$u_t + cu_x = \frac{1}{6}ch^2 u_{xxx}(\gamma^2 - 1) = eu_{xxx}, \quad e = \frac{1}{6}ch^2(\gamma^2 - 1).$$

Видно, что в соответствии с общими соображениями размерности схема второго порядка дает дифференциальное приближение без вторых производных, т. е. без диссипации. При этом появляются так называемые дисперсионные члены — члены с третьей производной. Рассмотрим простейшее решение вида  $\exp i(\omega t - kx)$ , где  $i$  — мнимая единица ( $i^2 = -1$ );  $\omega$  — частота;  $k$  — волновое число. Данное решение называется **плоской волной**. Подставим такое решение в дифференциальное приближение, тогда

$$i\omega - ikc = ik^3 e.$$

Отсюда получим **дисперсионное соотношение** вида

$$\omega = kc + k^3 e,$$

показывающее, какие плоские волны могут распространяться в данной системе. Любая плоская волна распространяется с **групповой скоростью**

$$v_g = \frac{d\omega}{dk} = c + 3k^2 e.$$

В исходном уравнении  $e = 0$ , поэтому все плоские волны, составляющие **волновой пакет**, двигаются с одной скоростью. Такое согласованное движение сохраняет исходную форму. В данной схеме при  $\gamma \neq 1$  справедливо неравенство  $e \neq 0$ . Поэтому разные плоские волны перемещаются с разной скоростью. Это явление называется **дисперсией**. Из выражения для  $v_g$  видно, что  $v_g < c$  (при  $\gamma < 1$ ), т. е. волны в численном решении отстают от волн в точном решении. И чем больше  $k$ , или меньше длина волны  $\lambda$ , тем больше отставание. Решение в случае, например, прямоугольного импульса имеет вид некой главной «головы» с длинным осциллирующим «хвостом».

**Определение 9.3.** Схема называется **диссипативной**, если она имеет ненулевую **аппроксимационную (численную) вязкость**  $d$ , и **бездиссипативной** в противном случае.

**Пример 9.5.** Для диссипативной схемы из примера 9.4 — явной схемы с левой разностью — имеем формальное дисперсионное соотношение вида

$$\tilde{\omega} - ikc = -k^2 d^2,$$

или

$$\omega = kc + ik^2 d^2,$$

откуда столь же формально получаем плоскую волну вида

$$\exp(\tilde{\omega}(wt - kx)) = \exp(-k^2 d^2 t) \exp(ik(ct - x)).$$

Отсюда видно, что скорость распространения возмущений для любого  $k$  одинакова. Но амплитуда возмущения уменьшается со временем. Это происходит тем быстрее, чем больше  $k$  (т. е. чем меньше длина волны  $\lambda$ , поскольку  $k = 2\pi/\lambda$ ). При  $d^2 < 0$ , или при  $\gamma > 1$ , имеет место не уменьшение, а рост амплитуды волны, т. е. неустойчивость.

## 9.4. Волновое уравнение

Рассмотрим начально-краевую задачу для **волнового уравнения**

$$\begin{aligned} u_{tt} - c^2 u_{xx} &= 0, \quad 0 < x < l, \quad t > 0; \\ u(x, 0) &= u_0(x), \quad u_t(x, 0) = v_0(x), \\ u(0, t) &= 0, \quad u(l, t) = 0. \end{aligned}$$

Будем искать численное решение, для чего введем *равномерные сетки* по времени и пространству и запишем *трехслойную разностную схему*

$$y_{\bar{t}\bar{t}} - c^2 (\sigma \hat{y}_{\bar{x}\bar{x}} + (1 - 2\sigma) y_{\bar{x}\bar{x}} + \sigma \check{y}_{\bar{x}\bar{x}}) = 0$$

на девятиточечном шаблоне.

В уравнения схемы входят решения  $\check{y}$  и  $y$  с известных временных слоев и неизвестное сеточное решение  $\hat{y}$ .

Очевидно, что при любом постоянном весе  $\sigma$  данная схема в силу своей симметричности относительно слоя  $t$  имеет порядок *аппроксимации* по крайней мере  $O(\tau^2 + h^2)$ . Специальный анализ показывает, что подбором  $\sigma$  можно добиться повышенного порядка аппроксимации. Пока мы этим заниматься не будем. Ограничимся изучением *устойчивости* данной разностной схемы.

Обозначим  $Ay = c^2 y_{\bar{x}\bar{x}}$ . Воспользуемся равенством

$$\sigma \hat{y} + (1 - 2\sigma)y + \sigma \check{y} = y + \sigma \tau^2 y_{\bar{t}\bar{t}}.$$

Тогда исходную схему можно записать в виде

$$(E - \sigma \tau^2 A)y_{\bar{t}\bar{t}} = Ay.$$

Используем для доказательства энергетический метод. Умножим скалярно уравнение схемы на

$$y_t^\circ = \frac{y_t + y_{\bar{t}}}{2}.$$

Так как

$$y_{\bar{t}\bar{t}} = \frac{y_t - y_{\bar{t}}}{\tau},$$

то

$$(y_{\bar{t}\bar{t}}, y_t^\circ) = \frac{1}{2} (\|y_{\bar{t}}\|^2)_t, \quad (Ay_{\bar{t}\bar{t}}, y_t^\circ) = -c^2 (y_{\bar{x}\bar{t}\bar{t}}, y_{\bar{x}t}^\circ) = -\frac{1}{2} c^2 (\|y_{\bar{x}\bar{t}}\|^2)_t.$$

Последнее равенство следует из формулы интегрирования по частям. Здесь и далее для сокращения записи знак нормы без специального индекса используется для гильбертовой нормы. В частности,

$$\|y_{\bar{x}\bar{t}}]\|^2 = (y_{\bar{x}\bar{t}}, y_{\bar{x}\bar{t}}].$$

Таким образом,

$$((E - \sigma\tau^2 A)y_{\bar{t}\bar{t}}, y_{\circ}) = \frac{1}{2}(\|y_{\bar{t}}\|^2 + c^2\sigma\tau^2\|y_{\bar{x}\bar{t}}]\|^2)_t.$$

Преобразуем выражение

$$(Ay, y_{\circ}) = -c^2(y_{\bar{x}}, y_{\circ}_{\bar{x}\bar{t}}],$$

используя формулу Грина, т. е. суммирование по частям. Так как для любой сеточной функции справедливы соотношения

$$y = \frac{1}{2}(y^{(0,5)} + \tilde{y}^{(0,5)} - \frac{1}{2}\tau^2 y_{\bar{t}\bar{t}}), \quad y_{\circ} = (y^{(0,5)})_{\bar{t}} = (\tilde{y}^{(0,5)})_t = \frac{y_t + y_{\bar{t}}}{2},$$

то

$$(Ay, y_{\circ}) = -c^2(y_{\bar{x}}, y_{\circ}_{\bar{x}\bar{t}}] = -\frac{1}{2}c^2\left(y_{\bar{x}}^{(0,5)} + \tilde{y}_{\bar{x}}^{(0,5)}, (y_{\bar{x}}^{(0,5)})_{\bar{t}}\right] + \\ + \frac{1}{8}c^2\tau^2(y_{\bar{x}\bar{t}\bar{t}}, y_{\bar{x}\bar{t}} + y_{\bar{x}\bar{t}}] = -\frac{1}{2}c^2(\|\tilde{y}_{\bar{x}}^{(0,5)}]\|^2)_t + \frac{1}{8}\tau^2c^2(\|y_{\bar{x}\bar{t}}]\|^2)_t.$$

В результате имеем равенство

$$\frac{1}{2}(\|y_{\bar{t}}\|^2 + \sigma\tau^2c^2\|y_{\bar{x}\bar{t}}]\|^2)_t = \left(-\frac{1}{2}c^2\|\tilde{y}_{\bar{x}}^{(0,5)}]\|^2 + \frac{1}{8}c^2\tau^2\|y_{\bar{x}\bar{t}}]\|^2\right)_t,$$

т. е.

$$\|\hat{y}\|_*^2 = \|y\|_*^2,$$

где

$$\|y\|_*^2 = \|y_{\bar{t}}\|^2 + c^2\left(\sigma - \frac{1}{4}\right)\tau^2\|y_{\bar{x}\bar{t}}]\|^2 + c^2\|\tilde{y}_{\bar{x}}^{(0,5)}]\|^2.$$

При  $\sigma \geq 1/4$  величина  $\|y\|_*$  является величиной типа нормы, точнее полунормы — величины типа нормы, для которой несправедливо первое условие нормы (см. определение 1.6). Рассмотрим другие параметры, для чего запишем неравенство типа вложения из метода разделения переменных:

$$\|y_{\bar{x}\bar{t}}]\|^2 \leq \frac{4}{h^2}\|y_{\bar{t}}\|^2.$$

Отсюда

$$\|y\|_*^2 \geq \left( \frac{h^2}{4} + \left( \sigma - \frac{1}{4} \right) \tau^2 c^2 \right) \|y_{\bar{x}\bar{t}}\|^2 + c^2 \|\check{y}_{\bar{x}}^{(0,5)}\|^2.$$

Правая часть неотрицательна при

$$c^2 \left( \sigma - \frac{1}{4} \right) \tau^2 + \frac{h^2}{4} \geq 0,$$

значит,

$$\sigma \geq \frac{1}{4} - \frac{h^2}{4c^2\tau^2}.$$

Отсюда видно, что явная схема с  $\sigma = 0$  устойчива при  $\tau \leq h/c$ , т. е. при выполнении *критерия Куранта*. Это неудивительно, так как волновое уравнение (уравнение колебаний) можно записать в виде

$$u_{\xi\eta} = 0, \quad \xi = x - ct, \quad \eta = x + ct,$$

откуда

$$u = \varphi(x - ct) + \psi(x + ct),$$

где  $\varphi, \psi$  — функции, определяемые начальными и граничными данными  $u_0(x), v_0(x)$ . Поэтому и условие устойчивости имеет вид, сходный с условием устойчивости для *уравнения переноса*.

При выполнении условия устойчивости справедливо неравенство

$$\|y\|_*^2 \geq c^2 \|\check{y}_{\bar{x}}^{(0,5)}\|^2.$$

Из него следует, что с учетом начальных и граничных условий данной задачи величина  $\|y\|_*$  действительно является нормой. А именно, если  $\|y\|_* = 0$ , то  $\|\check{y}_{\bar{x}}^{(0,5)}\| = 0$ . Из неравенства вложения  $\|z\|_C \leq \sqrt{l} \|z_{\bar{x}}\|$  (см. 7.2) получаем  $\check{y}^{(0,5)} = 0$ . Это равенство справедливо в силу однородных граничных условий первого рода. Однородные начальные условия дают  $y \equiv 0$ , следовательно,  $\|y\|_*$  — норма.

Из изложенного следует сходимость данной трехслойной разностной схемы со скоростью по крайней мере  $O(\tau^2 + h^2)$  при выполнении условия устойчивости

$$\sigma \geq \frac{1}{4} - \frac{h^2}{4c^2\tau^2}.$$

Устойчивость данной схемы по правой части мы рассматривать не будем.

Отметим, что расчет по данной схеме может проводиться только при известных значениях  $y$  на нулевом и первом временных слоях.

Решение на нулевом слое задано начальным условием, а на первом рассчитывается некоторым дополнительным способом по начальным решению и скорости. Наличие двух предыдущих слоев, данные на которых определяют решение на новом слое, делает естественным появление норм, содержащих полусумму величин с двух последовательных слоев.

**Пример 9.6.** Рассмотрим в качестве примера вопрос о повышении порядка аппроксимации данной трехслойной схемы за счет выбора веса. Рассматриваемую схему можно переписать в более простом для исследования виде

$$y_{\bar{t}\bar{t}} - c^2 y_{\bar{x}\bar{x}} - c^2 \tau^2 \sigma y_{\bar{x}\bar{x},\bar{t}\bar{t}} = 0.$$

Вычислим погрешность аппроксимации на точном решении, для которого справедливы равенства

$$u_{tttt} - c^4 u_{xxxx} = 0, \quad u_{xxtt} - c^2 u_{xxxx} = 0.$$

Используем их:

$$\begin{aligned} \psi_h &= u_{tt} + \frac{1}{12} \tau^2 u_{tttt} - c^2 u_{xx} - c^2 \frac{1}{12} h^2 u_{xxxx} - c^2 \tau^2 \sigma u_{xxtt} + O(\tau^4 + h^4) = \\ &= \left( \frac{1}{12} \tau^2 c^4 - \frac{1}{12} h^2 c^2 - c^4 \tau^2 \sigma \right) u_{xxxx} + O(\tau^4 + h^4). \end{aligned}$$

Отсюда получаем, что схема с весом

$$\sigma = \frac{1}{12} - \frac{h^2}{12c^2\tau^2}$$

имеет погрешность аппроксимации  $\psi_h = O(\tau^4 + h^4)$ .

**Замечание 9.7.** Волновое уравнение не обязательно решать с помощью трехслойных разностных схем. Можно применять и двухслойные схемы, если модифицировать исходную задачу и записать ее в виде системы уравнений. Это можно сделать разными способами.

1. Простейший вариант состоит во введении скорости по правилу  $v = u_t$ . Тогда получаем систему уравнений

$$u_t = v, \quad v_t = c^2 u_{xx}.$$

2. Весьма распространенным вариантом является введение потенциала скоростей  $v = \int_0^x u_t d\xi$  и сведение исходного уравнения второго

порядка по обеим переменным к системе уравнений акустики первого порядка:

$$u_t = v_x, \quad v_t = c^2 u_x.$$

Такие задачи могут решаться с помощью двухслойных разностных схем. Это, в частности, снимает вопрос о расчете решения на первом временном слое, где оно не задано начальными данными.

В случае трехслойных схем такой расчет выполнить несложно по начальным условиям для решения и для скорости с учетом самого уравнения, позволяющего вычислить ускорение в начальный момент времени. Все это вместе дает возможность использовать расчетную формулу типа формулы Тейлора для нахождения решения на первом слое.

## 9.5. Библиографические комментарии

Разностные методы решения уравнений гиперболического типа рассматриваются практически во всех руководствах, указанных в качестве основных в списке литературы. Так, дополнительные сведения о предмете данной главы можно найти в [6, 12, 14, 22, 25, 53, 80, 84, 110, 113, 134, 137, 139, 149, 151, 154, 155, 176].

В монографии [139] подробно рассмотрены и трехслойные схемы, устойчивость которых исследована в [150].

Первоначальные сведения о дифференциальных приближениях разностных схем можно найти в [80], а систематическое изложение данного метода — в [190].

Интерес к гиперболическим уравнениям особенно велик, так как к ним относятся столь важные уравнения, как уравнения магнитной гидродинамики, газовой динамики и многие другие, описывающие явления с участием конвективного переноса. Поэтому существует много специализированной литературы, посвященной данному предмету: [16, 97, 98, 135, 144, 155, 181, 182].

Далее (во второй части) приведены еще две главы, в которых рассмотрены численные решения гиперболических уравнений. Представлены разностные схемы для линейного и квазилинейного уравнений переноса и системы уравнений газовой динамики. Там же дана и дополнительная библиография.

# 10. ЧИСЛЕННОЕ РЕШЕНИЕ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ

Представлены разностные схемы для решения задачи Дирихле для уравнения Пуассона, в том числе схема повышенного порядка аппроксимации. Решена задача Штурма — Лиувилля для разностного оператора Лапласа в двумерном случае. Рассмотрен вопрос о нахождении решения разностных задач, соответствующих уравнению Лапласа, в том числе метод счета на установление решений параболических уравнений для нахождения решений стационарных задач. Приведены и обоснованы продольно-поперечная и локально-одномерная схемы для нахождения решения счетом на установление. Представлены проекционные методы для решения эллиптических уравнений — методы Ритца и Галеркина. Введены аппроксимации оператора Лапласа в криволинейных координатах (цилиндрических и сферических).

## 10.1. Решение задачи Дирихле для уравнения Пуассона

Простейшим эллиптическим уравнением является *уравнение Пуассона* (или *уравнение Лапласа* при нулевой правой части) в двумерном случае, для которого поставлена *задача Дирихле*, т. е. граничные условия первого рода. Рассмотрим ее:

$$\begin{aligned} u_{x_1 x_1} + u_{x_2 x_2} &= -f(x_1, x_2), \quad (x_1, x_2) \in G; \\ u|_{\Gamma} &= \mu(x), \quad x \in \Gamma. \end{aligned}$$

Пусть  $G$  — прямоугольник,

$$G = \{(x_1, x_2): 0 < x_1 < l_1, 0 < x_2 < l_2\},$$

$\Gamma$  — его граница,  $\Gamma = \partial G$ .

Введем сетку (рис. 10.1):

$$\bar{\omega}_h = \{(x_{1,i}, x_{2,j}): x_{1,i} = i h_1, x_{2,j} = j h_2, i = 0, 1, \dots, N_1, j = 0, 1, \dots, N_2\},$$

где  $h_1, h_2 = \text{const}$ , и проведем дискретизацию уравнения, записав разностную схему на шаблоне типа «крест» (рис. 10.2):

$$\begin{aligned} y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} &= -f = -\varphi; \\ y|_{\Gamma_h} &= \mu = \nu, \end{aligned}$$

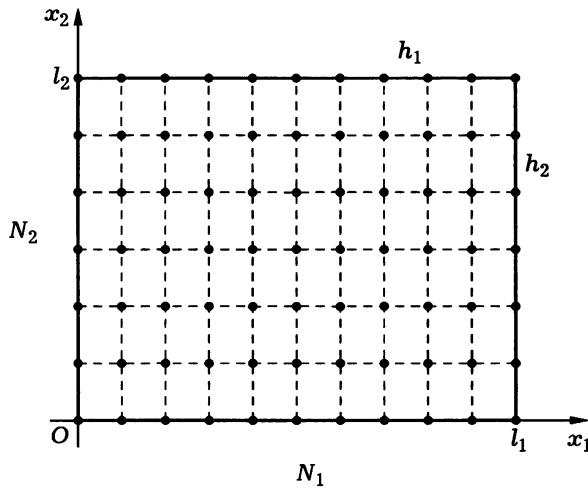


Рис. 10.1

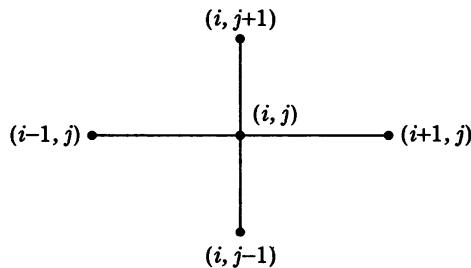


Рис. 10.2

где  $\varphi_{ij} = f(x_{1,i}, x_{2,j})$  в  $G_h$ ,  $\nu_{ij} = \mu(x_{1,i}, x_{2,j})$  на  $\Gamma_h$ . Видно, что угловые точки  $(0,0)$ ,  $(0,l_1)$ ,  $(0,l_2)$ ,  $(l_1,l_2)$  не участвуют в определении сеточного решения внутри области.

Данную схему можно представить в каноническом виде

$$Ay = \sum_{S'} By + F$$

для проверки применимости *принципа максимума* (см. 7.6)

$$\left( \frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{ij} = \frac{y_{i+1,j} + y_{i-1,j}}{h_1^2} + \frac{y_{i,j+1} + y_{i,j-1}}{h_2^2} + \varphi_{ij}.$$

Отсюда видно, что во всех внутренних узлах сетки, для которых записано данное сеточное уравнение, выполнено условие положительности коэффициентов:

$$A > 0, \quad B > 0, \quad D = A - \sum_{S'} B \geqslant 0.$$

Но то же самое выражение можно записать и в граничных узлах, где  $A = 1$ ;  $S'(x) = \emptyset$  — пустое множество и, следовательно,  $D = 1$ . В данной схеме выполнены все условия применимости принципа максимума для любых шагов  $h_1, h_2$ . Справедливы и все следствия из него: единственность решения разностной задачи, его знакопределеннность в зависимости от знаков  $\mu, f$ . В частности (см. теорему 7.5), мы имеем устойчивость решения разностной задачи по граничным данным.

Исследуем аппроксимацию исходной задачи данной разностной схемой. Очевидно, что на решениях с четвертыми непрерывными в  $\bar{G}$  производными имеем  $\psi_h = O(h_1^2 + h_2^2)$ .

Исследуем устойчивость решения. Пусть  $y = \bar{y} + \tilde{y}$ , где  $\bar{y}$  — решение однородной разностной схемы с неоднородными граничными условиями, а  $\tilde{y}$  — решение неоднородной разностной схемы с однородными граничными условиями. Согласно теореме 7.3, получаем оценку  $\|\tilde{y}\|_C \leq \|\nu\|_C$ . Рассмотрим сеточную функцию

$$Y = k(l_1^2 + l_2^2 - x_1^2 - x_2^2), \quad k > 0,$$

т. е.  $Y \geq 0$  при  $(x_1, x_2) \in \bar{G}$ . Подставим  $Y$  в разностную схему и воспользуемся равенством

$$x_{1,\bar{x}_1 x_1}^2 = \frac{(x_1 + h_1)^2 - 2x_1^2 + (x_1 - h_1)^2}{h_1^2} = 2.$$

Тогда разностная схема в канонической форме записи будет иметь вид

$$-Y_{\bar{x}_1 x_1} - Y_{\bar{x}_2 x_2} = 2 \cdot 2k = 4k, \quad Y|_{\Gamma_h} = \tilde{\mu} \geq 0.$$

Выберем  $k = \|\varphi\|_C/4$ . В этом случае справедлива теорема сравнения (см. теорему 7.4):

$$\|\tilde{y}\|_C \leq \|Y\|_C \leq \frac{l_1^2 + l_2^2}{4} \|\varphi\|_C.$$

В итоге имеем оценку решения разностной схемы

$$\|y\|_C \leq \|\tilde{y}\|_C + \|\bar{y}\|_C \leq \|\nu\|_C + \frac{l_1^2 + l_2^2}{4} \|\varphi\|_C.$$

Пусть  $y = z + u_h$ , где  $u_h$  — проекция точного решения на сетку. Тогда  $A_h z = \psi_h$  и из той же самой оценки имеем

$$\|z\|_C \leq \frac{l_1^2 + l_2^2}{4} \|\psi_h\|_C = O(h_1^2 + h_2^2).$$

Таким образом, данная разностная схема сходится со вторым порядком по обеим переменным.

## 10.2. Разностная схема для уравнения Пуассона повышенного порядка точности

Рассмотрим снова задачу Дирихле для уравнения Пуассона в прямоугольнике. В 10.1 была исследована схема

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi,$$

имеющая погрешность аппроксимации  $O(h_1^2 + h_2^2)$ . Рассмотрим более подробно невязку данной схемы на решении исходной задачи:

$$\psi_h = -f - u_{\bar{x}_1 x_1} - u_{\bar{x}_2 x_2} = -f - \Delta u - \frac{h_1^2}{12} u_{x_1^4} - \frac{h_2^2}{12} u_{x_2^4} + O(h_1^4 + h_2^4),$$

если у решения уравнения Пуассона  $u$  существуют шестые производные, непрерывные в  $\bar{G}$ .

Преобразуем  $\psi_h$ , воспользовавшись свойствами точного решения:

$$u_{x_1^4} + u_{x_1^2 x_2^2} = -f_{x_1^2}, \quad u_{x_1^2 x_2^2} + u_{x_2^4} = -f_{x_2^2}.$$

Тогда

$$u_{x_1^4} = -f_{x_1^2} - u_{x_1^2 x_2^2}, \quad u_{x_2^4} = -f_{x_2^2} - u_{x_1^2 x_2^2}.$$

В итоге получим

$$\psi_h = -f - \Delta u + \frac{h_1^2}{12} f_{x_1^2} + \frac{h_2^2}{12} f_{x_2^2} + \frac{1}{12} (h_1^2 + h_2^2) u_{x_1^2 x_2^2} + O(h_1^4 + h_2^4).$$

Выберем схему

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 \bar{x}_2 x_1 x_2} = -\varphi = -f - \frac{h_1^2}{12} f_{\bar{x}_1 x_1} - \frac{h_2^2}{12} f_{\bar{x}_2 x_2}.$$

Проведенное выше исследование структуры невязки  $\psi_h$  показывает, что погрешность аппроксимации данной схемы равна  $O(h_1^4 + h_2^4)$ . Оператор данной разностной схемы определен на девятиточечном *шаблоне типа «ящик»*, в котором дополнительные точки определяются добавленным членом  $y_{\bar{x}_1 \bar{x}_2 x_1 x_2}$  (рис. 10.3).

Нетрудно представить разностную схему в виде системы линейных уравнений стандартного для применения *принципа максимума*. Опуская выкладки, ограничимся результатом: принцип максимума справедлив для данной разностной схемы при выполнении условий

$$\frac{1}{\sqrt{5}} \leq \frac{h_1}{h_2} \leq \sqrt{5}.$$

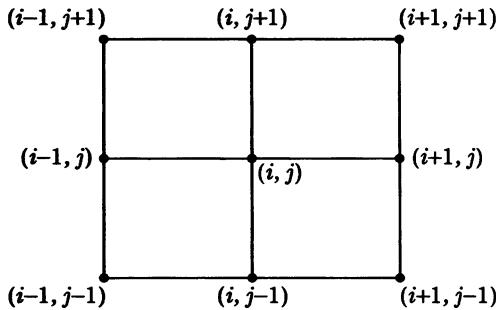


Рис. 10.3

При этом выполнено условие положительности коэффициентов. Тогда, используя теорему 7.4, получим оценку (см. 10.1) для нормы ошибки  $z = y - u_h$ , равную

$$\|z\|_C \leq \frac{l_1^2 + l_2^2}{4} \|\psi_h\|_C = O(h_1^4 + h_2^4).$$

Таким образом, сеточное решение сходится к точному решению (если последнее имеет шестые непрерывные производные в  $\bar{G}$ ) с четвертым порядком по  $h$ .

### 10.3. Собственные функции разностного оператора Лапласа и их применение

#### 10.3.1. Разностная задача

**Штурма — Лиувилля в двумерном случае**

Рассмотрим следующую **двумерную задачу Штурма — Лиувилля на собственные значения**:

$$y_{\bar{x}_1 \bar{x}_1} + y_{\bar{x}_2 \bar{x}_2} + \lambda^2 y = 0 \text{ в } G_h, \quad y|_{\Gamma_h} = 0,$$

где  $G = (0, l_1) \times (0, l_2)$  — прямоугольник. Нетрудно видеть, что нетривиальное решение этой задачи на *равномерной* по обеим переменным *прямоугольной сетке* имеет вид

$$y_{k_1 k_2} = \mu_{k_1 k_2} = \frac{2}{\sqrt{l_1 l_2}} \sin \frac{\pi k_1 x_1}{l_1} \sin \frac{\pi k_2 x_2}{l_2},$$

$$\lambda_{k_1 k_2}^2 = \frac{4}{h_1^2} \sin^2 \frac{\pi k_1 h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi k_2 h_2}{2l_2},$$

$$k_1 = 1, 2, \dots, N_1 - 1, \quad k_2 = 1, 2, \dots, N_2 - 1,$$

т. е. данная задача и в разностном случае допускает разделение переменных.

Решение выполняется аналогично алгоритму 7.5. В результате получаем

$$\frac{9}{l_1^2} + \frac{9}{l_2^2} \leq \lambda^2 \leq \frac{4}{h_1^2} + \frac{4}{h_2^2}.$$

Аналогичным образом имеем следующие оценки:

$$\left( \frac{9}{l_1^2} + \frac{9}{l_2^2} \right) \|y\|^2 \leq (A_h y, y) \leq \left( \frac{4}{h_1^2} + \frac{4}{h_2^2} \right) \|y\|^2,$$

где  $A_h y = -y_{\bar{x}_1 x_1} - y_{\bar{x}_2 x_2}$ . Следовательно, получим неравенство

$$\gamma_1 E \leq A \leq \gamma_2 E,$$

в котором  $\gamma_1 = 9/l_1^2 + 9/l_2^2$ ,  $\gamma_2 = 4/h_1^2 + 4/h_2^2$  — постоянные энергетической эквивалентности единичного оператора и разностного оператора Лапласа. Это неравенство свидетельствует, в частности, о положительной определенности разностного оператора Лапласа. Данные собственные функции являются *ортонормированными*, и любая сеточная функция, обращающаяся в нуль на границе, может быть представлена в виде разложения по системе  $\mu_k = \mu_{k_1 k_2}(x_1, x_2)$ .

Рассмотрим в качестве примера применения такого разложения решение задачи Дирихле для уравнения Пуассона

$$\begin{aligned} u_{x_1 x_1} + u_{x_2 x_2} &= -f(x_1, x_2), \quad (x_1, x_2) \in G; \\ u|_{\Gamma} &= \mu(x), \quad x \in \Gamma. \end{aligned}$$

Ее разностная аппроксимация имеет вид

$$Ly = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi, \quad y|_{\Gamma_h} = \nu.$$

Считаем, что область  $G$  — прямоугольник, сетка — равномерная. Введем функцию  $\bar{y}$ , равную  $\nu$  в граничных узлах сетки и нулю во внутренних. Ищем решение в виде  $y = \bar{y} + \bar{\varphi}$ . Получаем для  $\bar{y}$  задачу

$$L\bar{y} = -\bar{\varphi} = -\varphi - Ly, \quad \bar{y}|_{\Gamma_h} = 0.$$

Так как правая часть  $\bar{\varphi}$  задана только во внутренних узлах сетки, можно считать, что на границе выполнено равенство  $\bar{\varphi} = 0$ . Тогда

$$\bar{y} = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} a_{k_1 k_2} \mu_{k_1 k_2}(x_1, x_2), \quad \bar{\varphi} = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} b_{k_1 k_2} \mu_{k_1 k_2}(x_1, x_2),$$

где  $b_{k_1 k_2} = (\bar{\varphi}, \mu_{k_1 k_2})$ .

После подстановки  $\bar{y}$  и  $\bar{\varphi}$  в схему имеем

$$-\sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} \lambda_{k_1 k_2}^2 a_{k_1 k_2} \mu_{k_1 k_2}(x_1, x_2) = -\sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} b_{k_1 k_2} \mu_{k_1 k_2}(x_1, x_2),$$

откуда  $a_{k_1 k_2} = b_{k_1 k_2} / \lambda_{k_1 k_2}^2$ , т. е. решение задачи найдено:

$$y = \bar{y} + \bar{y}, \quad \bar{y} = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} \frac{b_{k_1 k_2}}{\lambda_{k_1 k_2}^2} \mu_{k_1 k_2}(x_1, x_2).$$

### 10.3.2. Численное нахождение решения разностной задачи

Рассмотрим возможность нахождения решения построенных разностных схем для эллиптических уравнений, в частности, для уравнения Пуассона. Для этого нужно решить систему линейных алгебраических уравнений (СЛАУ)

$$Ay = \varphi,$$

где  $Ay = -y_{\bar{x}_1 x_1} - y_{\bar{x}_2 x_2}$  во внутренних узлах сетки.

При этом обязательно должно быть выполнено условие равенства нулю сеточной функции на границе области, что позволяет исключить из системы граничные значения. Таким образом, в рассматриваемой системе фигурируют только значения сеточной функции в строго внутренних узлах сетки. В противном случае оператор  $A$  не будет *самосопряженным* даже в случае граничных условий первого рода.

Простейшими способами решения СЛАУ уравнений  $Ay = f$  являются явные стационарные *итерационные методы* вида

$$\frac{y^{k+1} - y^k}{\tau} + Ay^k = f.$$

Рассмотрим возможность их применения к данному случаю.

Начнем с метода *простой итерации*. Из 1.7.2 следует, что оптимальное значение итерационного параметра  $\tau = \frac{2}{\gamma_1 + \gamma_2}$ , т. е. в данном случае

$$\tau = \frac{2}{\left(\frac{9}{l_1^2} + \frac{9}{l_2^2}\right) + \left(\frac{4}{h_1^2} + \frac{4}{h_2^2}\right)}.$$

Это выражение очень схоже с выражением временного шага, при котором решение явной схемы для уравнения теплопроводности является

устойчивым. Ранее, в 1.6.3 мы также доказывали, что метод простых итераций может сходиться лишь при выполнении условия  $\tau < 2/\gamma_2$ .

Отметим, что для сходимости *метода Якоби*

$$Dy^{k+1} + Cy^k = f,$$

где  $D$  — диагональ  $A$ ,  $A = D + C$ , т. е. в стандартной форме метода вида

$$D(y^{k+1} - y^k) + Ay^k = f,$$

необходимо строгое диагональное преобладание, которого нет в случае разностного оператора Лапласа.

Метод последовательной верхней релаксации  $A = A_1 + D + A_2$ , где  $A_1$  — нижняя треугольная матрица;  $D$  — диагональ;  $A_2$  — верхняя треугольная матрица, имеет вид

$$\frac{D + \omega A_1}{\omega}(y^{k+1} - y^k) + Ay^k = f.$$

Этот метод сходится при  $0 < \omega < 2$ . *Метод Зейделя*, являющийся частным случаем метода последовательной верхней релаксации при  $\omega = 1$ , сходится.

*Двухслойные итерационные методы вариационного типа* также сходятся. При этом погрешность итерационного решения за  $m$  итераций уменьшается в  $\rho^{-m}$  раз, где

$$\rho = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2} = O\left(\frac{h^2}{l^2}\right).$$

Соответственно и *число обусловленности* данной СЛАУ

$$M_A = \frac{\lambda_{\max}}{\lambda_{\min}} \approx \frac{\gamma_2}{\gamma_1}$$

(используется спектральная норма матрицы), поэтому во всех оценках сходимости итераций фигурирует данное отношение.

Рассмотренный метод простой итерации позволяет использовать для нахождения решения стационарного уравнения  $-\Delta u = f$  уравнение  $v_t - \Delta v = f$  с одними и теми же граничными условиями и произвольными начальными данными для  $v$ . Поясним это на примере исходной задачи. Пусть  $w = u - v$  — разность искомого и нестационарного решений. Тогда, так как  $u_t = 0$ ,

$$w_t - \Delta w = 0,$$

$$w|_{t=0} = w_0(x_1, x_2),$$

$$w|_{\Gamma} = 0.$$

Умножим скалярно уравнение для  $w$  на решение  $w$  и с учетом граничных условий получим энергетическое тождество

$$\frac{1}{2}(\|w\|_2^2)_t + \|\operatorname{grad} w\|_2^2 = 0.$$

*Оператор Лапласа* обладает полным набором собственных функций. Используя его так же, как и в дискретном случае, получим **неравенство типа вложения**:

$$\|\operatorname{grad} w\|_2^2 \geq \lambda_{\min}^2 \|w\|_2^2, \quad \lambda_{\min}^2 = \frac{\pi^2}{l_1^2} + \frac{\pi^2}{l_2^2}.$$

Подставим неравенство типа вложения в энергетическое тождество и разделим на  $\|w\|_2$ . В результате получим дифференциальное неравенство

$$\|w\|_{2t} + \lambda_{\min}^2 \|w\|_2 \leq 0.$$

Домножив это неравенство на экспоненту от  $\lambda_{\min}^2 t$ , запишем эквивалентное неравенство в виде

$$(e^{\lambda_{\min}^2 t} \|w\|_2)_t \leq 0.$$

Его интегрирование с учетом начальных данных дает

$$\|w\|_2 \leq e^{-\lambda_{\min}^2 t} \|w_0\|_2.$$

Таким образом, характерное время выхода решения на стационарный режим составляет

$$T \approx \frac{3 \div 5}{\lambda_{\min}^2} = \frac{3 \div 5}{\frac{\pi^2}{l_1^2} + \frac{\pi^2}{l_2^2}}.$$

**Определение 10.1.** Процедура нахождения стационарного решения путем поиска установившегося решения нестационарной задачи называется **счетом на установление (методом установления)**.

Определим условия, при которых можно вести счет на установление, решая параболическую задачу. Для этого рассмотрим схему с весами для уравнения теплопроводности в пространственно-двумерном случае

$$y_t = y_{\bar{x}_1 x_1}^{(\sigma)} + y_{\bar{x}_2 x_2}^{(\sigma)} + \varphi.$$

Проводя в точности те же действия, что и в 8, с учетом двумерности задачи получим следующее условие устойчивости в  $L_2$ :

$$|\rho_k| \leq 1, \quad \rho_k = \frac{1 - \lambda_k \tau(1 - \sigma)}{1 + \sigma \tau \lambda_k^2},$$

откуда

$$\sigma \geq \frac{1}{2} - \frac{1}{\lambda_k^2 \tau},$$

или

$$\sigma \geq \frac{1}{2} - \frac{1}{\left(\frac{4}{h_1^2} + \frac{4}{h_2^2}\right)\tau}$$

(для данной задачи коэффициент теплопроводности равен единице).

Таким образом, явная схема с весом  $\sigma = 0$  устойчива в норме  $L_2$  при

$$\tau \leq \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right)^{-1} = \frac{h_1^2 h_2^2}{2(h_1^2 + h_2^2)}.$$

*Принцип максимума* применим к данной схеме при выполнении условия

$$\frac{1}{\tau} - \frac{2(1-\sigma)}{h_1^2} - \frac{2(1-\sigma)}{h_2^2} \geq 0,$$

т. е.

$$\tau \leq \tau_0 = \frac{1}{2(1-\sigma)} \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)^{-1},$$

что при  $\sigma = 0$  дает в точности то же самое ограничение.

Нетрудно видеть, что оптимальный для решения системы линейных алгебраических уравнений шаг  $\tau$  практически равен шагу  $\tau_0$  (так как  $h_1 \ll l_1$ ,  $h_2 \ll l_2$ ). Поэтому описанные алгоритмы мало чем отличаются друг от друга.

**Замечание 10.1.** В данной главе мы рассматриваем только один способ нахождения решения разностных схем, соответствующих эллиптическим уравнениям, — метод установления. Это не означает отрицания тех методов, которые представлены в 1. Все описанные в 1 алгоритмы (с необходимыми оговорками об условиях) решения СЛАУ применимы и при решении эллиптических задач.

Для решения полученных разностных задач может использоваться и прямой метод типа метода Гаусса. Получаемые при этом матрицы являются, как правило, разреженными. Их хранение и операции с ними требуют применения специальных алгоритмов, коротко описанных в 1. Существуют варианты метода Гаусса, позволяющие эффективным образом использовать разреженность матрицы и за счет перестановок строк и столбцов добиваться решения больших задач за время, вполне сопоставимое с временными затратами итерационных методов.

На сетках с постоянными шагами в прямоугольной области весьма эффективным может быть применение быстрого преобразования

Фурье. Возможно использование и многих других прямых методов решения СЛАУ, возникающих при дискретизации эллиптических задач.

Выявленная выше аналогия между решением систем уравнений итерационными методами и решением параболических уравнений позволяет применять метод установления. Он может быть реализован разными способами. В частности, знание границ спектра оператора позволяет использовать *метод Ричардсона с чебышевским набором параметров* при условии их соответствующего упорядочения. В этом варианте получается алгоритм, внешне совпадающий с алгоритмом нахождения решения параболического уравнения с временными шагами, которые задаются специальным образом.

#### 10.4. Экономичные разностные схемы для решения уравнения теплопроводности в многомерном случае

В 10.3 было показано, что решение эллиптического уравнения, например, уравнения Пуассона, может быть получено в виде предела при  $t \rightarrow +\infty$  решения нестационарной задачи для уравнения теплопроводности. Поэтому далее рассмотрим алгоритм решения многомерного уравнения теплопроводности.

При нахождении решения счетом на установление желательно расчет вести с максимально возможным шагом по времени  $\tau$ . При этом использовать явную схему для двумерного случая можно лишь при выполнении условия

$$\tau \leq \frac{1}{2} \left( \frac{1}{h_1^2} + \frac{1}{h_2^2} \right)^{-1}.$$

Очевидно, что в трехмерном случае в правой части соответствующего неравенства появится еще слагаемое  $1/h_3^2$ . Следовательно,  $\tau = O(h^2)$ , что приводит к большому числу временных слоев для нахождения установившегося решения.

Обойти ограничение на устойчивость можно при использовании неявных схем ( $\sigma \neq 0$ ). Но при этом для нахождения решения на новом временном слое потребуется решить систему линейных алгебраических уравнений (СЛАУ), мало чем отличающуюся от системы для исходного эллиптического уравнения.

Рассмотрим в прямоугольнике  $G = [0, l_1] \times [0, l_2]$ ,  $t > 0$ , следующую задачу:

$$\begin{aligned} u_t &= u_{x_1 x_1} + u_{x_2 x_2} + f, \\ u|_{t=0} &= u_0(x_1, x_2), \\ u|_{\Gamma} &= \mu. \end{aligned}$$

Будем решать эту задачу численно, пытаясь избавиться от ограничения  $\tau = O(h^2)$ .

**Определение 10.2.** *Разностные схемы*, в которых можно использовать шаг  $\tau = O(h)$  и число действий в которых для перехода со слоя на слой есть величина  $O(N^p)$ , где  $N$  — число точек по каждой пространственной переменной,  $p$  — размерность пространства (число переменных), называются **экономичными**.

Отметим, что  $N^p$  — число точек пространственной сетки. Использование метода Гаусса для решения СЛАУ требует около  $N^{3p}$  действий. Это и ведет к разработке специальных алгоритмов.

Введем для удобства записи операторы

$$\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}.$$

#### 10.4.1. Продольно-поперечная схема

Эта схема называется также *схемой переменных направлений*. Для ее построения используем шаблон, изображенный на рис. 10.4, введя промежуточный (через  $\tau/2$ ) временной слой. Отнесем к нему сеточные функции  $\bar{y}$ ,  $\bar{\varphi}$  и составим схему:

$$\begin{aligned}\frac{2}{\tau}(\bar{y} - y) &= \Lambda_1 \bar{y} + \Lambda_2 y + \bar{\varphi}, \\ \frac{2}{\tau}(\hat{y} - \bar{y}) &= \Lambda_1 \bar{y} + \Lambda_2 \hat{y} + \bar{\varphi}.\end{aligned}$$

Таким образом, вычисление решения  $\hat{y}$  на новом временном слое состоит из двух последовательных шагов.

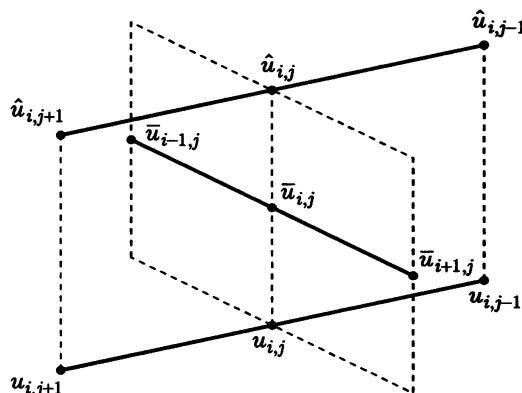


Рис. 10.4

1. Сначала для любого  $x_2$  решается полностью неявное разностное уравнение теплопроводности

$$\frac{2}{\tau}(\bar{y} - y) = \Lambda_1 \bar{y} + \varphi_1,$$

где  $\varphi_1 = \Lambda_2 y + \bar{\varphi}$  — известная функция. Это решение может быть найдено с помощью обычной прогонки.

2. Далее с помощью прогонки для любого  $x_1$  решается полностью неявное разностное уравнение теплопроводности

$$\frac{2}{\tau}(\hat{y} - \bar{y}) = \Lambda_2 \hat{y} + \varphi_2,$$

где  $\varphi_2 = \Lambda_1 \bar{y} + \bar{\varphi}$  — известная функция. В результате получаем, что необходимое для нахождения решения число действий равно  $O(N^p)$ ,  $p = 2$ .

Исследуем *аппроксимацию* данной разностной схемы. Для этого вычтем уравнения двух последовательных шагов схемы друг из друга:

$$\bar{y} = \frac{1}{2}(\hat{y} + y) + \frac{\tau}{4}\Lambda_2(y - \hat{y}),$$

откуда получим следующее выражение для суммы уравнений схемы:

$$\frac{1}{\tau}(\hat{y} - y) = \frac{1}{2}\Lambda_2(y + \hat{y}) + \Lambda_1 \bar{y} + \bar{\varphi} = \frac{1}{2}(\Lambda_1 + \Lambda_2)(\hat{y} + y) - \frac{\tau}{4}\Lambda_1\Lambda_2(\hat{y} - y) + \bar{\varphi},$$

т. е.

$$y_t = \Lambda y^{(0,5)} - \frac{\tau^2}{4}\Lambda_1\Lambda_2 y_t + \bar{\varphi}, \quad y_t = \frac{1}{\tau}(\hat{y} - y), \quad \Lambda = \Lambda_1 + \Lambda_2.$$

Все слагаемые данного уравнения нами уже изучались. Поэтому на решениях исходной задачи, имеющих третью непрерывные производные по  $t$ , четвертые — по  $x$  и пятые производные вида  $u_{x_1^\alpha x_2^\beta t}$ ,  $\alpha + \beta = 4$ , невязка данной разностной схемы равна  $\psi_h = O(h_1^2 + h_2^2 + \tau^2)$ . Это очевидно и потому, что данная разностная схема отличается от симметричной (по  $t$ ) схемы слагаемым порядка  $O(\tau^2)$ .

Исследуем устойчивость данной схемы по начальным данным, для чего запишем ее в каноническом виде  $By_t + Ay = f$ , воспользовавшись заменой  $\hat{y} = y + \tau y_t$ . Отсюда

$$B = E + \frac{1}{2}\tau A + \frac{\tau^2}{4}\Lambda_1\Lambda_2, \quad A = -\Lambda.$$

Считаем, что граничные условия являются однородными (если это не так, то исковую функцию  $y$  нужно изменить так, чтобы это было

выполнено). Тогда операторы  $A$ ,  $-\Lambda_1$ ,  $-\Lambda_2$  являются самосопряженными и положительно определенными. Следовательно, выполнены все условия теоремы 7.10. Устойчивость имеет место, если справедливо неравенство  $B \geq \tau/2A$ . Из положительности операторов  $E$ ,  $-\Lambda_1$ ,  $-\Lambda_2$  получаем, что искомое неравенство выполнено для любого  $\tau$ . Таким образом, схема устойчива по начальным данным в энергетической норме  $\|y\|_A^2 = (Ay, y)$ . Ограничимся исследованием устойчивости в этой норме. Использование метода разделения переменных приводит в данном случае также к безусловной устойчивости, в том числе и по правой части, в норме  $L_2$ . Отсюда получаем, что данная разностная схема сходится к решению исходной дифференциальной задачи со скоростью  $O(\tau^2 + h_1^2 + h_2^2)$  в норме  $L_2$ .

#### 10.4.2. Локально-одномерная схема

Продольно-поперечная схема на задачи с  $p \geq 3$  непосредственно не обобщается вследствие возникающих несимметричности и условной устойчивости. Имевшаяся в двумерном случае симметричность давала равные (по модулю) ошибки с разными знаками на двух последовательных шагах, компенсировавшие друг друга.

Рассмотрим так называемую *локально-одномерную схему* также с использованием промежуточных (дробных) слоев. Эта схема имеет лишь суммарную аппроксимацию, а на промежуточных слоях она не аппроксимирует исходное дифференциальное уравнение. Однако ошибки аппроксимации при суммировании гасят друг друга. Так что решение на «целом» слое оказывается приближением точного.

Рассмотрим уравнение

$$u_t = \sum_{i=1}^p u_{x_i x_i} + f.$$

Аппроксимируем это уравнение, используя симметричную неявную схему

$$y_t = \sum_{i=1}^p \Lambda_i y^{(0,5)} + \varphi,$$

( $\Lambda_i$  — те же операторы, что и ранее). Ошибка аппроксимации такой схемы равна  $O(\tau^2 + h^2)$  при правильном выборе  $\varphi$ .

Наряду с исходной схемой построим локально-одномерную схему. Для этого между слоями  $t$  и  $\hat{t}$  введем  $p+1$  промежуточных слоев с шагами  $\tau/p$  между ними. Первый слой соответствует моменту времени  $t$ , последний с номером  $p+1$  — моменту времени  $\hat{t}$ . На каждом

таком слое с номером  $\alpha$  суммарный оператор в левой части заменим оператором  $r\Lambda_\alpha$ . Обозначим решение на промежуточных шагах через  $w_\alpha$ ,  $\alpha = 1, 2, \dots, p$ . Тогда  $w_\alpha$  является решением следующей разностной задачи:

$$\begin{aligned} \frac{1}{\tau}(\hat{w}_\alpha - w_\alpha) &= \frac{1}{2}\Lambda_\alpha(\hat{w}_\alpha + w_\alpha) + \varphi_\alpha, \quad \alpha = 1, 2, \dots, p; \\ w_1 &= y, \quad w_2 = \hat{w}_1, \quad \dots, \quad w_p = \hat{w}_{p-1}, \quad \hat{w}_p = \hat{y}. \end{aligned}$$

Очевидно, что для любого  $p$  соответствующее уравнение является одномерным, решаемым методом обычной прогонки. Остальные независимые переменные участвуют в нем только в качестве параметров. Поэтому и схема называется локально-одномерной.

Ясно, что отдельное уравнение из представленной системы не аппроксимирует исходную задачу. Но при этом для любого  $\alpha$  вычисление  $\hat{w}_\alpha$  по симметричной неявной схеме является безусловно устойчивым в  $L_2$  и допускает решение при любом  $\tau$ , в том числе при  $\tau = O(h)$ . Очевидно, что требуемое для нахождения  $\hat{y}$  число действий равно  $O(N^p)$ , как и для каждой прогонки.

Для исследования аппроксимации сравним рассматриваемую схему с исходной. Для этого, опустив правые части, запишем ее уравнения в виде

$$\begin{aligned} \left(E - \frac{1}{2}\tau\Lambda_\alpha\right)\hat{w}_\alpha &= \left(E + \frac{1}{2}\tau\Lambda_\alpha\right)w_\alpha, \\ w_\alpha &= \hat{w}_{\alpha-1}, \quad \alpha = 2, 3, \dots, p, \quad w_1 = y. \end{aligned}$$

Операторы второй производной по разным переменным перестановочны. Разностные операторы  $\Lambda_\alpha$  тоже попарно перестановочны. Отсюда получаем

$$\left(E - \frac{1}{2}\tau\Lambda_p\right)\hat{w}_p = \left(E + \frac{1}{2}\tau\Lambda_p\right)\hat{w}_{p-1},$$

следовательно,

$$\left(E - \frac{1}{2}\tau\Lambda_{p-1}\right)\left(E - \frac{1}{2}\tau\Lambda_p\right)\hat{w}_p = \left(E + \frac{1}{2}\tau\Lambda_p\right)\left(E + \frac{1}{2}\tau\Lambda_{p-1}\right)\hat{w}_{p-2},$$

и т. д. В результате запишем

$$\prod_{k=1}^p \left(E - \frac{1}{2}\tau\Lambda_k\right)\hat{w}_p = \prod_{k=1}^p \left(E + \frac{1}{2}\tau\Lambda_k\right)w_1.$$

Раскроем произведение, опустив при этом члены порядка выше  $O(\tau^2)$ . Используя равенства  $w_1 = y$ ,  $\hat{w}_p = \hat{y}$ , получим

$$\left(E - \frac{1}{2}\tau\Lambda + \frac{1}{4}\tau^2 \sum_{i,j} \Lambda_i \Lambda_j + O(\tau^3)\right)\hat{y} = \left(E + \frac{1}{2}\tau\Lambda + \frac{1}{4}\tau^2 \sum_{i,j} \Lambda_i \Lambda_j + O(\tau^3)\right)y,$$

т. е.

$$y_t = \frac{1}{2} \sum_{\alpha=1}^p \Lambda_\alpha (\hat{y} + y) - \frac{\tau^2}{4} \sum_{i,j} \Lambda_i \Lambda_j y_t + O(\tau^3).$$

Отсюда видно, что данная схема обладает так называемой «*суммарной аппроксимацией*  $\psi_h = O(\tau^2 + h^2)$ , так как она отличается от исходной слагаемыми порядка  $O(\tau^2)$  при условии существования у точного решения непрерывных производных вида  $u_{x_i^p x_j^q t}$  достаточно высокого порядка.

Как уже указывалось, локально-одномерная схема является безусловно устойчивой, откуда следует сходимость приближенного решения к точному со скоростью  $O(\tau^2 + h^2)$ .

Анализ опущенных правых частей показывает, что для обеспечения аппроксимации они должны удовлетворять условию

$$\varphi = \sum_{\alpha=1}^p \varphi_\alpha.$$

Итак, представлены две схемы из так называемого класса *экономичных разностных схем*. Их достоинство состоит в безусловной устойчивости и применимости экономичных одномерных алгоритмов типа *прогонки* для нахождения решения за малое число действий. Недостатком является несимметричность алгоритмов, которая при решении неустойчивых задач может сильно искажить получаемое численное решение.

## 10.5. Проекционные методы решения эллиптических уравнений

До сих пор мы рассматривали конечно-разностные методы. Рассмотрим теперь алгоритмы, основанные на вариационных и проекционных принципах. Для определенности опять будем рассматривать задачу *Дирихле* с однородными граничными условиями для *уравнения Пуассона* в прямоугольнике в двумерном случае.

### 10.5.1. Метод Ритца

В линейном пространстве бесконечно дифференцируемых финитных в  $G$  функций введем норму

$$\|u\|_{H^1}^2 = \int_G (\operatorname{grad} u)^2 dx_1 dx_2.$$

Пополним это пространство. Получим пространство  $H^1$ , являющееся гильбертовым.

Рассмотрим задачу о нахождении минимума функционала  $\Phi(u)$ :

$$\Phi(u) = \int_G ((\operatorname{grad} u)^2 - 2uf) dx_1 dx_2.$$

Допустим, что решение этой задачи  $\bar{u}$  обладает вторыми непрерывными производными. Тогда для произвольного параметра  $\lambda$  и функции  $v \in H^1$  на функциях  $u = \bar{u} + \lambda v$  функционал  $\Phi(u)$  принимает большие значения:

$$\begin{aligned} \Phi(u) = \int_G & \left( (\operatorname{grad} \bar{u})^2 + 2\lambda(\operatorname{grad} \bar{u}, \operatorname{grad} v) + \right. \\ & \left. + \lambda^2(\operatorname{grad} v)^2 - 2\bar{u}f - 2\lambda vf \right) dx_1 dx_2. \end{aligned}$$

Из того факта, что  $\Phi(u)$  как функция параметра  $\lambda$  принимает минимальное значение при  $\lambda = 0$ , следует равенство

$$\frac{d\Phi(u)}{d\lambda} \Big|_{\lambda=0} = 0,$$

т. е.

$$\int_G ((\operatorname{grad} \bar{u}, \operatorname{grad} v) - vf) dx_1 dx_2 = 0.$$

Отсюда для функции  $\bar{u}$ , имеющей вторые непрерывные производные и обращающейся в нуль на границе, получаем

$$\int_G (-\Delta \bar{u} - f)v dx_1 dx_2 = 0, \quad v \in H^1,$$

что возможно только тогда, когда

$$-\Delta \bar{u} = f.$$

Это равенство можно получить, если выбрать  $v = \Delta \bar{u} + f$ . Таким образом, решение  $\bar{u}$ , доставляющее минимум функционалу и имеющее непрерывные вторые производные, является решением исходной задачи. При этом необязательно, чтобы решение  $\bar{u}$ , дающее минимум функционалу  $\Phi(u)$ , имело непрерывные вторые производные. Поэтому решение задачи минимизации функционала называют **обобщенным решением** краевой задачи

$$-\Delta u = f, \quad u|_{\Gamma} = 0.$$

Данный метод решения краевой задачи называется **методом Ритца**.

Будем искать приближенное решение задачи следующим образом. Заменим бесконечномерное пространство функций  $H^1$  на конечномерное пространство  $H_h^1$  с помощью сетки

$$\bar{\omega}_h = \{x_{1,i} = ih_1, i = 0, 1, \dots, N_1, x_{2,j} = jh_2, j = 0, 1, \dots, N_2\}.$$

Разобьем каждый прямоугольник на два треугольника диагональю, выполнив простейшую триангуляцию области (рис. 10.5). В качестве

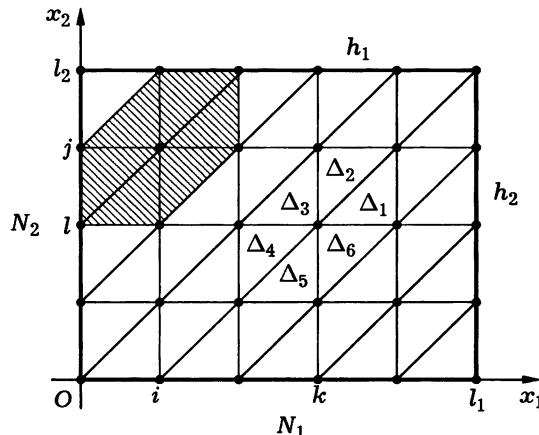


Рис. 10.5

$H_h^1$  берем пространство конечных элементов, соответствующих данной триангуляции:

$$y = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} y_{ij} \varphi_{ij}(x_1, x_2).$$

Здесь  $y_{ij}$  — значения приближенного решения в точках сетки;  $\varphi_{ij}$  — кусочно-линейные по  $x_1, x_2$  в пределах треугольника (конечного элемента) базисные функции, равные единице в точке  $(i, j)$  и нулю в остальных (рис. 10.6). Вследствие нулевого граничного условия сумма не содержит слагаемых с  $i = 0, N_1, j = 0, N_2$ .

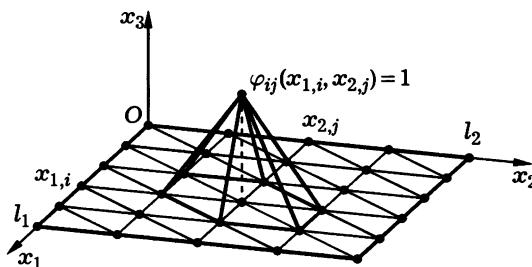


Рис. 10.6

Условие минимума функционала дает равенство

$$\frac{\partial \Phi}{\partial y_{kl}} \left( \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} y_{ij} \varphi_{ij} \right) = 0.$$

Отсюда получаем

$$\begin{aligned} \frac{\partial}{\partial y_{kl}} \int_G \left( \left( \sum_{i,j} y_{ij} \frac{\partial \varphi_{ij}}{\partial x_1} \right)^2 + \left( \sum_{i,j} y_{ij} \frac{\partial \varphi_{ij}}{\partial x_2} \right)^2 - 2f \sum_{i,j} y_{ij} \varphi_{ij} \right) dx_1 dx_2 = \\ = 2 \int_G \left( \sum_{i,j} y_{ij} \frac{\partial \varphi_{ij}}{\partial x_1} \frac{\partial \varphi_{kl}}{\partial x_1} + \sum_{i,j} y_{ij} \frac{\partial \varphi_{ij}}{\partial x_2} \frac{\partial \varphi_{kl}}{\partial x_2} - f \varphi_{kl} \right) dx_1 dx_2 = 0. \end{aligned}$$

Необходимо вычислить входящие в последнее равенство величины (см. рис. 10.5). Так как  $\partial \varphi_{kl} / \partial x_1$ ,  $\partial \varphi_{kl} / \partial x_2$  могут быть отличны от нуля только в треугольниках  $\Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5, \Delta_6$ , то полученное равенство, т. е. уравнение для  $y_{ij}$ , содержит  $y_{ij}$  только при  $i = k-1, k, k+1$ ,  $j = l-1, l, l+1$ .

Для подтверждения этого приведем значения градиента  $\varphi_{kl}$  в указанных треугольниках:

$$\text{grad}(\varphi_{kl}) = \begin{cases} \left( -\frac{1}{h_1}, 0 \right) & \text{в } \Delta_1; \\ \left( 0, -\frac{1}{h_2} \right) & \text{в } \Delta_2; \\ \left( \frac{1}{h_1}, -\frac{1}{h_2} \right) & \text{в } \Delta_3; \\ \left( \frac{1}{h_1}, 0 \right) & \text{в } \Delta_4; \\ \left( 0, \frac{1}{h_2} \right) & \text{в } \Delta_5; \\ \left( -\frac{1}{h_1}, \frac{1}{h_2} \right) & \text{в } \Delta_6. \end{cases}$$

Обозначим шестиугольник, составленный из данных треугольников, через  $G_{kl}$ .

Пусть

$$\mathcal{J}_{kl}^1(i, j) = \int_{G_{kl}} \frac{\partial \varphi_{ij}}{\partial x_1} \frac{\partial \varphi_{kl}}{\partial x_1} dx_1 dx_2 = \int_{\Delta_1 + \Delta_3 + \Delta_4 + \Delta_6} \frac{\partial \varphi_{ij}}{\partial x_1} \frac{\partial \varphi_{kl}}{\partial x_1} dx_1 dx_2,$$

так как в треугольниках  $\Delta_2$  и  $\Delta_5$  выполнено равенство

$$\frac{\partial \varphi_{kl}}{\partial x_1} = 0.$$

Поскольку функции формы  $\varphi$  представляют собой линейные многочлены в пределах каждого треугольника, а сетка является прямоугольной, производные вычисляются без труда. В результате получаем следующие значения коэффициентов:

$$\mathcal{J}_{kl}^1(k, l) = \frac{2h_2}{h_1}, \quad \mathcal{J}_{kl}^1(k+1, l) = \mathcal{J}_{kl}^1(k-1, l) = -\frac{h_2}{h_1},$$

$$\mathcal{J}_{kl}^1(i, j) = 0 \text{ для всех остальных } i, j.$$

Аналогично

$$\mathcal{J}_{kl}^2(i, j) = \int_G \frac{\partial \varphi_{ij}}{\partial x_2} \frac{\partial \varphi_{kl}}{\partial x_2} dx_1 dx_2 = \int_{\Delta_2 + \Delta_3 + \Delta_5 + \Delta_6} \frac{\partial \varphi_{ij}}{\partial x_2} \frac{\partial \varphi_{kl}}{\partial x_2} dx_1 dx_2,$$

так как в треугольниках  $\Delta_1$  и  $\Delta_4$  выполнено равенство  $\partial \varphi_{kl} / \partial x_2 = 0$ . Отсюда

$$\mathcal{J}_{kl}^2(k, l) = \frac{2h_1}{h_2}, \quad \mathcal{J}_{kl}^2(k, l+1) = \mathcal{J}_{kl}^2(k, l-1) = -\frac{h_1}{h_2},$$

$$\mathcal{J}_{kl}^2(i, j) = 0 \text{ для всех остальных } i, j.$$

В результате имеем систему линейных алгебраических уравнений (СЛАУ) для  $y_{kl}$ :

$$\left( \frac{2h_1}{h_2} + \frac{2h_1}{h_2} \right) y_{kl} - \frac{h_1}{h_2} (y_{k,l-1} + y_{k,l+1}) - \frac{h_2}{h_1} (y_{k-1,l} + y_{k+1,l}) = h_1 h_2 g_{kl},$$

$$g_{kl} = \frac{1}{h_1 h_2} \int_G f \varphi_{kl} dx_1 dx_2.$$

Если в последнем интеграле вычислить его значение приближенно по формуле типа центральных прямоугольников, то получим

$$g_{kl} \approx \frac{1}{h_1 h_2} f_{kl} \int_G \varphi_{kl} dx_1 dx_2 = f_{kl},$$

так как объем под графиком функции  $\varphi_{kl}$  — пирамидой единичной высоты, площадь основания которой  $3h_1 h_2$ , — равен  $h_1 h_2$ .

В результате, разделив уравнение системы на  $h_1 h_2$ , получим хорошо знакомое разностное уравнение

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -g.$$

Отличие от разностной схемы состоит только в способе аппроксимации правой части. Однако в случае сетки, построенной произвольным образом, полученная СЛАУ может быть такой, предугадать и записать которую с помощью разностных методов весьма сложно.

### 10.5.2. Метод Галеркина

Рассмотрим чуть измененную задачу

$$-\Delta u = f$$

в прямоугольнике  $G = (0, l_1) \times (0, l_2)$  с границей  $\partial G$  (рис. 10.7) с граничными условиями

$$\left( \frac{\partial u}{\partial \vec{n}} + \alpha u \right) \Big|_{\Gamma} = 0, \quad u \Big|_{\partial G \setminus \Gamma} = 0,$$

где  $\Gamma = \{(x_1, x_2): x_1 = l_1\}$ ;  $\vec{n}$  — вектор внешней нормали. Допустим, что решение задачи существует. Умножим уравнение на функцию  $\varphi$  с кусочно-непрерывными производными, равную нулю на  $\partial G \setminus \Gamma$ . Получим так называемое *интегральное тождество*:

$$\int_G (\operatorname{grad} u, \operatorname{grad} \varphi) dx_1 dx_2 + \int_{\Gamma} \alpha u \varphi dx_2 = \int_G f \varphi dx_1 dx_2.$$

Нетрудно видеть, что классическое решение исходной задачи, имеющее квадратично суммируемые производные, удовлетворяет этому тождеству. Обратное неверно: функция  $u$ , удовлетворяющая данному интегральному тождеству для любой  $\varphi$  из указанного класса, не обязана иметь вторые производные и тем самым удовлетворять исходному уравнению в обычном смысле.

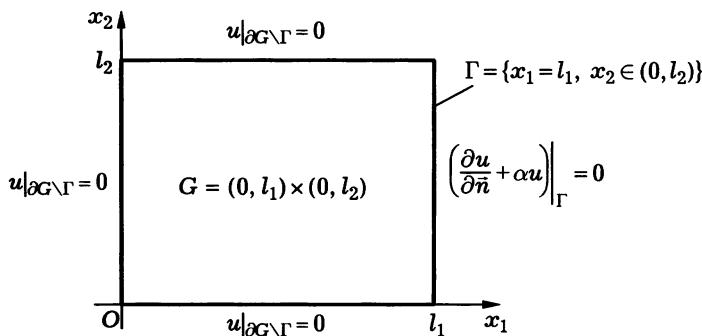


Рис. 10.7

Функцию  $u$ , удовлетворяющую интегральному тождеству и имеющую интегрируемые с квадратом производные, называют **обобщенным решением** исходной задачи.

Отметим, что определений обобщенных решений очень много. Выбор нужного зависит от специфики задачи. Необходимость введения понятия обобщенного решения диктуется тем, что классическое решение часто отсутствует, в то время как модель требует наличия решения, которое может быть и не столь гладким, как классическое. В то же время достаточно гладкое обобщенное решение оказывается классическим.

Поступим так же, как и в методе Ритца: проведем триангуляцию и найдем решение в виде

$$y = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} y_{ij} \varphi_{ij}(x_1, x_2).$$

В данной сумме присутствуют, вообще говоря, ненулевые значения решения на границе  $x_1 = l_1$ , соответствующие  $i = N_1$ . Потребуем, чтобы интегральное тождество было выполнено для любой  $\varphi \in H_h^1$ , где  $H_h^1$  — функции указанного в **10.5.1** вида.

Описываемый метод решения задачи называют **методом Галеркина**.

В результате имеем СЛАУ

$$\begin{aligned} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} y_{ij} \left( \int_G (\operatorname{grad} \varphi_{ij}, \operatorname{grad} \varphi_{kl}) dx_1 dx_2 + \int_{\Gamma} \alpha \varphi_{ij} \varphi_{kl} dx_2 \right) &= \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} \int_G f \varphi_{kl} dx_1 dx_2, \quad 1 \leq k \leq N_1, \quad 1 \leq l \leq N_2 - 1. \end{aligned}$$

Рассмотрим вид полученных уравнений.

Пусть  $k \neq N_1$ , тогда  $\varphi_{kl}(l_1, x_2) = 0$  и полученное уравнение имеет тот же вид, что и аналогичное уравнение метода Ритца, т. е.

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -g, \quad g_{kl} = \frac{1}{h_1 h_2} \int_G f \varphi_{kl} dx_1 dx_2.$$

Для индекса  $k = N_1$ , соответствующего границе  $x_1 = l_1$ , получим следующее уравнение:

$$-\frac{1}{2} y_{\bar{x}_2 x_2, N_1, j} + \frac{1}{h_1} (y_{\bar{x}_1, N_1, j} + \alpha_{-1} y_{N_1, j-1} + \alpha y_{N_1, j} + \alpha_{+1} y_{N_1, j+1}) = g_{N_1, j},$$

где

$$\alpha_{-1} = \frac{1}{h_2} \int_G \alpha \varphi_{N_1,j} \varphi_{N_1,j-1} dx_2;$$

$$\alpha = \frac{1}{h_2} \int_G \alpha \varphi_{N_1,j} \varphi_{N_1,j} dx_2;$$

$$\alpha_{+1} = \frac{1}{h_2} \int_G \alpha \varphi_{N_1,j} \varphi_{N_1,j+1} dx_2.$$

Рассматриваемый метод позволяет достаточно просто решать задачи с граничными данными смешанного типа. Нетрудно видеть, что в данной простой ситуации методы Галеркина и Ритца (при небольшой модификации функционала) дают одинаковые решения. Как и в методе Ритца, в случае произвольной сетки СЛАУ, полученная методом Галеркина, может быть такой, которую предугадать и записать с помощью разностных методов весьма сложно.

Отметим, что прогресс в области применения треугольных сеток в значительной степени определяется наличием программ триангуляции сложных областей и программ для решения больших СЛАУ. Конечные элементы сложного строения могут обеспечить и повышенный порядок аппроксимации.

## 10.6. Оператор Лапласа в криволинейных координатах и его разностная аппроксимация

До сих пор мы рассматривали только декартовы координаты, в которых составляющие оператора Лапласа слагаемые по каждой из координат могут аппроксимироваться одинаковым способом. В случае цилиндрических или сферических координат ситуация иная. В связи с широким применением таких систем координат представим методы построения *разностных аппроксимаций оператора Лапласа* в указанных случаях.

Рассмотрим оператор Лапласа

$$Au = \Delta u = \operatorname{div} \operatorname{grad} u$$

в некоторой криволинейной системе  $Ox_1x_2x_3$ .

Введем прямоугольную в рассматриваемой системе координат разностную сетку в области  $\Omega$ , в которой решается задача. Точки сетки пронумеруем индексами  $i, j, k$ . Обозначим через  $h_{\alpha,l} = x_{\alpha,l} - x_{\alpha,l-1}$ ,  $l = i, j, k$ , шаги сетки по направлению  $x_\alpha$ ,  $\alpha = 1, 2, 3$ . Выберем прямоугольную разностную ячейку  $\omega_{i,j,k}$  с центром в точке  $(i, j, k)$  размером

$\hbar_{1,i} \times \hbar_{2,j} \times \hbar_{3,k}$ . Считаем, что координаты точек ячейки  $\omega_{i,j,k}$  таковы, что  $x_{\alpha,l} \in [x_{\alpha,l-1/2}, x_{\alpha,l+1/2}]$ , т. е.  $\hbar_{\alpha,l} = x_{\alpha,l+1/2} - x_{\alpha,l-1/2}$ . В свою очередь, точки с полуцелыми индексами  $x_{\alpha,l+1/2} \in [x_{\alpha,l}, x_{\alpha,l+1}]$  находятся, вообще говоря, необязательно в середине отрезка. Здесь  $\alpha = 1, 2, 3$ ,  $l = i, j, k$ . Нормали к граням ячейки по направлению совпадают с направлениями координатных осей.

Получим разностную аппроксимацию интегро-интерполяционным методом. Для этого проинтегрируем  $Au = \Delta u$  по рассматриваемой ячейке. Тогда, согласно *теореме Гаусса — Остроградского*, получим следующее интегральное соотношение, справедливое для различных систем координат:

$$\begin{aligned} \int_{\omega_{i,j,k}} Au dV &= \int_{\omega_{i,j,k}} \operatorname{div} \nabla u dV = \oint_{\partial\omega_{i,j,k}} \frac{\partial u}{\partial n} dS_n = \\ &= \int_{\partial\omega_{i,j,k,1,i+1/2}} (\nabla u)_{1,i+1/2} dS_{n,1,i+1/2} - \int_{\partial\omega_{i,j,k,1,i-1/2}} (\nabla u)_{1,i-1/2} dS_{n,1,i-1/2} + \\ &+ \int_{\partial\omega_{i,j,k,2,j+1/2}} (\nabla u)_{2,j+1/2} dS_{n,2,j+1/2} - \int_{\partial\omega_{i,j,k,2,j-1/2}} (\nabla u)_{2,j-1/2} dS_{n,2,j-1/2} + \\ &+ \int_{\partial\omega_{i,j,k,3,k+1/2}} (\nabla u)_{3,k+1/2} dS_{n,3,k+1/2} - \int_{\partial\omega_{i,j,k,3,k-1/2}} (\nabla u)_{3,k-1/2} dS_{n,3,k-1/2}. \end{aligned}$$

В этом выражении  $dV$  — дифференциал объема;  $\nabla$  — оператор Гамильтона,  $\nabla \equiv \operatorname{grad}$ ;  $dS_n$  — проекция дифференциала площади на внешнюю нормаль  $n$ ;  $\partial\omega_{i,j,k}$  — граница ячейки  $\omega_{i,j,k}$ ;  $\partial\omega_{i,j,k,\alpha,l\pm1/2}$  — часть границы указанной ячейки, нормаль к которой по направлению совпадает с направлением координатной оси  $\alpha$  ( $\alpha = 1, 2, 3$ ,  $l = i, j, k$ ).

Пусть  $V_{i,j,k}$  — объем ячейки  $\omega_{i,j,k}$ . Разделим полученное интегральное соотношение на  $V_{i,j,k}$  и результат будем использовать для получения разностной аппроксимации оператора Лапласа.

Рассмотрим конкретные варианты систем координат. Построим для них аппроксимации оператора Лапласа на основе полученного интегрального соотношения.

### 10.6.1. Цилиндрические координаты

Рассмотрим цилиндрическую систему координат  $(r, \varphi, z)$ , где  $x = r \cos \varphi$ ,  $y = r \sin \varphi$ ,  $(x, y, z)$  — декартовы координаты. В этой системе координат дифференциал объема  $dV = r dr d\varphi dz$ , вектор дифференциала площади  $dS = (rd\varphi dz, drdz, rdrd\varphi)$ .

Запишем выражение для лапласиана:

$$\Delta u = \operatorname{div} \operatorname{grad} u = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \varphi^2} + \frac{\partial^2 u}{\partial z^2}.$$

Компоненты градиента равны

$$\operatorname{grad} u = \left( \frac{\partial u}{\partial r}, \frac{1}{r} \frac{\partial u}{\partial \varphi}, \frac{\partial u}{\partial z} \right).$$

Объем ячейки  $V_{h,i,j,k} = 0,5(r_{i+1/2}^2 - r_{i-1/2}^2)h_{\varphi,j}h_{z,k}$ .

Аппроксимируя интегральное соотношение (см. 10.6), получим следующий результат:

$$\begin{aligned} V_{h,i,j,k} A_{h,i,j,k} y &= \\ &= \frac{y_{i+1,j,k} - y_{i,j,k}}{h_{r,i+1}} r_{i+1/2} h_{\varphi,j} h_{z,k} - \frac{y_{i,j,k} - y_{i-1,j,k}}{h_{r,i}} r_{i-1/2} h_{\varphi,j} h_{z,k} + \\ &+ \frac{1}{r_i} \frac{y_{i,j+1,k} - y_{i,j,k}}{h_{\varphi,j+1}} h_{r,i} h_{z,k} - \frac{1}{r_i} \frac{y_{i,j,k} - y_{i,j-1,k}}{h_{\varphi,j}} h_{r,i} h_{z,k} + \\ &+ \frac{y_{i,j,k+1} - y_{i,j,k}}{h_{z,k+1}} \frac{r_{i+1/2}^2 - r_{i-1/2}^2}{2} h_{\varphi,j} - \frac{y_{i,j,k} - y_{i,j,k-1}}{h_{z,k}} \frac{r_{i+1/2}^2 - r_{i-1/2}^2}{2} h_{\varphi,j}. \end{aligned}$$

Полученное выражение можно заметно упростить, если использовать равномерную сетку или, например, формулу  $r_i = \frac{1}{2}(r_{i+1/2} + r_{i-1/2})$ . Можно также упростить выражения для объема элемента ячейки сетки и площадей ее поверхностей.

Видно, что построенная аппроксимация сохраняет такое важное свойство исходного оператора, как *самосопряженность* при соответствующих ограничениях на граничные условия. Применение *интегро-интерполяционного метода* автоматически обеспечивает *консервативность*.

Очевидно, что приведенную *аппроксимацию* можно использовать только тогда, когда знаменатель в ней не обращается в нуль, т. е. во внутренних ячейках, где  $r_i \neq 0$ .

Рассмотрим вопрос об аппроксимации условий в центре системы координат ( $r = 0$ ). Известно, что данная точка является особой и для обеспечения ограниченности решения в нуле и во всей области необходимо требовать выполнения условия  $r \partial u / \partial r \rightarrow 0$  при  $r \rightarrow 0$ .

Интегральное тождество (см. 10.6) позволяет тривиальным образом аппроксимировать данное условие. Для этого нужно построить сетку так, чтобы первая по направлению  $r$  точка имела ненулевую радиальную координату, например отстояла от центра на половину соседнего по радиусу шага сетки. Тогда в этой точке имеет место в точности та же по форме разностная аппроксимация, в которой формально положено  $r_{i-1/2} = 0$ , т. е. отсутствует второй член выражения.

Отметим, что возможно использование сетки, в которой одна из точек имеет координату  $r = 0$ . Очевидно, что решение в такой точке не зависит от угла. Аппроксимация может быть получена тем же способом, однако в качестве элементарной ячейки, по которой производится интегрирование, необходимо выбрать цилиндр.

### 10.6.2. Сферические координаты

Рассмотрим сферическую систему координат  $(r, \varphi, \vartheta)$ , которая с декартовыми координатами  $x, y, z$  связана уравнениями  $x = r \sin \vartheta \cos \varphi$ ,  $y = r \sin \vartheta \sin \varphi$ ,  $z = r \cos \vartheta$ . В этой системе координат дифференциал объема равен  $dV = r^2 \sin \vartheta dr d\varphi d\vartheta$ , вектор дифференциала площади равен  $dS = (r^2 \sin \vartheta d\varphi d\vartheta, r dr d\vartheta, r \sin \vartheta dr d\varphi)$ .

Выражение для лапласиана имеет вид

$$\Delta u = \operatorname{div} \operatorname{grad} u = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin^2 \vartheta} \frac{\partial^2 u}{\partial \varphi^2} + \frac{1}{r^2 \sin \vartheta} \frac{\partial}{\partial \vartheta} \left( \sin \vartheta \frac{\partial u}{\partial \vartheta} \right),$$

компоненты градиента равны

$$\operatorname{grad} u = \left( \frac{\partial u}{\partial r}, \frac{1}{r \sin \vartheta} \frac{\partial u}{\partial \varphi}, \frac{1}{r} \frac{\partial u}{\partial \vartheta} \right).$$

Объем ячейки  $V_{h,i,j,k} = \frac{1}{3} (r_{i+1/2}^3 - r_{i-1/2}^3) \hbar_{\varphi,j} \sin \vartheta_k \hbar_{\vartheta,k}$ .

Аппроксимируя интегральное соотношение (см. 10.6), получаем следующий результат:

$$\begin{aligned} V_{h,i,j,k} A_{h,i,j,k} y &= \frac{y_{i+1,j,k} - y_{i,j,k}}{h_{r,i+1}} r_{i+1/2}^2 \hbar_{\varphi,j} \sin \vartheta_k \hbar_{\vartheta,k} - \\ &- \frac{y_{i,j,k} - y_{i-1,j,k}}{h_{r,i}} r_{i-1/2}^2 \hbar_{\varphi,j} \sin \vartheta_k \hbar_{\vartheta,k} + \frac{1}{\sin \vartheta_k} \frac{y_{i,j+1,k} - y_{i,j,k}}{h_{\vartheta,j+1}} \hbar_{r,i} \hbar_{\vartheta,k} - \\ &- \frac{1}{\sin \vartheta_k} \frac{y_{i,j,k} - y_{i,j-1,k}}{h_{\varphi,j}} \hbar_{r,i} \hbar_{\vartheta,k} + \frac{y_{i,j,k+1} - y_{i,j,k}}{h_{\vartheta,k+1}} \hbar_{r,i} \sin \vartheta_{k+1/2} \hbar_{\varphi,j} - \\ &- \frac{y_{i,j,k} - y_{i,j,k-1}}{h_{\vartheta,k}} \hbar_{r,i} \sin \vartheta_{k-1/2} \hbar_{\varphi,j}. \end{aligned}$$

Можно сделать те же замечания относительно свойств разностного оператора Лапласа в сферической системе координат, что и для цилиндрической системы координат.

В отличие от цилиндрической системы координат особыми точками для оператора в сферической системе координат являются не только центральная точка, в которой  $r = 0$ , но и точки, в которых  $\sin \vartheta = 0$ .

Известно, что для нахождения единственного ограниченного решения задачи для уравнения Лапласа в такой системе координат необходимо потребовать выполнения условий  $r^2 \partial u / \partial r \rightarrow 0$  при  $r \rightarrow 0$  и  $\sin \vartheta \partial u / \partial \vartheta \rightarrow 0$  при  $\sin \vartheta \rightarrow 0$ .

Аппроксимация данных условий не составляет труда и может быть выполнена точно так же, как и в цилиндрической системе координат с очевидными изменениями. При этом первую по радиусу точку необходимо отделить от нуля, равно как и точки по углу  $\vartheta$  — первую и последнюю — от своих границ.

**Замечание 10.2.** Описанный алгоритм построения разностной аппроксимации оператора Лапласа в цилиндрической и сферической системах координат применим и к другим системам. Принципиальным здесь является использование интегрального тождества.

**Замечание 10.3.** Записанные соотношения тривиальным образом обобщаются на случай эллиптических дивергентных операторов вида

$$Au = \operatorname{div}(\varkappa(\operatorname{grad} u))$$

с произвольным коэффициентом  $\varkappa$ , в том числе, возможно, зависящим и от решения.

## 10.7. Библиографические комментарии

Разностные методы решения уравнений эллиптического типа рассматриваются практически во всех основных руководствах, указанных в списке литературы. Так, дополнительные сведения можно найти в [6, 12, 14, 22, 25, 80, 84, 110, 113, 134, 137, 139, 141, 149, 151, 154, 155, 176].

Мы мало уделяем внимания столь важному методу решения уравнений математической физики, как метод конечных элементов. Его описание, теорию и практику применения можно найти в следующих книгах: [11, 23, 45, 60, 66, 72, 73, 93, 114, 117, 122, 157, 158, 165, 166, 189]. Метод конечных элементов в особенности успешно применяется именно для решения эллиптических краевых задач.

Аналогичное замечание можно сделать и по поводу вариационных методов. Они используются чаще всего для решения эллиптических

уравнений. При этом численный алгоритм обычно строится на основе некоторого вариационного принципа, имеющего с точки зрения законов природы фундаментальный характер. Описание, теорию и практику применения вариационных методов можно найти в книгах [17, 18, 25, 61, 62, 117, 137].

В [138] представлен оригинальный метод численного решения задач математической физики — метод разностных потенциалов, который изначально возник для решения эллиптических краевых задач, а сейчас применяется для решения многих других задач. Он во многом соединяет в себе достоинства разностных методов численного решения и методов типа граничных интегральных уравнений.

Методы, в которых решение сложной задачи разбивается на совокупность решения более простых задач, подобно продольно-поперечно-му алгоритму и локально-одномерной схеме, также не ограничиваются описанными вариантами. Более подробно эти методы рассмотрены в [14, 110, 111, 147, 192].

# 11. ЧИСЛЕННОЕ РЕШЕНИЕ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

Рассмотрены основные свойства *интегральных уравнений* и методы их численного решения. Для уравнений второго рода описаны разностный метод решения, метод *последовательных приближений*, метод замены ядра на вырожденное, метод Галеркина. На примере *уравнения Фредгольма первого рода* описан алгоритм регуляризации. Выведено уравнение Эйлера для соответствующего функционала. Качественно пояснено действие регуляризации. Приведены примеры *некорректно поставленных задач*.

## 11.1. Корректно поставленные задачи

**Определение 11.1.** *Интегральным* называется *уравнение*, в котором неизвестная функция стоит под знаком интеграла.

**Пример 11.1.** Уравнение

$$\int_a^b K(x, \xi, u(\xi)) d\xi = F(x, u(x))$$

является интегральным и, вообще говоря, нелинейным. #

Интегральные уравнения встречаются очень часто при описании самых разнообразных процессов. Можно даже услышать утверждения, не лишенные оснований, что интегральные уравнения являются первичными, а дифференциальные — вторичными. Действительно, исходные уравнения механики сплошной среды — интегральные уравнения балансов вещества, импульса, энергии. Из них в предположении существования нужных производных следуют дифференциальные уравнения. Огромное количество решений, содержащих какие-либо особенности (разрывы решения или его производных), может быть описано только с помощью интегральных уравнений, например *ударные волны* и другие разрывные решения. Кроме того, часто исходную дифференциальную задачу можно свести к решению интегрального уравнения.

В некоторых отношениях интегральные уравнения имеют преимущества:

- 1) накладывают меньшие требования на гладкость решения и, следовательно, описывают более широкий класс решений;
- 2) часто выражают основные физические законы сохранения и баланса;
- 3) содержат в себе всю нужную информацию о решении. Например,

$$u_t = a^2 u_{xx} + f(x, t, u), \quad u|_{t=0} = u|_{\Gamma} = 0,$$

откуда

$$u(x, t) = \int_0^t \int_Q G(x, \xi, t - \tau) f(\xi, \tau, u(\xi, \tau)) d\xi d\tau,$$

где  $G$  — функция Грина данной задачи;

4) переход от одномерной ситуации к многомерной осуществляется в них лишь увеличением размерности области интегрирования. В то же время переход от обыкновенных дифференциальных уравнений (ОДУ) к дифференциальным уравнениям в частных производных связан с принципиальными усложнениями.

Однако у интегральных уравнений есть и недостатки. Главными из них являются следующие:

- 1) получаемая при их численном решении система линейных алгебраических уравнений (СЛАУ), как правило, имеет заполненную матрицу, что затрудняет ее обращение;
- 2) очень часто ядра интегральных уравнений являются сингулярными, что затрудняет их дискретизацию. Сингулярность означает наличие особенности в ядре уравнения, делающей интеграл несобственным и расходящимся. Обычно расходящийся интеграл понимается в смысле главного значения по Коши.

Мы ограничимся рассмотрением линейных задач. Пример — **интегральное уравнение Фредгольма второго рода**

$$u(x) - \lambda \int_a^b K(x, \xi) u(\xi) d\xi = f(x), \quad x \in [a, b].$$

Ядро  $K$  этого уравнения задано в квадрате  $[a, b] \times [a, b]$ .

Если ядро  $K$  отлично от нуля только в половине указанного квадрата — треугольнике  $0 \leq \xi \leq x$ , — то данное уравнение превращается

в уравнение Вольтерра второго рода

$$u(x) - \lambda \int_a^x K(x, \xi) u(\xi) d\xi = f(x), \quad x \in [a, b].$$

Если в левых частях представленных уравнений отсутствует  $u(x)$ , то это уравнения первого рода. Они будут рассмотрены в 11.2, так как являются примерами некорректно поставленных задач.

Перечислим ряд свойств записанных уравнений.

1. Для обоих уравнений можно поставить **задачу Штурма — Лиувилля** отыскания собственных функций  $\varphi_i$  и собственных значений  $\lambda_i$ :

$$u(x) - \lambda \int_a^b K(x, \xi) u(\xi) d\xi = 0, \quad x \in [a, b].$$

Если ядро вещественно и симметрично, т. е.  $K(x, \xi) = K(\xi, x)$ , то существует по крайней мере одна собственная функция и одно собственное значение. Все собственные значения такого оператора действительны, а собственные функции, соответствующие различным собственным значениям, ортогональны.

2. Если  $\lambda \neq \lambda_i$  ( $\lambda_i$  — собственное число), то уравнение имеет единственное решение. Для симметричного ядра оно может быть представлено в виде разложения Шмидта по системе собственных функций  $\varphi_i$  (см. теорему Гильберта — Шмидта из теории интегральных уравнений).

3. Гладкость решения неоднородного уравнения определяется гладкостью правой части и ядра.

4. Если  $\lambda = \lambda_i$ , т. е. параметр совпадает с одним из собственных значений, то при одних правых частях решение вообще не существует, а при других — существует и неединственно.

5. Множество собственных значений уравнения Вольтерра второго рода пусто: для любого значения  $\lambda$  однородное уравнение всегда имеет только нулевое решение. Поэтому неоднородное уравнение всегда имеет решение, и притом единственное.

### 11.1.1. Разностный метод численного решения

Простейший способ численного решения интегрального уравнения состоит во введении сетки  $\bar{\omega}_h$  на  $[a, b]$ , замене интеграла квадратурной формулой и решении полученной системы алгебраических уравнений.

Например, если решается уравнение

$$\int_a^b K(x, \xi, u(\xi)) d\xi = F(x, u(x)), \quad x \in [a, b],$$

то при замене

$$\int_a^b \Phi d\xi \approx \sum_{i=1}^N c_i \Phi_i$$

получим

$$\sum_{i=1}^N c_i K(x_j, \xi_i, y(\xi_i)) = F(x_j, y(x_j)), \quad 1 \leq j \leq N.$$

Здесь  $y(x_j) = y_j$  — искомое сеточное решение. Полученная система является в общем случае нелинейной и требует для своего решения, например, применения метода Ньютона.

Для линейного уравнения Фредгольма второго рода

$$y_j - \lambda \sum_{i=1}^N c_i K_{ji} y_i = f_j, \quad 1 \leq j \leq N.$$

Эта СЛАУ имеет заполненную матрицу, поэтому естественнее всего для решения такой системы использовать метод Гаусса.

Мы не будем заниматься обоснованием сходимости указанного разностного метода. Отметим лишь, что теория Фредгольма разрешимости интегральных уравнений практически совпадает с аналогичной теорией для СЛАУ (см., в частности, альтернативу Фредгольма).

На практике чаще всего сходимость решения проверяют путем последовательных расчетов на сгущающихся сетках. Перед этим, вообще говоря, необходимо решить задачу определения собственных значений. Однако до конца она, как правило, не решается. Поэтому приходится опытным путем определять, не находится ли параметр задачи вблизи спектра, т. е.  $\lambda \approx \lambda_i^N$  или нет. В первом случае никакой сходимости численного решения скорее всего не будет.

Для уравнения Вольтерра второго рода получим СЛАУ

$$y_j - \lambda \sum_{i=1}^j c_i K_{ji} y_i = f_j, \quad 1 \leq j \leq N,$$

матрица которой является треугольной. Она решается за один ход метода Гаусса без каких-либо сложностей. При этом решение существует и единственno для любого  $\lambda$ .

Качество получаемого численного решения (при параметре  $\lambda$ , не совпадающем ни с каким собственным значением  $\lambda_i$ ) определяется числом точек, выбором квадратурной формулы и видом ядра  $K(x, \xi)$ . Влияние последнего особенно велико, если ядро  $K$  является сингулярным. При выборе квадратурной формулы можно использовать все те квадратурные формулы, которые были рассмотрены в 4.

### 11.1.2. Метод последовательных приближений

Запишем интегральное уравнение Фредгольма второго рода в виде

$$u(x) = \lambda A u + f(x),$$

где

$$A u(x) = \int_a^b K(x, \xi) u(\xi) d\xi.$$

**Теорема 11.1.** Решение интегрального уравнения Фредгольма второго рода может быть найдено методом *последовательных приближений*, т. е. последовательным вычислением

$$y^{k+1} = \lambda A y^k + f,$$

при условии ограниченности ядра  $K$  и малости  $\lambda$ .

◀ Пусть

$$y^{k+1}(x) = \lambda \int_a^b K(x, \xi) y^k(\xi) d\xi + f(x) \quad \text{и} \quad z^{k+1} = u - y^{k+1}.$$

Тогда для последовательности  $\{z_k\}$  ошибок последовательных приближений имеем уравнение

$$z^{k+1}(x) = \lambda \int_a^b K(x, \xi) z^k(\xi) d\xi,$$

откуда следует оценка

$$\|z^{k+1}\|_C \leq |\lambda| \|K\|_C \|z^k\|_C (b-a).$$

Следовательно, при  $q = |\lambda| \|K\|_C (b-a) < 1$  оператор  $A$  является сжимающим, а последнее неравенство выполнено для достаточно малых  $\lambda$ . Сходимость итераций является равномерной. ►

**Следствие 11.1.** Для уравнения Вольтерра второго рода метод сходится равномерно по  $x$  при любом значении  $\lambda$ .

◀ Имеем оценку ошибки

$$|z^1| \leq |\lambda| \|K\|_C (x-a) \|z^0\|_C, \quad |z^2| \leq \frac{1}{2} (|\lambda| \|K\|_C (x-a))^2 \|z^0\|_C$$

и так далее, откуда

$$|z^k| \leq \frac{1}{k!} (|\lambda| \|K\|_C (x-a))^k \|z^0\|_C,$$

т. е.

$$\|z^k\|_C \leq \frac{1}{k!} (|\lambda| \|K\|_C (b-a))^k \|z^0\|_C.$$

Правая часть этого неравенства стремится к нулю при  $k \rightarrow \infty$  для любого  $\lambda$ , что и доказывает утверждение следствия. ►

**Замечание 11.1.** Так как  $q = O(\lambda)$ , то из вида оценки ошибки следует, что метод последовательных приближений эквивалентен представлению решения в виде ряда по степеням  $\lambda$ :

$$u(x) = \sum_{k=0}^{\infty} \lambda^k u_k(x).$$

Если подставить такое представление решения в уравнение и приравнять коэффициенты при равных степенях  $\lambda$ , то получим

$$u_0 = f, \quad u_{k+1} = \int_a^b K(x, \xi) u_k(\xi) d\xi, \quad k = 0, 1, \dots,$$

т. е.

$$y^k(x) = \sum_{i=0}^k \lambda^i u_i(x).$$

**Пример 11.2.** Решим уравнение

$$u(x) - \lambda \int_0^{\pi/2} (\sin x \cos \xi) u(\xi) d\xi = 1.$$

В соответствии с описанным алгоритмом

$$u_0 = 1,$$

$$u_1(x) = \sin x \int_0^{\pi/2} \cos \xi d\xi = \sin x,$$

$$u_2(x) = \sin x \int_0^{\pi/2} \cos \xi \sin \xi d\xi = \frac{1}{2} \sin x,$$

$$u_3(x) = \frac{1}{4} \sin x$$

и так далее, откуда

$$u(x) = 1 + \left( \lambda + \frac{1}{2} \lambda^2 + \frac{1}{4} \lambda^3 + \dots \right) \sin x = 1 + \frac{\lambda}{1 - \lambda/2} \sin x = 1 + \frac{2\lambda}{2 - \lambda} \sin x.$$

Последний ряд сходится при  $|\lambda| < 2$ .

### 11.1.3. Замена ядра вырожденным

**Определение 11.2.** Ядро интегрального оператора Фредгольма называется **вырожденным**, если

$$K(x, \xi) = \sum_{i=1}^N A_i(x) B_i(\xi).$$

Ядро уравнения Вольтерра вырожденным не бывает, так как в случае вырожденности оно должно было бы равняться нулю (в силу условия  $K(x, \xi) = 0$  при  $\xi > x$ ).

Для вырожденного ядра

$$u(x) - \lambda \sum_{i=1}^N A_i(x) \int_a^b B_i(\xi) u(\xi) d\xi = f(x),$$

откуда

$$u(x) = f(x) + \sum_{i=1}^N \lambda c_i A_i(x).$$

Для неизвестных коэффициентов  $c_i$  получаем СЛАУ

$$\begin{aligned} \sum_{i=1}^N \lambda c_i A_i(x) - \sum_{j=1}^N \sum_{i=1}^N \lambda^2 c_i A_j(x) \int_a^b B_j(\xi) A_i(\xi) d\xi = \\ = \lambda \sum_{i=1}^N A_i(x) \int_a^b B_i(\xi) f(\xi) d\xi. \end{aligned}$$

Пусть функции  $A_i(x)$  линейно независимы и  $\lambda \neq 0$ . Тогда

$$\sum_{j=1}^N \left( \delta_{ij} - \lambda \int_a^b A_j(\xi) B_i(\xi) d\xi \right) c_j = \int_a^b B_i(\xi) f(\xi) d\xi, \quad i = 1, 2, \dots, N.$$

Это — СЛАУ, решение которой дает точное решение исходной задачи.

Отсюда и возникает следующий алгоритм: заменить *ядро*  $K$  на *приближенное*, являющееся вырожденным, для которого можно найти точное решение. Для некоторой полной системы функций, например системы Фурье,

$$K(x, \xi) = \sum_{i=1}^{\infty} A_i(x) B_i(\xi).$$

Далее делаем замену

$$K(x, \xi) \approx \sum_{i=1}^N A_i(x) B_i(\xi),$$

после чего выполняем описанный алгоритм. Обоснование метода мы не приводим.

**Пример 11.3.** Запишем то же уравнение, что и в примере 11.2:

$$u(x) - \lambda \int_0^{\pi/2} (\sin x \cos \xi) u(\xi) d\xi = 1.$$

Ядро данного уравнения вырожденное, поэтому

$$u(x) = f(x) + c\lambda \sin x = 1 + c\lambda \sin x.$$

Следовательно,

$$1 + c\lambda \sin x - \lambda \int_0^{\pi/2} \sin x \cos \xi d\xi - c\lambda^2 \sin x \int_0^{\pi/2} \sin \xi \cos \xi d\xi = 1.$$

Отсюда

$$c - 1 - \frac{\lambda c}{2} = 0, \quad c = \frac{2}{(2 - \lambda)},$$

в результате решение примет вид

$$u(x) = f(x) + \frac{2\lambda}{2 - \lambda} \sin x.$$

#### 11.1.4. Метод Галеркина (метод моментов)

Будем искать приближенное решение в виде разложения по некоторой *полной системе функций*

$$y(x) = f(x) + \sum_{k=1}^N c_k \varphi_k(x).$$

Далее подставим такое представление искомого решения в уравнение и потребуем ортогональности невязки всем функциям  $\varphi_k(x)$  используемой системы.

Получим СЛАУ

$$\sum_{j=1}^N a_{ij} c_j = b_i, \quad 1 \leq i \leq N,$$

где

$$a_{ij} = \int_a^b \varphi_i(\xi) \varphi_j(\xi) d\xi - \lambda \int_a^b \int_a^b K(x, \xi) \varphi_i(x) \varphi_j(\xi) dx d\xi.$$

Отсюда видно, что если система функций  $\varphi_i$  ортонормирована, то данная система и метод эквивалентны приближенной замене ядра на специальное вырожденное:

$$K(x, \xi) \approx \sum_{i=1}^N \varphi_i(x) \psi_i(\xi), \quad \psi_i(\xi) = \int_a^b K(x, \xi) \varphi_i(x) dx.$$

Обоснования данного метода мы также делать не будем.

## 11.2. Некорректные задачи

Напомним, что **задача** называется **некорректно поставленной** (по Адамару), если не выполнено хотя бы одно из условий корректности постановки: решение существует, единственно и непрерывно зависит от входных данных.

Принципиально важно, что данное определение имеет смысл лишь при точном указании пространств, в которых ставится задача. В данном случае для операторного уравнения первого рода

$$Au = f, \quad u \in U, \quad f \in F.$$

Условия корректности постановки задачи:

- 1) решение задачи  $u \in U$  существует для всех  $f \in F$ , т. е. областью значений оператора  $A$  является  $F$ ;
- 2) любым элементом  $f \in F$  решение определяется однозначно, т. е. на  $F$  существует обратный оператор  $A^{-1}$ ;
- 3) имеет место непрерывная зависимость  $u$  от  $f$ , т. е. оператор  $A^{-1}$  непрерывен.

Задачи, рассматриваемые в классической математической физике, удовлетворяют условиям корректности постановки при естественном выборе пространств  $U, F$ . Поэтому существовало мнение, что задачи, которые не удовлетворяют условиям корректности, являются «неверными» и их некорректность — это следствие неверности постановки задачи. В особенности этому способствовало нарушение устойчивости решения. Действительно, в большинстве задач науки и техники в качестве входных данных используется экспериментальная информация, заведомо содержащая ошибки измерений. Неустойчивость фактически означает невозможность получения физически осмыслиенного решения. Однако дальнейшее развитие прикладной математики показало, что неустойчивые задачи возникают в значительном количестве приложений, так что корректно поставленные задачи совсем не исчерпывают множество постановок задач, правильно отражающих физические явления.

Типичными примерами некорректных задач являются **интегральные уравнения Фредгольма и Вольтерра первого рода**:

$$\int_a^b K(x, \xi)u(\xi) d\xi = f, \quad \int_a^x K(x, \xi)u(\xi) d\xi = f.$$

Покажем неустойчивость решения уравнения Фредгольма. Пусть  $du = \exp(i\omega x)$ , где  $i$  — мнимая единица ( $i^2 = -1$ ). Найдем соответ-

ствующее такому изменению решения изменение правой части для ограниченных ядра  $K$  и его производной  $\partial K / \partial \xi$ :

$$\delta f = \int_a^b K(x, \xi) e^{i\omega\xi} d\xi = \frac{1}{i\omega} K(x, \xi) e^{i\omega\xi} \Big|_a^b - \frac{1}{i\omega} \int_a^b \frac{\partial}{\partial \xi} K(x, \xi) e^{i\omega\xi} d\xi = O\left(\frac{1}{\omega}\right).$$

Таким образом, при большом значении параметра  $\omega$  возмущение правой части есть величина порядка  $O(\omega^{-1})$ , т. е. малому приращению  $\delta f$  соответствует возмущение решения  $O(1)$ , что и свидетельствует о некорректности постановки задачи.

Наиболее известным является уравнение Вольтерра первого рода

$$\int_a^x K(x, \xi) u(\xi) d\xi = f(x) - f(a),$$

возникающее при дифференцировании функции  $f(x)$  (в этом случае  $K \equiv 1$ ). Например, если  $\delta f = \varepsilon \exp\left(\frac{ix}{\varepsilon^2}\right)$ , то  $\delta u = \frac{i}{\varepsilon} \exp\left(\frac{ix}{\varepsilon^2}\right)$ , что свидетельствует о неустойчивости решения при малом  $\varepsilon$ .

Для обоих типов уравнений решение существует не для всех правых частей. Например, если

$$K(x, \xi) = \sum_{i=1}^N A_i(x) B_i(\xi),$$

то правая часть  $f$  в случае уравнения Фредгольма должна представлять собой линейную комбинацию функций  $A_i(x)$ . В случае задачи о вычислении производной правая часть  $f$  должна быть обязательно дифференцируемой.

### 11.2.1. Предпосылки метода регуляризации

Общим итогом рассмотрения уравнений первого рода является тот факт, что малые возмущения правой части могут привести не только к большим возмущениям решения, но и к отсутствию решения задачи вообще. Поскольку погрешности входных данных есть всегда (хотя бы при численном решении), возникает необходимость иной постановки задачи для нахождения решения некорректно поставленных задач.

Принципиально данный вопрос решается путем применения теоремы А.Н. Тихонова. Пусть  $\bar{U}$  — подпространство метрического пространства  $U$ ,  $F_A$  — его образ в метрическом пространстве  $F$  при отображении с помощью оператора  $A$ .

**Теорема 11.2 (А.Н. Тихонова).** Пусть оператор  $A$  непрерывен на  $\bar{U}$  и при каждом  $\bar{f} \in F_A$  уравнение  $Au = \bar{f}$  имеет единственное решение  $\bar{u} \in \bar{U}$ . Тогда, если  $\bar{U}$  — компакт (в метрике  $U$ ), то обратное отображение  $\bar{u} = A^{-1}\bar{f}$  непрерывно в метрике пространства  $U$ .

Доказательство теоремы мы приводить не будем.

Здесь компактом в некоторой метрике называется множество, из всякой последовательности элементов которого можно выделить подпоследовательность, сходящуюся в той же метрике к элементу этого множества.

Таким образом, данная теорема показывает, что при условии поиска решения задачи на компакте обратный оператор является непрерывным, если имеют место непрерывность самого оператора и единственность решения. При этом все рассматривается только на данном компакте.

В результате появилось новое определение корректности.

**Определение 11.3.** Пусть в пространстве  $U$  дополнительно выделено некоторое подпространство  $\bar{U} \in U$ . Тогда *задача* решения операторного уравнения  $Au = f$  называется **корректно поставленной по Тихонову** (или **условно-корректной**), при условиях:

- 1) априори известно, что решение  $u$  существует и принадлежит  $\bar{U}$ ;
- 2) решение единственно;
- 3) бесконечно малым вариациям  $f$ , не выводящим решение  $u$  из подпространства  $\bar{U}$ , соответствуют бесконечно малые вариации решения  $u$ . Соответствующее множество  $\bar{U}$ , на образе которого  $A\bar{U}$  оператор  $A^{-1}$  существует и непрерывен, называется **множеством корректности**.

Таким образом, при теоретическом исследовании корректности по Тихонову не доказываются теоремы существования решения. Его существование и принадлежность заданному подмножеству корректности постулируются в самой постановке задачи исходя из дополнительных физических или каких-либо иных соображений. Единственность решения задачи можно установить лишь для решений из множества  $\bar{U}$ , а не всего исходного пространства.

В условии 3 непрерывная зависимость обратного оператора предполагается только на множестве  $A\bar{U}$ . Таким образом, устойчивость решения задачи обеспечивается сужением класса возможных решений до множества  $\bar{U}$ , или, что то же самое, сужением множества возможных правых частей до  $A\bar{U}$ .

Понятно, что данное определение появилось как следствие теоремы А.Н. Тихонова. Решение операторного уравнения первого рода является

устойчивым, если оно принадлежит компакту (множеству корректности), а входные данные задачи — правая часть уравнения  $f$  — являются образом этого компакта при отображении оператором  $A$ . Однако ниоткуда не следует, что информация, полученная, например, экспериментально, будет такой, что окажется принадлежащей этому образу.

В связи с этим появилось понятие о множестве практической эквивалентности. Большинство некорректно поставленных задач являются результатом обработки экспериментальной информации, которая заранее известна. Как правило, можно считать заданной погрешность  $\delta$  входной информации:  $\|f - \tilde{f}\| \leq \delta$ . Здесь  $f$  — известная величина. Отсюда понятно, что нет смысла различать решения, дающие один и тот же результат в пределах заданной точности.

**Определение 11.4.** Множество  $U_\delta \in U$ , такое, что  $\|f - Au\| \leq \delta$  при  $u \in U_\delta$ , называется **множеством практической эквивалентности**.

Введение понятия эквивалентных решений делает возможным видоизменение исходной задачи таким образом, чтобы полученная постановка обладала нужными свойствами.

### 11.2.2. Понятие регуляризующего оператора и пример регуляризации операторного уравнения первого рода

Рассмотрим общую схему. Пусть требуется найти решение некорректно поставленной задачи

$$A[x, u(\xi)] = f(x), \quad u(\xi) \in U, \quad f(x) \in F.$$

Здесь  $A$  — некоторый оператор;  $U$  и  $F$  — нормированные пространства. Считаем, что для произвольной правой части  $f$  решение может не существовать, но имеются такие  $\tilde{f} \in F$ , что существуют решения данной задачи  $\bar{u} \in U$ .

Изменим задачу, введя в оператор дополнительные члены с малым параметром  $\alpha$ :

$$A_\alpha[x, u_\alpha(\xi)] = f(x),$$

Обозначим решение последней задачи  $u_\alpha$ .

**Определение 11.5.** Оператор  $A_\alpha$  называется **регуляризирующим**, если:

1) задача с параметром  $\alpha$  является корректно поставленной для любой  $f \in F$  при любом  $\alpha > 0$  (не слишком большом);

2) для любого  $\varepsilon > 0$  существуют функции  $\alpha(\delta)$  и  $\delta(\varepsilon)$ , такие, что при  $\|f - \bar{f}\|_F < \delta(\varepsilon)$  выполнено  $\|u_{\alpha(\varepsilon)} - \bar{u}\|_U < \varepsilon$ .

При этом функции  $\alpha(\varepsilon)$ ,  $\delta(\varepsilon)$  зависят, вообще говоря, и от  $\bar{f}$ .

Следовательно, задача для оператора  $A_\alpha$  является корректно поставленной и ее решение  $u_\alpha$  мало (при малых  $\alpha$ ) отличается от нужного нам решения  $\bar{u}$ .

Отметим, что исходная задача может быть преобразована в корректно поставленную путем выбора правых частей  $f$  из  $\bar{F}$ . Например, если в задаче вычисления производной взять правую часть  $f$  из класса непрерывно дифференцируемых функций, то задача станет корректной: малость  $\delta f$ , т. е. вариации  $f$ , означает малость  $\|\delta f'\|_C$ , а значит, и малость вариации решения. Однако это неконструктивно, так как в реальных задачах численного дифференцирования  $f$  — в лучшем случае лишь непрерывная функция.

Очевидно также, что регуляризующих операторов может быть много, что дает возможность выбора.

Заметим, что процедура перевода задачи нахождения решения уравнения Вольтерра первого рода из класса некорректности в класс корректности на первый взгляд выглядит совсем просто: нужно продифференцировать исходное уравнение первого рода. Тогда формально будет получено уравнение Вольтерра второго рода, задача нахождения решения которого в принципе является корректно поставленной. Однако в правой части такого уравнения будет стоять производная правой части исходного уравнения. Следовательно, использованная замена в действительности включила в себя процедуру нахождения производной правой части, которая является некорректной. Ясно, что таким способом сделать задачу корректной невозможно.

Как показано в 11.2.1, задачу можно сделать устойчивой, если решение искать на компакте. На этой основе возникло понятие квазирешения.

**Определение 11.6.** *Квазирешением* исходной задачи называется функция  $u$  из множества корректности  $\bar{U}$ , являющаяся решением следующей вариационной задачи:

$$\inf \{ \|Au - f\|, u \in \bar{U} \}.$$

При определенных условиях алгоритм нахождения квазирешения является регуляризующим для исходной задачи. Заметим, что малого параметра в явном виде здесь нет. Однако он присутствует в погрешности входной информации.

Понятие квазирешения и обоснование такого метода нахождения решения исходной задачи являются теоретической основой, в частности, многочисленных вариантов метода подбора решения. Такие методы широко распространены.

Другой метод построения *регуляризующего алгоритма* представлен ниже.

Рассмотрим уравнение Фредгольма первого рода:

$$A[x, u(\xi)] = \int_a^b K(x, \xi) u(\xi) d\xi = f(x), \quad u \in U, \quad f \in F, \quad A : U \mapsto F.$$

Считаем, что однородное уравнение имеет только тривиальное решение. Поэтому отображение  $A : U \mapsto F$  однозначно и для  $f \in F$  решение либо существует и единствено, либо не существует. Ядро  $K(x, \xi)$  задано на  $[c, d] \times [a, b]$

Исходная задача может быть сформулирована в виде

$$\int_c^d (A[x, u(\xi)] - f(x))^2 dx \rightarrow \min.$$

Рассмотрим ее модификацию

$$M[\alpha, f(x), u(\xi)] = \int_c^d (A[x, u(\xi)] - f(x))^2 dx + \alpha \Omega_n[u(\xi)] \rightarrow \min,$$

где

$$\Omega_n[u(\xi)] = \int_a^b d\xi \sum_{k=0}^n p_k(\xi) \left( \frac{\partial^k u(\xi)}{\partial \xi^k} \right)^2 —$$

*стабилизатор Тихонова*  $n$ -го порядка;  $p_k(\xi) \geq 0$  — непрерывные функции. Обычно выбирают  $p_k(\xi) = 1$ , если нет конкретных оснований для другого выбора.

Для функций класса  $U$ , для которых  $M[\alpha, f(x), u(\xi)]$  имеет смысл, естественным образом вводится норма  $\|u\|_U^2 = \Omega_n[u]$ , а соответствующее пространство называется *пространством Соболева*  $W_2^n$ . Считаем, что правая часть  $f \in L_2(c, d)$ .

**Теорема 11.3 (А.Н. Тихонова).** Задача определения минимума функционала  $M[\alpha, f, u]$  имеет решение для любых  $\alpha$  и  $f$ , соответствующий алгоритм является регуляризующим для исходной задачи.

Доказательство данной теоремы мы приводить не будем.

Отметим лишь, что доказательство основано на том, что функции из  $U$  имеют суммируемые с квадратом  $n$ -е производные, поэтому ограниченное по норме  $U$  множество функций (что соответствует минимизирующей последовательности для функционала  $M$ ) является компактом в  $L_2$ , т. е. из него можно выбрать сходящуюся подпоследовательность. Смысл регуляризации и состоит в поиске решения задачи не на всем максимально широком пространстве, а лишь на его компактной части.

Одной из главных сложностей *метода регуляризации* является выбор параметра  $\alpha$ . Если он мал, то задача будет слабо устойчива, если велик, то ее решение будет сильно отличаться от искомого. Обычно в прикладных задачах правая часть  $f$  известна с некоторой точностью  $\|\delta f\| \sim \delta$ . Поэтому, как правило, выбирают такой параметр  $\alpha$ , чтобы невязка решения

$$r = \left( \int_c^d (A[x, u_\alpha(\xi)] - f(x))^2 dx \right)^{1/2}$$

являлась величиной порядка  $\delta$ , т. е.  $r \sim \delta$ . Очевидно, что искать такой параметр  $\alpha$ , чтобы невязка  $r$  была меньше  $\delta$ , бессмысленно, но и много больше  $\delta$  тоже нельзя. Поэтому на практике решение проводят с некоторым набором значений  $\alpha$ , из которого выбирают удовлетворяющее данному критерию.

Второй сложностью метода регуляризации является выбор  $n$ . При больших  $n$  решение получается очень сглаженным. Поэтому обычно полагают  $n = 1$ .

Запишем уравнение Эйлера, которому удовлетворяет решение  $u_\alpha$  задачи поиска  $\min M[\alpha, f, u]$ :

$$\alpha \sum_{k=0}^n \int_a^b p_k(\xi) [u^{(k)}(\xi)]^2 d\xi + \int_c^d dx \left( \int_a^b K(x, \xi) u(\xi) d\xi - f(x) \right)^2 \rightarrow \min.$$

Если  $u$  — решение данной задачи, то первая вариация функционала должна быть равной нулю. Отсюда получаем

$$\begin{aligned} & \alpha \sum_{k=0}^n \int_a^b p_k(\xi) u^{(k)}(\xi) \delta u^{(k)}(\xi) d\xi + \\ & + \int_c^d dx \left( \int_a^b K(x, \xi) u(\xi) d\xi - f(x) \right) \int_a^b K(x, \xi) \delta u(\xi) d\xi = 0. \end{aligned}$$

Проведем дальнейшие выкладки лишь для  $n = 1$ . Для  $n > 1$  они аналогичны. Выполняя интегрирование по частям, запишем

$$\begin{aligned} \alpha p_1(\xi) u'(\xi) \delta u \Big|_a^b + \alpha \int_a^b (p_0(\xi) u(\xi) - (p_1(\xi) u'(\xi))'_\xi) \delta u(\xi) d\xi + \\ + \int_c^d dx \int_a^b K(x, \xi) u(\xi) d\xi \int_a^b K(x, \eta) \delta u(\eta) d\eta = \\ = \int_c^d f(x) dx \int_a^b K(x, \xi) \delta u(\xi) d\xi. \end{aligned}$$

Потребуем выполнения условий  $p_1(\xi) u'(\xi) \Big|_a = p_1(\xi) u'(\xi) \Big|_b = 0$ . Тогда полученное условие равенства нулю первой вариации трансформируется в интегродифференциальное уравнение

$$\alpha(p_0 u - (p_1 u')') + \int_a^b Q(x, \xi) u(\xi) d\xi = \Phi(x),$$

где

$$Q(x, \xi) = \int_c^d K(\eta, \xi) K(\eta, x) d\eta, \quad \Phi(x) = \int_c^d K(\xi, x) f(\xi) d\xi.$$

Для нахождения  $u$  необходимо численно решить полученное уравнение, применив описанные ранее методы.

Поясним качественно действие регуляризации. Пусть правая часть  $\Phi$  уравнения получила приращение  $\beta \exp(i\omega x)$ . Тогда решение получит приращение  $\gamma \exp(i\omega x)$ . Оценим слагаемые по порядку величины  $\omega$ , считая параметр  $\omega$  большим. Подставим приращение в интегродифференциальное уравнение и в соответствии с приведенными выше выкладками получим

$$\gamma \left( \alpha(1 + \omega^2) + \frac{1}{\omega} \right) \sim \beta,$$

т. е.

$$\gamma \sim \frac{\beta}{\alpha(1 + \omega^2) + \frac{1}{\omega}}.$$

Данное соотношение демонстрирует эффект регуляризации. При  $\alpha = 0$  имеем  $\gamma \sim \omega \beta$ , т. е. приращение  $\gamma$  велико. Если  $n = 0$ , т. е. нет

слагаемого с  $p_1$ , то

$$\gamma \sim \frac{\beta}{\alpha + \frac{1}{\omega}}.$$

Следовательно, возмущения решения по порядку величины соответствуют возмущениям правой части, расчет становится устойчивым. Если  $n = 1$ , то  $\gamma \sim \beta/(\alpha\omega^2)$ , т. е. возмущения, соответствующие высокочастотным гармоникам, будут подавляться. Эффект такого подавления будет еще более сильным при большем  $n$ .

### 11.2.3. Примеры некорректно поставленных задач

Приведем ряд примеров прикладных задач, к которым применима описанная схема исследований.

1. *Сглаживание функции.* При этом нужно «решить» уравнение  $u = f$ . Уравнение Эйлера для соответствующего функционала имеет вид

$$-\alpha[(p_1 u')' - p_0 u] + u = f.$$

Задача имеет смысл, если  $f$  измерена в эксперименте и имеет большую случайную составляющую.

2. *Численное дифференцирование* функции. Нужно решить *уравнение Вольтерра первого рода*:

$$\int_a^x u(\xi) d\xi = f(x) - f(a).$$

Эта задача соответствует уже рассмотренной выше задаче для уравнения Фредгольма первого рода. В конкретном случае задачи численного дифференцирования ядро имеет вид

$$K(x, \xi) = \begin{cases} 1, & \xi \leq x; \\ 0, & \xi > x. \end{cases}$$

Подставим  $K(x, \xi)$  в соотношения, полученные при регуляризации уравнений Фредгольма первого рода:

$$Q(x, \xi) = b - \max(x, \xi) = \int_a^b K(\eta, x) K(\eta, \xi) d\eta,$$

$$\Phi(x) = \int_a^b (f(\xi) - f(a)) K(\xi, x) d\xi = \int_x^b (f(\xi) - f(a)) d\xi.$$

В результате получим уравнение Эйлера

$$-\alpha((p_1 u')' - p_0 u) + \int_a^x (b-x)u(\xi)d\xi + \int_x^b (b-\xi)u(\xi)d\xi = \int_x^b (f(\xi) - f(a))d\xi,$$

$$u'(a) = u'(b) = 0.$$

**3. Суммирование ряда Фурье** с неточно заданными коэффициентами.

Рассмотрим в качестве примера-пояснения ряд Фурье вида

$$f(x) = \sum_{n=1}^{\infty} a_n \cos nx, \quad x \in [0, 2\pi],$$

и возмутим его коэффициенты на величину  $\delta a_n = \varepsilon/n$ . Тогда сумма ряда окажется возмущенной на величину

$$\delta f(x) = \varepsilon \sum_{n=1}^{\infty} \frac{\cos nx}{n},$$

причем такая сумма, например, при  $x = 0$  не существует. Однако квадратичная норма возмущения вектора  $\delta a$  коэффициентов ряда Фурье

$$\|\delta a\|_2^2 = \varepsilon \sum_{n=1}^{\infty} \frac{1}{n^2} = \varepsilon \frac{\pi^2}{6},$$

т. е. за счет выбора малого параметра  $\varepsilon$  ее можно сделать сколь угодно малой.

**4. Обращение преобразования Лапласа** функции.

**5. Решение плохо обусловленных** систем линейных алгебраических уравнений (СЛАУ).

**6. Решение задачи Коши для уравнения Лапласа.**

Рассмотрим задачу Коши для уравнения Лапласа вида

$$\Delta u = 0, \quad (x, y) \in [0, \pi] \times [0, Y]; \quad u(x, 0) = \varphi(x), \quad \left. \frac{\partial u}{\partial y} \right|_{(x, 0)} = \psi(x),$$

при следующих входных данных:  $\varphi(x) = 0$ ,  $\psi(x) = \frac{\sin nx}{n}$ . Тогда получим решение

$$u(x, y) = \frac{\sin nx \sin ny}{n^2}.$$

Из вида  $u(x, y)$  следует неустойчивость решения: равномерная норма входных данных при  $n \rightarrow \infty$  стремится к нулю, в то время как аналогичная норма решения задачи стремится к бесконечности.

Очевидно, что представленные примеры некорректно поставленных задач соответствуют в основном поиску решения обратных задач: по известному результату действия некоторого оператора требуется найти функцию, на которую он действовал. Например, в задаче Коши для уравнения Лапласа таким результатом является значение нормальной производной решения на нижней границе. При этом некорректные задачи имеют место тогда, когда оператор прямой задачи является «очень хорошим», чаще всего вполне непрерывным (иногда называется компактным), т. е. оператором, переводящим ограниченное множество в компактное. Такой оператор переводит достаточно широкое множество во множество, значительно более узкое, которое иногда является подмножеством исходного. Очевидно, что обратный оператор действует наоборот. Ярким примером таких операторов являются операторы прямого и обратного преобразования Лапласа. Прямое преобразование (при его некотором расширенном истолковании) даже  $\delta$ -функцию Дирака переводит в константу, а обратное преобразование может любую константу перевести в  $\delta$ -функцию.

### 11.3. Библиографические комментарии

Приведенный в данной главе материал является традиционным. Мы, в частности, в основном следуем [80]. Дополнительные подробности можно найти во многих руководствах по численным методам, приведенных в списке литературы.

Отметим отдельно [18, 26, 96, 117], причем в [26] даны и тексты программ для ЭВМ.

Теория решения некорректно поставленных задач уже превратилась в самостоятельный раздел прикладной математики и математической физики. Представленное в данной главе описание метода регуляризации может быть изучено глубже с помощью работ [47, 76, 101, 118, 145, 168, 169], а также [25, 176].

Существует специфический класс задач, которые могут быть решены с использованием теории некорректно поставленных задач — это задачи оптимального управления. Очень много полезной информации о таких задачах содержится в [177].

## **Часть II**

# **Избранные вопросы теории и практики численных методов**

## 12. МЕТОДЫ ТРИАНГУЛЯЦИИ ПРОСТРАНСТВЕННЫХ ОБЛАСТЕЙ

Дан обзор существующих методов трехмерной *дискретизации (триангуляции)* пространственных областей, приведена их классификация, описаны методы оценки качества сетки. Рассмотрены *прямые методы*, представлены *шаблоны дискретизации параллелепипеда, шара и цилиндра*. Описан метод переноса сеток из простых областей в более сложные с помощью *изопараметрических отображений*. Рассмотрены *итерационные методы* трехмерной дискретизации пространственных областей: *методы граничной коррекции, методы на основе критерия Делоне и методы исчерпывания*. Приведены варианты алгоритмов для каждого из указанных методов. Обсуждены особенности построения сеток в сложных областях.

### 12.1. Методы триангуляции и оценка качества сетки

При решении различных задач математического моделирования широко применяют *проекционно-сеточные методы*. Для их использования необходимо предварительно построить так называемую сетку, т. е. некое топологическое множество точек (узлов), связанных между собой ребрами — отрезками прямых (а в некоторых случаях и кривых) линий таким образом, что исходная область разбивается на элементы определенной формы. При этом в качестве элементов сетки, если речь идет о геометрически сложных областях, обычно используются геометрические симплексы — треугольники в двумерном и тетраэдры в трехмерном случае. Процесс построения сетки обычно называется *дискретизацией* или *триангуляцией* (даже если речь идет о трех измерениях).

Развитие вычислительной техники способствовало значительному прогрессу в области численных методов вообще и методов триангуляции в частности. Разработаны новые более эффективные методы, значительно более ресурсоемкие. В то же время под многие эмпирические методы триангуляции подведена теоретическая база.

В настоящее время двумерная триангуляция (без адаптации к решению) хорошо изучена: разработаны и теоретически обоснованы эффективные и надежные методы построения и оптимизации сеток; свойства

элементов-треугольников надежно установлены. Вместе с тем проблема трехмерной триангуляции еще далека от окончательного решения: большая часть методов теоретически не обоснована, а многие задачи вообще не решены.

На первый взгляд подобное различие кажется странным, ведь обычно математические методы, разработанные для случая двух измерений, легко переносятся на случай трех и более измерений. К сожалению, трехмерное пространство обладает рядом особенностей, которые затрудняют подобный перенос.

Например, плоскость можно элементарным образом заполнить правильными треугольниками, а трехмерное пространство заполнить правильными тетраэдрами нельзя. Это основное препятствие на пути создания качественных трехмерных сеток. Поскольку в качестве элементов невозможно использовать правильные тетраэдры, приходится обходиться их подобиями, что негативно сказывается на аппроксимационных свойствах сетки.

Более того, любой треугольник можно разбить на треугольники, подобные ему (на четыре, девять, шестнадцать и т. д.). Тетраэдр в общем случае нельзя разбить на подобные тетраэдры. Это является основным препятствием на пути использования методов дробления, эффективно применяющихся в двумерном случае.

Любой многоугольник на плоскости можно ребрами разбить на непересекающиеся треугольники. В общем случае произвольный многогранник нельзя разбить на непересекающиеся тетраэдры, не используя дополнительных вершин. Эта проблема является краевогольной, поскольку существенно усложняет использование практических алгоритмов триангуляции.

Помимо указанных сложностей теоретического плана, существуют сложности практического характера [199, 222]. Например, невозможность (а точнее говоря, крайне высокая сложность) осуществления контроля человеком над процессом триангуляции. Если в случае плоской области всегда можно вывести результат на дисплей для последующей его коррекции оператором, то для трехмерных областей это выполнить чрезвычайно трудно. Ввиду этого на методы трехмерной триангуляции накладываются дополнительные требования по надежности работы и правильности построения. Кроме того, следует учитывать и значительное увеличение потребляемых ресурсов из-за большего числа пространственных измерений (и соответственно количества элементов сетки).

Все **методы триангуляции** по принципу построения можно разбить на два класса: **прямые** методы и **итерационные** методы. По-

строительство сетки прямыми методами проводится за один этап, причем ее топология (иначе говоря, граф связей между узлами) и координаты всех узлов известны изначально. Построение сетки итерационным методом проводится последовательно: на каждом шаге добавляется один или несколько элементов, причем изначально неизвестны ни координаты узлов, ни топология сетки. Кроме того, координаты узлов и топология могут меняться в процессе построения.

Итерационные методы обладают достаточной универсальностью, поэтому, в отличие от прямых, они могут быть использованы для триангуляции областей довольно произвольного вида. За эту универсальность приходится расплачиваться существенно большим потреблением ресурсов и более трудоемкой реализацией метода в конкретном алгоритме.

В настоящее время разработано большое количество программных пакетов на основе того или иного итерационного метода, реализующих построение сеток (частично или полностью) в автоматическом режиме. В основном эти пакеты коммерческие, что вполне оправдано с учетом затрачиваемых на их создание усилий.

Сетки, построенные итерационными методами, как правило, неструктурированы и неоднородны. Неструктурированность обусловлена тем, что топология сетки формируется в процессе построения, и поэтому естественно может варьироваться даже в пределах одной подобласти. По этой же причине однородность если и может возникнуть, то только случайно.

Поскольку перед построением сетки ничего нельзя сказать о ее будущей структуре, нельзя гарантировать и ее качества. Часто построенную сетку можно существенно улучшить с помощью одного из многочисленных методов оптимизации [227, 232, 268]. Этой возможностью обычно не пренебрегают в связи с тем, что время, затрачиваемое на оптимизацию, как правило, существенно меньше времени, затрачиваемого на построение.

Настоящая глава основана на материалах [43, 44].

### 12.1.1. Классификация методов

Главными преимуществами прямых методов являются скорость работы и надежность. Сетка строится практически мгновенно при минимальной затрате ресурсов и с минимальным риском ошибки. В то же время эти методы применимы только для областей определенной геометрической конфигурации, поэтому они не являются универсальными.

Итерационные методы, напротив, универсальны и, как правило, применимы для областей довольно произвольной формы. Именно поэтому итерационные методы в основном и используются в автоматических программных комплексах. Недостатком этих методов являются ресурсоемкость, низкая скорость работы по сравнению с прямыми методами и меньшая надежность.

Прямых методов разработано немного ввиду ограниченной возможности их использования. Условно они могут быть разделены на два тесно связанных вида: *методы на основе шаблонов* и *методы отображения (изопараметрические)*.

Методы на основе шаблонов подразумевают разбиение областей заданного вида (параллелепипед, шар, цилиндр и т. д.). Соответственно для каждого вида области используется свой шаблон, т. е. принцип размещения узлов и установки связей между ними.

Методы отображения являются своего рода попыткой перекинуть мостик между областями строгой геометрической формы и областями произвольного вида. Если можно построить взаимно однозначное отображение между заданной областью и какой-либо простой геометрической формой, то, разбив последнюю, можно отобразить полученную сетку на исходную область. Очевидным недостатком этого метода является искажение сетки при отображении, которое может существенно снизить качество триангуляции.

*Сетки*, полученные прямыми методами, являются *структурированными*, т. е. их топология полностью определяется некоторым набором правил. Это означает, что зная только индексы узла, можно определить все соседние узлы, а также вычислить их координаты. Это важное свойство позволяет существенно экономить компьютерные ресурсы.

Итерационные методы благодаря своей универсальности получили наибольшее развитие. Разработаны три основных вида итерационных методов: *методы граничной коррекции*, методы на основе *критерия Делоне* и *методы исчерпывания*.

Методы граничной коррекции являются самыми быстрыми из итерационных методов, но, к сожалению, имеют ряд неискоренимых недостатков. Построение сеток этими методами осуществляется в два этапа. На первом этапе производится триангуляция некой простой суперобласти, полностью включающей в себя заданную область. Как правило, эта суперобласть представляет собой параллелепипед, триангуляция которого осуществляется сравнительно просто на основе одного из многочисленных шаблонов. На втором этапе все узлы полученной сетки, лежащие вблизи границы заданной области, проецируются на

поверхность границы, а узлы, лежащие вне заданной области, удаляются. Чтобы компенсировать неизбежные геометрические искажения элементов сетки вблизи границ, часто дополнительно проводят этап оптимизации сетки, что в итоге позволяет получить достаточно хорошие результаты.

Очевидно, что данные методы нельзя применять для дискретизации областей с заданной триангуляцией границ. Это существенное ограничение, а также другие сложности снижают популярность метода, сводя на нет его основное преимущество — высокую скорость работы.

Сущность методов исчерпывания заключается в последовательном вырезании из заданной области фрагментов тетраэдрической формы до тех пор, пока вся область не окажется исчерпанна. В англоязычной литературе этот метод получил название *advancing front*, что также хорошо отражает идею метода. Исходными данными на каждой итерации является фронт, т.е. триангуляция границы еще не исчерпанной части области. Каждый треугольник этой триангуляции является основанием исключаемого из области тетраэдра, причем на каждой итерации может быть исключен либо один тетраэдр, либо сразу целый слой тетраэдров. После исключения тетраэдра (-ов) фронт обновляется, после чего происходит переход к следующей итерации.

Методы исчерпывания универсальны и могут быть использованы для областей достаточно произвольной формы (даже для несвязных областей), что объясняет их популярность. В частности, именно эти методы используются в программном комплексе **ANSYS**. Вместе с тем следует отметить их высокую ресурсоемкость и низкую скорость работы.

Методы на основе критерия Делоне часто называют просто методами Делоне, хотя это не совсем корректно, поскольку Б.Н. Делоне никаких методов не разрабатывал, он лишь предложил простой и эффективный критерий, используемый при установке связей между узлами. В основе этих методов лежит размещение в заданной области узлов и последующая расстановка между ними связей согласно критерию Делоне (либо иному схожему критерию).

В двумерном случае этот подход получил наибольшую популярность, поскольку он позволяет быстро и эффективно конструировать сетки с априори высоким качеством триангуляции. Однако при переходе к трехмерному случаю исследователи столкнулись с рядом проблем, затрудняющих использование этого критерия. Тем не менее эти методы получили достаточно хорошее развитие и пользуются заслуженной популярностью.

### 12.1.2. Оценка качества сетки

Сравнение эффективности различных методов дискретизации невозможно без введения некоего критерия качества построенной сетки. Поскольку сетка строится не ради самого построения, а ради решения некоторой задачи, следует увязать этот критерий с аппроксимационными свойствами сетки. Согласно теории, эти свойства в основном зависят от формы элементов [195]. В частности, в оценку погрешности аппроксимации конечным элементом  $\omega$ , как правило, входит величина

$$\frac{R(\omega)}{\text{diam}(\omega)}, \quad (12.1)$$

где  $R(\omega)$  — радиус шара, вписанного в конечный элемент  $\omega$ ;  $\text{diam}(\omega)$  — диаметр конечного элемента  $\omega$  [189]. Поскольку прямое нахождение (12.1) весьма трудоемко, на практике применяют различные альтернативные оценки. Существует большое количество таких оценок, некоторые наиболее популярные приведены в табл. 12.1 (см. также [253]). В формулах табл. 12.1 использованы следующие обозначения и параметры, определенные для заданного тетраэдра:  $R_c$  — радиус описанной сферы;  $R_i$  — радиус вписанной сферы;  $L_{\max}$  — максимальная длина ребра;  $L_{\min}$  — минимальная длина ребра;  $S_i$ ,  $i = \overline{1, 4}$ , — площадь  $i$ -й грани;  $V$  — объем;  $\bar{L}$  — среднее арифметическое длин ребер;  $\tilde{L}$  — среднее геометрическое длин ребер;  $\delta$  — максимальный двугранный угол;  $\vartheta$  — минимальный телесный угол.

Таблица 12.1

Величина	Интервал возможных значений	Оптимальное значение
$\beta = R_c/R_i$	$[1, +\infty)$	3,0
$\sigma = L_{\max}/R_i$	$[1, +\infty)$	4,898979...
$\omega = R_c/L_{\max}$	$[1/2, +\infty)$	0,612375...
$\tau = L_{\max}/L_{\min}$	$[1, +\infty)$	1,0
$k = V^4 \left( \sum_{i=0}^3 S_i^2 \right)^{-3}$	$(0, 1]$	$4,57247410^{-4}$
$\alpha = \bar{L}^3/V$	$[1, +\infty)$	8,4852816...
$\gamma = \tilde{L}^3/V$	$[1, +\infty)$	8,4852816...
$\delta$	$[\arccos 1/3, \pi]$	$\arccos 1/3 \approx 1,23095$
$\vartheta$	$(0, \pi/2]$	$\pi/2$

Однако оптимальной с точки зрения точности полноты оценки качества сетки и удобства нахождения является следующая оценка [189, 264]:

$$\mu = \frac{V}{abc}, \quad (12.2)$$

где  $V$  — объем тетраэдра;  $abc$  — наибольшее из произведений длин тройки ребер, выходящих из одной вершины. Далее мы будем использовать именно эту оценку.

Поскольку значение  $\mu$  имеет порядок десятых и сотых долей единицы, для наглядности его удобно разделить на аналогичную характеристику правильного тетраэдра, которая составляет  $\sqrt{2}/12 \approx 0,118$ . Назовем это отношение *аппроксимационной характеристикой* (AX) элемента. Возможные значения AX лежат в пределах от нуля до единицы; чем ближе значение к единице, тем лучше.

Для качественного анализа сетки наиболее важными являются минимальное и среднее значения AX: первое характеризует качество аппроксимации, второе свидетельствует об общем качестве сетки. Для геометрически сложных областей достаточно хорошим результатом является среднее значение AX, равное  $1/2$ .

### 12.1.3. Особенности построения сеток в сложных областях

Одной из распространенных задач дискретизации является *триангуляция сложных областей*, т. е. областей, на которые наложены различные дополнительные ограничения, обусловленные, например, изменением свойств материала либо некими конструкционными особенностями моделируемого объекта. Как правило, ограничения, накладываемые на область, а точнее говоря, на сетку в этой области, носят характер запрета на пересечение ребрами сетки некоторых заданных поверхностей. Например, если речь идет о композитном материале, ребра сетки не должны пересекать границу между включением и матрицей. Фактически это равнозначно тому, что каждый конечный элемент сетки должен состоять строго из одного материала. Это ограничение необходимо учитывать при построении сетки.

Другими ограничениями, встречающимися чрезвычайно редко, являются ограничения в виде заданной кривой, которая не должна пересекать грани элементов сетки, т. е. должна быть аппроксимирована непрерывной цепочкой ребер.

Таким образом, условно можно выделить два типа сложных областей:

- 1) области, состоящие из непересекающихся замкнутых подобластей;
- 2) области с внутренними ограничениями в виде поверхностей или кривых.

Примерами областей первого типа служат модели композитных материалов (рис. 12.1, показано шарообразное включение в кубической ячейке композита), примерами области второго типа служат модели, предназначенные для исследования развития трещин в материале (рис. 12.2, показан фрагмент плоскости, моделирующий трещину в трехмерной среде). Отметим также, что в двумерном случае типичным примером областей второго типа являются геодезические карты, где узлами сетки служат замеры высот, а линиями-ограничениями — контуры рек, оврагов и озер.

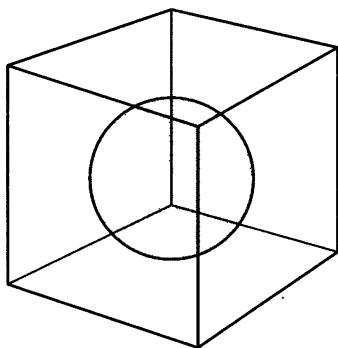


Рис. 12.1

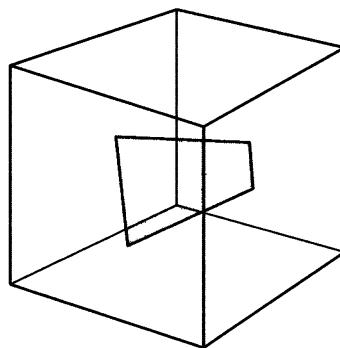


Рис. 12.2

По сути, дискретизация сложных областей первого типа заключается в независимой триангуляции каждой подобласти с обеспечением согласования сеток на границах. Из такой постановки задачи вытекает следующая схема решения:

- 1) триангуляция границ смежных подобластей;
- 2) дискретизация подобластей исходя из заданной триангуляции границ.

Триангуляция каждой подобласти может быть выполнена наиболее подходящим методом. В частности, для областей простой геометрической формы (параллелепипеды, шары, цилиндры, призмы) лучше использовать прямые методы. Для геометрически сложных областей подходящими являются методы исчерпывания, так как результат триангуляции границы дает фактически необходимые исходные данные для этих методов, т. е. начальный фронт.

Для сложных областей второго типа обычно используются методы на основе критерия Делоне, так как они позволяют предварительно размещать узлы на заданных ограничениях (поверхностях/кривых), либо методы граничной коррекции. Эти методы с некоторыми дополнительными условиями можно применять и для триангуляции областей первого типа.

Заметим также, что встречаются сложные области комбинированного типа, например одна из подобластей может включать в себя ограничения в виде поверхностей или кривых. Поэтому указанное разграничение областей на типы весьма условно.

## 12.2. Прямые методы

Главными преимуществами *прямых методов* являются высокая скорость работы, надежность и простота реализации; основным недостатком — ограниченная область применения. Фактически применение прямых методов эффективно только для триангуляции самых простых областей — шара, параллелепипеда, цилиндра и т. п. Однако нередко такие области являются частью некоторых сложных областей, и использование прямых методов вместо *итерационных* в этом случае позволяет существенно экономить машинные ресурсы и время. Кроме того, прямые методы могут быть эффективны также для триангуляции геометрически сложных областей заданного типа, что, однако, требует индивидуального подхода к каждой задаче. При этом, даже несмотря на то, что алгоритм получается адаптированным под заданную область, не всегда можно гарантировать хорошее качество сетки.

Сетки, построенные прямыми методами, могут быть использованы и в итерационных методах. В первую очередь это касается *методов граничной коррекции*. Размещение узлов в *методах на основе критерия Делоне* нередко осуществляется с помощью одного из прямых алгоритмов (с последующей коррекцией).

Таким образом, несмотря на все ограничения, прямые методы находят свое применение в дискретизации пространственных областей.

Важной особенностью сеток, построенных с помощью прямых методов, является их структурированность. Структурированные сетки имеют четкую топологию и позволяют вводить особую индексацию вершин. Таким образом, зная только индексы узла, можно легко определить все соседние узлы (иными словами, инцидентные ему узлы, т. е. все узлы, с которыми у него есть общее ребро), а также вычислить его координаты.

В качестве примера рассмотрим так называемую кубическую сетку (рис. 12.3), т. е. сетку, полученную разбиением исходного параллелепипеда на равные кубы (слово «куб» здесь употребляется исключительно в топологическом смысле, поскольку ребра у такого куба необязательно строго равны). Если размеры кубов  $h_x, h_y, h_z$ , и они ориентированы по осям координат, то узел с индексами  $i, j, k$  имеет координаты

$$(O_x + ih_x, O_y + jh_y, O_z + kh_z),$$

а соседними являются узлы с индексами

$$(i \pm 1, j, k), \quad (i, j \pm 1, k), \quad (i, j, k \pm 1).$$

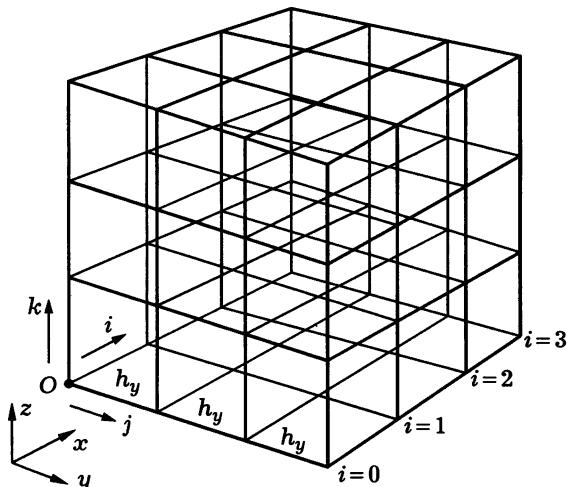


Рис. 12.3

Таким образом, нет необходимости хранить в памяти ни координаты узла, ни список соседних узлов. Аналогичную систему индексации можно использовать и для других структурированных сеток. Помимо экономии ресурсов, структурированность сетки заметно облегчает написание алгоритмов, в том числе алгоритмов сгущения и оптимизации.

### 12.2.1. Методы на основе шаблонов

**Шаблоном** называют некоторый принцип размещения узлов и установки связей между ними. Каждый шаблон применим только к областям заданного вида. Благодаря таким ограничениям сетки, построенные на основе шаблона, часто обладают высоким качеством.

Вследствие четкой структурированности сетки, построенные на основе шаблонов, как правило, невозможно стыковать друг с другом. Решать эту проблему можно путем использования специальных «переходных» областей, в которых сетка строится либо на основе специальных шаблонов, либо с помощью одного из итерационных методов.

**Триангуляция параллелепипеда.** Параллелепипед — это самая простая и довольно часто встречающаяся область для триангуляции. Для нее существует несколько различных шаблонов, основой которых является описанная выше кубическая сетка.

При «классическом» подходе [189, 198, 232, 263 и др] область разбивается на кубы, а затем каждый куб разбивается на пять (рис. 12.4, *a*) или шесть (рис. 12.4, *b*) тетраэдров путем вставки диагональных и внутренних ребер. У каждого из этих вариантов есть свои достоинства и недостатки. Перечислим их основные особенности.

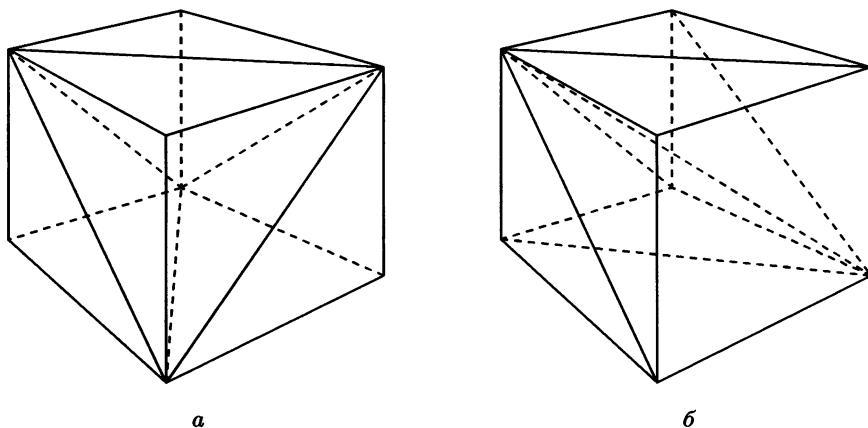


Рис. 12.4

**Шаблон 1 (разбиение на пять тетраэдров):**

- сетка неоднородна, узлы имеют резко различающееся число соседних узлов (одна половина имеет по 6 соседних узлов, а другая — по 18);
- достаточно хорошее качество сетки (средняя AX получающихся тетраэдров — 0,7);
- большая сложность построения и согласования (из-за необходимости чередовать направления диагональных ребер).

**Шаблон 2 (разбиение на 6 тетраэдров):**

- сетка однородна, все внутренние узлы имеют одинаковое число соседних узлов (по 14);

- относительно низкое качество сетки (средняя АХ получающихся тетраэдров — 0,3);

- простота построения и согласования сетки на границе.

Существуют и лучшие шаблоны. Рассмотрим два из них. В обоих этих шаблонах используется идея вставки внутрь каждого элемента кубической сетки дополнительного узла, который соединяется ребрами с вершинами куба, в результате чего исходный параллелепипед разбивается на два вида элементов:

- 1) граничные — в виде четырехугольной пирамиды (т. е. пирамиды, основанием которой является квадрат);
- 2) внутренние — в виде «объемного ромба» (октаэдра), составленного из двух четырехугольных пирамид, соединенных основаниями.

На рис. 12.5, а показан исходный элемент сетки с дополнительным узлом, на рис. 12.5, б — получающийся в результате дополнительный ромбовидный элемент, состоящий из двух четырехугольных пирамид, соединенных основаниями.

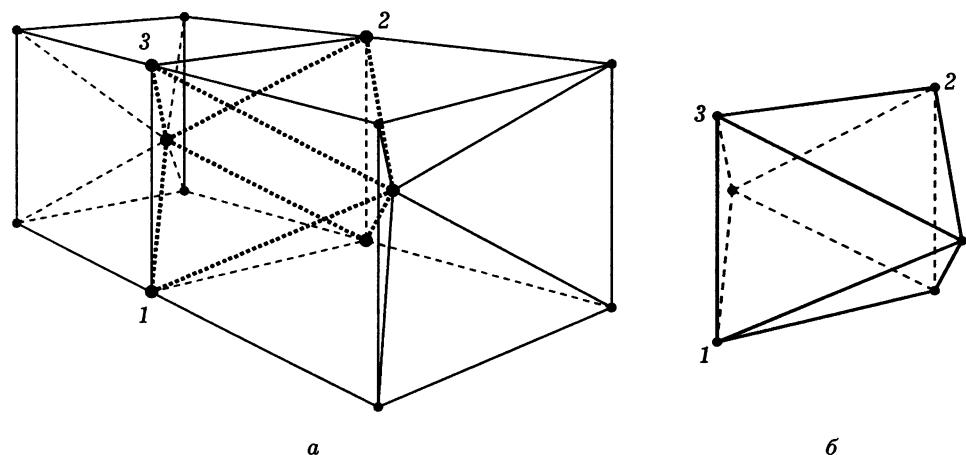


Рис. 12.5

Чтобы разбить граничные пирамидальные элементы, достаточно вставить диагональное произвольно ориентированное ребро. При этом получаются два одинаковых тетраэдра с АХ, равной примерно 0,5.

Разбить внутренние ромбовидные элементы можно различными способами. Рассмотрим такие способы.

Шаблон 3: вставка диагонального ребра между узлами кубической сетки (рис. 12.6).

Шаблон 4: вставка ребра между дополнительными узлами (рис. 12.7).

И в том, и в другом случае получаются 4 одинаковых тетраэдра. Однако отметим и существенные различия между шаблонами.

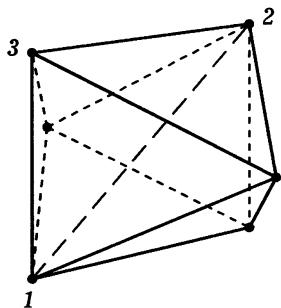


Рис. 12.6

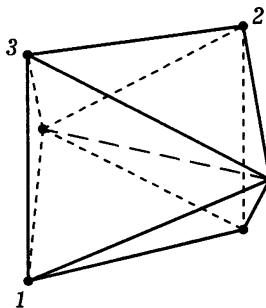


Рис. 12.7

**Шаблон 3:**

- два варианта вставки ребер;
- неоднородность сетки (дополнительные узлы имеют всего по 8 соседних узлов);
  - полная однородность элементов, т. е. все элементы одинаковы;
  - хорошее качество сетки (все тетраэдры имеют АХ, равную примерно 0,5);

**Шаблон 4:**

- единственный вариант вставки ребер (для внутренних элементов);
  - приграничные элементы отличаются от внутренних;
  - подавляющая однородность сетки (все внутренние узлы, за исключением дополнительных приграничных узлов, имеют по 14 соседних узлов);
  - очень высокое качество сетки (приграничные тетраэдры имеют АХ, равную примерно 0,5, внутренние — примерно 0,9).

Разумеется, можно использовать и комбинированный шаблон, в котором из трех возможных вариантов вставки ребра выбирается наилучший (например, по критерию максимизации минимальной АХ четырех получающихся тетраэдров). Такой алгоритм оправдан в случае сильного геометрического искажения области, например, когда область представляет собой не параллелепипед, а криволинейную шестигранную призму.

Недостатком шаблонов 3 и 4 является чуть большая, по сравнению с шаблонами 1 и 2, сложность построения сетки. К достоинствам шаблонов 3 и 4 стоит отнести произвольность направления диагональных ребер на границе, что облегчает согласование триангуляции при дискретизации сложных областей. Кроме того, эти шаблоны позволяют легко осуществлять локальное сгущение сетки с помощью специального переходного шаблона.

В завершение заметим, что по совокупности достоинств и недостатков оптимальным является использование шаблона 4.

**Триангуляция цилиндра.** Прежде чем приступить к триангуляции цилиндра, обратимся к триангуляции круга. Существует несколько различных подходов к решению этой задачи. Один из вариантов предполагает разбиение круга на несколько четырехугольных областей, которые затем разбиваются на квадраты. Окончательно полученные квадраты делятся на два треугольника. Разбиение круга на пять четырехугольных подобластей с их последующим разбиением на квадраты и треугольники показано на рис. 12.8.

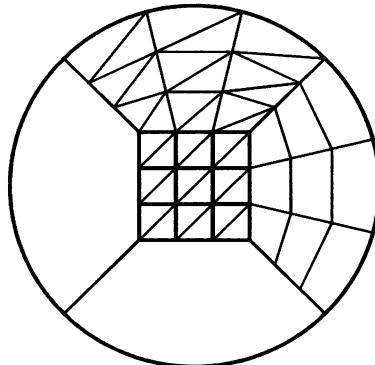


Рис. 12.8

Очевидно, что полученная таким образом сетка не обладает ни однородностью, ни качеством. Существует шаблон, который позволяет получить на круге однородные сетки высокого качества. Оказывается, что для этого достаточно делить круг не на четырехугольные, а на треугольные секторы. Причем количество секторов может варьироваться от 4 до 7 (меньшее или большее приведет уже к значительнымискажениям сетки). Узлы внутри секторов размещаются с помощью концентрических окружностей, причем на каждой последующей окружности количество узлов увеличивается на число секторов, т. е. в каждый сектор добавляется по узлу. Если распределять узлы на отрезках дуг равномерно, то полярные координаты узлов можно вычислять по индексам с помощью формул вида

$$\rho_{i,j} = R_i, \quad \varphi_{i,j} = \frac{2\pi j}{iN}, \quad i = 1, 2, \dots, \quad j = 1, 2, \dots, iN,$$

где  $i$  — номер окружности;  $j$  — номер узла;  $R_i$  — радиус окружности;  $N$  — количество секторов.

На рис. 12.9 и 12.10 приведены примеры использования этого шаблона при  $N = 4$  и  $6$  соответственно.

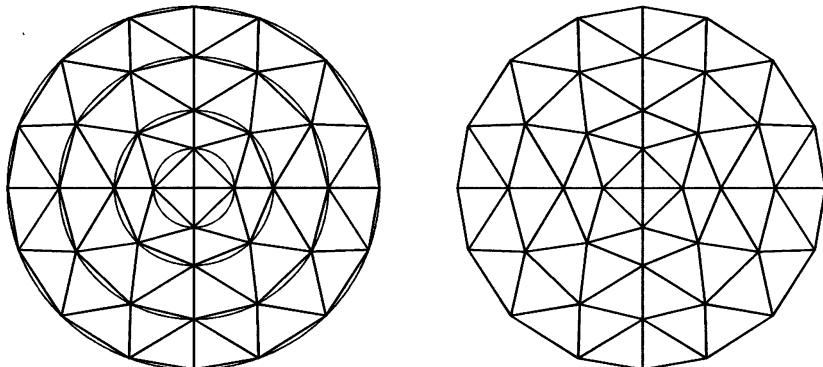


Рис. 12.9

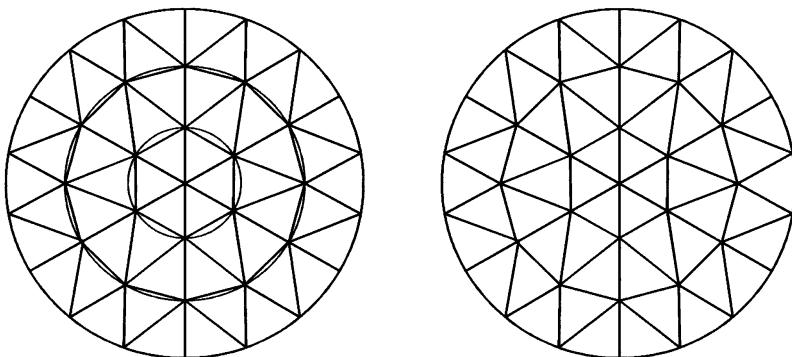


Рис. 12.10

Построенные сетки обладают неплохим качеством, а также подавляющей однородностью (все узлы, кроме граничного и центрального, имеют одинаковое число соседних узлов), а сетка, построенная на основе 6 секторов — даже полной однородностью.

Ввиду очевидных достоинств рассматриваемого шаблона в дальнейшем будет использоваться именно он.

Триангуляцию цилиндра разумнее всего проводить путем разбиения его на слои. Каждый слой будет представлять собой тонкий цилиндр («блин»), причем триангуляция обоих его оснований должна быть идентичной. Соединив ребрами соответствующие друг другу узлы на разных основаниях, можно получить так называемую призматическую сетку (рис. 12.11). Эти элементы обычно называются клиновидными или клиньями и также могут использоваться в качестве конечных элементов.

Таким образом, исходная задача сведена к разбиению конечного элемента — пятигранной призмы — на тетраэдры. Как и в случае с

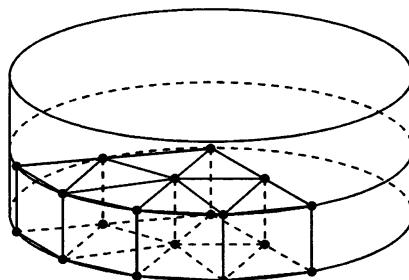


Рис. 12.11

кубической сеткой, это можно делать несколькими способами. Самый простой способ — разбиение каждой призмы на три тетраэдра с помощью диагональных ребер. При этом, однако, необходимо тщательно согласовывать направление вставляемых ребер, чтобы не столкнуться с невозможностью разбиения многогранника на непересекающиеся тетраэдры без использования дополнительных узлов. В данном конкретном случае это означает, что при некоторых комбинациях диагональных ребер призма не будет разбиваться на тетраэдры (рис. 12.12).

Однако существует лучший способ. В нем используется вставка дополнительного узла внутрь призмы. Но после соединения дополнительного узла с вершинами призмы будут получены элементы уже трех типов: четырехугольные пирамиды на внешней («круглой») границе цилиндра, объемные ромбы (октаэдры) между призмами на одном слое и собственно готовые тетраэдры на основаниях цилиндров-«блинов» (рис. 12.13).

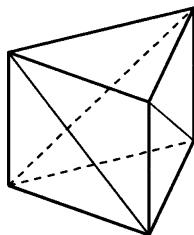


Рис. 12.12

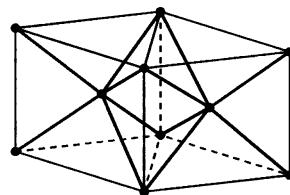


Рис. 12.13

Окончательно полученные элементы разбивают на тетраэдры: граничные пирамиды — вставкой диагональных ребер, внутренние ромбы — вставкой ребра между дополнительными узлами. При удачном подборе параметров итоговая сетка будет обладать неплохим качеством (среднее значение AX равно примерно 0,5). К сожалению, однородности добиться не удается (дополнительные узлы будут иметь всего

по 11 соседних узлов, а базовые — по 18), хотя четкая структурированность по-прежнему будет иметь место.

**Триангуляция шара.** Один из самых популярных методов триангуляции сферических областей основан на варианте *метода дробления*. Как и в случае круга, шар разбивается на сектора, но уже тетраэдрической формы. Для этого можно использовать либо три взаимно перпендикулярные плоскости, проходящие через центр шара (рис. 12.14), либо вписанный в шар икосаэдр (правильный двадцатигранник, гранями которого являются равносторонние треугольники). Второй случай предпочтительней из-за существенно меньших геометрических искажений сетки при ее построении (рис. 12.15).

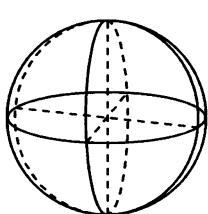


Рис. 12.14

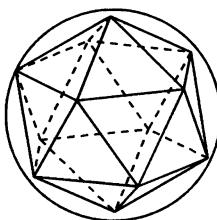


Рис. 12.15

**Метод дробления** заключается в последовательном измельчении некоторой уже построенной грубой сетки путем дробления ее элементов. В данном случае необходимо раздробить сектора, имеющие форму тетраэдров с одной неплоской гранью. Особенностью этой задачи является использование одинакового шаблона для каждого сектора, что автоматически обеспечивает согласование триангуляции между секторами ввиду очевидной симметрии.

Рассмотрим два возможных шаблона дробления тетраэдра.

Шаблон 1: разбиение тетраэдра на 8 частей (рис. 12.16). Для этого каждое его ребро делится на две части дополнительным узлом; через эти узлы проводят плоскости, отсекающие 4 тетраэдра с исходными вершинами. Оставшийся в итоге объемный ромб (октаэдр) делится на

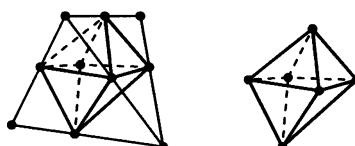


Рис. 12.16

4 тетраэдра вставкой внутреннего ребра. Из трех возможных вариантов проведения ребра выбирается оптимальный.

Шаблон 2: разбиение тетраэдра на 27 частей (рис. 12.17). Для этого необходимо разделить каждое его ребро на три равные части и вставить по дополнительному узлу в центр каждой грани. Каждая грань при этом разбивается на 9 треугольников. Соединив ребрами узлы, лежащие в центрах граней, получим разбиение исходного тетраэдра на 11 тетраэдров и 4 объемных ромба, каждый из которых затем оптимальным образом разбивается на 4 тетраэдра.

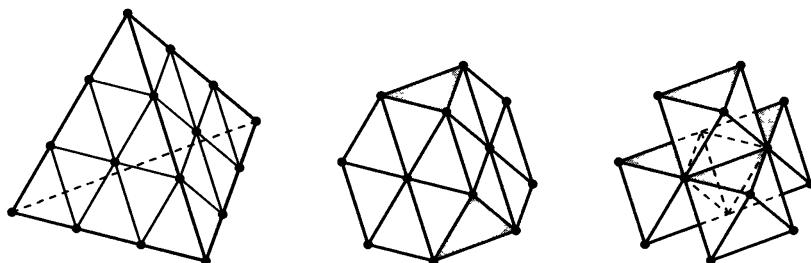


Рис. 12.17

Процесс дробления можно продолжать итерационно до тех пор, пока элементы не достигнут необходимого размера. Сложность, с которой можно столкнуться при использовании шаблона 2 для триангуляции секторов, заключается в наличии «кривой» (выпуклой) грани. Однако этой сложности можно избежать, если использовать не декартовы, а сферические координаты. В этом случае нужные координаты новых узлов будут получаться автоматически, без использования каких-либо процедур отображения. Рассмотрим данный алгоритм на примере двумерной области (рис. 12.18: *a* — декартовы координаты; *b* — полярные координаты; полюс в точке 1). Приведем соответствующие формулы.

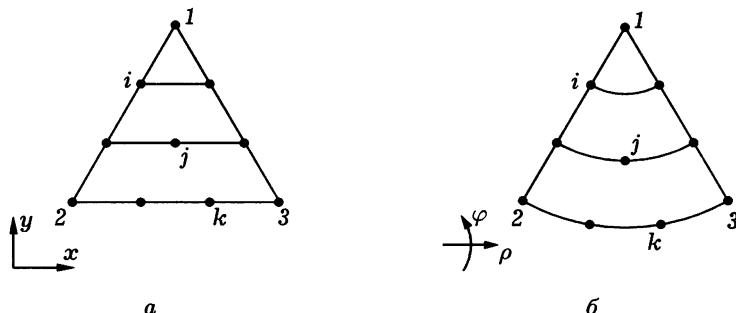


Рис. 12.18

Для декартовой системы координат:

$$\begin{aligned}x_i &= \frac{2}{3}x_1 + \frac{1}{3}x_2, & y_i &= \frac{2}{3}y_1 + \frac{1}{3}y_2, \\x_j &= \frac{1}{3}x_1 + \frac{1}{3}x_2 + \frac{1}{3}x_3, & y_j &= \frac{1}{3}y_1 + \frac{1}{3}y_2 + \frac{1}{3}y_3, \\x_k &= \frac{1}{3}x_2 + \frac{2}{3}x_3, & y_k &= \frac{1}{3}y_2 + \frac{2}{3}y_3 = y_2 = y_3.\end{aligned}\quad (12.3)$$

Для полярной системы координат:

$$\begin{aligned}\varphi_i &= \varphi_2, & \rho_i &= \frac{1}{3}\rho_2, \\ \varphi_j &= \frac{1}{2}\varphi_2 + \frac{1}{2}\varphi_3, & \rho_j &= \frac{1}{3}\rho_2 + \frac{1}{3}\rho_3, \\ \varphi_k &= \frac{1}{3}\varphi_2 + \frac{2}{3}\varphi_3, & \rho_k &= \frac{1}{3}\rho_2 + \frac{2}{3}\rho_3 = \rho_2 = \rho_3.\end{aligned}\quad (12.4)$$

Заметим, что при использовании метода дробления на основе 20 секторов сетка, которая образуется на границе сферы, будет состоять из правильных треугольников. Это весьма благотворно сказывается на качестве аппроксимации граничных условий [229]. Общее качество дискретизации шара методом дробления также довольно высоко, хотя и снижается с каждым шагом дробления. Для трех шагов дробления (при дроблении на 27 тетраэдров) средняя АХ будет около 0,5 (при этом линейные размеры элементов составляют примерно 1/50 диаметра шара). К сожалению, сетки, построенные на основе метода дробления, неоднородны и, как правило, не имеют четкой топологической структуры при оптимальном подборе вставляемых ребер.

### 12.2.2. Методы отображения

Методы отображения основаны на возможности построения взаимно однозначного отображения между областями различной геометрической формы. Таким образом, используя оператор отображения, можно перенести сетку из более простой области на заданную.

Существенным недостатком этих методов является неизбежное ухудшение качества сетки из-за геометрических искажений, возникающих при отображении. Вместе с тем даже сложные операции отображения требуют сравнительно небольших затрат ресурсов, так как при отображении меняются только координаты узлов, а связи остаются неизменными.

Поэтому в качестве отображаемых сеток («образов») имеет смысл использовать только сетки, построенные на шаблонах. Более универсальные итерационные методы разумнее применять сразу для заданной области.

Как правило, для отображения сеток используют два типа преобразований — «простейшие» аффинные, позволяющие только растягивать или сжимать сетку, и более универсальные изопараметрические, позволяющие отображать сетки даже в криволинейные области (рис. 12.19).

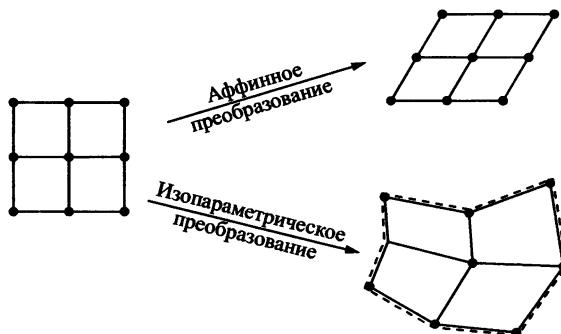


Рис. 12.19

Напомним, что аффинным называется линейное преобразование координат

$$x'_i = a_i x_i + b_i, \quad (12.5)$$

т. е. это преобразование сжатия/растяжения и сдвига. Оно позволяет приводить некоторую геометрическую форму (квадрат, треугольник, параллелепипед и т. д.) к так называемому «стандартному» виду, примеры которого квадрат, треугольник, круг (рис. 12.20).

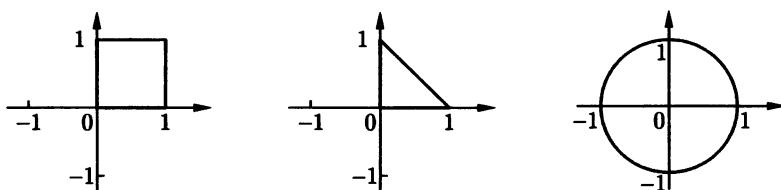


Рис. 12.20

Обычно это необходимо делать потому, что многие программные библиотеки, а также справочные таблицы (например, коэффициентов кубатурных формул) ориентированы именно на стандартные формы. Таким образом, в методах триангуляции аффинные преобразования, как правило, играют лишь незначительную вспомогательную роль.

Большее значение имеют *изопараметрические преобразования*. Заметим, что они нашли широкое применение не только в методах отображения, но и при решении задач на основе криволинейных элементов [189].

Сущность изопараметрического преобразования заключается в следующем: задается некая система внутренних *координат*, называемых *барицентрическими*, которая однозначным образом связывает положение любой точки данной геометрической формы (треугольник, квадрат, тетраэдр и т. д.) с определенным множеством базисных точек, также принадлежащих данной геометрической форме. В качестве таких точек обычно выбираются углы, середины сторон и т. п. Таким образом, изменив положение базисных точек, можно легко определить и новое положение всех остальных точек, используя их барицентрические координаты.

Проиллюстрируем сказанное на простейшем примере. Рассмотрим некоторый невырожденный треугольник с вершинами  $a_1, a_2, a_3$  (вершина  $a_i$  имеет координаты  $(a_{i1}, a_{i2})$ ) и определим его барицентрические координаты. Для каждой точки  $x = (x_1, x_2)$  этого треугольника барицентрические координаты  $\lambda_1, \lambda_2, \lambda_3$  вводятся как решение системы:

$$\sum_{i=1}^3 a_{ij} \lambda_i = x_j, \quad j = 1, 2; \quad \sum_i \lambda_i = 1. \quad (12.6)$$

Поскольку определитель этой системы равен удвоенной площади треугольника, система имеет единственное решение для каждой точки  $x$ . Заметим также, что, так как  $\lambda_i(a_j) = \delta_{ij}$ , функции  $\lambda_i(x)$ ,  $i = 1, 2, 3$ , являются базисными для элемента-треугольника.

Таким образом мы связали декартовы координаты каждой точки  $x$  треугольника с декартовыми координатами его вершин:

$$x_j = \sum_{i=1}^3 \lambda_i a_{ij}, \quad j = 1, 2. \quad (12.7)$$

Пусть необходимо отобразить сетку, построенную в треугольнике  $a_1, a_2, a_3$ , на треугольник  $b_1, b_2, b_3$ . Для этого сначала нужно найти барицентрические координаты всех узлов сетки. При этом нет необходимости решать систему (12.6), вместо этого можно воспользоваться геометрической интерпретацией (рис. 12.21). Оказывается,

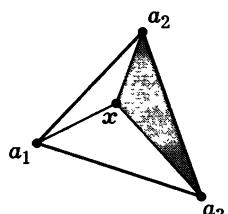


Рис. 12.21

барицентрические координаты легко определяются через отношения площадей треугольников:

$$\lambda_1 = \frac{S(x, a_2, a_3)}{S(a_1, a_2, a_3)}, \quad \lambda_2 = \frac{S(x, a_3, a_1)}{S(a_1, a_2, a_3)}, \quad \lambda_3 = \frac{S(x, a_1, a_2)}{S(a_1, a_2, a_3)}.$$

Сами площади элементарным образом рассчитываются методами векторной алгебры. Нахождение обратной матрицы системы линейных алгебраических уравнений (СЛАУ) (12.6), как правило, также не представляет каких-либо трудностей, поэтому в любом случае затраты вычислительных ресурсов на этом этапе минимальны.

После того как барицентрические координаты узла найдены, его новое положение вычисляется простой подстановкой в формулу (12.7) координат вершин  $b_i$  вместо координат  $a_i$ .

В рассмотренном примере прямолинейный треугольник отображается в прямоугольный треугольник. В этом случае вполне можно обойтись и аффинным преобразованием. Рассмотрим теперь отображение прямоугольной области в некоторую криволинейную.

На рис. 12.22 показана исходная прямоугольная область  $a_1a_2a_3a_4$ , которую необходимо изопараметрически отобразить в область  $b_1b_2b_3b_4$  с двумя криволинейными границами.

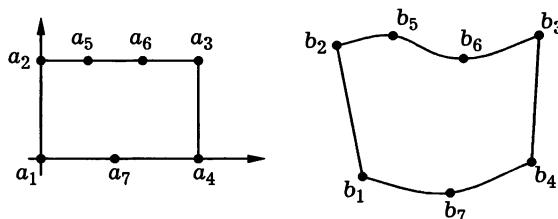


Рис. 12.22

Поскольку у области имеются криволинейные границы, необходимо использовать дополнительные базисные вершины, чтобы передать эту кривизну. В данном случае достаточно двух вершин для верхней кривой и одной вершины для нижней кривой. Отрезки  $a_2a_3$  и  $a_1a_4$  при этом необходимо дополнительными узлами разбить на равные части; узлы  $b_5$ ,  $b_6$ ,  $b_7$  размещаются более произвольно, но на заданных кривых и желательно на примерно равных интервалах. Теперь необходимо построить систему барицентрических координат или, что фактически то же самое, систему базисных функций. В этой задаче проявляется одно замечательное свойство изопараметрического отображения — его базисные функции можно легко выразить через функции другого базиса этой же области. В данном конкретном случае систему из 7 нужных

базисных функций можно выразить через 4 базисные функции прямоугольника  $a_1a_2a_3a_4$ .

Пусть имеется система базисных функций  $\{L_i\}$ ,  $i = 1, 2, 3, 4$ ,  $L_i(a_j) = \delta_{ij}$ . Для прямоугольника шириной  $C$  и высотой  $H$  (см. рис. 12.22) она может иметь, например, следующий вид [189]:

$$\begin{aligned} L_1 &= \left(1 - \frac{x_1}{C}\right) \left(1 - \frac{x_2}{H}\right), \quad L_2 = \left(1 - \frac{x_1}{C}\right) \frac{x_2}{H}, \\ L_3 &= \frac{x_1}{C} \frac{x_2}{H}, \quad L_4 = \frac{x_1}{C} \left(1 - \frac{x_2}{H}\right). \end{aligned}$$

По сути, это отношения площадей прямоугольников.

Тогда набор базисных функций  $\{\lambda_i(x)\}$ ,  $i = 1, 2, \dots, 7$ , выражается через  $\{L_i\}$  следующим образом:

$$\begin{aligned} \lambda_1 &= L_1(2L_1 - 1), \quad \lambda_2 = \frac{1}{2}L_2(3L_2 - 2)(3L_2 - 1), \\ \lambda_3 &= \frac{1}{2}L_3(3L_3 - 2)(3L_3 - 1), \quad \lambda_4 = L_4(2L_4 - 1), \\ \lambda_5 &= \frac{9}{2}L_2L_3(3L_2 - 1), \quad \lambda_6 = \frac{9}{2}L_2L_3(3L_3 - 1), \quad \lambda_7 = 4L_1L_4. \end{aligned} \quad (12.8)$$

Несложно проверить, что для данного набора также выполняется свойство  $\lambda_i(a_j) = \delta_{ij}$ ,  $i, j = 1, 2, \dots, 7$ .

Таким образом, алгоритм вычисления новых декартовых координат узла  $x$  выглядит следующим образом:

- 1) поиск прямым вычислением барицентрических координат  $L_i$ ,  $i = 1, 2, 3, 4$ ;
- 2) вычисление по (12.8) барицентрических координат  $\lambda_i$ ,  $i = 1, 2, \dots, 7$ ;
- 3) вычисление новых декартовых координат узла  $x$  по формуле

$$\tilde{x}_j = \sum_{i=1}^7 \lambda_i b_{ij}, \quad j = 1, 2.$$

Подводя итог, заметим, что указанный метод без каких-либо особенностей переносится на случай трех измерений. Несмотря на кажущуюся сложность, метод изопараметрических отображений является мощным инструментом, сравнительно простым и удобным в использовании.

## 12.3. Методы граничной коррекции

В самом названии этих методов содержится их основная идея. Наложив на заданную область некоторую уже построенную сетку, можно отсечь от этой сетки все выходящие за пределы нужной области фраг-

менты, а затем скорректировать положение узлов, лежащих вблизи границы, так, чтобы они попали в углы, на ребра и на грани области.

Таким образом, алгоритм разбивается на два этапа: построение первичной сетки и ее коррекция. Поскольку эти этапы практически не связаны друг с другом, рассмотрим их отдельно.

### 12.3.1. Построение первичной сетки

Самый простой подход к решению этой задачи заключается в использовании одного из методов на основе шаблонов. В этом случае в качестве области для построения первичной сетки выбирают одну из подходящих геометрических форм, для которых разработаны шаблоны. Эта область должна полностью включать в себя заданную. Как правило, в качестве такой суперобласти используют параллелепипед, хотя в некоторых случаях (например, радиальной симметрии) удобнее использовать цилиндр.

Триангуляция на основе шаблонов подробно рассмотрена выше. Заметим, что построенная таким образом итоговая сетка получается близкой к равномерной, т. е. линейные размеры ее элементов примерно равны. Это обусловлено тем, что при построении первичной сетки не используется никакой информации о геометрии заданной области.

Существует алгоритм, который позволяет учитывать особенности геометрии области и таким образом строить аддативные сетки, адаптированные к геометрии, а не к конкретной задаче. Этот алгоритм разработан в 80-х годах XX столетия и получил название octree (его двумерный вариант называется quadtree). С тех пор предложено несколько его разновидностей [263, 278, 279, 295], рассмотрим одну из них.

Идея алгоритма заключается в следующем: исходная область помещается в кубическую сетку, элементы которой последовательно дробятся на более мелкие кубы до тех пор, пока размеры получаемых в итоге кубических ячеек не достигнут желаемых значений; при этом каждый куб дробится только в том случае, если его грани пересекаются границей исходной области либо если внутри куба целиком оказываются особенности геометрии типа отверстия или полости. Таким образом удается добиться «естественному» увеличения плотности узлов вблизи границ области и ее «особенных» участков. Чтобы избежать значительных перепадов размеров элементов, дополнительно вводят ограничение на степень раздробленности соседних элементов — она не должна отличаться более чем на единицу (рис. 12.23: идея алгоритма quadtree для двумерного случая).

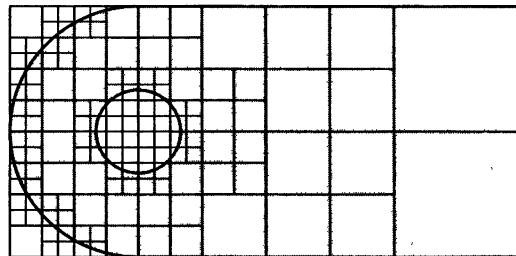


Рис. 12.23

Следующим этапом метода граничной коррекции является построение треугольной (тетраэдрической) сетки на основе полученного разбиения на квадраты (кубы). Поскольку возможных вариантов размещения узлов на ребрах и гранях квадратов (кубов) в такой сетке немного (для квадрата с учетом отражения и поворота — всего шесть), для каждого варианта используется свой заранее заданный шаблон. На рис. 12.24 и 12.25 показаны два набора таких шаблонов: классический (см. рис. 12.24) и разработанный авторами [43, 44]. Последний основан на вставке дополнительного узла (см. рис. 12.25) и позволяет получить сетки лучшего качества из подобных элементов. На рис. 12.26 приве-

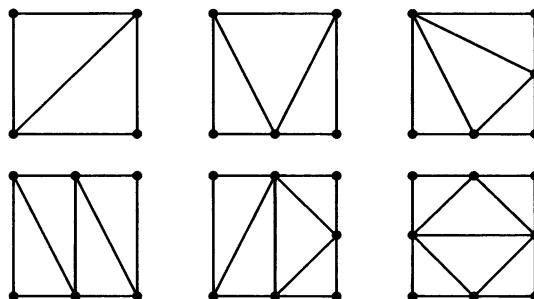


Рис. 12.24

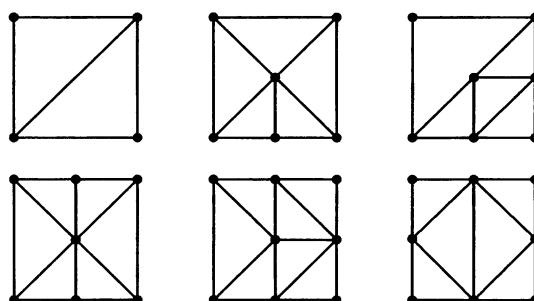


Рис. 12.25

дена сетка, построенная в результате применения указанных наборов шаблонов: выше оси симметрии сетка, построенная по классическим шаблонам, а ниже — по улучшенным.

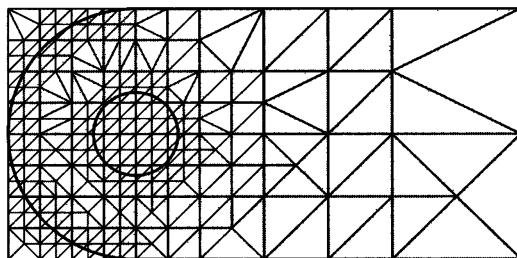


Рис. 12.26

Заметим, что описанный алгоритм без каких-либо особенностей переносится на трехмерный случай, поэтому дополнительно рассматривать его не будем. Следует также обратить внимание на то, что полученная сетка не является структурированной, хотя при дальнейшем выполнении алгоритма это не играет никакой роли, так как при граничной коррекции свойство структурированности в любом случае утрачивается.

### 12.3.2. Коррекция первичной сетки

Пусть имеется некоторая сетка, наложенная на заданную область. Необходимо отсечь у нее все лишнее и скорректировать положение узлов, лежащих вблизи границы области. Это непростая задача. Ее решение определяется способом задания самой области. Основная трудность при реализации метода граничной коррекции — необходимость добиться того, чтобы «ребра» границы области были аппроксимированы ребрами сетки, а во все «углы» границы непременно попали узлы сетки. Если граница области является гладкой, таких проблем не возникнет, однако нас интересует более общий случай, в котором граница может иметь ряд различных особенностей (быть несвязной, невыпуклой, а также состоять из фрагментов поверхностей, пересекающихся под достаточно острыми углами, в том числе иметь конусообразные выпуклости).

Если в качестве входных данных задаются все угловые точки, кривые и составляющие границу сплайны (сплайнами в данном случае называются фрагменты поверхностей), то проблема решается достаточно просто. Такая ситуация типична, когда область импортируется из CAD-системы. В этом случае задается массив особых точек,

в которые входят все угловые и другие характерные точки. В них обязательно должны размещаться узлы триангуляции. Если в заданной области есть ребра, которым не принадлежат никакие угловые точки (например, линия пересечения цилиндра с призмой), то в указанный массив следует добавить произвольную точку, лежащую на этом ребре. Процесс граничной коррекции можно разбить на этапы.

1. Для каждого элемента массива особых точек находят ближайший узел первичной сетки и передвигают его с сохранением всех связей.

2. Для каждого элемента массива особых точек и для каждого ребра границы, в которое входит эта особая точка (таких ребер может и не быть, если, например, точка является вершиной конусообразного выступа), анализируют множество соседних узлов передвинутого в эту точку узла первичной сетки и находят узел, расположенный ближе всех по отношению к ребру границы. Этот узел проецируют на ребро и среди множества соседних узлов вновь находят узел, лежащий ближе всех к ребру, исключая узел, спроектированный на предыдущей итерации. Процедура проецирования и поиска подходящих соседних узлов повторяется до тех пор, пока найденный узел не окажется элементом множества особых точек, т. е. ребро границы окажется полностью аппроксимированным цепочкой ребер первичной сетки. Обработанное ребро исключается из дальнейшего рассмотрения в рамках данного этапа алгоритма. По окончании этого этапа во всех углах границы области будут помещены узлы сетки, а все ребра окажутся аппроксимированы цепочками ребер. Далее остается только скорректировать положение узлов вблизи сплайнов; это один из самых простых этапов.

3. Для каждого сплайна рассматривают ребра первичной сетки, пересекающие этот сплайн. Находят точку пересечения ребра с поверхностью и в нее перемещают тот конец ребра, который лежит к ней ближе. Следует делать именно так, а не проецировать ближайший узел на поверхность (что на первый взгляд кажется лучшим вариантом), поскольку в этом случае спроектированный узел может оказаться за пределами заданной области или попасть внутрь другого тетраэдра. Если ребро делится сплайном точно пополам, для выбора перемещаемого узла проводится дополнительный анализ на основе, например, качества получающихся при этом тетраэдров.

4. В итоге проводят отсечение всех фрагментов первичной сетки, оставшихся за пределами заданной области. Иначе говоря, удаляют все узлы и ребра первичной сетки, лежащие вне заданной области.

Метод граничной коррекции обладает достаточно высокой скоростью работы, что является его главным достоинством, и сравнительной простотой реализации. Но у него есть и недостатки. Во-первых, его фактически невозможно использовать для триангуляции областей с заданной триангуляцией границы. Во-вторых, этот метод ненадежен, поэтому построенные с его помощью сетки необходимо проверять на правильность структуры. В-третьих, эти сетки обладают априори низким качеством элементов вблизи границы, поэтому для них столь же необходим и этап оптимизации (следует заметить, что при этом, как правило, качество сетки удается существенно улучшить). В-четвертых, метод граничной коррекции обладает низкой «чувствительностью», поэтому при недостаточно малом шаге триангуляции некоторые особенности области могут быть потеряны. В этом смысле алгоритм octree обладает лучшими свойствами по сравнению с другими вариантами.

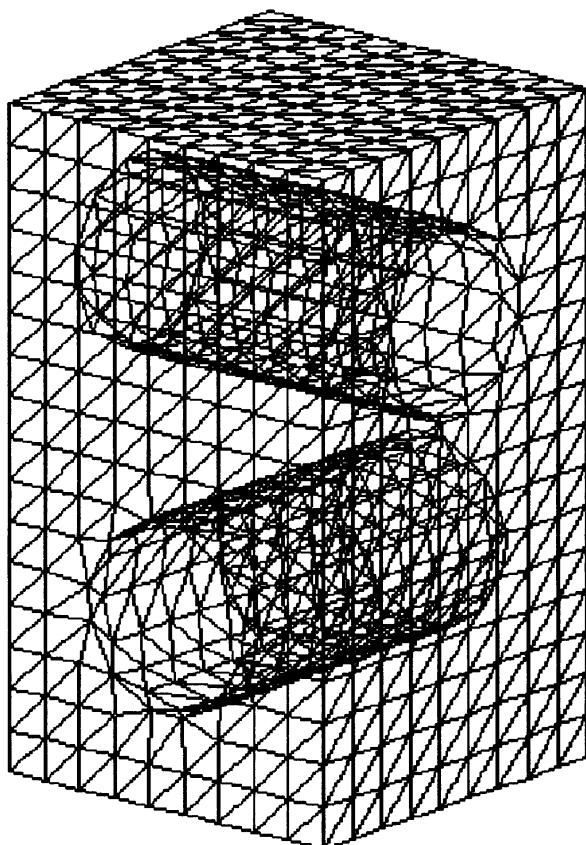


Рис. 12.27

Несмотря на указанные недостатки, метод граничной коррекции может быть успешно применен для триангуляции сложных областей, включающих в себя ограничения типа внутренних поверхностей и ребер. Поскольку при реализации алгоритма внутренние ребра ничем не отличаются от внешних, то данный метод не требует даже модификации самой программы.

Типичные ошибки, с которыми сталкиваются при реализации метода граничной коррекции, — это «слипание» узлов (когда два узла первичной сетки оказываются перемещенными в одну точку), что приводит к появлению вырожденных тетраэдров, а также к появлению тетраэдров с нулевым объемом (все четыре узла оказываются в одной плоскости). Избежать этого можно в результате удаления лишних вершин и последующего обновления связей.

Пример использования метода граничной коррекции — триангуляция призмы с двумя цилиндрическими отверстиями (рис. 12.27, показаны только граничные узлы «видимых» сплайнов).

## 12.4. Методы на основе критерия Делоне

Напомним, что такое критерий Делоне: треугольная сетка на плоскости удовлетворяет критерию Делоне, или является *триангуляцией Делоне*, если внутри окружности, описанной вокруг любого треугольника, не попадают никакие другие узлы этой сетки. Этот термин также употребляется и по отношению к треугольнику сетки: треугольник удовлетворяет критерию Делоне, или условию «пустой» окружности, если критерию Делоне удовлетворяет сетка, составленная только из самого треугольника и соседних с ним треугольников [159, 198].

На рис. 12.28 показана сетка, удовлетворяющая критерию Делоне (рис. 12.28, *a*) и не удовлетворяющая ему (рис. 12.28, *б*).

Триангуляция Делоне обладает рядом интересных свойств [159, 244, 263], в частности наибольшей суммой минимальных углов всех своих треугольников, а также наименьшей суммой радиусов описанных

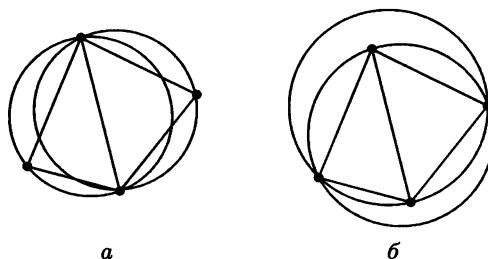


Рис. 12.28

вокруг треугольников окружностей среди всех возможных сеток на той же системе точек. Поскольку значения минимальных углов явным образом фигурируют в некоторых оценках качества аппроксимации [195], можно сказать, что триангуляция Делоне для заданного набора точек в определенном смысле является оптимальной.

Алгоритмы построения сеток на основе критерия Делоне предложены в [290]. За короткое время эти идеи были существенно развиты, и в настоящий момент проблема двумерной триангуляции методами на основе критерия Делоне фактически является закрытой. Разработаны быстрые и эффективные алгоритмы построения и оптимизации сеток, в том числе и в сложных (двумерных) областях [159, 160].

При попытках обобщить идеи алгоритмов на случай трех измерений обнаружился ряд проблем. Во-первых, выяснилось, что критерий Делоне в трехмерном случае не работает: сетка, удовлетворяющая критерию Делоне, не максимизирует минимальные углы. Хотя при этом следует отметить, что свойство минимизации радиусов описанных сфер сохраняется. Более того, как показывает работа [266], это справедливо и для  $n$ -мерного случая. Во-вторых, обнаружилось, что в пространстве не работают также и алгоритмы последовательного улучшения. Остановимся на этом подробнее.

В двумерном случае существует простой метод приведения произвольной триангуляции к триангуляции Делоне. Идея основана на том факте, что пару треугольников, не удовлетворяющих критерию Делоне, можно заменить на пару дуальных к ним треугольников, которые всегда удовлетворяют критерию. Это достигается перестановкой внутреннего ребра четырехугольника, образованного треугольниками (см. рис. 12.28). Операцию (так называемый *флип*) продолжают итерационно для каждой пары треугольников, не удовлетворяющих критерию, до тех пор, пока такие треугольники существуют.

У двумерного флипа есть и трехмерный аналог, но он основан на другой идеи. Оказывается, почти любые два соседних тетраэдра можно превратить в три. Для этого достаточно вставить внутрь шестигранника, образованного тетраэдрами, внутреннее ребро (рис. 12.29), причем сделать это можно единственным образом. Мы говорим «почти» потому, что если любые три вершины этого шестигранника лежат на одной прямой либо четыре его вершины лежат в одной плоскости, то эта операция приводит к образованию вырожденных тетраэдров. Операция замены двух тетраэдров тремя (см. рис. 12.29) и наоборот называется *трейд*. Как и флип, трейд позволяет заменять элементы, не удовлетворяющие критерию Делоне, на элементы, ему удовлетворяющие.

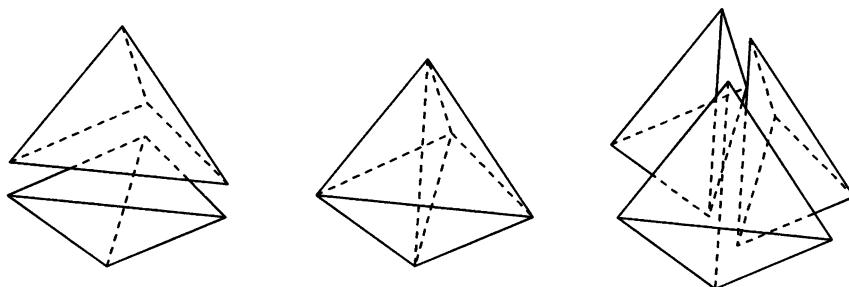


Рис. 12.29

Попытки использовать трейд для приведения произвольной трехмерной сетки к триангуляции Делоне закончились неудачно, поскольку такие алгоритмы очень часто «зацикливались» в локальных оптимумах. Вместе с тем не так давно получен важный теоретический результат: доказано, что если к существующей триангуляции Делоне добавить новые тетраэдры (разбив один из внутренних или присоединив тетраэдр к внешней грани), то полученную сетку можно гарантированно привести к триангуляции Делоне с помощью последовательных трейдов [244]. Этот факт чрезвычайно важен при построении триангуляций Делоне с ограничениями.

#### 12.4.1. Построение триангуляции Делоне на заданном наборе точек

Исходными данными для построения триангуляции Делоне является некоторый набор точек, которые должны стать узлами будущей триангуляции. Очевидным достоинством такого подхода является очень точный контроль над размерами элементов сетки — фактически эти размеры определяются плотностью размещения узлов. Увеличивая плотность размещения узлов в особых местах области, можно автоматически добиться локального сгущения сетки вблизи таких мест. Кроме того, если область является сложной, можно столь же легко обеспечить размещение узлов на поверхностях и ребрах ограничений.

Что касается самого способа предварительного размещения узлов, то существует множество вариантов. Заметим, что следует категорически воздерживаться от принципа случайного размещения узлов, поскольку при этом весьма вероятно появление точек, лежащих очень близко друг к другу, что, в свою очередь, неминуемо влечет появление в сетке очень коротких ребер и, соответственно, тетраэдров предельно низкого качества.

Самый простой и часто используемый метод размещения узлов основан на принципах *граничной коррекции*. Исходная область помеща-

ется в некоторую суперобласть, заполняемую узлами в соответствии с заданной плотностью размещения узлов, затем узлы, лежащие вблизи границы области, проецируются на нее, а узлы, лежащие вне области, удаляются.

Существуют и другие, более сложные методы. Так, метод под названием «упаковка пузырьков» основан на физической аналогии заполнения области мыльными пузырьками. Модель этого процесса, разумеется, сильно упрощена — до уровня сил отталкивания между центрами пузырьков. Возникающие при этом дополнительные довольно значительные затраты на решение системы дифференциальных уравнений компенсируются оптимальным размещением узлов, позволяющим получить сетки с априори высоким качеством триангуляции [280]. Используются и другие методы размещения узлов [196, 203, 208, 269, 281].

Вернемся к задаче построения триангуляции Делоне на заданном наборе точек. Существует множество таких алгоритмов. Почти все они происходят из двумерных [159]. Используемые в них геометрические обоснования универсальны и годятся для любого числа измерений.

Рассмотрим один такой алгоритм. Заметим, что при его использовании сетку имеет смысл хранить именно как список элементов-тетраэдров, а не как узлы со списком соседних узлов.

Алгоритм состоит из следующих шагов.

1. Формирование множества  $U$  — набора заданных узлов.
2. Создание так называемой суперструктурой, представляющей собой произвольный выпуклый многогранник с треугольными гранями, такой, что все заданные узлы лежат внутри него. Вершинами многогранника могут быть как элементы множества  $U$ , так и дополнительные узлы. В дальнейшем до определенного этапа с этими узлами обращаются как и со всеми остальными. В качестве суперструктуры может быть использован тетраэдр.
3. Формирование множества  $G$  узлов сетки, куда переносятся все узлы множества  $U$ , использованные как вершины суперструктуры (если такие есть).
4. Если в качестве суперструктуры использован тетраэдр, то выполняется переход к п. 5; в противном случае на основе узлов суперструктуры формируется триангуляция Делоне. Если в качестве суперструктуры использован правильный многогранник (октаэдр или икосаэдр), то это можно сделать следующим образом: выбрав произвольный узел из множества  $U$ , перенести его в множество  $G$  и путем вставки ребер между этим узлом и всеми вершинами многогранника сформировать сетку из  $n$  тетраэдров,  $n$  — число граней, равное 8 или 20 соответственно. Эта сетка будет являться триангуляцией Делоне [243].

5. Поиск для всех тетраэдров сетки центров и радиусов описанной сферы.

6. Выбор произвольного узла  $q$  из множества  $U$  и перенос его в множество  $G$ , затем удаление всех тетраэдров, по отношению к которым узел  $q$  попадает внутрь описанной сферы. Таким образом, в сетке образуется полость в виде многогранника, в общем случае имеющего звездную форму. При этом любой луч, исходящий из узла  $q$ , должен пересекать границу этого многогранника в единственной точке. Если обнаруживаются тетраэдры, по отношению к которым узел  $q$  лежит в плоскости одной из граней (это возможно в случае неоднозначности триангуляции Делоне, если, например, пять или больше точек лежат на одной сфере), то их тоже необходимо удалить. Отметим, что фактически ребро (или грань) удаляется только в том случае, если удаляются все смежные с ним тетраэдры, при этом ребра и грани суперструктуры не удаляются никогда. Новые тетраэдры образуются путем вставки ребер между узлом  $q$  и вершинами этого многогранника. Доказано, что в результате получается триангуляция Делоне [244].

7. Нахождение для новообразованных тетраэдров центра и радиуса описанной сферы.

8. Если множество  $U$  непустое, то осуществляется переход к п. 6, в противном случае — к п. 9.

9. Удаление из сетки всех тетраэдров, среди вершин которых есть вспомогательные узлы, использовавшиеся для построения суперструктуры. В результате получается сетка, построенная только на заданных узлах множества  $G$ .

Двумерный аналог этого процесса проиллюстрирован на рис. 12.30–12.39.

На рис. 12.30 показан заданный изначально набор точек.

На рис. 12.31 представлено построение суперструктуры (квадрата); один из заданных узлов используется как вершина квадрата.

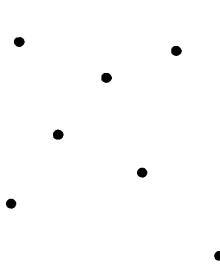


Рис. 12.30

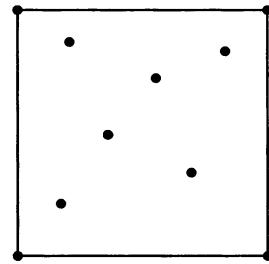


Рис. 12.31

На рис. 12.32 показана исходная триангуляция Делоне. Поскольку квадрат является правильным многоугольником, для построения триангуляции можно использовать любой из заданных узлов.

Далее необходимо выбрать новый (произвольный) узел и удалить все треугольники, по отношению к которым выбранный узел лежит внутри описанной окружности (рис. 12.33). После этого следует соединить новую вершину ребрами с углами образовавшегося многоугольника (рис. 12.34).

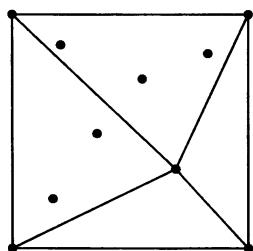


Рис. 12.32

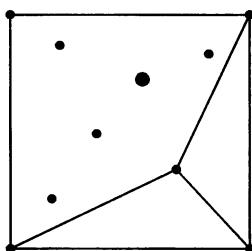


Рис. 12.33

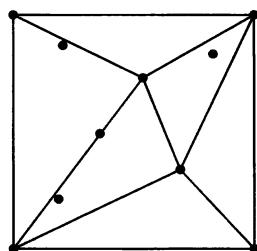


Рис. 12.34

Затем выбирают следующий узел и проводят итерационное повторение процедуры, пока все заданные узлы не будут использованы (рис. 12.35–12.38).

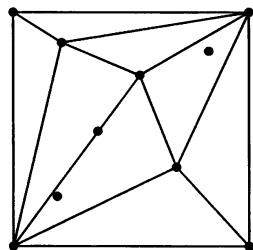


Рис. 12.35

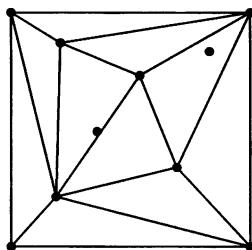


Рис. 12.36

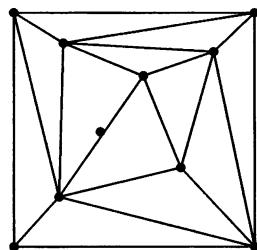


Рис. 12.37

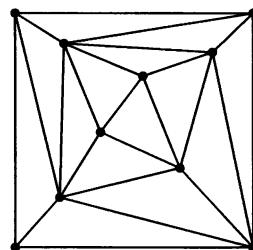


Рис. 12.38

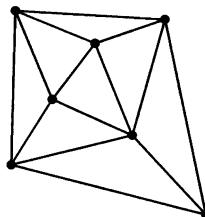


Рис. 12.39

Последний шаг алгоритма заключается в удалении из сетки всех треугольников, среди вершин которых имеются вспомогательные узлы суперструктурь (рис. 12.39).

Описанный алгоритм позволяет гарантированно строить триангуляцию Делоне для произвольного набора точек, причем граница сетки будет представлять собой в общем случае невыпуклый многогранник с треугольными гранями, опирающимися на узлы, наиболее удаленные от центра триангуляции. К сожалению, на практике приходится иметь дело с областями, представляющими собой более сложные геометрические формы.

В принципе описанный алгоритм можно использовать для любой области, если в качестве входных данных задавать не только набор точек, но и некую грубую начальную триангуляцию Делоне заданной области (т. е. сразу перейти к п. 5). В то же время, если такая триангуляция уже существует, проще получить итоговую сетку методами дробления, измельчив имеющиеся тетраэдры до нужных размеров, что будет гораздо быстрее и с учетом последующей минимальной оптимизации обеспечит примерно такое же качество. Данный вариант может быть использован, если необходимо обеспечить локальное сгущение сетки в нужных местах области (так как сетки, полученные методами дробления, являются равномерными). Однако остается нерешенным вопрос о построении начальной грубой сетки, что само по себе является отдельной задачей. Поэтому большее значение имеет задача построения триангуляции Делоне с ограничениями.

#### 12.4.2. Триангуляции Делоне с ограничениями

*Триангуляцией Делоне с ограничениями* в виде поверхностей называют сетку, обладающую следующим свойством: внутрь сферы, описанной вокруг любого тетраэдра этой сетки, не попадают никакие другие узлы сетки, видимые вершинам этого тетраэдра. Считается, что точка  $a$  видима точке  $b$  и наоборот, если отрезок  $[a, b]$  не пересекает никаких поверхностей ограничений. Общепринятого определения триангуляции Делоне с ограничениями, если ограничения представлены в том числе и отрезками, до сих пор нет.

В двумерном случае проблема построения триангуляции Делоне с ограничениями давно и успешно решена, причем самыми разными способами [160]. Что касается трехмерного случая, то к настоящему времени разработано лишь несколько алгоритмов, и все они базируются на одном и том же принципе. Их идея заключается в первоначальном построении в заданной области триангуляции Делоне без ограничений

и последующем восстановлении поверхностей и линий ограничений путем локальной перестройки сетки. Различие между алгоритмами состоит в способе этой перестройки [196, 198, 245, 253].

Рассмотрим один из алгоритмов, основанный на идее работы [240]. В ней доказана возможность построения триангуляции Делоне с ограничениями путем ретриангulationи только тех тетраэдров, которые пересекаются поверхностями ограничений, для случая пересечения тетраэдра произвольным выпуклым многогранником. Для ясности упростим алгоритм, введя дополнительные несложные требования к входным данным, а именно: потребуем, чтобы поверхности ограничения были представлены либо плоскостями, либо слабо изогнутыми (радиус кривизны много больше линейных размеров элементов) сплайнами, а также чтобы все угловые точки поверхностей ограничения использовались на первом этапе для построения триангуляции Делоне. В результате все тетраэдры построенной триангуляции могут пересекаться только либо ребрами поверхностей ограничений, либо самими поверхностями. Поскольку вариантов таких пересечений немного, для каждого из них можно использовать заранее подготовленный шаблон ретриангulationи. Процесс можно еще более упростить, если предварительно все ребра ограничений будут аппроксимированы цепочкой ребер триангуляции. Тогда тетраэдры могут быть пересечены только плоскостями (сплайнами), а возможных вариантов такого пересечения всего три: плоскость не пересекает вершины (рис. 12.40, *a*); плоскость пересекает две вершины (рис. 12.40, *б*); плоскость пересекает только одну вершину (рис. 12.40, *в*).

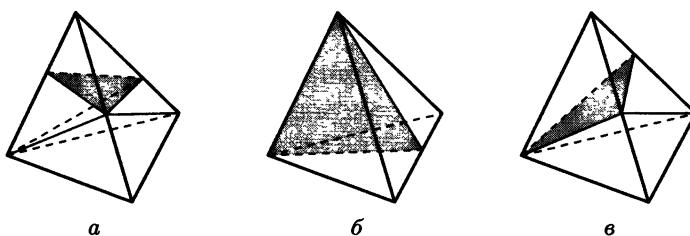


Рис. 12.40

Таким образом, алгоритм можно разбить на четыре шага:

- 1) построение триангуляции Делоне без ограничений;
- 2) восстановление ребер ограничений;
- 3) восстановление поверхностей ограничений;
- 4) отсечение лишних тетраэдров, оказавшихся вне границы заданной области.

Остановимся подробнее на каждом шаге.

1. Несмотря на кажущуюся простоту, этот шаг имеет несколько важных особенностей. Во-первых, все углы поверхностей ограничения (включая концы незакрепленных ребер) должны в обязательном порядке войти в набор узлов, по которым строится триангуляция. Во-вторых, поскольку многогранник, получающийся при использовании описанного выше алгоритма, не обязательно будет выпуклым, вполне возможно, что некоторые участки поверхностей ограничений окажутся вне этого многогранника. Чтобы не допустить этого, можно использовать слой дополнительных узлов, окружающий заданную область. На четвертом шаге все тетраэдры, образованные этими узлами, удаляются наряду с остальными тетраэдрами, оказавшимися за границей.

2. Восстановление ребер проводится с помощью дополнительных узлов, не входивших в изначальный набор. Заметим, что если в двухмерном случае всегда можно построить триангуляцию Делоне с ограничениями без использования дополнительных узлов, то в трехмерном это, как правило, невозможно.

Дополнительные узлы вставляются в точки пересечений ребер ограничений с гранями и ребрами тетраэдров. Вариантов пересечения ребра с тетраэдром всего шесть:

- 1) вершина–ребро (рис. 12.41, а);
- 2) вершина–грань (рис. 12.41, б);
- 3) ребро–ребро (рис. 12.41, в);
- 4) ребро–грань (рис. 12.41, г);

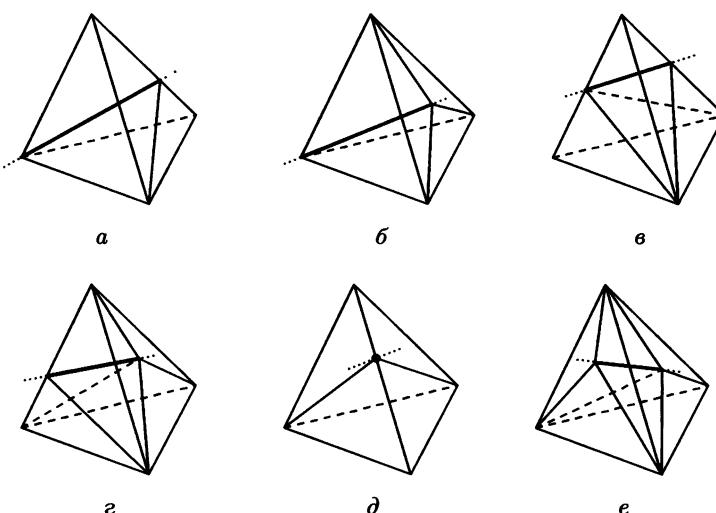


Рис. 12.41

- 5) ребро–вырожденный случай (рис. 12.41, *д*);
- 6) грань–грань (рис. 12.41, *е*).

Для каждого варианта используется свой шаблон разбиения тетраэдра. При этом необходимо согласовать триангуляцию на гранях соседних тетраэдров, поскольку некоторые случаи (см. рис. 12.41, *в*) допускают различные способы разбиения граней. Эта проблема решается установкой следующих двух правил: из всех вариантов всегда выбирается самое короткое ребро; если ребра равны, строится ребро от узла, например, с самым меньшим глобальным номером.

Вспомним, что тетраэдр может пересекаться не одним ребром, а несколькими. Теоретически ребер, пересекающих тетраэдр, может быть неограниченно много. Чтобы избежать ввода дополнительных шаблонов, достаточно обрабатывать каждое ребро отдельно. При обработке каждого ребра другие, еще не обработанные ребра, не учитываются, а после обработки ребро уже аппроксимировано цепочкой ребер триангуляции и естественным образом входит в сетку.

3. На предыдущем этапе все ребра ограничений аппроксимированы цепочками ребер сетки. Далее нужно добиться того, чтобы все поверхности ограничений были представлены множеством смежных граней. Это можно осуществить вставкой дополнительных вершин в точках пересечения поверхностей ограничения с ребрами сетки (см. рис. 12.40, *а–в*). Рассмотрим вопрос об однозначности разбиения граней. В случае если поверхность пересекает три ребра тетраэдра (см. рис. 12.40, *а*), ребра проводят так, чтобы нижняя часть элемента — пятигранник (усеченный тетраэдр) — не разбивалась на тетраэдры. Добиться этого можно, вставив дополнительный узел внутрь пятигранника. Тогда согласование ребер осуществляют по правилам, изложенным выше.

4. Удаление лишних тетраэдров, лежащих вне границы заданной области, не является сложным процессом. Можно сказать, что после их удаления граница области будет полностью аппроксимирована гранями и ребрами построенной триангуляции. То же самое относится и к внутренним ребрам, и к поверхностям ограничений, если таковые были.

На рис. 12.42 представлен результат описанного алгоритма триангуляции области, представляющей собой объединение икосаэдра и додекаэдра [210].

Качество сеток, построенных по данному алгоритму, находится на среднем уровне (тетраэдры у границ могут иметь очень плохую форму), поэтому обычно дополнительно прибегают к одному из методов оптимизации.

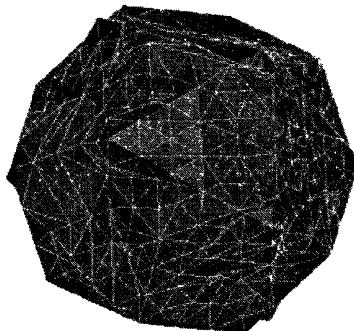


Рис. 12.42

В работах [244, 245] предложены другие варианты алгоритма, в которых не используются дополнительные точки. Эти варианты полностью основаны на локальных трансформациях, аналогичных трейду.

#### 12.4.3. Особенности технической реализации алгоритмов на основе критерия Делоне

Общий недостаток всех методов на основе критерия Делоне — крайне высокая чувствительность к точности машинных вычислений. Многие вычислительные процедуры, используемые в этих методах (нахождение центра и радиуса описанной окружности, проверка компланарности и коллинеарности векторов и т. п.) представляют собой *плохо обусловленные задачи*, а их интенсивное использование влечет накопление ошибок округления, что в итоге может привести к ошибкам в структуре сетки или зациклыванию алгоритма.

Типичная ошибка, допускаемая в реализации алгоритмов на основе критерия Делоне, заключается в использовании приближенных вычислений и операций, например, в использовании для сравнения не нуля ( $x > 0,0$ ), а заданной точности ( $x > \varepsilon$  или  $x > -\varepsilon$ ). Такой подход вполне оправдан и верен во многих других случаях, но только не в методах на основе критерия Делоне.

К сожалению, простое увеличение точности не дает существенных результатов. Использование машинных чисел с удвоенной точностью перестает быть эффективным уже в задачах средней сложности (сетки с несколькими тысячами узлов). Частично решить эту проблему можно, используя машинные числа с фиксированной запятой и раздельным хранением целой и десятичной частей числа (в виде целых чисел). Если число с фиксированной запятой имеет 10 десятичных разрядов, что соответствует 32-разрядному целому числу в современных ЭВМ, то

можно добиться точности вплоть до десятого знака, что является уже допустимым результатом [218].

Также возможно использование так называемой точной арифметики, модули для реализации которой уже разработаны для многих популярных прикладных языков программирования (C++, Фортран и др.). В точной арифметике все иррациональные числа представляются как набор предикатов (арифметических и алгебраических операторов) от рациональных/целых чисел и установленных констант (таких, как  $e$ ,  $\pi$  и т. д.), т. е., например, в точной арифметике  $\sqrt{3}$  — это действительно  $\sqrt{3}$ , а не 1,73205080756887729...

Недостатком обоих подходов является весьма существенное снижение скорости вычислений, так как ни описанная точная арифметика, ни числа с фиксированной запятой пока аппаратно не поддерживаются процессорами современных персональных компьютеров. Поэтому все эти операции приходится реализовывать на уровне подпрограмм, что приводит к дополнительным затратам ресурсов.

Иной путь улучшения ситуации — оптимизация процедур вычисления. Так, например, авторы [210] сумели создать устойчивый алгоритм построения триангуляции Делоне, сведя все геометрические операции к двум предикатам: проверке условия пустой сферы и нахождению положения точки относительно заданной плоскости.

Для проверки условия пустой окружности/сферы требуется найти радиус и центр описанной окружности, а затем проверить каждый подозрительный узел на условие: расстояние от узла до центра описанной окружности больше радиуса. Однако, если указанный центр найден недостаточно точно, то и гарантировать выполнение или невыполнение условия пустой окружности/сферы тоже нельзя. Реальные ситуации, при которых подозрительный узел находится вблизи окружности, достаточно часты.

Триангуляция с ограничениями является более сложной процедурой: добавляется необходимость определять положение точки относительно прямой/плоскости (справа, слева или на ней), а также проверять, пересекается ли один отрезок другим. В этом случае точность очень важна, так как ход алгоритма напрямую зависит от ответа на эти вопросы.

## 12.5. Метод исчерпывания

Идея *метода исчерпывания* предложена в [258], а его трехмерная реализация разработана в [256, 257]. Этот метод вот уже многие годы успешно используется в программном комплексе ANSYS для дискретизации достаточно произвольных трехмерных областей.

Общая идея этого метода заключается в последовательном изъятии из заданной области фрагментов тетраэдрической формы до тех пор, пока вся область не окажется «исчерпаной». Впрочем, этот процесс на самом деле имеет мало общего с практической реализацией. Английское название «advancing front» (продвижение фронта), пожалуй, лучше отражает сущность метода.

Основой всех вариантов этого метода является некоторая триангуляция границы заданной области, причем не грубая, а наиболее полно соответствующая требованиям разработчика, поскольку в дальнейшем эта триангуляция не будет меняться. Заметим, что подобное требование к начальным данным может быть как положительной стороной метода, так и отрицательной. Если никаких ограничений на границу области не накладывается, то это приводит к необходимости ее отдельной триангуляции, что само по себе является самостоятельной задачей. Если же необходимо обеспечить согласование триангуляции в сложной области, либо триангуляция поверхности задана изначально, то использование метода исчерпывания будет самым удачным выбором.

Триангуляция границы является тем самым фронтом, упоминаемым в английском названии. Используя какой-либо треугольник, принадлежащий фронту, на его основе можно построить тетраэдр, причем четвертой вершиной тетраэдра будет либо другая вершина, принадлежащая фронту, либо дополнительный узел, помещаемый внутрь заданной области. При изъятии полученного тетраэдра из фронта может быть удалено от одной (случай вставки дополнительного узла) до четырех граней и одновременно добавлено от одной до трех новых граней. Таким образом, текущий фронт дискретизации постепенно продвигается в пространстве. Обновив данные о фронте, можно вновь изъять тетраэдр, снова обновить фронт и так далее, пока вся область не окажется исчерпаной.

Заметим, что процесс исчерпывания применим для дискретизации как внутренней области, так и внешней, например, для дискретизации пространства вокруг модели самолета в задаче аэродинамики (рис. 12.43, [277]).

Несмотря на кажущуюся простоту идеи, реализация алгоритма исчерпывания содержит ряд тонкостей (заметим, что реализация алгоритмов исчерпывания в двумерном случае намного проще). Самый сложный шаг любого варианта исчерпывания заключается в проверке правильности построенного тетраэдра. Необходимо удостовериться, что этот новый тетраэдр не пересекается ни с какими уже существующими. Причем на каждой итерации алгоритма эта процедура с различными параметрами может вызываться от 5 до 20 раз (иногда

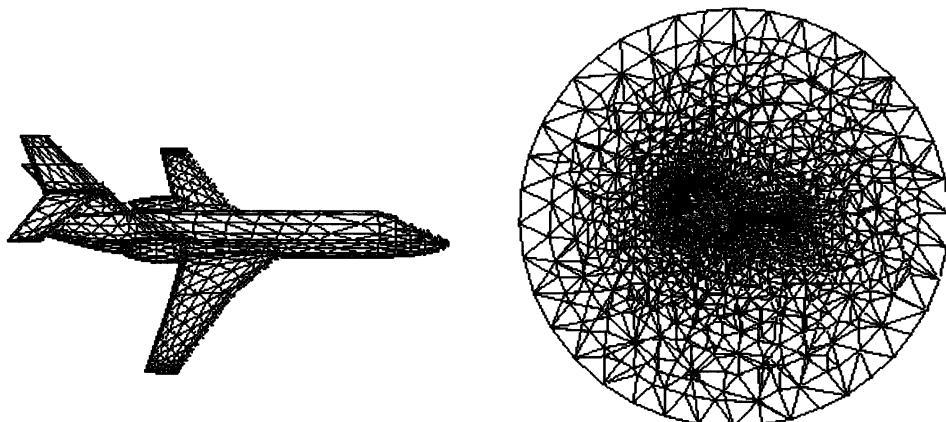


Рис. 12.43

и больше). От того, каким образом реализована эта проверка, зависит эффективность алгоритма.

Другая сложность порождена краеугольной проблемой трехмерной дискретизации. Довольно часто в ходе реализации метода образуются области, которые невозможно дискретизировать, не используя дополнительные внутренние узлы (в плоском случае любой многоугольник можно разбить на треугольники только внутренними ребрами).

Стоит также помнить, что в ходе реализации метода фронт может разбиться на несвязанные фрагменты.

Следующей особенностью метода исчерпывания является необходимость контроля над объемом и/или линейными размерами получающихся тетраэдров. В отличие от методов на основе критерия Делоне, в которых линейные размеры тетраэдров фактически определяются плотностью предварительно размещенных узлов, в методе исчерпывания размеры тетраэдров могут сильно различаться.

Появление близко расположенных тетраэдров с резко различными размерами влечет дальнейшее расхождение размеров новых тетраэдров, в результате чего обычно появляются тетраэдры плохой формы (сильно вытянутые, приплюснутые и т. п.) и низкого качества. Чтобы не допустить этой нежелательной тенденции, обычно используют специальную контрольную функцию  $f_V(x, y, z)$ , которая задает желаемый объем тетраэдра в данной точке пространства (для равномерных сеток это будет константа). При этом выбирается тот новый тетраэдр, объем которого ближе всех к значению этой функции в центре тетраэдра.

Существует еще одна сложность, связанная с выбором грани фронта для построения тетраэдра. В качестве этой грани желательно использовать грань, вершинами которой являются узлы с наименьшими зна-

чениями внутренних телесных углов. В то время как вычисление пла-парных и двугранных углов тривиально, вычисление телесных углов представляет собой более трудоемкую задачу (особенно если учесть, что каждый телесный угол в данном случае может быть сформирован различным числом граней) [256–258, 263, 268].

В рассмотренной реализации метода исчерпывания на каждой итерации из области изымается один элемент. В работе [265] предложен другой вариант реализации, в котором за один раз из области изымается сразу целый слой элементов, т. е. на каждой итерации треугольники строятся сразу для всех ребер/граней текущего фронта. Вопреки ожиданию, этот вариант не позволяет сколько-нибудь существенно ускорить процесс построения, так как все новые тетраэдры необходимо проверять на корректность, однако он избавляет от необходимости искать наиболее подходящие для построения элемента ребра или грани. Это скорее минус, чем плюс, так как из-за этой особенности описанный вариант менее надежен и может дать сбой в случае геометрически сложных областей. Вместе с тем в случае сравнительно простых областей он показывает неплохие результаты.

Сетки, построенные методом исчерпывания, как правило, обладают неплохим качеством, а последующая оптимизация положения узлов дает дополнительное повышение качества. Метод исчерпывания наиболее эффективен, если изначально задана триангуляция границы области. В этом и состоит его основное отличие от методов на основе критерия Делоне.

## 12.6. Оптимизация сеток

Под оптимизацией сетки понимают такое изменение структуры сетки и/или ее элементов, которое обеспечивает соответствие поставленной задаче и позволяет получить более точные результаты, либо получить эти результаты быстрее. Таким образом, главная цель улучшения сетки — это адаптация ее под поставленную задачу.

Увеличить точность решения можно многими путями, например:

- 1) используя аппроксимирующие функции/конечные элементы более высокого порядка (рассмотрение этого способа выходит за рамки данной главы);
- 2) оптимизируя сетку с целью улучшения аппроксимационных качеств ее элементов;
- 3) применяя локальное сгущение сетки в областях резкого изменения целевых функций или их производных.

Добиться улучшения формы элементов сетки можно двумя путями: перемещая внутренние узлы сетки; перестраивая сетку путем изменения связей между узлами. Разумеется, есть и комбинированные подходы.

Рассмотрим алгоритмы, получившие наибольшее распространение.

### 12.6.1. Оптимизация расположения узлов, или сглаживание сетки

Самый простой вариант оптимизации расположения узлов в англоязычной литературе носит название «Laplacian smoothing», т. е. «сглаживание по лапласиану», так как формула перемещения узла может быть получена из аппроксимации уравнения Лапласа конечными разностями [225]. Сущность алгоритма, разработанного в 60-е годы XX века, заключается в последовательном перемещении каждого внутреннего узла сетки в центр тяжести системы соседних (т. е. инцидентных ему) узлов. Эту процедуру продолжают итерационно определенное число шагов (обычно 5–10) либо до тех пор, пока максимальное смещение узлов не станет меньше наперед заданной величины.

Данный алгоритм с успехом применяется как в двумерном, так и в трехмерном случае (при этом он дает менее ощутимые результаты). Более сложные варианты алгоритма предполагают использование «весов» при вычислении центра тяжести. Эти «веса» соседних узлов могут зависеть, например, от длин инцидентных им ребер или объемов прилегающих тетраэдров либо определяться специальными глобальными функциями, принуждающими узлы сетки стягиваться к местам, в которых нужно обеспечить сгущение [221, 225, 228]. Существует вариант алгоритма с релаксацией [232]. В этом случае для узла  $P$  сначала находят узел  $\bar{P}$  как центр тяжести системы соседних узлов  $N_i$ :

$$\bar{P} = \frac{1}{n} \sum_{i=1}^n \rho_i N_i,$$

где  $\rho_i$  — «вес»  $i$ -го узла-соседа  $P$ . Затем определяют новое положение узла  $\tilde{P}$  с учетом параметра релаксации  $\omega$ :

$$\tilde{P} = \omega P + (1 - \omega) \bar{P}.$$

Считается, что вариант алгоритма с релаксацией более устойчив и надежен.

Основным недостатком любого подобного сглаживания является его ненадежность. Дело в том, что в некоторых случаях усреднение

положения узла может привести к появлению элементов очень плохой формы, либо вообще нарушить структуру сетки. Иначе говоря, нельзя гарантировать, что после сглаживания качество сетки будет улучшено, а не ухудшено [232]. Для борьбы с такими ситуациями используют так называемые «осторожные» алгоритмы. В этих алгоритмах перед тем, как переместить узел, сначала проверяют, не произойдет ли нарушение структуры сетки, а саму операцию переноса производят только в случае, если это приведет к улучшению качества вовлеченных в нее элементов.

Однако практика показывает, что подобные алгоритмы оптимизации весьма эффективны и с точки зрения отношения затрачиваемых усилий к получаемым результатам (улучшение качества элементов от 20 до 60%).

### 12.6.2. Оптимизация связей

Улучшить форму элементов можно за счет перестановки связей между узлами. Примером такой оптимизации является построение *триангуляции Делоне*. Как уже упоминалось ранее, любая триангуляция на плоскости может быть приведена к триангуляции Делоне с помощью серии *флипов*. Трехмерный аналог этой процедуры не столь универсален, но все равно может быть использован для улучшения качества сетки. *Трейд* позволяет менять два соседних тетраэдра на три путем вставки между ними дополнительного внутреннего ребра. Точно так же три тетраэдра, имеющие одно общее ребро, можно превратить в два, убрав это ребро. Авторы [226] развили идею трейда и предложили его вариант для случая  $n$  тетраэдров, имеющих общее ребро. В их алгоритме рассматривается множество узлов, инцидентных концам удаляемого ребра (таких узлов ровно  $n$ ). Эти  $n$  узлов представляют собой многоугольник в пространстве, который в данном случае можно рассматривать как многоугольник на плоскости. Этот многоугольник можно триангулировать числом способов, равным

$$K(n) = \frac{(2n - 4)!}{(n - 1)!(n - 2)!},$$

где  $K(n)$  — так называемое число Каталана. Так,  $K(3) = 1$ ,  $K(4) = 2$ ,  $K(5) = 5$ ,  $K(6) = 14$ ,  $K(7) = 42$  и т. д. Каждая триангуляция будет давать  $2K(n)$  тетраэдров. Таким образом, задача сводится к перебору возможных вариантов и выбору оптимального. Поскольку  $n$  обычно лежит в пределах от 3 до 6, это вполне осуществимо за реальное время.

Как показывает опыт, несмотря на определенную сложность в реализации, этот метод позволяет довольно эффективно улучшать качество сетки за счет приведения ее к триангуляции Делоне с ограничениями.

Резюмируя сказанное, заметим, что все алгоритмы оптимизации сетки не конкурируют, а взаимно дополняют друг друга. Согласно комплексному исследованию различных методов оптимизации сеток [227], алгоритмы оптимизации связей позволяют улучшать общее качество сетки, а алгоритмы сглаживания позволяют эффективно избавляться от слишком больших и слишком маленьких двугранных и телесных углов. Наилучшего результата можно добиться при последовательном применении обоих алгоритмов.

### 12.6.3. Сгущение сетки

Под сгущением сетки понимают локальное измельчение ее элементов в местах резкого изменения целевых функций или их производных. Такая операция может существенно улучшить точность решения, поскольку ошибка аппроксимации явным образом зависит от размеров конечных элементов.

Заметим, что существует ряд алгоритмов, которые позволяют строить сетки сразу со сгущением (quadtree/octree, алгоритм построения триангуляции Делоне [276] и др.), однако в общем случае сгущение сетки представляет собой отдельную процедуру, проводимую уже после ее построения.

Сгущение может осуществляться двумя путями: дроблением элементов либо добавлением в сетку новых узлов. При этом необходимо согласовать в сетке вновь созданные элементы со старыми, не измененными элементами, граничащими с ними (далее мы будем называть эти элементы переходными).

Самый простой способ такого согласования — дополнительное разбиение переходных элементов. Центральный треугольник дробится на четыре ему подобных, а в граничащие с ним элементы вставляются дополнительные ребра (в трехмерном случае ситуация совершенно аналогичная с тем различием, что дробятся не только ребра, но и грани). У этого способа есть существенный недостаток. Если продолжать измельчать центральный треугольник, это приведет к появлению вытянутых переходных элементов с очень плохими аппроксимационными свойствами.

Один из способов борьбы с этим нежелательным явлением предложен в работах [270, 271]. Этот способ назван бисекцией длиннейшего

ребра. Как следует из названия метода, его суть заключается в том, что на каждой итерации выбирается самое длинное ребро из области триангуляции, подлежащей сгущению, которое затем разбивается пополам вставкой дополнительного узла. Треугольники, которым принадлежит это ребро, также разбиваются пополам вставкой внутренних ребер. При этом алгоритм продолжает действовать до тех пор, пока все элементы не будут удовлетворять заданной функции распределения размеров.

Согласно описываемым исследованиям, в двумерном случае при использовании алгоритма бисекции минимальный угол уменьшается не более чем вдвое. Это вполне допустимо, если учитывать предельную простоту реализации алгоритма. Алгоритм бисекции можно использовать и в трехмерном случае. Точно так же выбирается самое длинное ребро из всех элементов, размеры которых превышают допустимые, и все тетраэдры, содержащие это ребро, разбиваются на два [270, 271].

Оценка ухудшения качества трехмерной сетки в результате применения алгоритма бисекции до сих пор не получена. Это связано в первую очередь с принципиально иным способом оценки качества тетраэдров по сравнению с оценкой качества треугольников. Совершенно очевидно, что в этом случае может иметь место ухудшение качества, т. е.  $A_X$ , более чем на 50% (в отличие от двумерного случая) [254]. Еще одним недостатком алгоритма является использование в общем случае большого числа бисекций для достижения желаемого распределения размеров.

Принципиально иной алгоритм локального сгущения предложен в [263]. Этот алгоритм основан на методе построения триангуляции Делоне, причем без ограничений. Идея его удивительно проста и изящна. Фактически в область, в которой необходимо обеспечить сгущение сетки, по одному добавляются дополнительные узлы, и на каждом таком шаге производится локальная перестройка сетки до триангуляции Делоне. Этот алгоритм эффективен и позволяет избежать значительных потерь качества при сгущении сеток [228, 233].

Все описанные алгоритмы сгущения в равной степени подходят как к структурированным, так и к неструктурированным сеткам. Свойство структурированности, однако, открывает дополнительные возможности в осуществлении сгущения. Если сетка строится по определенному шаблону, то можно особым образом подобрать и шаблон для осуществления локального сгущения.

## 13. МНОГОСЕТОЧНЫЕ МЕТОДЫ

Изложены основы *многосеточных методов* решения систем линейных алгебраических уравнений (СЛАУ) высокого порядка, появившихся при аппроксимации краевых задач математической физики. Кратко рассмотрены теоретические основы многосеточных методов и основные направления их развития.

### 13.1. Проблема решения больших сеточных задач

Возможности аналитических методов ограничены простейшими случаями, поэтому численные методы зачастую являются единственным способом отыскания приближенного решения краевых задач, сходимых в конечном итоге к системе линейных алгебраических уравнений (СЛАУ) высокого порядка с *плохо обусловленной матрицей* коэффициентов. Первоначально для решения подобных СЛАУ применялись как *прямые*, так и *итерационные методы*. Однако прямые методы более чувствительны к погрешностям округления и для их реализации требуется большой объем оперативной памяти и вычислительной работы. Поэтому использование итерационных методов оказалось эффективнее для решения разностных краевых задач на достаточно мелких вычислительных сетках.

Теоретический анализ показывает, что вычислительные усилия при численном решении некоторой краевой задачи на сетке, состоящей из  $N$  узлов, некоторым прямым или итерационным методом составят  $O(N^\alpha \lg^\beta N)$  арифметических операций при  $\alpha \geq 1$ ,  $\beta \geq 0$ . Поскольку сетка может состоять из нескольких миллионов узлов ( $N > 10^6$ ), разработке эффективных итерационных методов решения СЛАУ уделяется много внимания. Результаты асимптотического анализа различных алгоритмов приведены в [110].

Долгое время оставался нерешенным вопрос о возможности построения итерационного алгоритма, который позволял бы получать численное решение разностных краевых задач с минимальными вычислительными усилиями, т.е. с  $\alpha = 1$ ,  $\beta = 0$ . Для разработки подобного алгоритма понадобилась принципиально новая идея, которую предложил Р.П. Федоренко. Им опубликованы две работы, в которых использовалась высокая сходимость некоторых итерационных методов

для высокочастотных гармоник [179, 180]. В 1966 г. Н.С. Бахвалов доказал оптимальность метода по числу арифметических операций для достижения точности, согласованной с порядком аппроксимации [13]. Статья Г.П. Астраханцева [7], опубликованная в 1971 г., завершает начальный период в истории многосеточных методов.

Первые работы по многосеточным методам не привлекли широкого внимания. Спустя несколько лет А. Брандт опубликовал статьи [204, 205], после выхода которых метод получил признание, и количество публикаций по данной тематике стало стремительно расти.

Как оказалось, метод Зейделя ( $\alpha = 2, \beta = 1$ ) может быть применен для решения больших сеточных задач несмотря на результат анализа асимптотической скорости его сходимости [110]. Если решать разностную краевую задачу на нескольких сетках, то теоретически можно достичь оптимальной (неулучшаемой) скорости сходимости, выполняя  $O(N)$  арифметических операций.

К настоящему времени опубликовано несколько монографий по многосеточным методам, которые носят справочно-обзорный характер из-за огромного числа вариантов классических многосеточных методов. Весьма обширным источником информации является сайт [www.mgnet.org](http://www.mgnet.org).

## 13.2. Основы многосеточных методов

С прикладной точки зрения *методы Якоби и Зейделя* обладают рядом положительных качеств: позволяют решать разностные краевые задачи без хранения матрицы системы, не содержат проблемно-зависимых компонентов, легко программируются и распараллеливаются. Однако эти качества перечеркиваются их единственным недостатком — крайне медленной скоростью сходимости.

Долгое время считалось, что методы Якоби и Зейделя непригодны для решения разностных уравнений на мелких сетках. Однако они обладают интересной особенностью: на первых итерациях погрешность приближений к решению убывает гораздо быстрее, чем на последующих.

В качестве примера рассмотрим следующую задачу Дирихле в области  $G = (0, 1)$ :

$$\begin{aligned} \frac{d^2u}{dx^2} &= 8, \\ u(0) &= u(1) = 1. \end{aligned}$$

Точное решение  $u$  данной краевой задачи имеет вид

$$u(x) = 1 - 4x + 4x^2.$$

Заменяя обычную производную второго порядка  $u''$  разностной производной, получаем следующую систему линейных уравнений:

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = 8, \quad i = 2, 3, \dots, N,$$

$$y_1 = 1, \quad y_{N+1} = 1,$$

где  $y_i$  — значение приближенного решения в точке  $x_i = ih; h = 1/N$  — шаг сетки. Определим погрешность  $\mathcal{E}$  приближений к решению в виде

$$\mathcal{E} = \max_i |y_i - u_h(x_i)|.$$

Выполним несколько итераций метода Зейделя, начиная с нулевого начального приближения  $y_i^{(0)} = 0$  на сетках с  $N = 11, 21, \dots, 101$ . Типичное изменение погрешности приближений к решению краевой задачи в процессе выполнения первых итераций метода Зейделя показано на рис. 13.1. Видно, что эффект высокой скорости убывания погрешности на первых итерациях проявляется тем сильнее, чем меньше  $N$ .

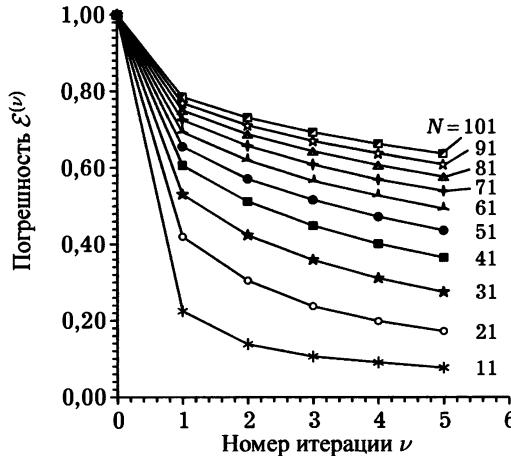


Рис. 13.1

Заметим, что условия вычислительного эксперимента специально подобраны таким образом, чтобы эффект был более наглядным (наибольшая погрешность начального приближения локализована вблизи границ отрезка). Тщательный теоретический анализ показал сложный характер сходимости метода Зейделя: разлагая погрешность в ряд Фурье, нетрудно убедиться, что метод Зейделя сходится неравномерно

на различных гармониках [14, 137, 176, 291]. При этом высокочастотные гармоники удаляются эффективно в течение первых итераций, а для удаления низкочастотных гармоник требуется большее число итераций. Поэтому зависимость необходимого числа итераций метода Зейделя от шага сетки  $\nu \sim h^{-2}$  справедлива при достаточно больших значениях  $\nu$ . Впервые на быструю сходимость метода Зейделя на первых итерациях обратил внимание Р.П. Федоренко и предложил использовать ее для создания высокоскоростных методов решения разностных краевых задач [179, 180].

В литературе по многосеточным методам итерационные алгоритмы, которые эффективно удаляют высокочастотные гармоники, называются *сглаживателями*. Такие алгоритмы в сочетании с выбранным способом упорядочивания неизвестных называются *сглаживающими процедурами*. Итерации сглаживающей процедуры называются *сглаживающими итерациями* или *сглаживанием*. Аналитический аппарат, используемый для оценки сглаживающих свойств того или иного итерационного метода, приводится в [291].

Прежде чем приступать к возможным вариантам алгоритмизации основной идеи Р.П. Федоренко, следует заметить, что задача о решении системы линейных алгебраических уравнений (СЛАУ)  $Ay = b$  итерационными методами допускает эквивалентную формулировку. Пусть выбрано некоторое начальное приближение  $y^{(0)}$  к решению СЛАУ  $Ay = b$ , подставляя которое в исходную систему, получаем

$$Ay^{(0)} = b - r^{(0)},$$

где  $r^{(0)}$  — вектор невязки  $r$ , соответствующий начальному приближению  $y^{(0)}$ . Добавим к начальному приближению  $y^{(0)}$  некоторую *поправку*  $c$ , такую, чтобы обнулить вектор *невязки*  $r^{(0)}$ :

$$A(y^{(0)} + c) = b.$$

Нетрудно видеть, что поправка  $c$  является разностью между искомым решением и начальным приближением, т. е.  $c = y - y^{(0)}$ . Тогда исходную задачу можно переформулировать в терминах поправки и невязки следующим образом:

$$Ac = r^{(0)} = b - Ay^{(0)},$$

т. е. отыскивается не само решение  $y$ , а разность между искомым решением  $y$  и некоторым приближением  $y^{(0)}$  к нему.

Системы линейных алгебраических уравнений  $Ay = b$  и  $Ac = r^{(0)}$  совершенно эквивалентны, поскольку различаются лишь правыми частями, да и то при  $y^{(0)} \neq 0$ . Тем не менее эквивалентная формулировка

обладает определенными преимуществами при использовании итерационных методов.

Рассмотрим итерационный метод вида

$$Ac^{(q)} = b - Ay^{(q-1)} = r^{(q-1)},$$

где  $q$  — номер итерации. Здесь поправка  $c^{(q)}$  является разностью между искомым решением и текущим приближением, т. е.  $c^{(q)} = u - y^{(q-1)}$ . После проведения каждой итерации необходимо пересчитать приближение к решению

$$y^{(q)} := c^{(q)} + y^{(q-1)}$$

и обнулить поправку

$$c^{(q+1)} := 0.$$

Здесь знак двоеточия с равенством означает присвоение правой части левой. Заметим, что поправка стремится к нулю по мере сходимости итераций:

$$q \rightarrow +\infty \Rightarrow y^{(q)} \rightarrow y \Rightarrow c^{(q)} \rightarrow 0.$$

Данное обстоятельство предоставляет некоторую свободу действий при вычислении поправки  $c$ , которую необходимо использовать для уменьшения общего объема вычислений, поскольку решение систем

$$Ac^{(q)} = r^{(q-1)}$$

и

$$Ay = b$$

требует одинаковых усилий. В многосеточных методах помимо СЛАУ  $Ac^{(q)} = r^{(q-1)}$  применяется вспомогательная система

$$\tilde{A}\tilde{c}^{(q)} = \tilde{r}^{(q-1)},$$

решение которой  $\tilde{c}^{(q)}$  близко к искомой поправке  $c^{(q)}$ , но для его отыскания нужно выполнить меньшую вычислительную работу. Поправка  $\tilde{c}^{(q)}$  вычисляется на специально построенных подсетках самой мелкой сетки в целях использования быстрой сходимости метода Зейделя на первых итерациях.

**Классические многосеточные методы** (КММ) являются первой попыткой алгоритмизации идеи Р.П. Федоренко для создания высокоэффективных алгоритмов. Сформулируем простейший (двухуровневый) вариант такого метода в абстрактной форме на примере краевой задачи

$$\mathfrak{L}u = -f, \quad u \Big|_{\partial G} = \vartheta,$$

где  $\mathfrak{L}$  — линейный эллиптический оператор; область  $G$  —  $n$ -мерный куб.

Предположим, что построена вычислительная сетка  $G_h^0$  в области  $G$ . Сетка  $G_h^0$  образует **нулевой сеточный уровень**, т. е. нулевым считается уровень с самой мелкой сеткой. В данном случае порядок нумерации уровней не имеет принципиального значения. Все, что связано с сеткой  $G_h^0$ , будет иметь верхний индекс 0. Запишем разностный аналог исходной краевой задачи на сетке  $G_h^0$  в матричной форме:

$$A^0 y^0 = b^0. \quad (13.1)$$

Выполним несколько сглаживающих итераций на сетке  $G_h^0$ . Обозначим через  $\hat{y}^0$  полученное приближение к решению  $y^0$  системы (13.1), которое удовлетворяет соотношению

$$A^0 \hat{y}^0 = b^0 - r^0,$$

где  $r^0$  — вектор невязки. Перепишем теперь СЛАУ (13.1) в терминах «поправка» — «невязка». Для этого обозначим через  $c^0$  разность между искомым решением  $y^0$  и полученным приближением  $\hat{y}^0$ , т. е.  $c^0 = y^0 - \hat{y}^0$ . Подставляя  $y^0 = \hat{y}^0 + c^0$  в (13.1), получим

$$A^0 y^0 = A^0(\hat{y}^0 + c^0) = b^0 \Rightarrow A^0 c^0 = b^0 - A^0 \hat{y}^0 = r^0. \quad (13.2)$$

Ранее уже отмечалось, что решение СЛАУ  $A^0 c^0 = r^0$  требует тех же вычислительных усилий, что и решение системы (13.1). Чтобы сократить объем вычислений, построим дополнительную сетку  $G_h^1$ , которая имеет меньшее число узлов, чем сетка  $G_h^0$ . Будем считать, что сетка  $G_h^1$  образует **первый сеточный уровень**, и все величины, связанные с сеткой  $G_h^1$ , будут иметь верхний индекс 1. **Сетки** типа  $G_h^1$  называются **грубыми**, а сетка  $G_h^0$  — самой **мелкой**.

Теперь необходимо на сетке  $G_h^1$  сконструировать задачу, решение которой будет близко к поправке  $c^0$ , удовлетворяющей системе (13.2). Основная трудность состоит в том, что сетки  $G_h^0$  и  $G_h^1$  имеют разное число узлов и  $h^0 < h^1$ . Пусть  $y^1$  — сеточная функция, определенная на сетке  $G_h^1$ , которая аппроксимирует поправку  $c^0$  на сетке  $G_h^0$ . Положим, что аналог системы (13.2) на сетке  $G_h^1$  может быть записан в виде

$$A^1 y^1 = b^1. \quad (13.3)$$

Матрица  $A^0$  в (13.1) получена в результате аппроксимации оператора  $\mathfrak{L}$  на сетке  $G_h^0$ , поэтому можно принять, что матрица  $A^1$  может быть получена в результате аппроксимации того же оператора  $\mathfrak{L}$  на сетке  $G_h^1$ . Выбор правой части системы уравнений (13.3) является нетривиальным: нельзя положить  $b^1 = r^0$ , поскольку вектор  $b^1$  имеет меньшее число компонент, чем вектор  $r^0$ . Поэтому для *проектирования* вектора

невязки  $r^0$  с сетки  $G_h^0$  на сетку  $G_h^1$  используют специальный *оператор сужения*  $\mathbb{R}$ , т. е.

$$b^1 = \mathbb{R}r^0.$$

Тогда СЛАУ (13.3) принимает вид

$$A^1 y^1 = \mathbb{R}r^0. \quad (13.4)$$

Решение системы (13.4) требует гораздо меньшего объема вычислений, чем решение системы (13.2), потому что сетка  $G_h^1$  имеет меньшее число узлов, чем сетка  $G_h^0$ .

Предположим, что система уравнений (13.4) решена и вектор  $y^1$  найден. Теперь необходимо пролонгировать вектор  $y^1$  с сетки  $G_h^1$  на сетку  $G_h^0$ . Опять же нельзя положить, что

$$c^0 = y^1,$$

поскольку вектор  $y^1$  имеет меньшее число компонент, чем вектор  $c^0$ . Поэтому для пролонгирования вектора  $\hat{y}^1$  с сетки  $G_h^1$  на сетку  $G_h^0$  используют специальный *оператор пролонгации*  $\mathbb{P}$ , т. е. полагают, что

$$c^0 = \mathbb{P}y^1.$$

Далее необходимо пересчитать полученное ранее на сетке  $G_h^0$  приближение к решению системы  $y^0$ , добавляя к нему поправку  $y^1$ , пролонгированную на сетку  $G_h^0$  с сетки  $G_h^1$ :

$$\hat{y}^0 := \hat{y}^0 + \mathbb{P}y^1$$

(равенство означает присвоение правой величины левой).

Наличие в многосеточном алгоритме операторов сужения и пролонгации неизбежно приводит к различным вычислительным погрешностям. Поэтому после пересчета приближения к решению  $\hat{y}^0$  необходимо выполнить еще несколько итераций Зейделя для удаления высокочастотных гармоник. На этом многосеточная итерация считается завершенной. Если критерий останова не достигнут, то вычисления повторяются в приведенном выше порядке до тех пор, пока не будет достигнут заданный критерий останова.

### 13.3. Классические многосеточные методы

*Классические многосеточные методы* (КММ) отличаются сложностью организации вычислений, поэтому основные положения этих методов демонстрируют на примере одномерных задач. Подобные

задачи не требуют применения многосеточных методов, однако в одномерном случае основные идеи и способы их реализации становятся наиболее наглядными [235, 291]. Далее подробно рассматривается трехуровневый алгоритм и приводится краткий обзор основных направлений развития КММ.

### 13.3.1. Пример одномерной задачи

Рассмотрим одномерную задачу Дирихле

$$\begin{aligned} u'' &= -f(x), \quad 0 < x < 1, \\ u(0) &= u_0, \quad u(1) = u_1 \end{aligned} \quad (13.5)$$

для иллюстрации основных компонентов КММ. Построим вычислительную сетку  $G_h^0$  для конечно-разностной аппроксимации краевой задачи (13.5) посредством разбиения единичного отрезка на  $N_x^0 = 2^m$  частей. Узлы  $x_i^0$  сетки  $G_h^0$  задаются соотношениями

$$x_i^0 = (i-1)h^0, \quad i = 1, 2, \dots, N_x^0 + 1,$$

где  $h^0 = 1/N_x^0 = 2^{-m}$  — шаг сетки. Вычислительная сетка для параметра  $m = 3$  (*мелкая* и две *грубые сетки*) показана на рис. 13.2. Все, что связано с сеткой  $G_h^0$ , будет иметь верхний индекс 0.

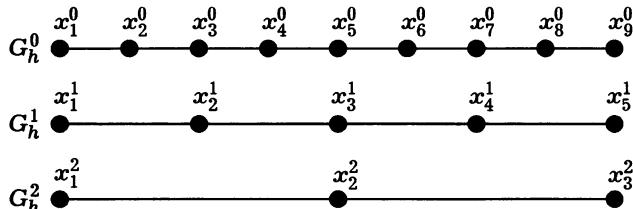


Рис. 13.2

Разностный аналог краевой задачи (13.5) на сетке  $G_h^0$  принимает вид

$$\begin{cases} y_1^0 = u(0), \\ \frac{y_{i-1}^0 - 2y_i^0 + y_{i+1}^0}{(h^0)^2} = -f_i^0, \quad i = 2, 3, \dots, N_x^0, \\ y_{N_x^0+1}^0 = u(1). \end{cases} \quad (13.6)$$

Предположим, что выбрано некоторое начальное приближение к решению системы алгебраических уравнений (13.6). Многосеточная итерация (рис. 13.3), может быть проведена в следующем порядке.

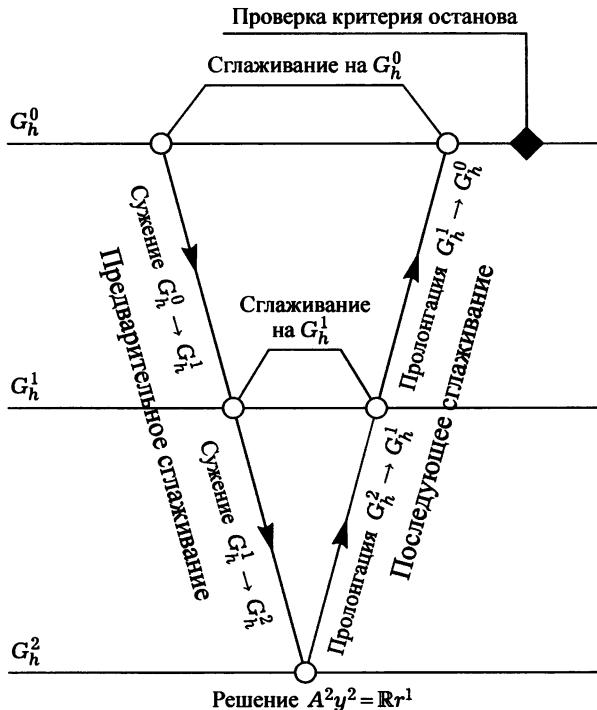


Рис. 13.3

**1. Сглаживание на сетке  $G_h^0$ .** Оно заключается в выполнении нескольких итераций симметричного варианта *метода Зейделя*:

$$\begin{cases} (\hat{y}^0)_i = \frac{1}{2} \left( (\hat{y}^0)_{i-1} + (y^0)_{i+1}^{(\nu-1)} + (h^0)^2 f_i^0 \right), & i = 2, 3, \dots, N_x^0, \\ (y^0)_i^{(\nu)} = \frac{1}{2} \left( (\hat{y}^0)_{i-1} + (y^0)_{i+1}^{(\nu)} + (h^0)^2 f_i^0 \right), & i = N_x^0, N_x^0 - 1, \dots, 2, \end{cases} \quad (13.7)$$

в процессе которых эффективно удаляются высокочастотные компоненты погрешности.

**2. Сужение  $G_h^0 \rightarrow G_h^1$ .** Обозначим через  $\hat{y}^0$  приближение к решению системы (13.6), полученное после выполнения нескольких итераций симметричного варианта метода Зейделя (13.7) на сетке  $G_h^0$ . Подставляя  $\hat{y}^0$  в (13.6), получаем

$$\begin{cases} \hat{y}_1^0 = u(0), \\ \frac{\hat{y}_{i-1}^0 - 2\hat{y}_i^0 + \hat{y}_{i+1}^0}{(h^0)^2} = -f_i^0 - r_i^0, & i = 2, 3, \dots, N_x^0, \\ \hat{y}_{N_x^0+1}^0 = u(1), \end{cases} \quad (13.8)$$

где  $r_i^0$  — компоненты вектора *невязки*. Согласно **13.2**, перепишем (13.8) в терминах «поправка» — «невязка». Для этого представим искомое решение разностной краевой задачи (13.6) в виде суммы полученного приближения  $\hat{y}^0$  и *поправки*  $c^0$ :

$$y_i^0 = \hat{y}_i^0 + c_i^0, \quad i = 1, 2, \dots, N_x^0 + 1. \quad (13.9)$$

Поправку  $c_i^0$  в (13.9) выберем из условия  $r_i^0 = 0$ , т. е. обнулим невязку в (13.8). Подстановка (13.9) в (13.8) приводит к следующей системе:

$$\begin{cases} c_1^0 + \hat{y}_1^0 = u(0), \\ \frac{(c_{i-1}^0 + \hat{y}_{i-1}^0) - 2(c_i^0 + \hat{y}_i^0) + (c_{i+1}^0 + \hat{y}_{i+1}^0)}{(h^0)^2} = -f_i^0, \quad i = 2, 3, \dots, N_x^0, \\ c_{N_x^0+1}^0 + \hat{y}_{N_x^0+1}^0 = u(1). \end{cases} \quad (13.10)$$

Сравнивая первые уравнения систем (13.8) и (13.10), нетрудно видеть, что  $c_1^0 = 0$ . Аналогично из последних уравнений систем (13.8) и (13.10) следует, что  $c_{N_x^0+1}^0 = 0$ . Тогда разностная краевая задача (13.8) в терминах «поправка» — «невязка» принимает вид

$$\begin{cases} c_1^0 = 0, \\ \frac{c_{i-1}^0 - 2c_i^0 + c_{i+1}^0}{(h^0)^2} = r_i^0 = -f_i^0 - \frac{\hat{y}_{i-1}^0 - 2\hat{y}_i^0 + \hat{y}_{i+1}^0}{(h^0)^2}, \quad i = 2, 3, \dots, N_x^0, \\ c_{N_x^0+1}^0 = 0. \end{cases} \quad (13.11)$$

Как уже упоминалось в **13.2**, решение системы линейных алгебраических уравнений (СЛАУ) (13.11) требует того же объема вычислений, что и решение исходной системы (13.6). Чтобы использовать быструю сходимость метода Зейделя на первых итерациях, построим вспомогательную грубую сетку  $G_h^1$  с шагом  $h^1 = 2h^0$  и количеством узлов  $N_x^1 = N_x^0/2 = 2^{m-1}$ . Возможны разные варианты построения вспомогательных сеток, одним из которых является удаление узлов с четными номерами из сетки  $G_h^0$ . Сетка  $G_h^1$  с  $N_x^1 = 4$  показана на рис. 13.2. Как и в **13.2**, все, что связано с сеткой  $G_h^1$ , будет иметь верхний индекс 1. Выбор числа  $N_x^0$  в виде  $2^m$  гарантирует, что граничные узлы сетки  $G_h^1$  будут расположены на границах единичного отрезка, т. е.  $x_1^1 = 0$  и  $x_{N_x^1+1}^1 = 1$ . Поэтому граничные условия Дирихле в данном случае аппроксимируются точно не только на сетке  $G_h^0$ , но и на сетке  $G_h^1$ .

Зададим на сетке  $G_h^1$  СЛАУ, решение  $y_i^1$  которой будет аппроксимировать поправку  $c_i^0$ , являющуюся, в свою очередь, решением (13.11).

Поскольку левая часть (13.11) получена в результате конечно-разностной аппроксимации одномерного оператора Лапласа на сетке  $G_h^0$ , то можно предложить следующий вид системы уравнений, связанной с сеткой  $G_h^1$ :

$$\begin{cases} y_1^1 = 0, \\ \frac{y_1^1 - 2y_2^1 + y_3^1}{(h^1)^2} = b_2^1, \\ \frac{y_2^1 - 2y_3^1 + y_4^1}{(h^1)^2} = b_3^1, \\ \frac{y_3^1 - 2y_4^1 + y_5^1}{(h^1)^2} = b_4^1, \\ y_5^1 = 0, \end{cases} \quad (13.12)$$

т. е. левая часть уравнений 2–4 системы (13.12) получена в результате конечно-разностной аппроксимации одномерного оператора Лапласа во внутренних узлах  $x_2^1$ ,  $x_3^1$  и  $x_4^1$  сетки  $G_h^1$ . Система (13.12) содержит только три неизвестных:  $y_2^1$ ,  $y_3^1$  и  $y_4^1$ , в то время как исходная система (13.11) содержит семь неизвестных:  $c_2^0$ ,  $c_3^0$ , …,  $c_8^0$ .

Теперь необходимо определить правую часть в уравнениях 2–4 системы (13.12), т. е. некоторым образом определить  $b_2^1$ ,  $b_3^1$  и  $b_4^1$  через компоненты вектора невязки  $r_2^0$ ,  $r_3^0$ , …,  $r_8^0$ . Простейшим способом задания  $b_2^1$ ,  $b_3^1$  и  $b_4^1$  является присвоение значений  $r_i^0$  в совпадающих узлах сеток  $G_h^0$  и  $G_h^1$ , т. е.

$$b_2^1 = r_3^0, \quad b_3^1 = r_5^0, \quad b_4^1 = r_7^0.$$

Однако в [235] показано, что подобное задание правой части в (13.12) не обеспечивает требуемой точности, если сеточная функция  $r_i^0$  сильно меняется на первых многосеточных итерациях. Поэтому обычно используют осреднение типа

$$b_2^1 = \frac{r_2^0 + 2r_3^0 + r_4^0}{4}, \quad b_3^1 = \frac{r_4^0 + 2r_5^0 + r_6^0}{4}, \quad b_4^1 = \frac{r_6^0 + 2r_7^0 + r_8^0}{4}, \quad (13.13)$$

как показано на рис. 13.4 (в одномерном случае). Подобное усреднение в КММ называется **оператором сужения**  $\mathbb{R}$ , и уравнения (13.13) условно записывают в виде

$$b^1 = \mathbb{R}r^0.$$

**3. Сглаживание на сетке  $G_h^1$ .** Оно состоит в выполнении нескольких итераций симметричного варианта метода Зейделя, которые

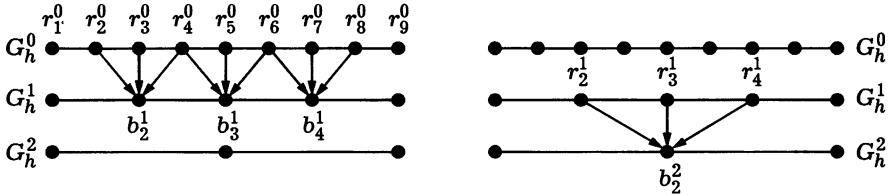


Рис. 13.4

применительно к системе (13.12) с правой частью (13.13) определяются как

$$\begin{cases} (\hat{y}^1)_i = \frac{1}{2} \left( (\hat{y}^1)_{i-1} + (y^1)_{i+1}^{(\nu-1)} - (h^1)^2 b_i^1 \right), & i = 2, 3, \dots, N_x^1, \\ (y^1)_i^{(\nu)} = \frac{1}{2} \left( (\hat{y}^1)_{i-1} + (y^1)_{i+1}^{(\nu)} - (h^1)^2 b_i^1 \right), & i = N_x^1, N_x^1 - 1, \dots, 2. \end{cases} \quad (13.14)$$

По мере сходимости многосеточного алгоритма  $y^1 \rightarrow 0$ , поэтому итерации (13.14) следует начинать с начального приближения  $(y^1)^{(0)} = 0$ . В процессе итераций (13.14) также удаляются высокочастотные компоненты погрешности.

**4. Сужение  $G_h^1 \rightarrow G_h^2$ .** Может оказаться, что сетка  $G_h^1$  является слишком мелкой, и на ней не удастся получить достаточно точную аппроксимацию поправки  $c^0$  за несколько итераций метода Зейделя. Поэтому необходимо построить еще одну грубую сетку  $G_h^2$  с шагом  $h^2 = 2h^1 = 4h^0$  посредством удаления узлов с четными номерами из сетки  $G_h^1$  (см. рис. 13.2).

Обозначим через  $\hat{y}^1$  приближение к решению системы (13.12) с правой частью (13.13), полученное после выполнения нескольких итераций симметричного варианта метода Зейделя (13.14) на сетке  $G_h^1$ . Подставляя  $\hat{y}^1$  в (13.12), получаем

$$\begin{cases} \hat{y}_1^1 = 0, \\ \frac{\hat{y}_{i-1}^1 - 2\hat{y}_i^1 + \hat{y}_{i+1}^1}{(h^1)^2} = b_i^1 - r_i^1, & i = 2, 3, \dots, N_x^1, \\ \hat{y}_5^1 = 0. \end{cases} \quad (13.15)$$

По аналогии с сужением  $G_h^0 \rightarrow G_h^1$ , перепишем (13.15) в терминах «поправка» — «невязка». Для этого представим искомое решение задачи (13.15) в виде суммы полученного приближения  $\hat{y}^1$  и поправки  $c^1$ :

$$y_i^1 = \hat{y}_i^1 + c_i^1, \quad i = 1, 2, \dots, N_x^1 + 1. \quad (13.16)$$

Сеточная функция  $c^1$  является поправкой к приближению  $\hat{y}^1$ , которое аппроксимирует поправку  $c^0$ . Подстановка (13.16) в (13.15) приводит к следующей системе:

$$\begin{cases} c_1^1 = 0, \\ \frac{c_{i-1}^1 - 2c_i^1 + c_{i+1}^1}{(h^1)^2} = r_i^1 = b_i^1 - \frac{\hat{y}_{i-1}^1 - 2\hat{y}_i^1 + \hat{y}_{i+1}^1}{(h^1)^2}, & i = 2, 3, \dots, N_x^1, \\ c_5^1 = 0. \end{cases} \quad (13.17)$$

По аналогии с переходом от (13.11) к (13.12) можно задать на сетке  $G_h^2$  следующую СЛАУ:

$$\begin{cases} y_1^2 = 0, \\ \frac{y_1^2 - 2y_2^2 + y_3^2}{(h^2)^2} = b_2^2, \\ y_3^2 = 0, \end{cases} \quad (13.18)$$

решение  $y^2$  которой будет аппроксимировать поправку  $c^1$ , являющуюся решением (13.17). В соответствии с рис. 13.4 правая часть  $b_2^2$  второго уравнения системы (13.18) может быть задана в виде

$$b_2^2 = \frac{r_2^1 + 2r_3^1 + r_4^1}{4}. \quad (13.19)$$

Решение системы (13.18) с правой частью (13.19) легко получить аналитически:

$$y_1^2 = 0, \quad y_2^2 = -\frac{(h^2)^2}{2} b_2^2 = -\frac{(h^2)^2}{2} \frac{r_2^1 + 2r_3^1 + r_4^1}{4}, \quad y_3^2 = 0. \quad (13.20)$$

На этом завершается первая половина многосеточной итерации, которая называется **предварительным сглаживанием** (см. рис. 13.3).

Вторая половина многосеточной итерации называется **последующим сглаживанием** (см. рис. 13.3).

**5. Пролонгация  $G_h^2 \rightarrow G_h^1$ .** Сеточная функция  $y^2$  (13.20), которая определена на сетке  $G_h^2$ , аппроксимирует поправку  $c^1$  на сетке  $G_h^1$ . Поскольку сетка  $G_h^2$  состоит из меньшего числа узлов, необходимо доопределить значения сеточной функции  $y^2$  в недостающих узлах посредством интерполяции. Простейшая (линейная) интерполяция позволяет получить следующую поправку  $c^1$  на сетке  $G_h^1$ :

$$c_1^1 = 0, \quad c_2^1 = \frac{1}{2}(y_1^2 + y_2^2), \quad c_3^1 = y_2^2, \quad c_4^1 = \frac{1}{2}(y_2^2 + y_3^2), \quad c_5^1 = 0. \quad (13.21)$$

Оператор, определяющий поправку на тонкой сетке по поправке на более грубой сетке последующего уровня, называется **оператором пролонгации**  $\mathbb{P}$ . При этом уравнения (13.21) условно записываются в виде

$$c^1 = \mathbb{P}y^2.$$

Одномерный оператор пролонгации  $\mathbb{P}$  схематично показан на рис. 13.5.

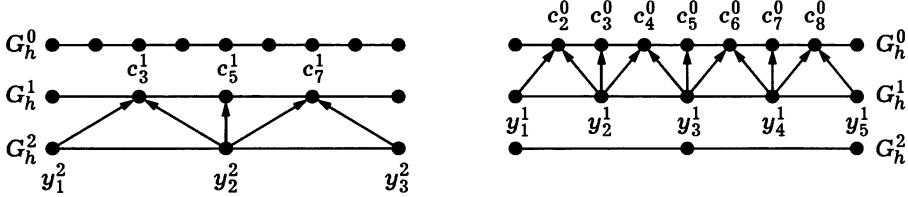


Рис. 13.5

**6. Сглаживание на сетке  $G_h^1$ .** Пролонгация неизбежно вносит вычислительную погрешность в поправку  $c^1$ . Чтобы удалить погрешность интерполяции, необходимо выполнить несколько итераций метода Зейделя (13.14). Для этого начальное приближение задают в виде

$$\hat{y}_i^1 := \hat{y}_i^1 + c_i^1, \quad i = 1, 2, \dots, N_x^1 + 1$$

(равенство означает присвоение правой величины левой), где  $\hat{y}^1$  — ранее полученное приближение к решению (13.12) на этапе предварительного сглаживания;  $c^1$  — поправка  $y^2$ , пролонгированная с сетки  $G_h^2$  на сетку  $G_h^1$ .

**7. Пролонгация  $G_h^1 \rightarrow G_h^0$ .** Сеточная функция  $y^1$  пролонгируется на сетку  $G_h^0$  аналогично тому, как функция  $y^2$  пролонгируется на сетку  $G_h^1$  (см. рис. 13.5).

**8. Сглаживание на сетке  $G_h^0$ .** Чтобы удалить погрешность интерполяции, внесенную в поправку  $c^0$ , необходимо выполнить несколько итераций метода Зейделя (13.7). Для этого начальное приближение задают в виде

$$\hat{y}_i^0 := \hat{y}_i^0 + c_i^0, \quad i = 1, 2, \dots, N_x^0 + 1$$

(равенство означает присвоение правой величины левой), где  $\hat{y}^0$  — ранее полученное приближение к решению (13.6) на этапе предварительного сглаживания;  $c^0 = \mathbb{P}y^1$  — поправка  $y^1$ , пролонгированная с сетки  $G_h^1$ .

Далее проверяется критерий останова итераций, и в случае необходимости выполняется следующая многосеточная итерация. Приведенная выше последовательность вычислений (см. рис. 13.3) напоминает

букву V, поэтому она получила название V-цикла. В общем случае, если  $N_x^0 = 2^m$ , то в КММ используют  $m - 1$  грубых сеток.

Проиллюстрируем V-цикл на примере модельной краевой задачи

$$\begin{aligned} \frac{d^2u}{dx^2} &= 10e^x, \\ u(0) &= u(1) = 0, \end{aligned} \quad (13.22)$$

точное решение которой имеет вид

$$u(x) = 10(e^x + (1 - e)x - 1). \quad (13.23)$$

Используемые сетки  $G_h^0$ ,  $G_h^1$  и  $G_h^2$  показаны на рис. 13.2. Разностная аппроксимация задачи (13.22) записывается как (13.6), где  $y_1^0 = y_{N_x^0+1}^0 = 0$  и  $f_i = -10e^{x_i^0}$ .

Точное решение разностной краевой задачи (13.6) найдем методом прогонки [79, 80, 130, 139, 149, 172, 291]. Полученное решение, которое обозначим как  $\chi$ , будет служить для демонстрации сходимости многосеточных итераций. Благодаря решению  $\chi$  краевой задачи можно точно вычислить искомые поправки  $c^0$  и  $c^1$ . Точные значения поправок обозначим  $(c^0)$  и  $(c^1)$ . Численное решение задачи (13.22) начнем с нулевого начального приближения. Рассмотрим основные этапы многосеточной итерации (рис. 13.6).

**Сглаживание на сетке  $G_h^0$ .** На рис. 13.6, а показаны приближение  $\hat{y}^0$  (•) к решению  $\chi$  (\*) и поправка  $(c^0) = \chi - \hat{y}^0$  (○) после трех итераций метода Зейделя (13.7).

**Сглаживание на сетке  $G_h^1$ .** На рис. 13.6, б показаны поправка  $(c^0)$  (\*), приближение  $\hat{y}^1$  (•) к решению системы (13.12) с правой частью (13.13), полученное после трех итераций метода Зейделя (13.14), и поправка  $(c^1) = (c^0) - \hat{y}^1$  (○), которая вычислена в совпадающих узлах сеток  $G_h^0$  и  $G_h^1$ . Видно, что полученная сеточная функция  $\hat{y}^1$  (•) достаточно хорошо аппроксимирует искомую поправку  $(c^0)$  (\*).

**Пролонгация  $G_h^2 \rightarrow G_h^1$ .** На рис. 13.6, в показаны искомая поправка  $(c^1)$  (\*) и функция  $y^2$  (•), пролонгированная на сетку  $G_h^1$  с помощью линейной интерполяции в соответствии с (13.21). На рис. 13.6, г заметно, что добавление поправки  $y^2$ , пролонгированной с сетки  $G_h^2$  на сетку  $G_h^1$ , к полученному ранее приближению  $\hat{y}^1$  (○) позволило лучше аппроксимировать искомую поправку  $(c^0)$  (\*).

**Пролонгация  $G_h^1 \rightarrow G_h^0$ .** На рис. 13.6, д показаны искомая поправка  $(c^0)$  (\*) и функция  $\hat{y}^1$  (•), пролонгированная на сетку  $G_h^0$  с помощью линейной интерполяции (см. рис. 13.5).

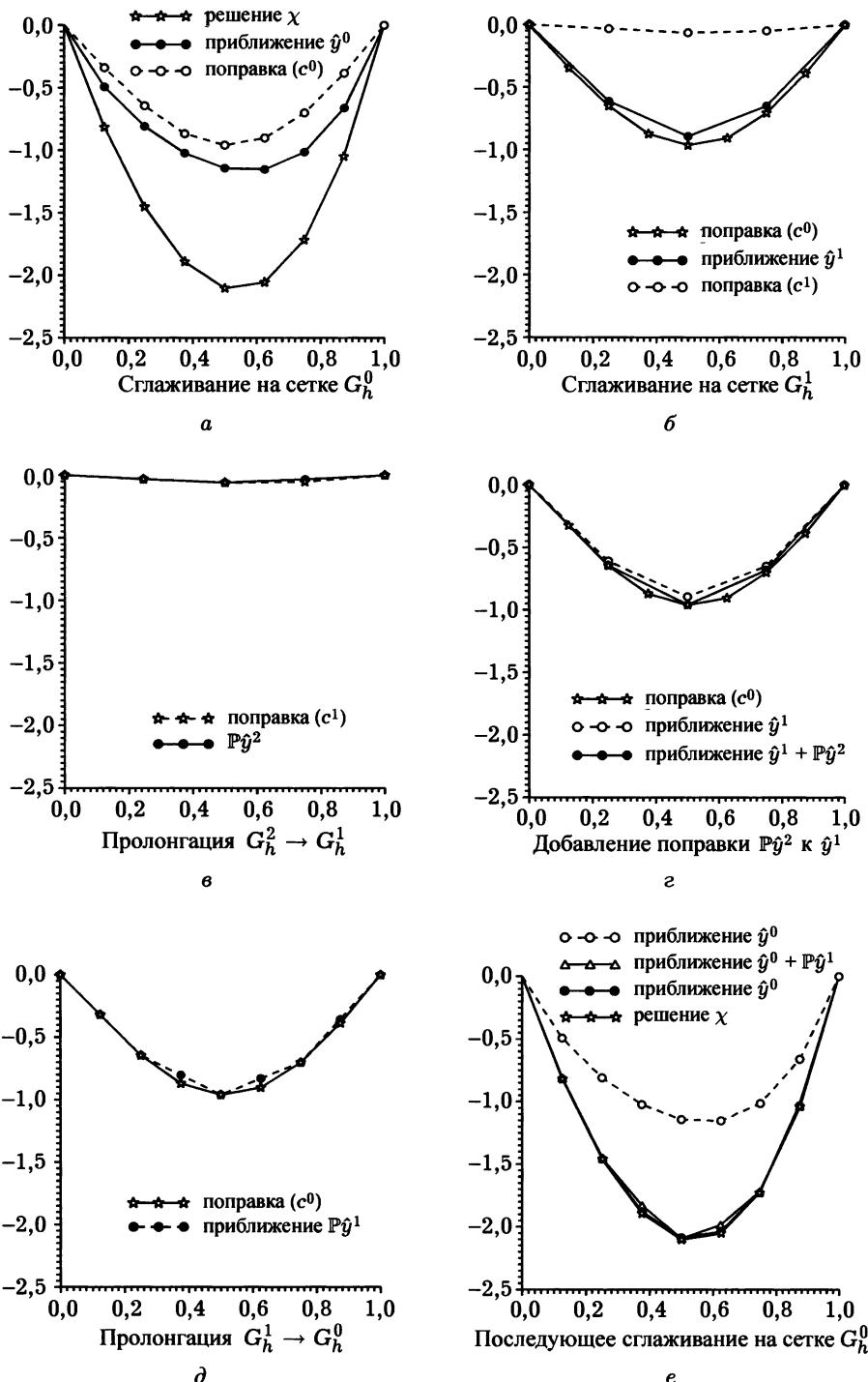


Рис. 13.6

**Сглаживание на сетке  $G_h^0$ .** На рис. 13.6, *e* показаны решение  $\chi$  ( $*$ ), полученное методом прогонки, приближение  $\hat{y}^0$  к решению СЛАУ (13.6) после предварительного сглаживания ( $\circ$ ), приближение  $\hat{y}^0 + \mathcal{P}y^1$  ( $\Delta$ ) и приближение  $\hat{y}^0$  к СЛАУ (13.6) после последующего сглаживания ( $\bullet$ ).

В КММ сглаживающие итерации на самой мелкой сетке требуют наибольших вычислительных усилий. Сравнивая приближения к решению, полученные после выполнения сглаживающих итераций на самой мелкой сетке до (см. рис. 13.6, *a*) и после (см. рис. 13.6, *e*) вычисления поправки на грубых сетках, нетрудно убедиться в высокой скорости сходимости рассмотренного варианта КММ для данной модельной задачи.

### 13.3.2. Основные направления развития КММ

Различные варианты КММ содержат проблемно-зависимые процедуры, совершенствование которых и определяет развитие многосеточных методов. Наиболее проблемно-зависимыми процедурами КММ являются интерполяция, связанная с необходимостью доопределения поправки в недостающих узлах сетки, и сглаживающая процедура.

Для иллюстрации влияния погрешностей интерполяции на точность вычисления поправки вернемся к модельной задаче (13.22). Для простоты ограничимся только первой многосеточной итерацией.

Данный вычислительный эксперимент состоит в выполнении следующих действий:

- 1) построение самой *мелкой сетки*  $G_h^0$  с  $N_x^0 = 10$ ;
- 2) построение *грубой сетки*  $G_h^1$  с  $N_x^1 = 5$ ;
- 3) аппроксимация краевой задачи на сетках  $G_h^1$  и  $G_h^0$ ;
- 4) решение разностной краевой задачи на сетке  $G_h^1$  методом прогонки;
- 5) пролонгация полученного решения с сетки  $G_h^1$  на сетку  $G_h^0$  при помощи линейной интерполяции;
- 6) выполнение трех итераций симметричного варианта метода Зейделя на самой мелкой сетке  $G_h^0$  для удаления погрешности интерполяции и высокочастотных гармоник.

Вычислительные сетки  $G_h^0$  и  $G_h^1$ , а также решение, полученное методом прогонки на сетке  $G_h^1$ , и приближение к решению, полученное на сетке  $G_h^0$  после трех итераций Зейделя, показаны на рис. 13.7, который иллюстрирует влияние погрешности интерполяции на точность вычисления поправки.

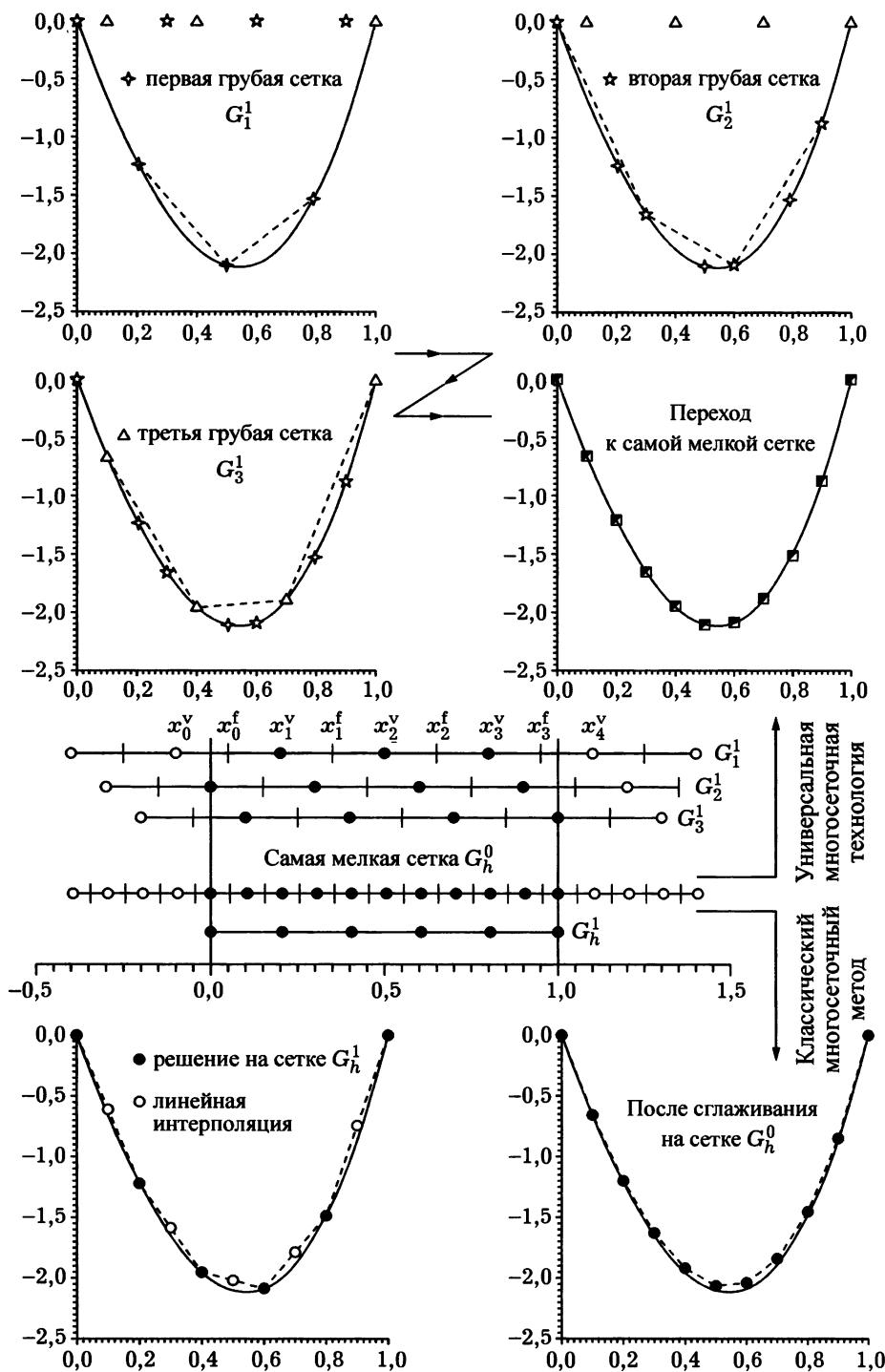


Рис. 13.7

Значения погрешности  $\mathcal{E} = \max_i |y_i - u_h(x_i)|$  решения на грубой сетке  $G_h^1$  и приближения к решению на мелкой сетке  $G_h^0$  приведены в табл. 13.1 (последний столбец, первое и второе число соответственно). Согласно полученным результатам, погрешность интерполяции приводит к возрастанию погрешности приближений к решению при переходе от грубой сетки к мелкой. При реализации КММ интерполяция вносит дополнительную погрешность в вычисленную поправку, а сглаживающая процедура удаляет погрешность интерполяции и высокочастотные гармоники. Баланс этих процедур определяет скорость сходимости КММ.

Таблица 13.1

УМТ		КММ	
Сетка	Погрешность $\mathcal{E}$	Сетка	Погрешность $\mathcal{E}$
1	$2,17 \cdot 10^{-2}$	1	$3,48 \cdot 10^{-3}$
2	$4,49 \cdot 10^{-3}$	—	—
3	$9,41 \cdot 10^{-3}$	—	—
1	$3,84 \cdot 10^{-3}$	1	$5,03 \cdot 10^{-2}$

*Примечание.* УМТ — универсальная многосеточная технология.

Одно из основных направлений развития КММ — уменьшение погрешности интерполяции. Чтобы этого добиться, можно, во-первых, использовать более точную интерполяцию (многочлены высокого порядка, сплайны и т. д.), однако это помогает не во всех случаях. Например, если коэффициенты уравнения не являются гладкими функциями, то интерполяция многочленами или сплайнами через точки (линии или поверхности) разрыва коэффициентов приводит к слишком большой погрешности и, как следствие, к медленной скорости сходимости. Решение краевых задач с разрывными коэффициентами с использованием КММ требует применения специальных операторов пролонгации [291].

В многомерном случае иногда практикуется построение дополнительных сеточных уровней посредством последовательного увеличения шага сетки в каждом пространственном направлении для упрощения интерполяции и уменьшения ее погрешности.

Во-вторых, уменьшить погрешность интерполяции можно путем воздействия на свойства интерполируемой функции (поправки). Поскольку по мере сходимости многосеточных итераций поправка стремится к нулю, то к нулю стремится и погрешность ее интерполяции. Поэтому имеет смысл перейти к более сложным, чем V-цикл, многосеточным циклам (рис. 13.8). Вычислительные затраты на проведение

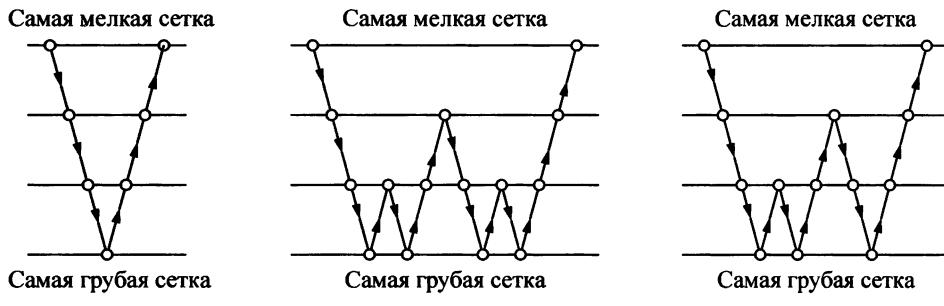


Рис. 13.8

сглаживающих итераций на грубых сетках малы по сравнению с затратами на проведение итераций на самой мелкой сетке. Поэтому вычислительные затраты, необходимые для реализации W- и F-циклов, не значительно больше, чем затраты на реализацию V-цикла, но точность вычисления поправки, которая в конечном итоге будет пролонгирована на самую мелкую сетку, ожидается гораздо большей. Обобщением V-, W- и F-циклов являются *адаптивные циклы*, в которых переход к более грубой или более мелкой сетке происходит в зависимости от погрешности полученного приближения.

Наконец, можно уменьшить влияние погрешности интерполяции на точность вычисления поправки с помощью более эффективных сглаживателей, которые лучше удаляют высокочастотные гармоники и погрешности интерполяции, чем метод Зейделя. В последнее время все чаще стали использовать различные варианты метода неполной факторизации и сопряженных градиентов. Однако более эффективные сглаживатели сложнее распараллелить, чем метод Якоби или Зейделя.

При реализации КММ применяются в основном два способа построения грубых сеток (рис. 13.9). Один из них основан на удалении узлов  $x^v$  (см. рис. 13.9, а), второй — на объединении разностных ячеек, ис-

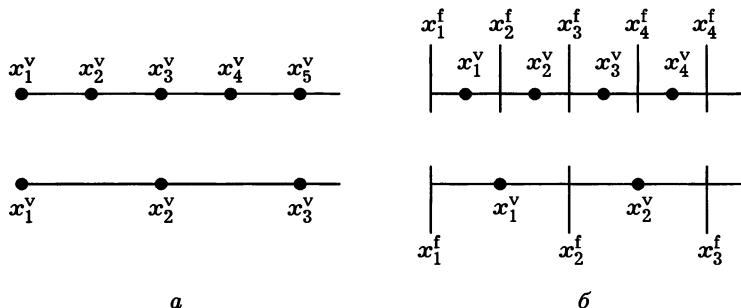


Рис. 13.9

пользуемых для аппроксимации интегро-интерполяционным методом, грани которых  $x^f$ , и удалении узлов  $x^v$  (см. рис. 13.9, б). Построение грубых сеток при реализации КММ зависит от решаемой задачи.

Конструкция вспомогательных СЛАУ на грубых сетках КММ также допускает различные варианты. Вариант КММ, рассмотренный в 13.3.1, применим только для дифференциальных уравнений с постоянными коэффициентами. Возможные варианты КММ для решения дифференциальных уравнений с переменными коэффициентами рассматриваются в [291].

Оригинальной разновидностью КММ являются **каскадные многосеточные методы**, в которых отсутствует предварительное сглаживание [202]. Вычисления ведутся в одном направлении: от грубых сеток к мелким. Основная идея состоит в получении более точного приближения на грубых сетках в целях уменьшения вычислительных затрат на мелких сетках.

Фактически каждая краевая задача приводит к появлению новых вариантов КММ. На первый взгляд, решение простейших задач не должно встречать никаких затруднений. Действительно, КММ являются самыми быстрыми методами решения задачи Дирихле для уравнения Пуассона на равномерной сетке с одинаковыми шагами по всем пространственным направлениям ( $h_x = h_y$ ). Однако если шаги вычислительной сетки различны ( $h_x \neq h_y$ ), то эффективность КММ резко снижается. Теоретический анализ показывает, что при определенном соотношении шагов сетки  $h_x/h_y$  способность метода Зейделя удалять высокочастотные гармоники и погрешности интерполяции существенно уменьшается [291]. Для подобных задач предложен специальный вариант КММ, который получил название Frequency Decomposition Multigrid Algorithm [236]. Следует подчеркнуть, что в данном варианте КММ используются четыре сетки на каждом уровне для уменьшения погрешности интерполяции.

Варианты КММ, предназначенные для решения нелинейных задач, получили обобщенное название FAS (Full Approximation Storage) [204].

КММ представляют собой совокупность проблемно-адаптированных алгоритмов, которая не может быть представлена в фиксированном виде в связи с многообразием способов интерполяции, многосеточных циклов, построений грубых сеток, сглаживающих процедур и других факторов [236]. Теоретически КММ обладают неулучшаемой скоростью сходимости, однако столь высокая скорость достигается только при оптимальной адаптации процедур к данной задаче на данной сетке. Отсутствие четких критериев оптимальности приводит к необходимости проведения дополнительных вычислительных экспериментов, чем

и объясняется множество публикаций по многосеточным методам. В настоящее время выработаны определенные рекомендации, например операторы сужения и пролонгации должны выбираться из условия

$$m_{\mathbb{R}} + m_{\mathbb{P}} > 2m,$$

где  $m_{\mathbb{R}}$  и  $m_{\mathbb{P}}$  определяются как наивысшая степень многочленов плюс единица, которые интерполируются точно с помощью операторов  $\mathbb{R}$  и  $\mathbb{P}$ ;  $2m$  — порядок решаемого дифференциального уравнения [235, 291]. Но часто при решении прикладных задач очень трудно понять без дополнительных вычислительных экспериментов, достигнута неулучшаемая скорость сходимости или нет.

Практически сразу после признания идеи Р.П. Федоренко западные математики осознали, что КММ является оптимальным для разнообразных частных случаев. Поэтому возникла потребность в разработке варианта КММ, который был бы эффективен для широкого класса задач. В одной из первых работ в данном направлении [219] был предложен вариант КММ, устроенный по принципу «черного ящика». В этой работе, как и в аналогичных работах, основные усилия были направлены на последовательное совершенствование отдельных процедур КММ, поэтому они изначально были обречены на неудачу [220]. Не сразу пришло понимание того, что вычислительный алгоритм, основанный на идеи Р.П. Федоренко и эффективный для широкого класса задач, должен быть основан на принципиально новых процедурах.

В заключение отметим, что наиболее элегантная алгоритмизация КММ достигается посредством рекурсивной формулировки [291]. Поэтому алгоритмический язык **Pascal**, поддерживающий рекурсию, до недавнего времени был единственным способом программной реализации КММ. Были предложены и нерекурсивные алгоритмы, но они носили скорее декларативный характер. Только в последние годы появились компьютерные программы для КММ, реализованные на **Fortran 90** и **C++**, которые тоже поддерживают рекурсию.

Способы применения КММ в автономных программах приводятся в [291]. Чаще всего у пользователя есть возможность выбора многосеточного цикла, сглаживающей процедуры, числа сглаживающих итераций и т. д. Достигнутая при этом скорость сходимости не оценивается.

Многосеточные методы, применяемые для решения сеточных задач, появляющихся при использовании метода конечных элементов, изложены в монографиях [126, 189]. Более подробный обзор КММ содержится в [163].

### 13.4. Универсальная многосеточная технология

Перспектива решать краевые задачи с минимальными вычислительными затратами выглядела очень заманчиво, но первоначально развитие *многосеточных методов* пошло по традиционному для 80-х годов XX века пути адаптации отдельных компонентов алгоритма к решаемой задаче. Достаточно быстро удалось разработать высокоеффективные многосеточные алгоритмы для решения *уравнения Пуассона* на равномерных сетках, однако усложнение решаемых краевых задач (нелинейность, анизотропия, разрывность коэффициентов и т. д.) быстро превратили КММ в труднообозримое семейство алгоритмов, практически не поддающееся формализации. Были предприняты многочисленные попытки разработать универсальный вариант КММ для решения широкого класса прикладных задач путем последовательного улучшения отдельных компонентов [219, 220].

Но уже начиная с середины 80-х годов XX века габариты и стоимость вычислительной техники стали стремительно уменьшаться, а производительность — возрастать. Массовый доступ к компьютерам получили инженеры, физики, химики и другие специалисты, которые не имели достаточной подготовки в области вычислительной математики, но у которых были свои задачи, зачастую связанные с необходимостью численного решения дифференциальных уравнений в частных производных.

Были разработаны автономные (т. е. устроенные по принципу «черного ящика») программы. При их использовании инженер только формулирует задачу, а детали вычислительного алгоритма ему могут быть даже неизвестны. Применительно к техническим приложениям работу автономных программ упрощенно можно представить следующим образом: конструктор проектирует некоторый узел с помощью графической программы, например *AutoCAD*. Затем геометрия узла переносится в вычислительный модуль, конструктор задает граничные и начальные условия, после чего проводит тепловой, прочностной, гидродинамический или другой расчет и анализирует результаты. Как правило, после анализа полученных результатов нужно внести изменения в конструкцию и повторить расчет. Обычно конструктор выполняет несколько подобных «итераций», чтобы получить оптимальную, с его точки зрения, конструкцию. Еще большую практическую ценность представляет возможность расчета машины в целом, а не только отдельных ее узлов, поскольку поэлементное моделирование часто связано с погрешностями постановки граничных условий. Уже сейчас такие программные продукты, как *ANSYS*, *STAR-CD*, *FLUENT*, *CFX*, *PHOENICS*

и др. получили широкое применение в НИИ и ОКБ для решения разнообразных прикладных задач.

Совершенствование автономных программ привело к необходимости развития новых алгоритмов, которые позволяют эффективно решать множество прикладных задач без контроля вычислительного процесса со стороны пользователя. Рассмотренные ранее КММ являются совокупностью проблемно-адаптированных алгоритмов. Для каждой краевой задачи можно предложить несколько вариантов КММ с различными проблемно-зависимыми процедурами. Вариант КММ с наиболее удачно адаптированными к решаемой краевой задачи проблемно-зависимыми процедурами теоретически обладает самой быстрой (неулучшаемой) скоростью сходимости. Необходимость адаптации процедур алгоритма при отсутствии четких критериев оптимальности делает практически невозможной формализацию КММ.

Далее будут изложены элементы *универсальной многосеточной технологии* (УМТ), которая предназначена для использования в перспективных автономных программах [109]. В отличие от КММ, УМТ основана на адаптации краевых задач к вычислительному алгоритму, поэтому она состоит из двух частей: аналитической и вычислительной. Основное назначение аналитической части УМТ заключается в приведении краевых задач к виду, удобному для последующего численного решения унифицированным многосеточным методом. Индивидуальные особенности краевых задач проявляются не в выборе проблемно-зависимых процедур, а в способе адаптации краевых задач к УМТ.

Рассмотрим для примера следующую краевую задачу:

$$\frac{d}{dx} \left( k(x) \frac{du}{dx} \right) - q(x)u(x) = -f(x), \quad 0 < x < 1,$$

$$u(0) = u_0, \quad u(1) = u_1, \quad k(x) \geq \alpha > 0, \quad q(x) \geq 0,$$

для иллюстрации основных процедур УМТ. Аналитическая часть УМТ для решения подобных краевых задач состоит в представлении искомого решения  $u(x)$  в виде суммы двух функций  $c(x)$  и  $\hat{u}(x)$ , т. е.

$$u(x) = c(x) + \hat{u}(x).$$

В последующих многосеточных итерациях сеточный аналог функции  $\hat{u}(x)$  будет служить приближением к решению разностной краевой задачи, а сеточный аналог функции  $c(x)$  — поправкой, вычисляемой на грубых сетках.

Представление  $u(x) = c(x) + \hat{u}(x)$ , называемое  $\Sigma$ -модификацией решения  $u(x)$ , является одной из форм адаптации решаемых краевых задач к УМТ. Подстановка данного представления в исходную краевую задачу приводит к ее следующей  $\Sigma$ -модифицированной форме:

$$\frac{d}{dx} \left( k(x) \frac{dc}{dx} \right) - q(x)c(x) = r(x), \quad c(0) = u_0 - \hat{u}(0), \quad c(1) = u_1 - \hat{u}(1),$$

где правая часть  $r(x)$  имеет вид

$$r(x) = -\frac{d}{dx} \left( k(x) \frac{d\hat{u}}{dx} \right) + q(x)\hat{u}(x) - f(x).$$

Члены с поправкой  $c(x)$  переносятся в левую часть, а остальные — в правую.

На первый взгляд  $\Sigma$ -модификация решения  $u(x) = c(x) + \hat{u}(x)$ , используемая в УМТ, похожа на представление, которое применяется в КММ, но между ними существует два принципиальных отличия:

1)  $\Sigma$ -модификация осуществляется перед дискретизацией исходной краевой задачи для более точной формулировки разностных задач на грубых сетках и возможности гибкого изменения типа и (или) порядка аппроксимации. Левая и правая части  $\Sigma$ -модифицированной задачи могут аппроксимироваться разными способами.

2)  $\Sigma$ -модификация не является единственным способом адаптации краевых задач к УМТ. Представление искомого решения в виде произведения двух функций ( $\Pi$ -модификация) может оказаться более предпочтительной для ряда нелинейных задач.

$\Sigma$ - и  $\Pi$ -модификации являются частными случаями аддитивного и мультипликативного выделения особенностей решения.

Вычислительная часть УМТ связана с компьютерным счетом и состоит из построения мелкой и грубых сеток, аппроксимации модифицированной краевой задачи интегро-интерполяционным методом и решения полученных сеточных уравнений при помощи унифицированного *многосеточного метода*.

Рассмотрим этапы вычислительной части УМТ.

**Построение самой мелкой сетки.** Первый этап вычислительной части УМТ состоит в построении самой *мелкой сетки*  $G_1^0$  в области  $G = [0, 1]$  для последующей аппроксимации модифицированной краевой задачи интегро-интерполяционным методом. Сетка  $G_1^0$  состоит из двух множеств точек  $G^v(0;1)$  и  $G^f(0;1)$ , которые могут быть как узлами сетки, так и гранями контрольных объемов (разностных ячеек, используемых при записи балансовых соотношений).

**Построение грубых сеток.** Построение *грубых сеток* в УМТ осуществляется посредством удаления двух точек из множеств  $G^v(0;1)$  и  $G^f(0;1)$  самой мелкой сетки  $G_1^0$  (рис. 13.10). Далее аналогично строим еще одну грубую сетку  $G_2^1$ . Построение начнем со сдвигом на одну точку. Третья грубая сетка  $G_3^1$  строится таким же образом. Непосредственно из рис. 13.10 следуют основные особенности построения грубых сеток при реализации УМТ:

1) грубые сетки  $G_1^1$ ,  $G_2^1$  и  $G_3^1$  не имеют общих точек, т. е.

$$G_n^1 \cap G_m^1 = \emptyset, \quad n \neq m;$$

2) мелкая сетка  $G_1^0$  представима в виде объединения грубых сеток  $G_1^1$ ,  $G_2^1$  и  $G_3^1$ , т. е.

$$G_1^0 = \bigcup_{k=1}^3 G_k^1;$$

3) все сетки геометрически подобны, однако шаг грубых сеток  $G_1^1$ ,  $G_2^1$  и  $G_3^1$  в три раза больше, чем шаг сетки  $G_1^0$ ;

4) вне зависимости от способа определения сеточных функций на самой мелкой сетке каждый контрольный объем на сетках  $G_1^1$ ,  $G_2^1$  и  $G_3^1$  является объединением трех контрольных объемов на сетке  $G_1^0$ .

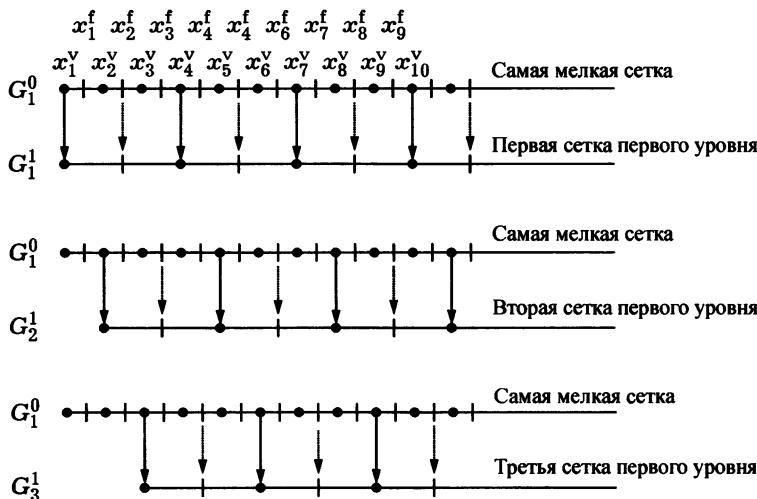


Рис. 13.10

Способ построения грубых сеток при реализации УМТ объединяет оба способа, используемые в КММ, он заключается как в удалении узлов, так и в объединении контрольных объемов (см. рис. 13.9).

Самая мелкая сетка  $G_1^0$  образует нулевой сеточный уровень, а три грубые сетки  $G_1^1$ ,  $G_2^1$  и  $G_3^1$  образуют первый сеточный уровень. Далее построение еще более грубых сеток осуществляется рекуррентным образом: каждая сетка  $G_i^L$ ,  $i = 1, 2, \dots, 3^L$ , уровня  $L$  рассматривается как самая мелкая сетка для трех грубых сеток  $G_j^{L+1}$ ,  $j = 1, 2, \dots, 3^{L+1}$ , следующего уровня  $L + 1$ . Девять еще более грубых сеток, полученных из трех сеток первого уровня, образуют второй сеточный уровень и т. д. (рис. 13.11). Построение грубых сеток завершается, когда на грубых сетках останется всего несколько узлов. В дальнейшем совокупность самой мелкой и всех грубых сеток будет называться многосеточной структурой.

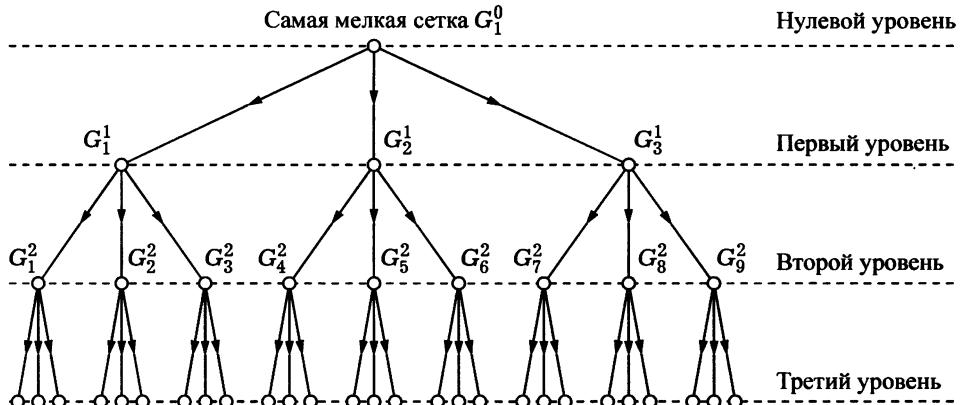


Рис. 13.11

Перечислим основные свойства грубых сеток УМТ, которые будут использованы для численного решения краевых задач.

**Свойство 1.** Каждый контрольный объем на сетках  $G_k^L$  можно представить в виде объединения  $3^L$  контрольных объемов на самой мелкой сетке  $G_1^0$ . В сочетании со свойством аддитивности определенного интеграла относительно подобластей это позволит существенно расширить класс краевых задач, решаемых унифицированным образом.

**Свойство 2.** Каждую сетку  $G_k^L$  ( $L \neq L^+$ ) можно представить в виде объединения трех соответствующих ей грубых сеток, что позволяет исключить интерполяцию. Как следствие, самую мелкую сетку  $G_1^0$  можно представить в виде объединения всех сеток одного уровня:

$$G_1^0 = \bigcup_{k=1}^{3^L} G_k^L, \quad L = 0, 1, \dots, L^+.$$

**Свойство 3.** Сетки одного уровня не имеют общих точек, т. е.

$$G_n^L \cap G_m^L = \emptyset, \quad n \neq m, \quad L = 1, 2, \dots, L^+,$$

что позволяет эффективно распараллеливать вычисления и экономно использовать память компьютера.

Многосеточные итерации УМТ схематично показаны на рис. 13.12. Нетрудно видеть, что УМТ занимает промежуточное положение между КММ и каскадными многосеточными методами. С одной стороны, в УМТ, как и в КММ, на каждой сетке выполняется несколько сглаживающих итераций для удаления высокочастотных гармоник. С другой стороны, в УМТ, как и в каскадных многосеточных методах, на уровнях с грубыми сетками выполняется дополнительная, по сравнению с КММ, вычислительная работа, чтобы уменьшить объем вычислений при сглаживании на уровнях с мелкими сетками.

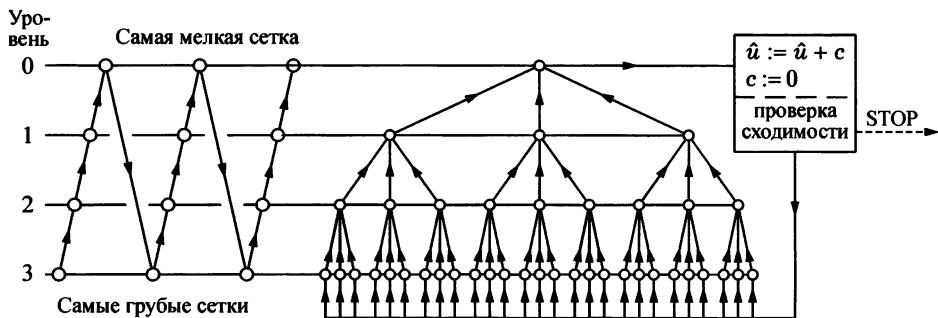


Рис. 13.12

В отличие от КММ, УМТ не содержит таких проблемно-зависимых процедур, как интерполяция и предварительное сглаживание, что позволяет применять ее к решению нелинейных краевых задач. Апроксимация модифицированных задач на многосеточной структуре осуществляется унифицированным образом с использованием свойств грубых сеток. В отличие от каскадных многосеточных методов, УМТ позволяет отыскивать численные решения краевых задач на адаптивных сетках.

Однако, как и каскадные многосеточные методы, УМТ не обладает оптимальной скоростью сходимости, поэтому по вычислительной эффективности уступает КММ при решении простейших краевых задач (см. рис. 13.7).

На рис. 13.7 приведены три грубые сетки, которые образуют первый уровень, и показаны численные решения, полученные на них методом прогонки. Далее показан процесс «сборки» полученных решений

в приближение к решению на самой мелкой сетке. В табл. 13.1 (второй столбец) приведены значения погрешности  $\mathcal{E} = \max |y_i - u(x_i)|$  решений, полученных на трех грубых сетках (первые три числа), и приближения к решению на самой мелкой сетке (последнее число), полученного после выполнения трех итераций Зейделя. Как следует из табл. 13.1, отсутствие погрешности интерполяции при реализации УМТ приводит к монотонному уменьшению погрешности  $\mathcal{E}$ , т. е. дополнительный объем вычислительной работы, выполненной на грубых сетках, позволил получить более точное приближение к решению на мелкой сетке.

## 14. ЧИСЛЕННОЕ РЕШЕНИЕ УРАВНЕНИЯ ПЕРЕНОСА

Описаны результаты разработки разностных схем для простейшего уравнения гиперболического типа, составляющего основу многих моделей механики сплошной среды. Представленные тесты и база данных об ошибках численного решения могут быть использованы при исследовании качественных и количественных характеристик как новых схем для *уравнения переноса*, так и, например, для *уравнений газовой динамики*.

### 14.1. Уравнение переноса: постановка задачи

*Уравнение переноса* — одно из фундаментальных уравнений математической физики [170], которое широко используется для описания движения сплошной среды. Оно является простейшим представителем класса уравнений, к которому относятся уравнения гидродинамики, магнитной и *газовой динамики*. Поэтому представляет интерес разработка численных методов решения этого уравнения и изучение их свойств.

В данной главе рассмотрены задачи Коши для *уравнений переноса* следующего вида:

$$\frac{\partial u}{\partial t} + \frac{\partial(au)}{\partial x} = 0, \quad x \in (-\infty, +\infty), \quad t \in (0, T], \quad (14.1)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0, \quad x \in (-\infty, +\infty), \quad t \in (0, T], \quad (14.2)$$

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v_1)}{\partial x} + \frac{\partial(\rho v_2)}{\partial y} = 0, \quad (x, y) \in \mathbb{R}^2, \quad t \in (0, T], \quad (14.3)$$

т. е. для одномерного *линейного* ( $a = \text{const}$ ), одномерного *квазилинейного* и *двумерного* линейного ( $v_1$  и  $v_2$  не зависят от  $\rho$ ) уравнений с финитными начальными данными (здесь они опущены) соответственно. Уравнение (14.2) формально соответствует (14.1) при  $a = u/2$ .

Одним из важнейших свойств точного решения задачи Коши для уравнения (14.1) является перенос начального профиля без искажений. *Разностные схемы*, аппроксимирующие такое уравнение, в той или иной мере искажают точное решение. Устойчивые разностные

схемы первого порядка аппроксимации, как правило, дают решение, «распывающееся» со временем по пространству. Схемы повышенного порядка аппроксимации в соответствии с общей теоремой [80] могут приводить к существенным качественным искажениям решения в виде явлений нормальной или аномальной дисперсии. В результате численное решение по прошествии некоторого времени «забывает» свой начальный профиль и перестает даже качественно соответствовать точному [5, 35, 38–41, 52, 54, 68, 74, 80, 200, 261 и др.]. Решения существенно различного начального вида со временем принимают близкие формы. Это влияет на качество и точность численного решения задачи.

На рис. 14.1 и 14.2 приведены примеры численного решения одной и той же задачи с использованием двух различных схем на некоторый момент времени. Точное решение представляет собой букву «М», переносимую без изменений. В обоих случаях графики решения приведены для чисел Куранта  $\gamma = 0,1; 0,5; 1,0; 1,1; 1,5$ . Результаты, представленные на рис. 14.1, получены расчетом по схеме с диссипацией; профиль решения в этом случае расплывается. При этом численное решение перемещается с правильной скоростью.

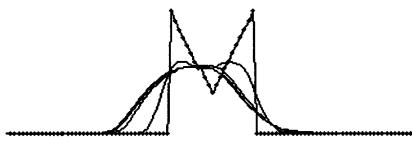


Рис. 14.1

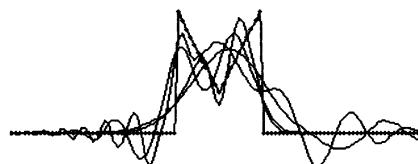


Рис. 14.2

Графики, приведенные на рис. 14.2, получены расчетом по схеме с дисперсией, поэтому показанное на нем численное решение рассыпалось на совокупность гармоник, каждая из которых перемещается со своей скоростью. Численное решение резко отличается от точного. При этом качество численного решения почти не зависит от параметров схем.

При выполнении условия  $\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} = 0$  (несжимаемая среда) перенос начального профиля без искажения справедлив и для двумерного случая (14.3).

Учет движения элемента сплошной среды приводит к уравнению переноса квазилинейного типа (14.2). Основным следствием нелинейности такого уравнения является «опрокидывание» начального профиля решения за конечное время и возникновение **ударных волн**. Изучение закономерностей распространения ударных волн представляет особый интерес в исследовании течений сплошной среды. Наличие разрывов в решении представляет определенные трудности, при этом точность численного решения зависит от качества воспроизведения этих разрывов.

Существует большое количество разностных схем для решения задач такого типа. Однако использование большинства из них приводит к «расплывчатому» решению и «размазанному» фронту ударных волн, а использование схем повышенного порядка аппроксимации может привести к искажениям численного решения, в том числе осцилляциям (см. рис. 14.2), т. е. немонотонности решения [80].

Качество численного решения можно повысить различными способами. Один из них связан с использованием неоднородных схем, в которых реализуются контроль за перемещением разрыва и изменение алгоритма вычислений в его окрестности. Однако основная проблема при использовании таких схем заключается в том, что появление разрывных решений возможно даже при гладких начальных условиях, в результате чего положение особых точек заранее неизвестно. К тому же число разрывов может меняться со временем и для каждого из них нужно перестроить алгоритм расчета отдельно. Поэтому особо важно исследование однородных схем повышенного порядка аппроксимации, а также построение новых квазимонотонных схем на основе однородных.

Целью данной главы является построение новых квазимонотонных разностных схем для численного решения задач (14.1)–(14.3) на основе анализа большого числа существующих методов численного решения. Для анализа качества известных и новых схем будем использовать систему «стандартных» тестов с удобной формой представления информации об ошибках численного решения задачи Коши. Такая система позволяет сравнить любую схему с известными схемами и единым образом определить ее характеристики. Полученная база данных об ошибках может быть полезна для разработки новых схем.

Далее в 14.2 приведены результаты сравнительного анализа как известных, так и предложенных конечно-разностных схем для решения задачи Коши для линейного одномерного уравнения переноса с финитными начальными данными. Описаны решения, полученные на основе системы тестов. Ошибка решения для выбранного метода зависит от начального условия, числа Куранта, заданной нормы и количества точек на носителе финитного начального условия. Значения ошибок являются базой для сравнительного анализа схем.

Материал 14.2 основан на работах [38–40], в которых приведены результаты сравнительного анализа 26 известных и новых, в основном двухслойных конечно-разностных схем для решения задачи (14.1) с финитными начальными условиями. Эти работы посвящены выработке и использованию системы стандартных тестов и удобных форм представления информации об ошибках.

В 14.2 представлен вывод новой монотонизированной схемы [38, 40], полученной путем введения «лимитеров» в известную схему К.И. Бабенко [8–10]. Анализ показал, что в широком диапазоне изменения чисел Куранта эта схема дает более высокую точность решения по сравнению с другими квазимонотонными схемами. Она лучшим образом сохраняет форму решения, в том числе и для разрывных немонотонных начальных условий.

Использованные схемы позволяют без расширения шаблона получить новые квазимонотонные схемы повышенного порядка аппроксимации на основе известных. Следуя [35, 36, 74], квазимонотонными здесь и далее называются разностные схемы, удовлетворяющие *принципу максимума* [139]. Данный принцип представляет собой совокупность условий, гарантирующих, в частности, достижение решением разностной задачи своего максимума (или минимума) на границе сеточной области. Это свойство сеточного решения служит многомерным «аналогом» свойства монотонности, которое определяется однозначно лишь в одномерном случае.

Отметим многообразие вариантов монотонизации: от первых вариантов [58, 178] до широко используемого алгоритма [19, 200, 201] и вариантов [35, 36, 74, 142]. В схемах, приведенных в этих работах (кроме [142]), используются данные о производной решения по пространству. Такие схемы могут быть условно отнесены к классу алгоритмов с «лимитерами». Особым является алгоритм монотонизации [54, 55, 234], построенный на знании области зависимости точного решения. Перечисленные работы далеко не исчерпывают возможные и уже используемые методы.

В 14.3 метод нелинейной монотонизации схемы К.И. Бабенко [38] перенесен на случай квазилинейного уравнения (14.2). Приведены результаты сравнения новой и ряда других известных конечно-разностных схем: явной и неявной схем с *левой разностью* [80, 139, 155], схемы *Лакса — Вендроффа* [64] и монотонизированной схемы «кабаре» [54, 55, 234]). Сравнение результатов, полученных по этим схемам, проведено на системе тестов, аналогичной использованной в 14.2 [38–40]. В завершении 14.3 приведены результаты численного решения (14.2) с помощью метода иного класса — разрывного метода Галеркина.

В 14.4 приведены результаты сравнительного анализа новой схемы и некоторых известных конечно-разностных схем для решения задачи (14.3) для линейного двумерного уравнения переноса с финитными начальными условиями. Результаты этого анализа являются непосредственным продолжением публикаций [38–40], посвященных одномерному уравнению. Для решения задачи (14.3) в двумерной области

применен алгоритм расщепления по координатам  $(x, y)$  [41, 68, 139]. При этом для решения одномерных задач используется одна из схем, рассмотренных в 14.2.2.

Данная глава содержит изложение результатов работ [5, 38–40, 42]. Итоги тестирования представлены в них как в графическом виде, так и в виде таблиц ошибок решения. Для определения ошибок численных решений использованы конечномерные аналоги норм пространств  $C$ ,  $L_1$ ,  $L_2$ ,  $W_1^1$ . Эти работы можно найти в электронном виде на сайтах ИПМ им. М.В. Келдыша РАН (<http://www.keldysh.ru/e-biblio>) и кафедры вычислительных методов факультета ВМиК МГУ им. М.В. Ломоносова (<http://old.cs.msu.su/vm/index.html>).

## 14.2. Линейное одномерное уравнение переноса

### 14.2.1. Постановка задачи для линейного одномерного уравнения переноса. Тестовые задачи

Рассмотрим задачу Коши (14.1) для линейного одномерного *уравнения переноса* с  $a = \text{const}$ ,  $a > 0$ , и *финитным начальным условием*  $u(x, 0) = u_0(x)$ , где функция  $u_0(x)$  равна нулю вне отрезка  $[l_1, l_2]$ , а на этом отрезке задана одним из следующих способов:

«левый треугольник»

$$u_0(x) = \frac{x - l_1}{l_2 - l_1}; \quad (14.4)$$

«прямоугольник»

$$u_0(x) = 1; \quad (14.5)$$

«косинус»

$$u_0(x) = \frac{1}{2} \left( 1 - \cos \frac{2\pi(x - l_1)}{l_2 - l_1} \right); \quad (14.6)$$

«зуб»

$$u_0(x) = \begin{cases} -\frac{2}{3}(l_{11} - l_1)(x - l_1) + 1, & x \in [l_1, l_{11}]; \\ \frac{1}{3}, & x \in [l_{11}, l_{22}]; \\ \frac{2}{3}(l_2 - l_{22})(x - l_2) + 1, & x \in (l_{22}, l_2]; \end{cases} \quad (14.7)$$

«M»

$$u_0(x) = \begin{cases} -\frac{2}{3}(l_{12} - l_1)(x - l_1) + 1, & x \in [l_1, l_{12}]; \\ \frac{2}{3}(l_2 - l_{12})(x - l_2) + 1, & x \in [l_{12}, l_2]; \end{cases} \quad (14.8)$$

«правый треугольник»

$$u_0(x) = \frac{l_2 - x}{l_2 - l_1}. \quad (14.9)$$

Задача Коши (14.1) имеет точное решение  $u(x, t) = u_0(x - at)$ . Это решение будем использовать в дальнейшем для контроля точности разностных схем. Точность будем определять через нормы разности точного и приближенного решений.

#### 14.2.2. Разностные схемы для линейного одномерного уравнения переноса

Для численного решения введем на плоскости  $(x, t)$  равномерную пространственно-временную сетку  $\omega_{ht} = \omega_h \times \omega_\tau$ , где

$$\omega_h = \{x_i = ih, i = 0, 1, \dots\}, \quad \omega_\tau = \{t_j = j\tau, j = 0, 1, \dots\}.$$

Через  $\gamma = at/h$  обозначим **число Куранта**, через  $h, \tau$  — шаги сетки по  $x$  и  $t$ . Здесь и далее будем использовать стандартные обозначения [139] для сеточных величин:  $y \equiv y_i^j$ ,  $\hat{y} \equiv y_i^{j+1}$ ,  $y_{+1} \equiv y_{+1}^j$ ,  $y_{-1} \equiv y_{-1}^j$ ,  $\hat{y}_{-1} \equiv y_{-1}^{j+1}$ ,  $\hat{y}_{+1} \equiv y_{+1}^{j+1}$ ,  $\check{y} = y_i^{j-1}$ , где верхний индекс — номер временного слоя, нижний индекс — номер узла по  $x$ . Решение на слое  $j$  считаем известным.

Рассмотрим на данной сетке наиболее известные разностные схемы или их производные, аппроксимирующие уравнение переноса (14.1). Ниже дано их краткое описание вместе с *первым дифференциальным приближением* [80, 190] — дифференциальным уравнением, более полно отражающим свойства разностного решения, чем исходное уравнение переноса. Приведены также условия устойчивости и порядок аппроксимации.

1. **Явная схема с левой разностью** [139, 155]:

$$\hat{y} = (1 - \gamma)y + \gamma y_{-1}. \quad (14.10)$$

Схема имеет первый порядок аппроксимации по времени и по пространству. Она устойчива, если выполнено условие  $\gamma \leq 1$ . При  $\gamma = 1$  применение схемы дает точное решение. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{1}{2}ah(1 - \gamma)u_{xx}.$$

2. **Схема Лакса — Вендроффа** [64, 132]:

$$\hat{y} = y - \gamma(F_p - F_l), \quad (14.11)$$

где

$$F_p = \frac{(y_{+1} + y) - \gamma(y_{+1} - y)}{2}, \quad F_l = \frac{(y + y_{-1}) - \gamma(y - y_{-1})}{2}.$$

Схема аппроксимирует исходную дифференциальную задачу со вторым порядком по времени и пространству. Она устойчива при  $\gamma \leq 1$ . При  $\gamma = 1$  схема дает точное решение. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{ah^2}{6}(\gamma^2 - 1)u_{xxx}.$$

3. *Схема с центральной разностью* [139, 155]:

$$\hat{y} = y + \frac{1}{2}\gamma(y_{+1} - y_{-1}). \quad (14.12)$$

Схема имеет первый порядок аппроксимации по времени и второй порядок по пространству, является безусловно неустойчивой как полусумма условно устойчивой схемы с левой разностью и безусловно неустойчивой схемы с правой разностью. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = -\frac{1}{2}ah\gamma u_{xx}.$$

4. *Схема Лакса* [132]:

$$\hat{y} = \frac{(y_{+1} + y_{-1}) - \gamma(y_{+1} - y_{-1})}{2}. \quad (14.13)$$

При выполнении условия устойчивости  $\gamma \leq 1$  и стремлении  $h^2$  к нулю быстрее, чем  $\tau$ , схема сходится. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{1}{2}ah\frac{1}{\gamma}(1 - \gamma^2)u_{xx}.$$

5. *Схема коррекции потоков SHASTA* [19, 64, 190, 200, 201]:

$$\hat{y} = \tilde{y} - \tilde{F}_p + \tilde{F}_m, \quad (14.14)$$

где

$$\tilde{y} = y - \frac{\gamma(y_{+1} + y) - \gamma(y + y_{-1})}{2} + (\nu(y_{+1} - y) - \nu(y - y_{-1}));$$

$$\tilde{F}_p = S \max \left\{ 0, \min [S(\tilde{y}_{+2} - \tilde{y}_{+1}), |F_p|, S(\tilde{y} - \tilde{y}_{-1})] \right\};$$

$$\nu = \frac{\gamma^2}{2} + \frac{1}{8}; \quad S = \text{sign}(\tilde{y}_{+1} - \tilde{y}); \quad F_p = \frac{1}{8}(\tilde{y}_{+1} - \tilde{y}).$$

Схема аппроксимирует решение со вторым порядком по времени и по пространству. Является нелинейной. Она устойчива при выполнении условия  $\gamma \leq \sqrt{7/12}$ .

6. *Схема с «лимитерами»* [35, 36]:

$$\hat{y} = y - \gamma(y - y_{-1}) - \frac{\gamma(\alpha_{+1/2}(y_{+1} - y) - \alpha_{-1/2}(y - y_{-1}))}{2}, \quad (14.15)$$

где

$$\alpha_{+1/2} = \alpha(R_{+1/2}), \quad R_{+1/2} = \frac{y - y_{-1}}{y_{+1} - y}$$

и

$$\alpha(R) = \begin{cases} 0, & R \leq 0; \\ \frac{(a + b(1 - \delta))R}{(a + b)(1 - \delta)}, & 0 < R < 1 - \delta; \\ \frac{a + bR}{a + b}, & |R - 1| \leq \delta; \\ \frac{(a + b(1 - \delta))R - 2a\delta}{(a + b)(1 - \delta)}, & 1 + \delta < R < 2; \\ 2, & R \geq 2. \end{cases}$$

Схема имеет первый порядок аппроксимации по времени. При  $0 < \delta < 1$ ,  $a = \text{const}$ ,  $b = \text{const}$ ,  $a + b \neq 0$  (здесь  $a$  и  $b$  — параметры схемы) схема имеет второй, а при  $a = 1/3$ ,  $b = 2/3$  — третий порядок аппроксимации по пространству. Она устойчива при выполнении условия  $\gamma \leq 0,5$ .

7. *Схема В.В. Рusanова* [119, 136]:

$$\hat{y} = y - \tilde{a}y_x^\circ - \tilde{\mu}y_{xx}^\circ - \tilde{\nu}y_{\bar{x}\bar{x}xx} + \tilde{\beta}y_{\bar{x}\bar{x}x}, \quad (14.16)$$

где

$$\begin{aligned} \tilde{a} &= (\beta_{30} + \beta_{32})a; \quad \tilde{\mu} = \beta_{21}\beta_{32}\tilde{a}^2\tau; \quad \tilde{\nu} = \frac{\omega_{32}h^4}{\tau}; \\ \tilde{\beta} &= \beta_{32}\beta_{21}\tilde{a}^3\tau^2 + \beta_{30}\vartheta_{31}\tilde{a}h^2; \\ \beta_{21} &= \frac{2}{3}; \quad \beta_{30} = \frac{1}{4}; \quad \beta_{32} = \frac{3}{4}; \quad \vartheta_{31} = -\frac{2}{3}. \end{aligned}$$

Схема имеет третий порядок аппроксимации по пространству при фиксированном параметре  $\gamma$ . При выполнении неравенств

$$-3 \leq 24\omega_{32} \leq \gamma^4 - 4\gamma^2$$

схема является устойчивой. При  $\gamma = 1$  и  $\omega_{32} = -0,125$  схема дает точное решение. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{ah^3}{\gamma} \left( \omega_{32} + \frac{1}{6}\gamma^2 - \frac{1}{24}\gamma^4 \right) u_{xxxx}.$$

**8. Явная схема с левой разностью 2-го порядка аппроксимации по  $x$ :**

$$\hat{y} = \left( 1 - \frac{3}{2}\gamma \right) y + \gamma \left( 2y_{-1} - \frac{1}{2}y_{-2} \right). \quad (14.17)$$

Схема имеет первый порядок аппроксимации по времени и второй порядок аппроксимации по пространству. Схема является безусловно неустойчивой. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = -\frac{ah\gamma}{2} u_{xx} + \frac{ah^2}{6} (2 + \gamma^2) u_{xxx}.$$

**9. Схема Кранка — Николсона [181]:**

$$\hat{y} = y - \frac{1}{4}\gamma(y_{+1} - y_{-1} + \hat{y}_{+1} - \hat{y}_{-1}). \quad (14.18)$$

Схема имеет второй порядок аппроксимации по времени и по пространству. Она безусловно устойчива. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{ah^2}{3} (1 + \gamma^2) u_{xxx}.$$

**10. Неявная схема с левой разностью [139, 155]:**

$$\hat{y} = \frac{\gamma}{1 + \gamma} \hat{y}_{-1} + \frac{1}{1 + \gamma} y. \quad (14.19)$$

Схема имеет первый порядок аппроксимации по времени и по пространству. Она безусловно устойчива. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{1}{2} ah(1 + \gamma) u_{xx}.$$

**11 и 12. Неявная схема типа Лакса — Вендроффа с весом ( $\sigma = 1$  или  $1/2$ ):**

$$\hat{y} = y - \frac{1}{2}\sigma\gamma\hat{y}_x^\circ - \frac{1}{2}(1 - \sigma)\gamma y_x^\circ + \frac{1}{2}\sigma\gamma^2\hat{y}_{\bar{x}x} + \frac{1}{2}(1 - \sigma)\gamma^2 y_{\bar{x}x}, \quad (14.20)$$

где

$$y_x^\circ = y_{i+1} - y_{i-1}; \quad y_{\bar{x}x} = y_{i-1} - 2y_i + y_{i+1}.$$

Схема имеет первый (при  $\sigma = 1$ ,  $\sigma = 1/2$ ) порядок аппроксимации по времени и второй порядок аппроксимации по пространству. Она безусловно устойчива.

Первое дифференциальное приближение схемы имеет следующий вид:

$$u_t + au_x = ah\gamma u_{xx} - \frac{ah^2}{6}(\gamma^2 - 1)u_{xxx}$$

при  $\sigma = 1$  и

$$u_t + au_x = \frac{ah\gamma}{2}u_{xx} - \frac{ah^2}{6}\left(1 + \frac{1}{2}\gamma^2\right)u_{xxx}$$

при  $\sigma = 1/2$ .

13. *Схема «кабаре»* [56]:

$$\hat{y} = (1 - 2\gamma)(y - y_{-1}) + \check{y}_{-1}. \quad (14.21)$$

Для определения решения на первом слое используется схема с левой разностью. Схема устойчива при  $\gamma \leq 1$ . Схема точна при  $\gamma = 1/2$  и  $\gamma = 1$ . Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = \frac{ah^2}{12}(1 - 3\gamma + 2\gamma^2)u_{xxx}.$$

14. *Схема «кабаре» с монотонизаторами* [55, 234]:

$$\hat{y} = \begin{cases} y^{(\max)}, & \tilde{y} > y^{(\max)}; \\ \tilde{y}, & y^{(\min)} \leq \tilde{y} \leq y^{(\max)}; \\ y^{(\min)}, & \tilde{y} < y^{(\min)}, \end{cases} \quad (14.22)$$

где

$$\begin{aligned} y^{(\max)} &= \max(y, y_{-1}); & y^{(\min)} &= \min(y, y_{-1}); \\ \tilde{y} &= (1 - \gamma)y + \gamma y_{-1} - \tau Q_{-1} + \alpha G + (1 - \alpha)G_{-1}; \\ Q_{-1} &= \frac{y_{-1} - y_{-1}^{-1}}{\tau} + a \frac{y - y_{-1}}{h}. \end{aligned}$$

Схема устойчива при  $\gamma \leq 1$ . Монотонизация осуществляется путем обрезания вычисляемого промежуточного решения по данным о решении из области зависимости на предыдущем слое. Возникающий дисбаланс компенсируется на следующем слое. При вычислении решения на первом слое  $G = 0$ , далее  $\hat{G} = \tilde{y} - \hat{y}$ .

15 и 16. Следующими схемами являются две **двухслойные асимметричные схемы**, полученные путем перестановки шаблона по  $x$  и  $t$  в схеме «кабаре». Такая замена возможна вследствие «равноправия»  $x, t$  в *уравнении переноса*.

Первая схема может быть записана следующим образом:

$$\hat{y}_{+1} = \frac{2-\gamma}{\gamma} y + y_{-1} - \frac{2-\gamma}{\gamma} \hat{y}. \quad (14.23)$$

Схема устойчива при  $\gamma \geq 1$ . При  $\gamma = 1$  и  $\gamma = 2$  схема дает точное решение. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = -\frac{ah^2}{12}(2 - 3\gamma + \gamma^2)u_{xxx}.$$

Вторая схема может быть записана в виде

$$\hat{y} = y + \frac{\gamma}{2+\gamma} \hat{y}_{-1} - \frac{\gamma}{2+\gamma} y_{+1}. \quad (14.24)$$

Схема безусловно устойчива. Ни при каком  $\gamma$ , отличном от нуля, не дает точного решения. Ее первое дифференциальное приближение имеет вид

$$u_t + au_x = -\frac{ah^2}{12}(2 + 3\gamma + \gamma^2)u_{xxx}.$$

Схема (14.24) представляет собой зеркально отраженную относительно оси ординат схему (14.23).

17 и 18. *Схема с «лимитерами-2»* в двух вариантах с различной локальной аппроксимацией [74]) может быть записана в виде

$$\begin{aligned} \hat{y} = y - \frac{1}{2}\gamma(1+\gamma)(y - y_{-1}) - \frac{1}{2}\gamma(1+\gamma)(y_{+1} - y) + \\ + \frac{1}{2}\gamma(\alpha_{+1/2}(y_{+1} - y) - \alpha_{-1/2}(y - y_{-1})), \end{aligned} \quad (14.25)$$

где

$$\alpha_{+1/2} = \alpha_{*0}(R_{+1/2}), \quad R_{+1/2} = \frac{y - y_{-1}}{y_{+1} - y}; \quad * = 2, 3;$$

$$\alpha_{20}(R) = \begin{cases} 1 - \gamma, & R \leq 0; \\ 1 - \gamma - \frac{2R}{\gamma}(1 - \gamma), & 0 < R \leq \frac{\gamma}{2 - \gamma}; \\ R - \gamma, & \frac{\gamma}{2 - \gamma} < R \leq \gamma; \\ 0, & \gamma < R \leq R^*; \\ -\frac{2}{\gamma}(1 - \gamma)(R - R^*), & R^* < R \leq R^* + \frac{\gamma(1 + \gamma)}{2(1 - \gamma)}; \\ -1 - \gamma, & R > R^* + \frac{\gamma(1 + \gamma)}{2(1 - \gamma)}; \end{cases} \quad (14.26)$$

$$\alpha_{30}(R) = \begin{cases} 1-\gamma, & R \leq 0; \\ 1-\gamma - \frac{2R}{\gamma}(1-\gamma), & 0 < R \leq \frac{\gamma}{2-\gamma}; \\ R-\gamma, & \frac{\gamma}{2-\gamma} < R \leq \frac{1+3\gamma}{4}; \\ \frac{1-R}{3}, & \frac{1+3\gamma}{4} < R \leq R^*; \\ -\frac{2}{\gamma}(1-\gamma)(R-R^*) + \frac{1-R^*}{3}, & R^* < R \leq R^{**}; \\ -1-\gamma, & R > R^{**}; \\ R^{**} = \frac{(6-7\gamma)R^* + \gamma(3\gamma+4)}{6(1-\gamma)}. \end{cases} \quad (14.27)$$

Схема с «лимитерами»  $\alpha_{20}$  имеет второй порядок аппроксимации по времени и по пространству, с «лимитером»  $\alpha_{30}$  — второй порядок аппроксимации по времени и третий по пространству:

Схема устойчива при выполнении условия  $\gamma < 1$ . Она получена на основе схемы Лакса — Вендроффа путем введения монотонизатора с ограничителями.

19. *Схема Р.П. Федоренко* [178]:

$$\hat{y} = y - \gamma(y - y_{-1}) - \frac{1}{2}\sigma(\gamma - \gamma^2)(y_{+1} - 2y + y_{-1}), \quad (14.28)$$

где

$$\sigma = \begin{cases} 1, & |y_{+1} - 2y + y_{-1}| < \lambda|y - y_{-1}|; \\ 0, & |y_{+1} - 2y + y_{-1}| \geq \lambda|y - y_{-1}|. \end{cases}$$

Схема при  $\lambda = 0$  превращается в схему с левой разностью первого порядка аппроксимации, при  $\lambda = \infty$  — в схему Лакса — Вендроффа второго порядка аппроксимации. При других значениях  $\lambda$  схема соединяет в себе свойства двух указанных схем. Она устойчива при  $\gamma \leq 1$ , при  $\gamma = 1$  ее применение дает точное решение. Схема является одной из первых схем с искусственными монотонизаторами.

20. *Схема со «сглаживанием»* [58]:

$$\hat{y} = \gamma y_{-1} + (1 - \gamma)y + \Delta_2 f(\xi), \quad (14.29)$$

где

$$f(\xi) = \begin{cases} 1, & \xi \leq 1; \\ 1/\xi, & \xi > 1; \end{cases} \quad \xi = \left| \frac{\Delta_2}{\Delta_1} \right|;$$

$$\Delta_1 = (y - y_{-1}) \min(\gamma, 1 - \gamma); \quad \Delta_2 = (1 - \gamma)y_{-1} - \frac{1 - \gamma}{1 + \gamma}\hat{y}_{-1} - \gamma \frac{1 - \gamma}{1 + \gamma}y.$$

В таком виде схема устойчива при  $\gamma < 1$ . В [58] приведен ее вариант, устойчивый при  $\gamma > 1$ . Схема построена путем монотонизации схемы «квадрат».

21. *Схема К.И. Бабенко («квадрат»)* [8–10, 176]:

$$\hat{y}_{+1} = y - \frac{1-\gamma}{1+\gamma}(\hat{y} - y_{+1}). \quad (14.30)$$

Схема безусловно устойчива. При  $\gamma = 1$  она дает точное решение задачи. Имеет второй порядок аппроксимации по времени и по пространству. Ее первое дифференциальное приближение

$$u_t + au_x = -\frac{ah^2}{12}(\gamma^2 - 1)u_{xxx}.$$

22. *Схема К.И. Бабенко* (записана для  $a = 1$ ) с коррекцией типа «лимитер» [38]:

$$\hat{y} = y + \frac{-2\gamma\tilde{y}_{\bar{x}} + (1-\gamma)(\mu_{-1} - 1)\tilde{y}_{-1,t}}{1 + \gamma + (1-\gamma)\mu}, \quad (14.31)$$

где

$$\begin{aligned} \tilde{y}_{\bar{x}} &= y - y_{-1}; & \tilde{y}_{-1,t} &= \hat{y}_{-1} - y_{-1}; \\ \mu_{-1} &= \mu(R_{-1}); & \mu &= \mu(R); \\ R_{-1} &= \frac{y_{\bar{x}}}{y_{-1,t}} = \gamma \frac{y - y_{-1}}{\hat{y}_{-1} - y_{-1}}; & R &= \gamma \frac{y_{+1} - y}{\hat{y} - y}. \end{aligned}$$

Зависимость коэффициента искусственной диффузии  $\mu(R)$  приведена в 14.2.3. Там же описан алгоритм монотонизации схемы. Схема устойчива при  $\gamma < 1$ .

23. *Схема «парабола»* второго порядка аппроксимации по  $x$  и  $t$  [38, 142]:

$$\hat{y} = y - \frac{\gamma}{2}(y_{-2} - 4y_{-1} + 3y) + \frac{\gamma^2}{2}(y_{-2} - 2y_{-1} + y). \quad (14.32)$$

Имеет второй порядок аппроксимации по времени и по пространству. Устойчива при  $\gamma \leq 2$ . При  $\gamma = 1$  и  $\gamma = 2$  дает точное решение. Ее первое дифференциальное приближение

$$u_t + au_x = \frac{ah^2}{6}(2 - 3\gamma + \gamma^2)u_{xxx}.$$

24. *Схема «парабола» с «лимитерами»* [38]:

$$\hat{y} = y - \frac{1}{2}\gamma(3 - \gamma - \eta_{-1/2})\tilde{y}_{\bar{x}} + \frac{1}{2}\gamma(1 - \gamma - \eta_{-3/2})\tilde{y}_{\bar{x},-1}, \quad (14.33)$$

где

$$\tilde{y}_{\bar{x}} = y - y_{-1}; \quad \tilde{y}_{\bar{x},-1} = y_{-1} - y_{-2};$$

$$R_{-3/2} = \frac{\tilde{y}_{\bar{x}}}{\tilde{y}_{\bar{x},-1}};$$

$$\eta = \eta(R); \quad \eta_{-1/2} = \eta(R_{-1/2}); \quad \eta_{-3/2} = \eta(R_{-3/2});$$

$$\eta(R) = \begin{cases} 1 - \gamma, & R \leq 0; \\ 1 - \gamma - 2R, & 0 < R \leq \frac{1-\gamma}{2}; \\ 0, & \frac{1-\gamma}{2} < R \leq R^*; \\ -2(R - R^*), & R^* < R \leq R_3; \\ 3 - \gamma - \frac{2}{\gamma}, & R > R_3 = R^* - \frac{1}{2}\left(3 - \gamma - \frac{2}{\gamma}\right). \end{cases}$$

Схема является устойчивой при  $\gamma \leq 1$ . При  $\gamma = 1$  схема передает точное решение. Схема построена путем монотонизации схемы «пара-бала».

25 и 26. Следующие две *схемы* получены из схем (14.23) и (14.24) путем их *монотонизации по данным из области зависимости* [38]. В случае (14.23) такая процедура приводит к схеме

$$\frac{\hat{y} - y}{\tau} + \frac{1}{2}\left(\frac{\hat{y}_{+1} - \hat{y}}{h} + \frac{y - y_{-1}}{h}\right) = -\alpha g_{-1} - (1 - \alpha)g, \quad (14.34)$$

где

$$\hat{y}_{+1} = \begin{cases} y_{\min}, & \hat{y}_{+1} \leq y_{\min}; \\ \hat{y}_{+1}, & y_{\min} < \hat{y}_{+1} < y_{\max}; \\ y_{\max}, & \hat{y}_{+1} \geq y_{\max}; \end{cases}$$

$$g = \frac{(y - \tilde{y})}{\tau}; \quad y_{\min} = \min(y, y_{-1}); \quad y_{\max} = \max(y, y_{-1}).$$

Обеспечить устойчивость вычислений при  $\gamma \in [1, 2]$  можно, если выбрать, например,

$$\alpha(\gamma) = \gamma^2 - \frac{5\gamma}{2} + 2.$$

Самый простой выбор коэффициента  $\alpha$  следующий:  $\alpha(\gamma) = 1/2$ .

В случае схемы (14.24) монотонизация приводит к схеме

$$\frac{\hat{y} - y}{\tau} + \frac{1}{2}\left(\frac{\hat{y} - \hat{y}_{-1}}{h} + \frac{y_{+1} - y}{h}\right) = -\beta g_{-1} - (1 - \beta)g. \quad (14.35)$$

Пусть сначала  $0 < \gamma < 1$  для фиксации области зависимости. Определим  $y_{\min}$ ,  $y_{\max}$  так же, как в схеме (14.34). Затем зададим коэффициент  $\beta$  в виде  $\beta = \gamma$  и

$$\hat{y} = \begin{cases} y_{\min}, & \hat{y} \leq y_{\min}; \\ \hat{y}, & y_{\min} < \hat{y} < y_{\max}; \\ y_{\max}, & \hat{y} \geq y_{\max}. \end{cases}$$

Пусть теперь  $1 \leq \gamma \leq 2$ . Тогда вместо соответствующих выражений, приведенных выше, имеем

$$y_{\min} = \min(y_{-2}, y_{-1}), \quad y_{\max} = \max(y_{-2}, y_{-1}),$$

$$\beta = \gamma - 1, \quad 1 - \beta = 2 - \gamma,$$

а правую часть заменим на  $-\beta g_{-2} - (1 - \beta)g_{-1}$ .

Все изученные разностные схемы являются двухслойными по времени. Исключением являются схемы (14.21) и (14.22), в которых используются три временных слоя. Отметим, что несмотря на обширность данного списка, он не содержит всех возможных схем даже на четырехточечном шаблоне. Число таких схем при использовании нелинейной монотонизации не ограничено.

На основе численных расчетов в [38–40] построены таблицы ошибок для различных алгоритмов, различных начальных условий и различных значений чисел Куранта. Эти таблицы содержат количественную информацию о свойствах схем в широком диапазоне изменения параметров задачи. На основе анализа таблиц ошибок построены таблицы схем, среди которых есть схемы, обеспечивающие минимальные ошибки для выбранных начальных условий. Анализ таблиц схем позволил выделить преимущественные схемы. Критерием отбора являлась точность численного решения.

Результаты расчетов задачи Коши для шести форм начального профиля получены при следующих значениях параметров:  $l = 520$ ,  $l_1 = 10$ ,  $l_2 = 30$ ,  $T = 400$ ,  $h = 1$ ,  $a = 1$ . Временной шаг  $\tau$  принимал одно из следующих значений: 4,0; 3,0; 2,5; 1,9; 1,5; 1,1; 0,9; 0,5; 0,25; 0,1. При таких параметрах число Куранта совпадает с временным шагом.

Решения получены по системе тестов в пространственно-временном прямоугольнике достаточно больших размеров. Его размеры выбраны из предположения, что время, за которое финитное начальное условие пробегает 20 своих длин, является достаточным для проявления свойств вычислительного метода.

### 14.2.3. Метод нелинейной монотонизации разностных схем для линейного одномерного уравнения переноса

Представим один из методов нелинейной монотонизации разностной схемы. Для простоты положим  $a = 1$ . Рассмотрим *схему К.И. Бабенко*, или «квадрат», в качестве базовой:

$$\hat{y} = y_{-1} - \frac{1-\gamma}{1+\gamma}(\hat{y}_{-1} - y). \quad (14.36)$$

Схема (14.36) имеет второй порядок аппроксимации по  $x, t$  и является немонотонной. При  $\gamma = 1$  ее применение дает точное решение. Схема является неявной, что не мешает находить решение на новом временном слое по явным формулам (14.36) *бегущим счетом* слева направо.

Схема (14.36) может быть записана в виде

$$y_t = -y_{\bar{x}} - \frac{1-\gamma}{1+\gamma}(y_{t,-1} + y_{\bar{x}}), \quad (14.37)$$

где

$$y_t = \frac{1}{\tau}(\hat{y} - y); \quad y_{\bar{x}} = \frac{1}{h}(y - y_{-1}).$$

Можно попытаться сделать ее монотонной, добавив слагаемые:

$$y_t = -y_{\bar{x}} - \frac{1-\gamma}{1+\gamma}(y_{t,-1} + y_{\bar{x}}) + \frac{1-\gamma}{1+\gamma}(\mu_{-1}y_{t,-1} + \mu y_{\bar{x}}), \quad (14.38)$$

где  $\mu$  — коэффициент искусственной диффузии. При этом величины  $y_{t,-1}$  и  $y_{\bar{x}}$  являются известными. Временная производная  $y_{t,-1}$  аппроксимирует  $-y_{\bar{x},-1}$  на решении уравнения переноса. Поэтому член с производной  $y_{t,-1}$  также дает искусственное диффузионное слагаемое, как и член с производной  $y_{\bar{x}}$ .

Выкладки [38] позволяют найти вид искусственной диффузии, обеспечивающей монотонность схемы (14.38). При этом  $\mu$  есть функция величины  $R = y_{\bar{x}}/y_{t,-1}$ . Однако алгоритм (14.38) не является консервативным: введенные слагаемые не компенсируют друг друга при суммировании. Для обеспечения консервативности (а это необходимое условие сходимости схемы [38]) нужен другой способ монотонизации.

Единственной возможностью обеспечить консервативность схемы и ее монотонность при повышенном порядке аппроксимации является

выбор коэффициента искусственной диффузии  $\mu$ , зависящего от неизвестного решения задачи на новом временном слое  $\hat{y}$ , и изменение добавленного в правую часть (14.38) диффузионного слагаемого. В этом случае монотонизованная схема будет нелинейной:

$$y_t = -y_{\bar{x}} - \frac{1-\gamma}{1+\gamma}(y_{t,-1} + y_{\bar{x}}) + \frac{1-\gamma}{1+\gamma}(\mu_{-1}y_{t,-1} - \mu y_t). \quad (14.39)$$

Перепишем (14.39) при  $y_{\bar{x}} \neq 0$  в форме

$$y_t + \varphi y_{\bar{x}} = 0, \quad (14.40)$$

где

$$\varphi = \frac{\frac{2}{1+\gamma} + \frac{1-\gamma}{1+\gamma}(1-\mu_{-1})\frac{y_{t,-1}}{y_{\bar{x}}}}{1 + \frac{1-\gamma}{1+\gamma}\mu}.$$

Схема (14.40) удовлетворяет условиям принципа максимума, если выполнены неравенства

$$\varphi \geq 0, \quad 1 - \gamma\varphi \geq 0. \quad (14.41)$$

Выполнение (14.41) можно обеспечить, если

$$\mu = \mu(R), \quad \mu_{-1} = \mu(R_{-1}), \quad R_{-1} = \frac{y_{\bar{x}}}{y_{t,-1}} = \gamma \frac{y - y_{-1}}{\hat{y}_{-1} - y_{-1}}.$$

Выберем  $\gamma \in (0, 1)$  аналогично другим схемам с «лимитерами». Такое ограничение вообще типично для гиперболических задач. При его выполнении за один временной шаг точка на характеристике перемещается не более чем на один пространственный шаг. Тогда достаточными условиями выполнения первого неравенства (14.41) являются следующие:

$$\begin{cases} \mu > -\frac{1+\gamma}{1-\gamma}; \\ \mu \leq 1 + \frac{2}{1-\gamma}R, \quad R > 0; \\ \mu \geq 1 + \frac{2}{1-\gamma}R, \quad R < 0. \end{cases} \quad (14.42)$$

Они приводят к тому, что второе неравенство (14.41) может быть выполнено лишь при

$$\mu \geq \mu^{**} > -\frac{1+\gamma}{1-\gamma}. \quad (14.43)$$

Тогда преобразования (14.41) в данной схеме приводят к следующим ограничениям на  $\mu$ :

$$\begin{cases} \mu \geqslant 1 - \frac{(1 + \mu^{**})R}{\gamma}, & R > 0; \\ \mu \leqslant 1 - \frac{(1 + \mu^{**})R}{\gamma}, & R < 0. \end{cases} \quad (14.44)$$

Сопоставление условий (14.44) и (14.42) показывает, что значение искусственной диффузии  $\mu = 1$ , дающее схему с левой разностью, входит в допустимую область изменения  $\mu$ , если потребовать выполнения условия  $\mu^{**} \geqslant -1$ . При этом наиболее просто множество допустимых  $\mu$  выглядит при  $\mu^{**} = -1$ . Видно, что при  $0 < \gamma < 1$  такое значение  $\mu^{**}$  удовлетворяет неравенству (14.43).

Полученные условия позволяют выбрать зависимость  $\mu(R)$  в виде, аналогичном искусственной диффузии [74]:

$$\mu(R) = \begin{cases} -1, & R \leqslant -R^* - \frac{1}{2}(1 - \gamma); \\ \frac{2}{1 - \gamma}(R + R^*), & -R^* - \frac{1}{2}(1 - \gamma) < R \leqslant -R^*; \\ 0, & -R^* < R \leqslant -\frac{1}{2}(1 - \gamma); \\ 1 + \frac{2}{1 - \gamma}R, & -\frac{1}{2}(1 - \gamma) < R \leqslant 0; \\ 1, & R > 0. \end{cases} \quad (14.45)$$

На рис. 14.3 показана зависимость коэффициента искусственной диффузии от величины  $R$ , на нем  $\circ$  —  $0,5(1 - \gamma) - R^*$ ;  $\bullet$  —  $R^*$ ;  $\square$  —  $0,5(1 - \gamma)$ . В (14.45) часть строк (1-я, 4-я, 5-я) совпадает с границами области допустимых значений  $\mu$  (14.42)–(14.44), 2-я строка обеспечивает переход от нулевого значения  $\mu$  к отрицательному по отрезку, параллельному отрезку, соответствующему 4-й строке (14.45). Средний отрезок зависимости сохраняет второй порядок аппроксимации (локальный) для линейных и близких к нему профилей. Параметр  $R^*$  в (14.45) подбирается по результатам тестирования. Чаще всего  $R^* = 1,2$ .

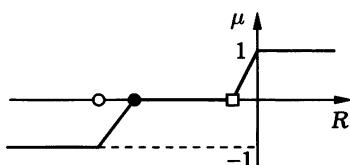


Рис. 14.3

Отметим, что в точном решении  $u_t = -u_x$ , что соответствует  $R = -1$ . Величина  $R$ , используемая в данной схеме, по существу, сходна со своим аналогом схемы [74], так как в монотонной явной схеме с левой разностью  $y_t = -y_{\bar{x}}$ , следовательно,  $R_{-1} = -y_{\bar{x}}/y_{\bar{x},-1}$ .

Принципиальное отличие полученной схемы от схемы (14.25) состоит в нелинейности решаемого уравнения относительно  $y_t$ . Однако анализ выражения (14.40) показывает, что его можно переписать так, чтобы в левой части стояла монотонная функция  $R$ . С помощью (14.45) она записывается в явном виде, позволяющем найти искомое значение  $R$  по аналитическим формулам, что, в свою очередь, дает возможность найти решение нелинейного алгебраического уравнения с недифференцируемой зависимостью от неизвестной величины без итераций.

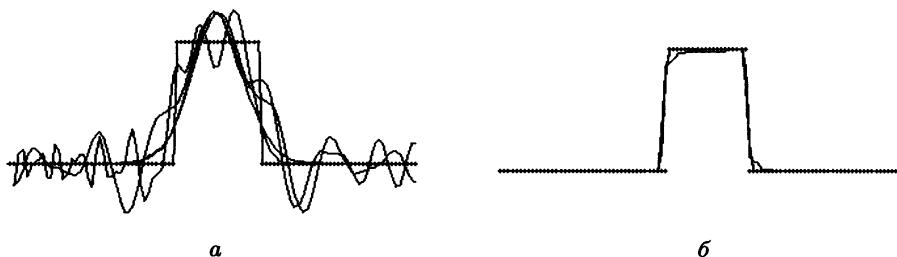


Рис. 14.4

На рис. 14.4 показан результат действия монотонизации для случая, когда точное решение имеет прямоугольный профиль. На рис. 14.4, *a* приведено решение исходной, немонотонизированной схемы при  $\gamma = 0,1; 0,5; 1,0; 1,1; 1,5$ , на рис. 14.4, *б* — решение, полученное по монотонизированной схеме К.И. Бабенко при  $\gamma = 0,1; 0,25; 0,5; 0,9$ . Точное тестовое решение представляет собой прямоугольник.

#### 14.2.4. Результаты расчетов для одномерного линейного уравнения переноса

Представим кратко зависимости ошибок численного решения от вида начального условия, выбранной разностной схемы и числа Куранта  $\gamma$ . Каждой из схем в [38–40] соответствует страница графической формы и страница ошибок. Точность разностной схемы зависит от выбранных норм, начального условия и числа Куранта  $\gamma$ . Лучшей схемой в [38–40] считается та, которая дает минимальную ошибку. Анализ таблиц ошибок позволяет составить таблицы схем. В таблицах представлены лучшие схемы для выбранной нормы и выбранного начального условия.

Анализ результатов тестирования показывает, что в списке лучших схем фигурируют в основном схемы (14.21)–(14.35) [40]. При  $\gamma < 1$  (интегрально по  $t$  и локально при  $t = T$ ) эти схемы дают меньшие значения ошибок, чем схемы (14.10)–(14.20). Только для условия (14.5) в нормах  $L_1$ ,  $L_2$  при  $\gamma = 0,5$  ошибки, полученные по схеме с «лимитерами» (14.15), меньше ошибок для схем (14.21)–(14.35).

Сравнительный анализ ошибок схем [40] показывает следующее. При  $0,5 \leq \gamma < 1$  предложенные в [38] схемы с «лимитерами» (схемы К.И. Бабенко и «парабола») в целом дают лучшие результаты, чем схема с «лимитерами-2» второго порядка в нормах  $L_1$ ,  $L_2$  для всех начальных условий. При  $\gamma = 0,5$  меньшую ошибку дает *схема К.И. Бабенко с лимитерами*, а при  $\gamma = 0,9$  — схема «парабола» с «лимитерами». При  $0 < \gamma < 0,5$  наоборот: ошибки для всех условий меньше в схеме с «лимитерами-2», чем в новых схемах. В нормах  $C$  и  $W_2^1$  нельзя назвать однозначно лучшую схему. При  $\gamma > 1$  схемы (14.23) и (14.24) могут конкурировать со схемами К.И. Бабенко и «парабола».

На рис. 14.5 показаны полученные численные решения для новой монотонизированной схемы «парабола» для различных начальных профилей и числа Куранта  $\gamma = 0,9$ . На рис. 14.5, *a* показаны численные решения на последовательные моменты времени, на рис. 14.5, *б* — точное решение. При  $\gamma = 0,9$  решение для начальных условий (14.5) и (14.6) передается без видимых изменений, для условия (14.7) характерно некоторое понижение амплитуды решения. При  $\gamma \leq 0,5$  амплитуда решения

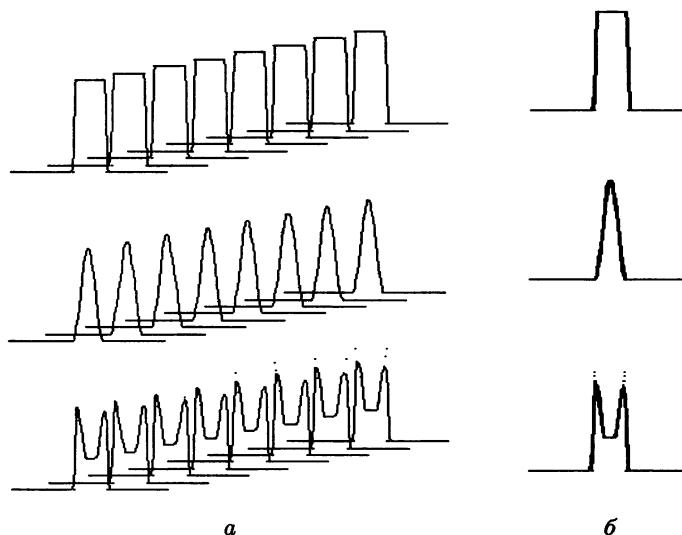


Рис. 14.5

уменьшается, особенно это заметно для начального условия (14.7). Увеличение  $\gamma$  повышает качество численного решения.

На рис. 14.6 приведены решения, полученные по монотонизированной схеме К.И. Бабенко. Содержание рис. 14.6 аналогично содержанию рис. 14.5. Для данной схемы при  $\gamma = 0,9$  и  $0,5$  решения для начальных условий (14.5) и (14.6) передаются без видимых изменений. Для условия (14.8) амплитуда численного решения чуть меньше, чем должна быть при точном переносе. При малых числах Куранта  $\gamma$  на гипотенузах треугольников начинается образование «ступенек», начальное условие (14.6) со временем превращается в прямоугольник. Качество решения улучшается с увеличением  $\gamma$ .

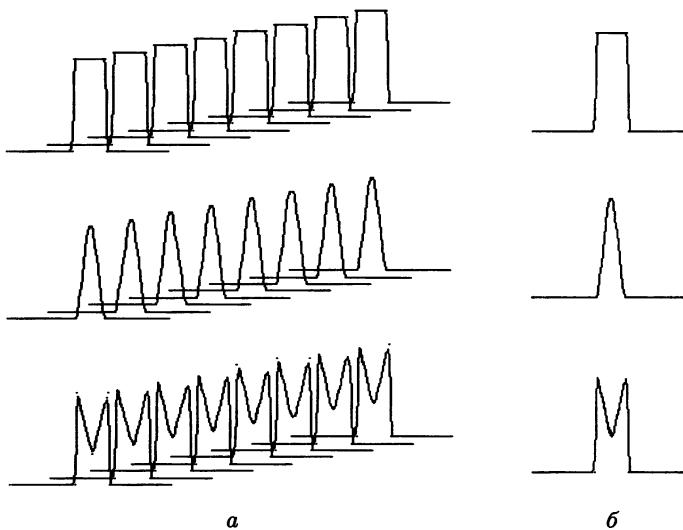


Рис. 14.6

Сравнение результатов схем с «лимитерами» (схемы К.И. Бабенко и «парабола») и схемы (14.35) показало, что первые две схемы дают более качественные решения при любом  $\gamma \in (0, 1)$ . Схема (14.35) лидирует по точности лишь в норме  $W_2^1$  для некоторых начальных условий и чисел Куранта  $\gamma$ . При  $\gamma > 1$  схема (14.34) дает более точное решение, чем схема (14.35). Численное решение, полученное по монотонизированной схеме «кабаре» [234] имеет более низкую точность и сильнее искажает начальный профиль, чем схемы с «лимитерами» [40]. Сравнение схемы (14.35) с монотонизированной схемой «кабаре» при  $\gamma < 1$  показывает, что схема (14.35) не дает лучших результатов во всех нормах одновременно. Она имеет преимущество в нормах  $C$  и  $W_2^1$  как интегрально, так и локально при  $t = T$ .

### 14.3. Одномерное квазилинейное уравнение

#### 14.3.1. Постановка задачи для квазилинейного одномерного уравнения переноса. Тестовые задачи

Будем рассматривать задачу Коши (14.2) для **квазилинейного уравнения переноса** с начальными условиями  $u(x, 0) = u_0(x)$ , где функция  $u_0(x)$  имеет один из следующих видов:

1) «треугольник»:

$$u_0(x) = \begin{cases} \frac{2(x - l_1)}{l_2 - l_1}, & x \in [l_1, \frac{l_1 + l_2}{2}); \\ \frac{2(l_2 - x)}{l_2 - l_1}, & x \in [\frac{l_1 + l_2}{2}, l_2]; \\ 0, & x \notin [l_1, l_2]; \end{cases}$$

2) «прямоугольник»:

$$u_0(x) = \begin{cases} 1, & x \in [l_1, l_2]; \\ 0, & x \notin [l_1, l_2]; \end{cases}$$

3) «левый треугольник»:

$$u_0(x) = \begin{cases} \frac{x - l_1}{l_2 - l_1}, & x \in [l_1, l_2]; \\ 0, & x \notin [l_1, l_2]; \end{cases}$$

4) «правый треугольник»:

$$u_0(x) = \begin{cases} \frac{l_2 - x}{l_2 - l_1}, & x \in [l_1, l_2]; \\ 0, & x \notin [l_1, l_2]; \end{cases}$$

5) «ступенька вниз»:

$$u_0(x) = \begin{cases} 1, & x \in (-\infty, l_1]; \\ 0, & x \in [l_1, +\infty); \end{cases}$$

6) «ступенька вверх»:

$$u_0(x) = \begin{cases} 0, & x \in (-\infty, l_1]; \\ 1, & x \in [l_1, +\infty). \end{cases}$$

Точные решения (они построены по алгоритмам и в соответствии с определением обобщенных решений [135]) для указанных условий таковы:

1) «треугольник»:

$$u(x,t) = \begin{cases} \frac{2(x-l_1)}{L+2t}, & x \in \left[l_1, \frac{l_2+l_1}{2} + t\right], \quad 0 < t \leq L/2; \\ \frac{2(l_2-x)}{l_2-l_1-2t}, & x \in \left[\frac{l_2+l_1}{2} + t, l_2\right], \quad 0 < t \leq L/2; \\ 0, & x \notin [l_1, l_2], \quad 0 < t \leq L/2; \\ \frac{2(x-l_1)}{L+2t}, & x \in \left[l_1, l_1 + \sqrt{\frac{L(L+2t)}{2}}\right], \quad t > L/2; \\ 0, & x \notin \left[l_1, l_1 + \sqrt{\frac{L(L+2t)}{2}}\right], \quad t > L/2, \end{cases} \quad (14.46)$$

где  $L = l_2 - l_1$ ;

2) «прямоугольник»:

$$u(x,t) = \begin{cases} \frac{x-l_1}{t}, & x \in [l_1, l_1 + t], \quad 0 < t \leq 2(l_2 - l_1); \\ 1, & x \in [l_1 + t, l_2 + t/2], \quad 0 < t \leq 2(l_2 - l_1); \\ 0, & x \notin [l_1, l_2 + t/2], \quad 0 < t \leq 2(l_2 - l_1); \\ \frac{x-l_1}{t}, & x \in [l_1, l_1 + \sqrt{2Lt}], \quad t > 2L; \\ 0, & x \notin [l_1, l_1 + \sqrt{2Lt}], \quad t > 2L, \end{cases} \quad (14.47)$$

где  $L = l_2 - l_1$ ;

3) «левый треугольник»:

$$u(x,t) = \begin{cases} \frac{x-l_1}{t+L}, & x \in [l_1, l_1 + \sqrt{L(t+L)}]; \\ 0, & x \notin [l_1, l_1 + \sqrt{L(t+L)}], \end{cases} \quad (14.48)$$

где  $L = l_2 - l_1$ ;

4) «правый треугольник»:

$$u(x,t) = \begin{cases} \frac{x-l_1}{t}, & x \in [l_1, l_1 + t], \quad 0 < t \leq L; \\ \frac{l_2-x}{L-t}, & x \in [l_1 + t, l_2], \quad 0 < t \leq L; \\ 0, & x \notin [l_1, l_2], \quad 0 < t \leq L; \\ \frac{x-l_1}{t}, & x \in [l_1, l_1 + \sqrt{Lt}], \quad t > L; \\ 0, & x \notin [l_1, l_1 + \sqrt{Lt}], \quad t > L, \end{cases} \quad (14.49)$$

где  $L = l_2 - l_1$ ;

5) «ступенька вниз»:

$$u(x, t) = \begin{cases} 1, & x \in (-\infty, l_1 + t/2]; \\ 0, & x \notin (-\infty, l_1 + t/2]; \end{cases} \quad (14.50)$$

6) «ступенька вверх»:

$$u(x, t) = \begin{cases} 0, & x \in (-\infty, l_1); \\ \frac{x - l_1}{t}, & x \in [l_1, l_1 + t]; \\ 1, & x \in (l_1 + t, +\infty). \end{cases} \quad (14.51)$$

Изменение начальных условий по сравнению с данными для линейной задачи (14.1) вызвано различием свойств решений задач (14.1) и (14.2), главным образом, опрокидыванием решений квазилинейного уравнения за конечное время.

Решения (14.46)–(14.51) использовались для контроля точности разностных схем. Точность определялась через нормы разности точного и приближенного решений.

Для численного решения поставленной задачи в [5] разработана новая конечно-разностная схема и приведено ее сравнение с явной и неявной схемами с левой разностью [139, 155], схемой Лакса — Вендроффа [64, 132], а также предложенной в [54, 55, 234] схемой прыжкового переноса (построена с помощью монотонизации схемы «кабаре» [56]). Предложенная в [5] новая схема рассмотрена в 14.3.2.

Аналогично [38–40] численное решение задачи (14.2) получено для указанных выше форм шести начальных профилей при следующих значениях параметров:  $l = 520$ ,  $l_1 = 0$ ,  $l_2 = 20$ ,  $T = 1000$ ,  $h = 1$ . Число Куранта  $\gamma = \max_{x,t} \frac{\tau|y(x,t)|}{h}$  ограничено сверху величиной  $\gamma_m$ , которая имеет значения 0,1; 0,25; 0,5; 0,9. Для неявной схемы с левой разностью также проведены расчеты при  $\gamma_m = 3$ . Таким образом, временной шаг  $\tau$  принимал те же значения, что и число Куранта.

### 14.3.2. Нелинейная монотонизация схемы К.И. Бабенко для квазилинейного одномерного уравнения переноса

Схему К.И. Бабенко («квадрат») вида (14.36) для квазилинейного одномерного уравнения переноса можно записать следующим образом:

$$y_t + \frac{1}{2}(y_{t,-1} - y_t) + \frac{1}{2}(y^2)_{\bar{x}} + \frac{1}{4}\left((\dot{y}^2)_{\bar{x}} - (y^2)_{\bar{x}}\right) = 0. \quad (14.52)$$

Пусть  $a = (\hat{y} + y)/2$ ,  $\gamma = a\tau/h$  — число Куранта, тогда (14.52) перепишем в виде

$$y_t + \frac{1}{2}(y^2)_{\bar{x}} + \frac{1}{2}y_{t,-1}(1 - \gamma_{-1}) - \frac{1}{2}y_t(1 - \gamma) = 0. \quad (14.53)$$

Проведем монотонизацию схемы, добавив в (14.53) слагаемые

$$-\frac{1}{2}\mu_{-1}y_{t,-1}(1 - \gamma_{-1}), \quad \frac{1}{2}\mu y_t(1 - \gamma),$$

где  $\mu$  — коэффициент искусственной диффузии. Поскольку временные производные  $y_t$ ,  $y_{t,-1}$  на точном решении квазилинейного уравнения перенося аппроксимируют производные  $-(y^2/2)_{\bar{x}}$ ,  $-(y^2/2)_{\bar{x},-1}$  соответственно, то введенные члены дают дополнительные диффузионные слагаемые. Таким образом, модифицированная схема примет вид

$$y_t + \frac{1}{2}(y^2)_{\bar{x}} + \frac{1}{2}y_{t,-1}(1 - \gamma_{-1})(1 - \mu_{-1}) - \frac{1}{2}y_t(1 - \gamma)(1 - \mu) = 0. \quad (14.54)$$

При  $\mu = 0$  схема (14.54) является исходной схемой (14.52), а при  $\mu = 1$  — схемой с левой разностью.

Запишем (14.54) в виде (подобно (14.40))

$$y_t + \varphi\left(\frac{y^2}{2}\right)_{\bar{x}} = 0,$$

где

$$\varphi = \frac{1 + \frac{(1 - \gamma_{-1})(1 - \mu_{-1})y_{t,-1}}{2\left(\frac{y^2}{2}\right)_{\bar{x}}}}{1 - \frac{1}{2}(1 - \gamma)(1 - \mu)}. \quad (14.55)$$

Предположим, что  $y \geq 0$  во всей пространственно-временной области. Тогда схема (14.55) удовлетворяет *принципу максимума* [139], если выполнены неравенства вида (14.41). Точное условие применимости принципа максимума отличается от второго условия (14.41): в нем фигурирует  $\tilde{a} = 1/2(y + y_{-1})$ . Однако из указанного условия (14.41) в предположении  $\hat{y} \geq 0$  следует применимость принципа максимума.

Положим  $\mu = \mu(R, \gamma)$ ,  $\mu_{-1} = \mu(R_{-1}, \gamma_{-1})$ , где параметр

$$R_{-1} = \frac{\left(\frac{y^2}{2}\right)_{\bar{x}}}{y_{t,-1}} = \frac{\tau}{h} \frac{y^2 - y_{-1}^2}{2(\hat{y}_{-1} - y_{-1})}.$$

Неравенства (14.41) выполнены, если функция  $\mu = \mu(R, \gamma)$  имеет вид (см. [74])

$$\mu(R, \gamma) = \begin{cases} 1, & R \geq 0; \\ 1 + \frac{2R}{(1-\gamma)}, & -\frac{1}{2}(1-\gamma) \leq R < 0; \\ 0, & -R^* \leq R \leq -\frac{1}{2}(1-\gamma); \\ \frac{2(R+R^*)}{(1-\gamma)}, & -R^* - \frac{1}{2}(1-\gamma) \leq R \leq -R^*; \\ -1, & R < -R^* - \frac{1}{2}(1-\gamma). \end{cases} \quad (14.56)$$

Введенная таким образом функция  $\mu$  обеспечивает монотонность схемы. В (14.56) 3-я строка соответствует отрезку, на котором схема сохраняет второй порядок аппроксимации для линейных и близких к ним профилей, а 4-я строка (14.56) обеспечивает переход  $\mu$  от нулевого значения к отрицательным, что соответствует появлению в схеме антидиффузионных слагаемых, приводящих к уменьшению «размазывания» разрывов.

Параметр  $R^*$  подбирался экспериментально. Для расчетов, как и в [38, 74], использовались значения  $R^* = 1, 2$ .

Уравнение (14.54) является нелинейным. Искомую величину  $\hat{y}$  можно вычислить по известному значению  $R$ . Значение  $R$  находится с помощью итераций.

### 14.3.3. Результаты расчетов для одномерного квазилинейного уравнения переноса

Представим кратко полученные в [5] численные результаты. Для каждой схемы, используемой для получения численного решения, в [5] приведены таблицы интегральных и локальных по времени ошибок для различных чисел Куранта.

С целью иллюстрации всех особенностей решения для каждого начального профиля представлены графики решения для нескольких промежуточных моментов времени. При этом каждый рисунок содержит графики точного решения и численных решений для различных чисел

Куранта. Такие рисунки приведены для предложенной монотонизованной схемы К.И. Бабенко, неявной схемы с левой разностью и схемы прыжкового переноса.

На рис. 14.7–14.9 приведены точные и численные решения, полученные по неявной схеме с левой разностью для чисел Куранта  $\gamma = 0,1; 0,5; 0,9$ , различных начальных профилей и различных  $t$ .

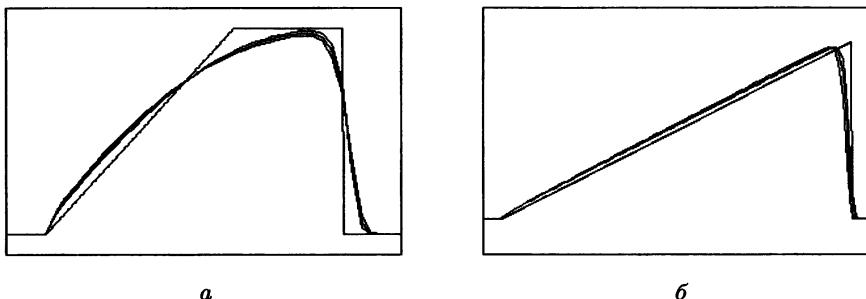


Рис. 14.7

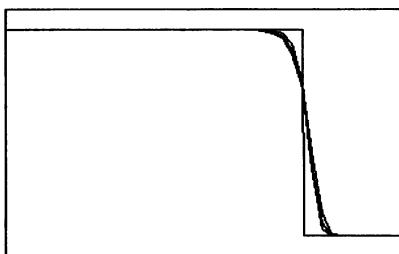


Рис. 14.8

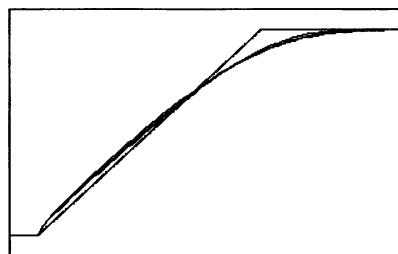


Рис. 14.9

На рис. 14.7 показано решение для моментов времени  $t = 18$  (*a*) и  $t = 160$  (*b*); начальным профилем является «прямоугольник».

На рис. 14.8 показано решение для момента времени  $t = 80$ ; начальным профилем является «ступенька вниз».

На рис. 14.9 показано решение для момента времени  $t = 40$ , начальным профилем является «ступенька вверх».

Численное решение, полученное по неявной схеме, еще хуже совпадает с точным, чем численное решение, полученное по явной схеме. Для неявной схемы характерно более сильное размазывание *разрывов* (от 7 до 15 шагов по пространству), которое усиливается с увеличением  $\gamma$ . Для численного решения характерно «выгибание» заднего фронта на начальных моментах времени, что усиливается при увеличении  $\gamma$ . С увеличением времени зависимость численного решения от числа Куранта ослабевает. Для начальных профилей (14.46)–(14.49) решение,

полученное по явной и неявной схемам с левой разностью, эволюционирует в волну треугольного профиля с распространяющейся вправо *ударной волной* и уменьшающейся амплитудой, что соответствует точному решению. Однако максимум амплитуды численного решения отстает от максимума точного решения. На рис. 14.7–14.9 численному решению, наиболее близкому к точному, соответствует  $\gamma = 0,1$ . Таким образом, особенности эволюции решения в виде образования и распространения ударных волн, а также взаимодействие ударных волн и **волн разрежения** схемами с левой разностью передаются плохо.

По сравнению со многими другими схемами решение, полученное с использованием *схемы прыжкового переноса* [54], лучше передает точное решение. Основным достоинством этой схемы является способность точно передавать разрывы типа ударных волн (рис. 14.10, приведено решение с начальным профилем «прямоугольник» при  $\gamma = 0,5; 0,9$  и  $t = 40$ ). Однако в силу специфики данной схемы решение меняется скачкообразно, поэтому точное положение ударных волн можно наблюдать только через определенные интервалы времени. Так, для начального профиля «ступенька вниз» (14.50) при  $\gamma = 0,5$  решение на каждом 4-м временном слое скачком смещается вправо на один шаг по  $x$ , догоняя точное решение. Таким образом, ошибка вычислений в эти моменты времени нулевая. Однако увеличение числа Куранта приводит к появлению дефектов численного решения в виде «ступенек» и неверных скоростей движения ударных волн. Из рис. 14.11 видно, что в случае начального условия (14.49) на волне сжатия независимо от числа Куранта (для приведенных численных решений  $\gamma = 0,5; 0,9$  и  $t = 5$ ) образуются «ступеньки».

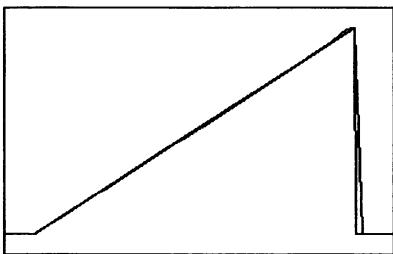


Рис. 14.10



Рис. 14.11

По сравнению с рассмотренными схемами монотонизованная схема К.И. Бабенко наилучшим (в смысле минимума норм ошибки) образом передает точное решение во всем изученном диапазоне изменения чисел Куранта. У схемы не наблюдается явлений дисперсии, расплывание решения незначительно — размазывание разрывов происходит на мень-

шее количество шагов (1–3 шага). Наблюдается небольшое смещение вправо левого фронта образующейся волны разрежения для начальных профилей (14.47), (14.49), (14.51). Рис. 14.12 ( $a$  —  $t = 18$ ,  $b$  —  $t = 160$ ; начальный профиль «прямоугольник»), рис. 14.13 ( $t = 80$ ; начальный профиль «ступенька вниз») и рис. 14.14 ( $t = 40$ ; начальный профиль «ступенька вверх») показывают, что численные решения при  $\gamma = 0,1; 0,5; 0,9$  практически не отличаются друг от друга. При этом решению, наиболее близкому к точному, соответствует  $\gamma = 0,1$ .

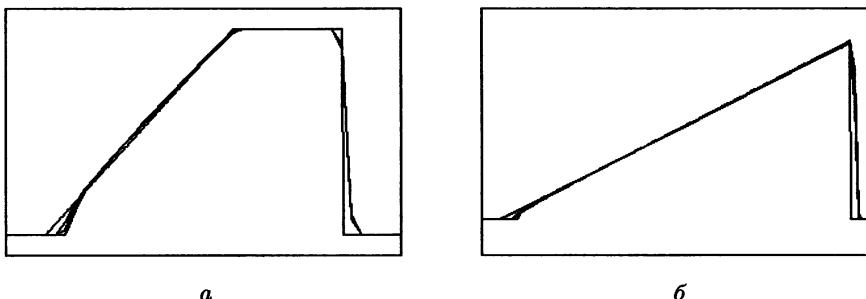


Рис. 14.12

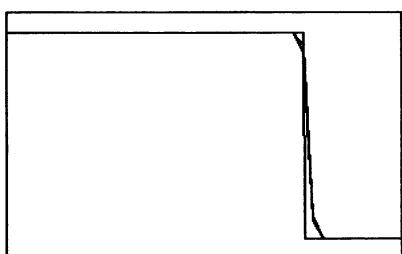


Рис. 14.13

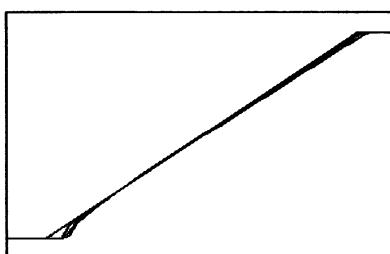


Рис. 14.14

Предложенная в [5] нелинейная монотонизованная схема К.И. Бабенко дает более высокое качество решения квазилинейного уравнения переноса во всем рассмотренном диапазоне чисел Куранта. Достоинством схемы является способность передавать разрывы решения с их размазыванием на наименьшее количество шагов по сравнению с другими рассмотренными схемами. Схема прыжкового переноса [54], несмотря на ее способность периодически точно воспроизводить численное решение, для некоторых начальных условий при определенных числах Куранта дает более низкое качество решения, что, в частности, связано с появлением ступенек, приводящих к сильному искажению решения.

#### 14.3.4. Решение квазилинейного уравнения переноса с помощью разрывного метода Галеркина

В завершение приведем результаты численного решения *квазилинейного уравнения переноса* с помощью метода иного класса: **разрывного метода Галеркина** (английская аббревиатура RKDG — Runge — Kutta discontinuous Galerkin method). Теория данного метода приведена в [211].

Представленные далее результаты опубликованы полностью в работе [42], посвященной тестированию разрывного метода Галеркина для численного решения квазилинейного уравнения переноса, а также его сравнению с другими известными схемами. Там же продолжены исследования, проведенные с использованием конечно-разностных методов для решения линейного и квазилинейного уравнений. Выполнено сравнение различных вариантов разрывного метода Галеркина и конечно-разностных схем, рассмотренных в [5]. Приведенная в [5] информация об ошибках численного решения позволяет сравнить качество различных схем. Сравнительный анализ ошибок численного решения показал, что метод RKDG дает более высокую точность решения в широком диапазоне изменения *чисел Куранта*, а также меньшее «размазывание» разрывов решений по сравнению с изученными конечно-разностными схемами.

Рассмотрим применение метода RKDG на примере задачи (14.2), записав ее в дивергентном виде:

$$u_t + (f(u))_x = 0, \quad x \in (0, l), \quad t \in (0, T]; \quad (14.57)$$

$$u(x, 0) = u_0(x), \quad x \in (0, l), \quad (14.58)$$

где  $f(u) = u^2/2$ . Здесь  $(0, l)$  — пространственный участок, которому принадлежит носитель финитных начальных данных и точного решения задачи на рассматриваемом временном промежутке.

На интервале  $(0, l)$  введем равномерную сетку с узлами  $\{x_{j+1/2}\}_{j=0}^N$ , обозначим  $I_j = (x_{j-1/2}, x_{j+1/2})$ ,  $h_j = x_{j+1/2} - x_{j-1/2}$ ,  $j = 1, 2, \dots, N$ .

Умножим уравнения (14.57) на произвольную гладкую функцию  $v(x)$  и проинтегрируем это произведение по  $I_j$ . После интегрирования по частям получим

$$\begin{aligned} & \int\limits_{I_j} \partial_t u(x, t) v(x) dx - \int\limits_{I_j} f(u(x, t)) \partial_x v(x) dx + \\ & + f(u(x_{j+1/2}, t)) v(x_{j+1/2}) - f(u(x_{j-1/2}, t)) v(x_{j-1/2}) = 0, \end{aligned} \quad (14.59)$$

$$\int\limits_{I_j} u(x, 0) v(x) dx = \int\limits_{I_j} u_0(x) v(x) dx.$$

Для каждого момента времени  $t \in [0, T]$  будем искать приближенное решение  $y$  как элемент конечномерного пространства.

$$V_h = V_h^k \equiv \left\{ v \in L^1(0, L) : v|_{I_j} \in P^k(I_j), j = 1, 2, \dots, N \right\},$$

где  $P^k(I_j)$  — пространство полиномов степени не выше  $k$  на интервале  $I_j$ .

Заменим гладкую функцию  $v$  произвольной пробной функцией, принадлежащей пространству  $V_h$ , а точное решение  $u$  — приближенным решением  $y$ . Функция  $y$  разрывна в точках  $x_{j+1/2}$ , поэтому необходимо заменить **физический поток**  $f(y(x_{j+1/2}, t))$  **численным потоком**  $H$ , зависящим от предельных значений функции  $y$  слева и справа от точки  $x_{j+1/2}$ , т. е.

$$H(y(x_{j+1/2}, t)) = H(y(x_{j+1/2}^-, t), y(x_{j+1/2}^+, t)).$$

Тогда уравнения (14.59) для всех  $j = 1, 2, \dots, N$  и  $v_h \in P^k(I_j)$  принимают следующий вид:

$$\int_{I_j} \partial_t y(x, t) v_h(x) dx - \int_{I_j} f(y(x, t)) \partial_x v_h(x) dx + \\ + H(y(x_{j+1/2}, t)) v_h(x_{j+1/2}^-) - H(y(x_{j-1/2}, t)) v_h(x_{j-1/2}^+) = 0, \quad (14.60)$$

$$\int_{I_j} y(x, 0) v_h(x) dx = \int_{I_j} u_0(x) v_h(x) dx.$$

Численный поток  $H$  должен быть монотонным и согласованным с физическим потоком  $f$ . Условия монотонности и варианты выбора формы потока приведены в [42, 211]. Часть из них описана в 15, а подробно для систем уравнений гиперболического типа подобные потоки представлены в [98].

Метод RKDG собственно и заключается в решении уравнений (14.60) методом Рунге — Кутты для различных видов монотонизованных потоков.

В работе [42] тестировался метод RKDG для кусочно-линейной и кусочно-квадратичной аппроксимаций, т. е. для  $k = 1, 2$ .

Для случая кусочно-квадратичной интерполяции по средним значениям с использованием так называемого TVDM-лимитера  $\Lambda\Pi_h^k$  (подробнее см. [42, 211]) графики точных и приближенных решений показаны на рис. 14.15–14.20. Все решения получены для числа Куранта  $\gamma = 0,5$ .

На рис. 14.15 показано решение для начального профиля «треугольник» при  $t = 6, 10, 70, 170$ .

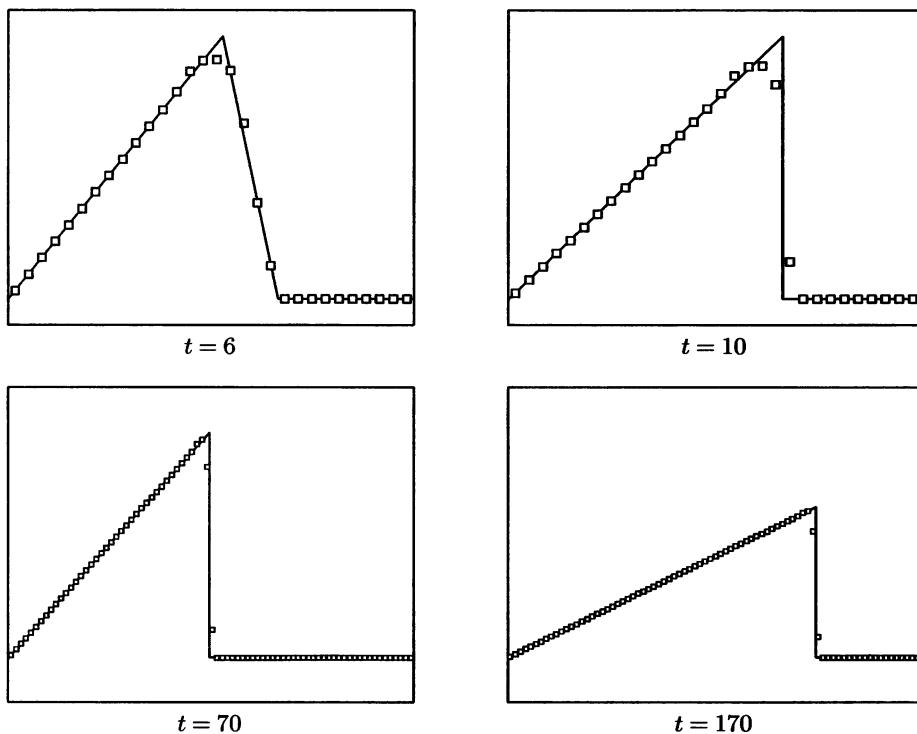


Рис. 14.15

Решение при начальном условии «прямоугольник» приведено на рис. 14.16 при  $t = 18, 40, 90, 160$ . Случаи «левого прямоугольника» при  $t = 60, 160$ , «правого прямоугольника» при  $t = 5, 20$ , «ступеньки вверх» при  $t = 10, 40$  и «ступеньки вниз» при  $t = 20, 80$  показаны соответственно на рис. 14.17–14.20.

Рассмотренный в работе [42] метод RKDG в большинстве случаев обеспечивает более высокое качество решения квазилинейного уравнения переноса по сравнению с другими известными методами [5]. «Размазывание» разрывов происходит на 2–3 шага.

Проведенное исследование показывает, что наиболее оптимальным монотонизатором является TVDM-«лимитер»  $\Lambda\Pi_h^1$ .

Следует отметить, что аппроксимация решения кусочно-линейными функциями оказывается в целом более эффективной, чем кусочно-квадратичная аппроксимация, так как она обеспечивает монотонность получаемой схемы.

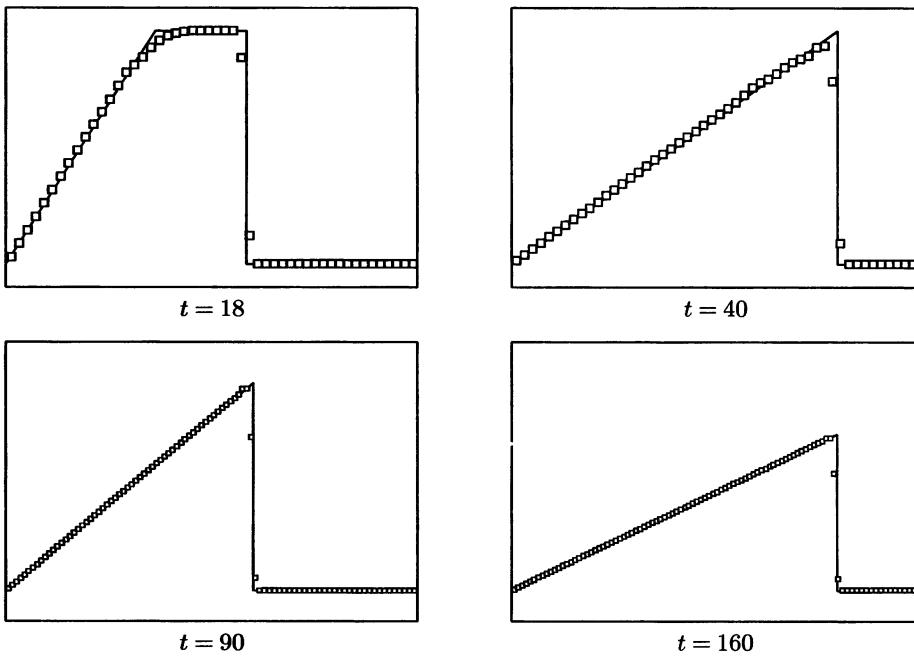


Рис. 14.16

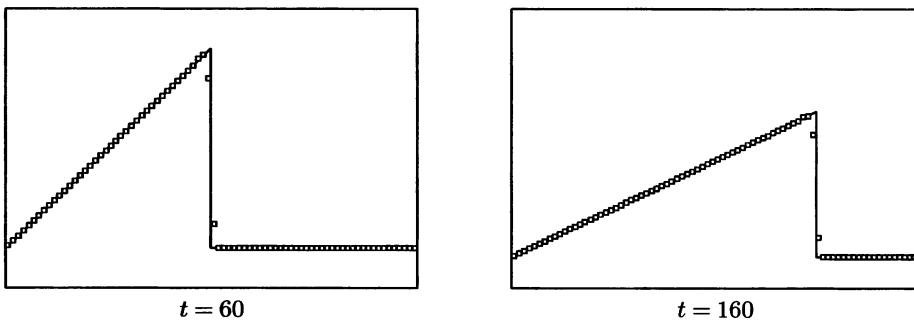


Рис. 14.17

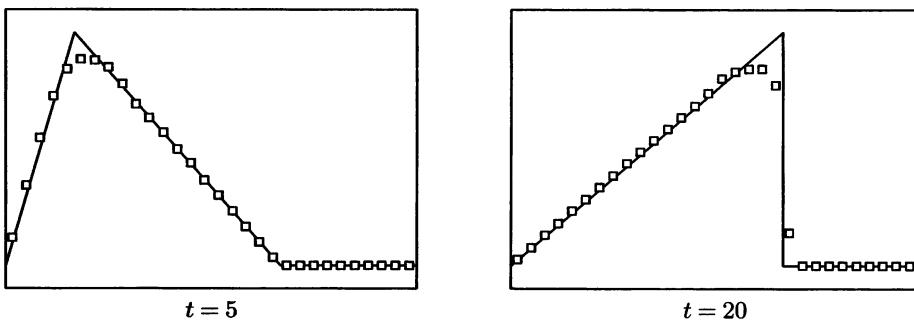


Рис. 14.18

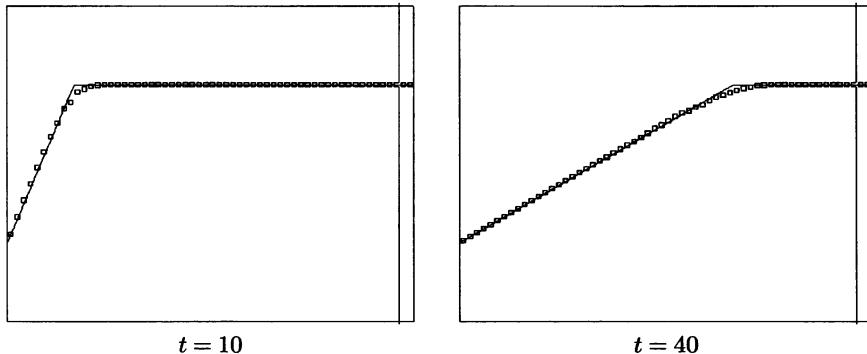


Рис. 14.19

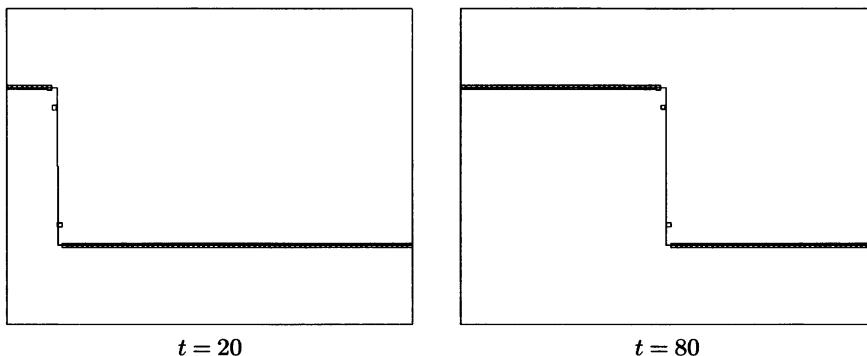


Рис. 14.20

#### 14.4. Двумерное линейное уравнение переноса

##### 14.4.1. Постановка задачи для линейного двумерного уравнения переноса. Тестовые задачи

Пусть в задаче Коши (14.3) для линейного двумерного уравнения переноса с финитным начальным условием

$$\rho(x, y, 0) = \rho_0(x, y) = \begin{cases} R(r, \varphi), & 0 \leq r \leq r_0; \\ 0, & r_0 < r, \end{cases}$$

где

$$r = \sqrt{(x - x_{00})^2 + (y - y_{00})^2};$$

$$\varphi = \begin{cases} \arccos \frac{x - x_{00}}{r}, & y \geq y_{00}; \\ 2\pi - \arccos \frac{x - x_{00}}{r}, & y < y_{00}; \end{cases} \quad r_0 = \text{const} > 0.$$

Скорости  $v_1(x, y, t)$ ,  $v_2(x, y, t)$  заданы следующим образом:

$$v_1(x, y, t) = \begin{cases} a_1(r)(y - y_0(t))r^{-1} \frac{da_2(t)}{dt} + \frac{dx_0(t)}{dt}, & 0 \leq r \leq r_1; \\ 0, & r_1 < r; \end{cases}$$

$$v_2(x, y, t) = \begin{cases} -a_1(r)(x - x_0(t))r^{-1} \frac{da_1(t)}{dt} + \frac{dy_0(t)}{dt}, & 0 \leq r \leq r_1; \\ 0, & r_1 < r. \end{cases}$$

Здесь  $x_0(0) = x_{00}$ ,  $y_0(0) = y_{00}$ ,  $a_2(0) = 0$ ,  $r_0 \leq r_1$ .

Хорошо видно, что при переносе в несжимаемой среде, в которой  $\partial v_1 / \partial x + \partial v_2 / \partial y = 0$ , начальный профиль решения сохраняется.

В данном случае задача (14.3) имеет следующее решение:

$$\rho(x, y, t) = \begin{cases} R(r(x, y, t), \varphi(x, y, t)), & 0 \leq r \leq r_0; \\ 0, & r_0 < r, \end{cases}$$

где

$$r(x, y, t) = \sqrt{(x - x_0(t))^2 + (y - y_0(t))^2};$$

$$\varphi(x, y, t) = a_1(r)a_2(t)r^{-1} + \begin{cases} \arccos \frac{x - x_0(t)}{r}, & y_0(t) \leq y; \\ 2\pi - \arccos x - x_0(t)r, & y < y_0(t). \end{cases}$$

Данный класс решений определен одной функцией  $R(r, \varphi)$ ,  $0 \leq r \leq r_0$ ,  $0 \leq \varphi \leq 2\pi$ , и четырьмя функциями  $a_1(r)$ ,  $0 \leq r \leq r_1$ ,  $a_2(t)$ ,  $y_0(t)$ ,  $x_0(t)$ ,  $t > 0$ , задающими поле скоростей. Функция  $R(r, \varphi)$  описывает начальный профиль. Функции  $x = x_0(t)$ ,  $y = y_0(t)$  определяют траекторию движения центра решения  $(x_0, y_0)$  в плоскости  $(x, y)$ . Функции  $a_1(r)$ ,  $a_2(t)$  определяют законы вращения профиля  $R(r, \varphi)$  относительно центра.

В работе [41] использованы три типа функций  $R(r, \varphi)$ :

«конус»:

$$R(r, \varphi) = H_c \left( 1 - \frac{r}{r_0} \right), \quad 0 \leq r \leq r_0;$$

«конус без сектора»:

$$R(r, \varphi) = \begin{cases} H_c \left( 1 - \frac{r}{r_0} \right), & \varphi \notin [\varphi_1, \varphi_2], \quad 0 \leq r \leq r_0; \\ 0, & \varphi \in (\varphi_1, \varphi_2), \end{cases}$$

где  $0 \leq \varphi_1 < \varphi_2 \leq 2\pi$ ;

буква «М»:

$$R(r, \varphi) = \begin{cases} 1, & (x, y) \in M; \\ 0, & (x, y) \notin M, \end{cases}$$

где  $M$  — объединение четырех множеств  $P_1, P_2, P_3, P_4$  на плоскости, определяемых следующими неравенствами:

$$P_1: x_1 \leq x \leq x_2, y_1 \leq y \leq y_5;$$

$$P_2: x_2 \leq x \leq x_3, y_3 - \frac{y_3 - y_2}{x_3 - x_2}(x - x_2) \leq y \leq y_5 - \frac{y_5 - y_4}{x_3 - x_2}(x - x_2);$$

$$P_3: x_3 \leq x \leq x_4, y_2 + \frac{y_3 - y_2}{x_4 - x_3}(x - x_3) \leq y \leq y_4 + \frac{y_5 - y_4}{x_4 - x_3}(x - x_3);$$

$$P_4: x_4 \leq x \leq x_5, y_1 \leq y \leq y_5.$$

Параметры  $x_i, y_i$  этой функции связаны неравенствами

$$x_1 < x_2 < x_3 < x_4 < x_5,$$

$$y_1 < y_2 < y_3 < y_4 < y_5.$$

Рассмотрены следующие закономерности движения центра решения: центр неподвижен; центр движется по прямой; центр движется по окружности радиусом  $r_2$ ,  $r_2 > r_0$ .

Предусмотрен ряд вариантов движения профиля относительно центра решения: фигура вращается как целое; периферия фигуры вращается быстрее центра ( $a_1(r)$  монотонно возрастает); периферия фигуры вращается медленнее центра ( $a_1(r)$  монотонно убывает).

При этом предполагается, что угловая скорость не зависит от времени, возрастает со временем, уменьшается со временем.

#### 14.4.2. Разностные схемы для численного решения линейного двумерного уравнения

Для численного решения задачи введем равномерную ортогональную по пространству пространственно-временную сетку.

Для решения двумерной задачи применим **метод расщепления** по координатам  $(x, y)$  [139]. При этом возникают одномерные задачи по каждому направлению. В [41, 68] проведена нелинейная монотонизация схемы К.И. Бабенко [8–10, 176] для одномерного *уравнения переноса* вида

$$\frac{\partial u}{\partial t} + \frac{\partial f(x, t, u)}{\partial x} = 0, \quad f(x, t, u) = a(x, t)u$$

с переменной скоростью  $a(x, t)$ . Хорошо известно [50], что для получения *корректной постановки* задачи для гиперболических уравнений на границах области необходимо задать столько граничных условий, сколько характеристик выходит из границы в рассматриваемую пространственную область. В связи с этим при переменной скорости возникают различные случаи. Эти случаи подробно разобраны в [41, 68] вплоть до получения конечных расчетных формул, с помощью которых и выполнены расчеты. Шаг по времени  $\tau$  выбран так, чтобы максимальное значение *числа Куранта*

$$\gamma = \max_{x,y} \left( \frac{\tau|v_1|}{h_x}, \frac{\tau|v_2|}{h_y} \right)$$

было ограничено сверху величиной  $\gamma_m$  ( $\gamma \leq \gamma_m$ ). Для всех приведенных расчетов  $\gamma_m = 0,25$ ,  $r_0 = 10$ . Пространственные шаги  $h_x$ ,  $h_y$  выбраны единичными. Область вычисления содержит  $200 \times 200$  узлов.

Для сравнения новой схемы [41, 68] с известными расчеты проведены также по *схеме с лимитерами* [35, 36], по *схеме с направленными разностями* и по *схеме П. Лакса* [64].

#### 14.4.3. Результаты расчетов для линейного двумерного уравнения переноса

Результаты численных расчетов представлены в [41] в виде рисунков и таблиц. На каждом рисунке продемонстрировано численное решение  $\rho = \rho(x, y, t)$  (справа) и его линии уровня (слева) на конечный момент времени  $t_{\text{fin}}$  для определенной схемы, определенного начального условия  $\rho_0(x, y)$ , поля скоростей  $v_1(x, y, t)$ ,  $v_2(x, y, t)$  и заданного  $\gamma_m$ . Также указаны максимальное  $u_{\text{max}}$  и минимальное  $u_{\text{min}}$  значения решения на конечный момент времени (рис. 14.21–14.27).

Численное решение, полученное по монотонизированной схеме К.И. Бабенко, на конечный момент времени «размазывается» на небольшое число узлов сетки (см. рис. 14.21–14.23). Наилучшим образом сохраняется форма начального профиля для третьего начального условия (буква «М»). Амплитуда численного решения практически неизменна. Для двух других начальных условий характерно некоторое понижение амплитуды численного решения.

Во всех трех случаях расчет проводился до  $t_{\text{fin}} = 480$  при  $\gamma_m = 0,25$ . Начальный профиль перемещался на 17 своих длин (без вращения), траектория движения представляла собой ромб. Рис. 14.21–14.23 соответствуют начальным профилям «конус», «конус без сектора» и «М».

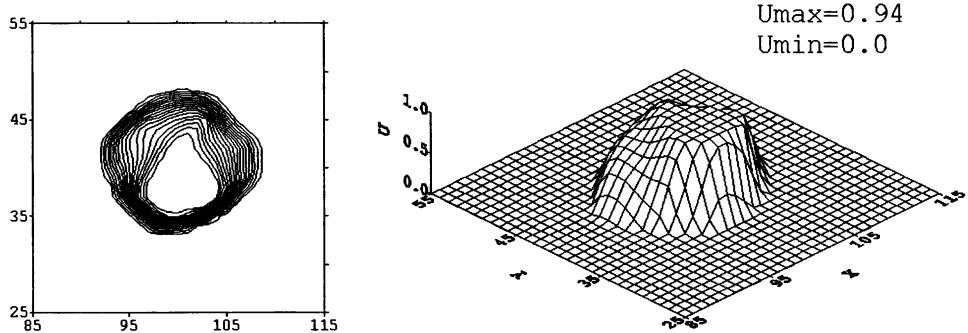


Рис. 14.21

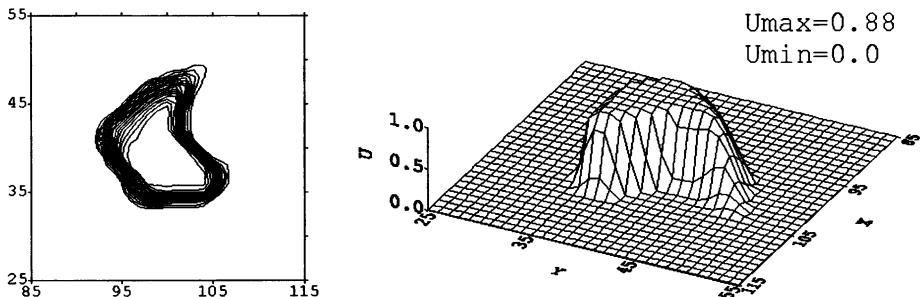


Рис. 14.22

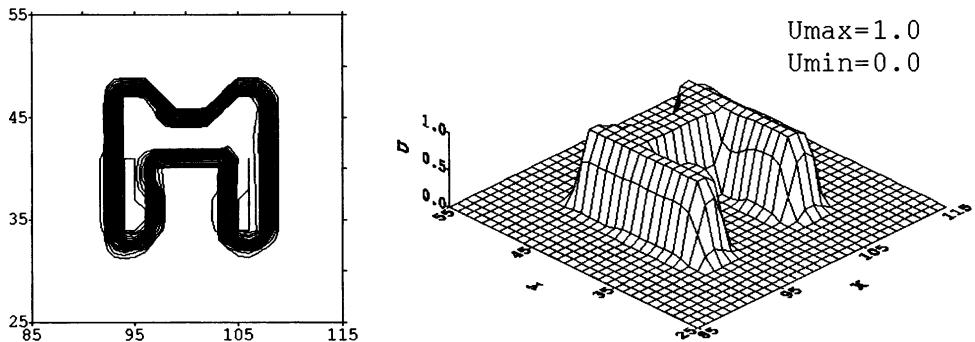


Рис. 14.23

Для численного решения, полученного по схеме с «лимитерами» [35, 36], характерно некоторое понижение амплитуды, сильнее выраженное, чем в случае схемы К.И. Бабенко (см. рис. 14.24–14.26). Линии уровня начальных профилей «конус» и «конус без сектора» приобретают более ярко выраженную «квадратную» форму по сравнению с начальной формой. Наилучшим образом передано численное решение задачи с начальным профилем «М».

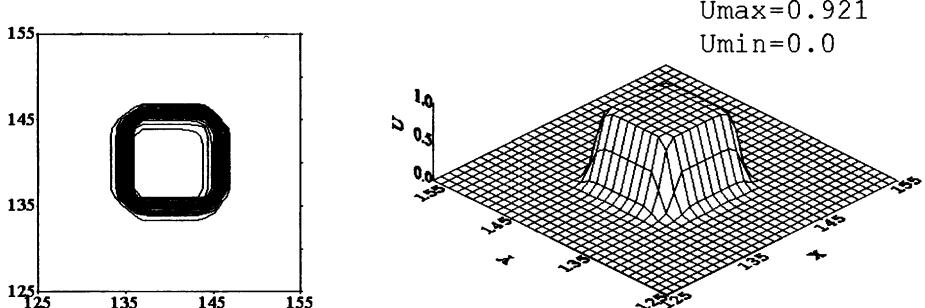


Рис. 14.24

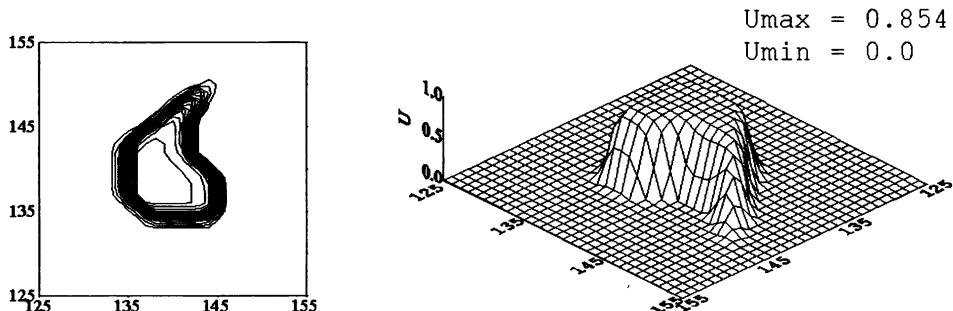


Рис. 14.25

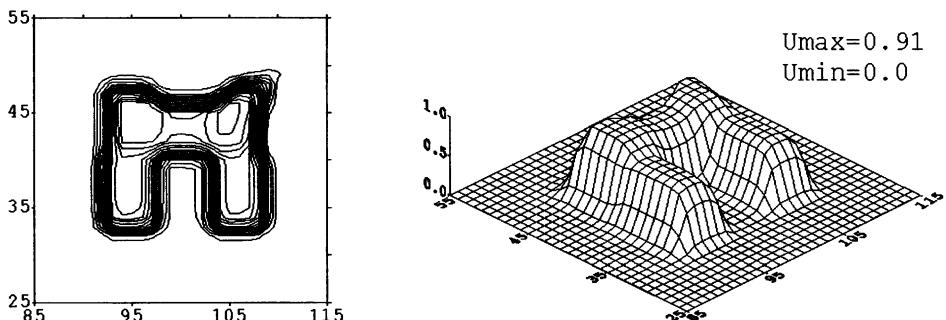


Рис. 14.26

В первых двух случаях (см. рис. 14.24 и 14.25) профиль перемещался без вращения на 9 своих длин ( $t_{\text{fin}} = 120$  при  $\gamma_m = 0,25$ ) по диагонали квадрата, образующего расчетную область.

В последнем случае (см. рис. 14.26) начальный профиль переместился без вращения на 17 своих длин ( $t_{\text{fin}} = 480$ ,  $\gamma_m = 0,25$ ) по траектории в виде ромба.

При любых формах поверхности начального условия и любых способах задания поля скоростей с течением времени численное решение,

полученное по схеме с направленными разностями, «размазывается» на достаточно большое количество интервалов разностной сетки и его амплитуда уменьшается. На финальный момент времени численное решение для любого допустимого значения  $\gamma$  приобретает характерный колоколообразный вид (рис. 14.27), не имеющий ничего общего с формой начального условия  $\rho_0(x, y)$ .

Решение, показанное на рис. 14.27, соответствует начальному профилю «М»,  $\gamma_m = 0,25$  и моменту времени  $t_{\text{fin}} = 120$ . Начальный профиль переместился без вращения на 9 своих длин.

Качество численного решения, полученного по схеме П. Лакса [64], близко по своим характеристикам к качеству решения, полученного по схеме с направленными разностями (начальный профиль «размазывается», амплитуда решения уменьшается). Кроме этого появляются характерные коротковолновые осцилляции. Сравнение норм ошибок показывает, что схема П. Лакса дает худший результат по всем показателям.

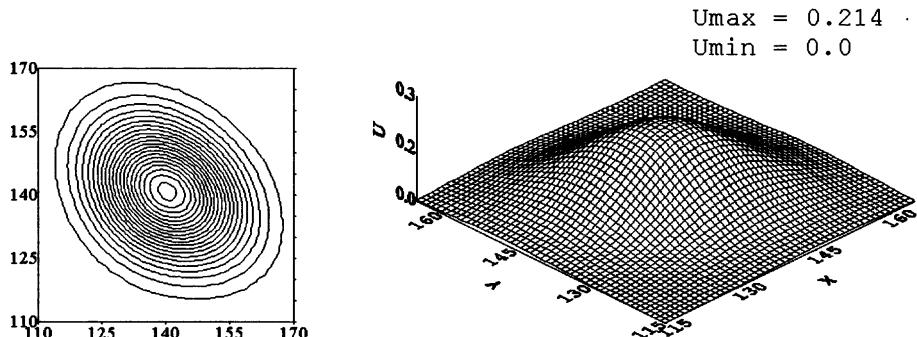


Рис. 14.27

Таким образом, монотонизованная схема К.И. Бабенко («квадрат») показала наилучшие результаты среди всех рассмотренных схем для данной задачи (при движении начального профиля без вращения), в особенности в случае негладкого начального профиля «М». Решения, полученные по схеме с направленными разностями и схеме П. Лакса, существенно уступают по точности решениям, полученным с помощью новой монотонизированной схемы К.И. Бабенко и схемы с «лимитерами».

# 15. ЧИСЛЕННОЕ РЕШЕНИЕ УРАВНЕНИЙ ГАЗОВОЙ ДИНАМИКИ

Кратко описаны уравнения газовой динамики. Представлены наиболее популярные современные конечно-разностные схемы, применяемые в вычислительной газовой динамике, проведено их теоретическое и численное исследование. При этом подробно описаны схемы Роу и Роу — Эйнфельдта — Ошера. Последняя хорошо зарекомендовала себя при решении трехмерных задач астрофизики.

## 15.1. Уравнения газовой динамики

*Уравнения газовой динамики* [8, 98, 132, 135, 155, 176 и др.], или *уравнения Эйлера*, в приближении сплошной среды имеют вид

$$\frac{\partial \rho}{\partial t} + \mathbf{v} \operatorname{grad} \rho + \rho \operatorname{div} \mathbf{v} = 0, \quad (15.1)$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} + \frac{1}{\rho} \operatorname{grad} P = \frac{1}{\rho} \mathbf{F}_e, \quad (15.2)$$

$$\frac{\partial \varepsilon}{\partial t} + \mathbf{v} \operatorname{grad} \varepsilon + \frac{P}{\rho} \operatorname{div} \mathbf{v} = 0. \quad (15.3)$$

Для замыкания к ним необходимо добавить *уравнение состояния*  $P = P(\rho, \varepsilon)$  (например, уравнение идеального газа  $P = (\gamma - 1)\varepsilon\rho$ ). В общем случае также необходимо добавить дополнительные уравнения и (или) члены в правую часть исходных уравнений, описывающие эффекты вязкости, горения, диффузационного переноса вещества и радиации, эволюцию магнитного поля и его влияние на течение вещества, т. е. магнитную гидродинамику, эволюцию поля излучения и его влияние на течение вещества, т. е. радиационную газовую динамику, и самогравитацию.

Здесь и далее  $\rho$  — плотность;  $\mathbf{v}$  — вектор скорости с декартовыми компонентами  $(u, v, w)$ ;  $P$  — давление;  $E = \varepsilon + \frac{\mathbf{v}^2}{2}$  — удельная полная энергия (на единицу массы, называемая также массовой плотностью);  $\varepsilon$  — удельная внутренняя энергия (на единицу массы);  $h^* = \varepsilon + \frac{P}{\rho} + \frac{\mathbf{v}^2}{2}$  — удельная полная энталпия (на единицу массы);  $\mathbf{F}_e$  — удельная внешняя сила (на единицу объема).

Следует отметить, что уравнения газовой динамики можно также вывести феноменологически, постулировав выполнение **законов сохранения** в интегральной форме. Так, из **закона сохранения массы** получаем уравнение

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0. \quad (15.4)$$

**Законы сохранения импульса и энергии** аналогично выводятся из интегральных соотношений, из которых следуют дифференциальные законы сохранения импульса

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \operatorname{div} \mathbf{T} = \mathbf{F}_e, \quad (15.5)$$

где

$$\mathbf{T} = \begin{pmatrix} \rho u^2 + P & \rho uv & \rho uw \\ \rho uv & \rho v^2 + P & \rho vw \\ \rho uw & \rho vw & \rho w^2 + P \end{pmatrix},$$

и **энергии**

$$\frac{\partial \rho E}{\partial t} + \operatorname{div}(\rho h^* \mathbf{v}) = \mathbf{F}_e \cdot \mathbf{v}. \quad (15.6)$$

Система уравнений (15.1)–(15.3) называется **уравнениями газовой динамики в недивергентной форме**, система уравнений (15.4)–(15.6) называется **уравнениями газовой динамики в консервативной (дивергентной) форме**. Дивергентные уравнения (15.4)–(15.6) могут быть преобразованы и записаны в **интегральной форме**:

$$\frac{\partial}{\partial t} \int_V dV = 0, \quad (15.7)$$

$$\frac{\partial}{\partial t} \int_V \rho dV + \int_{\Sigma} \rho (\mathbf{v} \cdot \mathbf{n}) ds = 0, \quad (15.8)$$

$$\frac{\partial}{\partial t} \int_V \rho \mathbf{v} dV + \int_{\Sigma} (\rho \mathbf{v} (\mathbf{v} \cdot \mathbf{n}) + P \mathbf{n}) ds = \int_V \mathbf{F}_e dV, \quad (15.9)$$

$$\frac{\partial}{\partial t} \int_V \rho E dV + \int_{\Sigma} \rho h^* (\mathbf{v} \cdot \mathbf{n}) ds = \int_V (\mathbf{F}_e \cdot \mathbf{v}) dV. \quad (15.10)$$

Эти уравнения не требуют предположения о дифференцируемости и даже непрерывности функций  $\rho$ ,  $\mathbf{v}$  и других и могут быть использованы для описания газодинамических течений с *ударными волнами* и *контактными разрывами*.

Остановимся на уравнении (15.7), которое выражает, что все интегральные соотношения записываются для объема с постоянным положением в пространстве. Это соответствует случаю так называемых **эйлеровых переменных**, когда геометрические координаты не меняются со временем ( $\partial \mathbf{r}/\partial t = 0$ ) и дифференцирование по  $t$  производится при постоянном радиус-векторе  $\mathbf{r}$ . Альтернативный подход, опирающийся на **лагранжевые переменные**, в котором интегральные соотношения записываются для «жидкой» частицы, движущейся вместе с веществом, будет рассмотрен ниже.

Рассмотрим еще одно важное понятие — **гиперболичность системы уравнений газовой динамики**. Для случая декартовых координат система уравнений газовой динамики в консервативной (дивергентной) форме (15.4)–(15.6) может быть записана как

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} + \frac{\partial \mathbf{H}}{\partial z} = \mathbf{R}.$$

Конкретные выражения для  $\mathbf{q}$ ,  $\mathbf{F}$ ,  $\mathbf{G}$ ,  $\mathbf{H}$  будут приведены в 15.2, здесь же только отметим, что система уравнений газовой динамики является гиперболической, т. е. матрицы Якоби  $\partial \mathbf{F}/\partial \mathbf{q}$ ,  $\partial \mathbf{G}/\partial \mathbf{q}$  и  $\partial \mathbf{H}/\partial \mathbf{q}$  имеют действительные собственные числа и могут быть приведены к диагональному виду. Это означает, что данные матрицы имеют полный набор *линейно независимых собственных векторов*. Свойство гиперболичности принципиально важно для построения конечно-разностных схем, используемых при решении газодинамических уравнений [50, 98, 135].

Рассмотрим *уравнения газовой динамики в лагранжевых переменных*, когда геометрические координаты меняются со временем вместе с движением вещества  $D\mathbf{r}/Dt = \mathbf{v}$ . Неизменной в этом случае остается метка «жидкой» частицы, в качестве которой можно принять, например,  $\mathbf{r}_0$  — положение «жидкой» частицы при  $t = 0$ . Оператор  $D/Dt$  называется лагранжевой производной по времени, т. е. частной производной по  $t$  при постоянном  $\mathbf{r}_0$ . Рассмотрим для простоты случай, когда внешние силы отсутствуют. Законы сохранения в интегральной форме могут быть выведены из (15.7)–(15.10) с помощью соотношений (см., например, [124, 155])

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla), \quad \frac{D\Delta}{Dt} = \Delta \operatorname{div} \mathbf{v},$$

где  $\Delta$  — детерминант якобиана преобразования от  $\mathbf{r}_0$  к  $\mathbf{r}$ ,  $\Delta = \left| \frac{\partial \mathbf{r}}{\partial \mathbf{r}_0} \right|$ . В результате получаем уравнения газовой динамики в лагранжевых

переменных в интегральной форме:

$$\frac{D}{Dt} \int_V \rho dV = \frac{D}{Dt} \int_M dm = 0, \quad (15.11)$$

$$\frac{D}{Dt} \int_V dV = \frac{D}{Dt} \int_M \rho^{-1} dm = \frac{D}{Dt} \int_V \Delta dV_0 = \int_M \rho^{-1} \operatorname{div} \mathbf{v} dm, \quad (15.12)$$

$$\begin{aligned} \frac{D}{Dt} \int_V \rho \mathbf{v} dV &= \frac{D}{Dt} \int_M \mathbf{v} dm = \\ &= \int_V \left( -\operatorname{div} \mathbf{T} + (\mathbf{v} \cdot \nabla)(\rho \mathbf{v}) + \rho \mathbf{v} \operatorname{div} \mathbf{v} + \mathbf{F}_e \right) dV = \\ &= \int_V (-\operatorname{grad} P + \mathbf{F}_e) dV = \int_M \rho^{-1} (-\operatorname{grad} P + \mathbf{F}_e) dm, \end{aligned} \quad (15.13)$$

$$\begin{aligned} \frac{D}{Dt} \int_V \rho E dV &= \frac{D}{Dt} \int_M E dm = \int_V (-\operatorname{div}(P \mathbf{v}) + (\mathbf{F}_e \cdot \mathbf{v})) dV = \\ &= \int_M \rho^{-1} (-\operatorname{div}(P \mathbf{v}) + (\mathbf{F}_e \cdot \mathbf{v})) dm. \end{aligned} \quad (15.14)$$

Здесь  $dV_0$  — дифференциал объема при  $t = 0$ ,  $dm = \rho dV$ . Из интегральных уравнений (15.11)–(15.14) следуют дифференциальные уравнения газовой динамики в лагранжевых переменных в дивергентной форме:

$$\frac{D}{Dt} \rho^{-1} - \rho^{-1} \operatorname{div} \mathbf{v} = 0, \quad (15.15)$$

$$\frac{D}{Dt} \mathbf{v} + \rho^{-1} \operatorname{grad} P = \rho^{-1} \mathbf{F}_e, \quad (15.16)$$

$$\frac{D}{Dt} E + \rho^{-1} \operatorname{div}(P \mathbf{v}) = \rho^{-1} (\mathbf{F}_e \cdot \mathbf{v}), \quad (15.17)$$

к которым необходимо прибавить уравнение движения эйлеровых переменных  $\mathbf{r}$

$$\frac{D\mathbf{r}}{Dt} = \mathbf{v}$$

и уравнение состояния. Уравнение энергии часто записывают в недивергентном виде

$$\frac{D\varepsilon}{Dt} + \frac{P}{\rho} \operatorname{div} \mathbf{v} = \frac{D\varepsilon}{Dt} + P \frac{D}{Dt} \rho^{-1} = 0. \quad (15.18)$$

Приведем дополнительно *изотермические уравнения газовой динамики* Поскольку температура является постоянной, то, например, из закона Менделеева — Клапейрона  $T = P/R\rho$  следует уравнение

состояния  $P = R\rho T_0 \equiv c_T^2 \rho$ , где  $R$  — газовая постоянная. В результате *законы сохранения массы и импульса* описываются системой изотермических уравнений

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0, \quad \frac{\partial \rho \mathbf{v}}{\partial t} + \operatorname{div} \mathcal{T} = \mathbf{F}_e,$$

где

$$\mathcal{T} = \begin{pmatrix} \rho u^2 + c_T^2 \rho & \rho uv & \rho uw \\ \rho uv & \rho v^2 + c_T^2 \rho & \rho vw \\ \rho uw & \rho vw & \rho w^2 + c_T^2 \rho \end{pmatrix}.$$

Уравнение энергии в этом случае не используется, так как энергия не сохраняется вследствие излучения и не входит в уравнения состояния. На практике часто вместо изотермических уравнений решают газодинамические уравнения, включающие уравнение состояния идеального газа, в котором отношение теплоемкостей  $\gamma$  берут очень близким к единице (например,  $\gamma = 1,001$ ). Согласно известной формуле из статистической физики  $\gamma = (\alpha + 2)/\alpha$ , близость  $\gamma$  к единице соответствует очень большому числу  $\alpha$  степеней свободы частиц, из которых состоит газ. В этом случае, если газ нагревается (например, ударной волной), то нагрев распространяется по всем степеням свободы и оказывает лишь незначительное влияние на давление и температуру.

## 15.2. Разностная схема Рой — Эйнфельдта — Ошера

*Уравнения газовой динамики* не могут быть решены аналитически за исключением некоторых специальных случаев, которых чрезвычайно немного. Следовательно, для их решения необходимо использовать численные методы, наиболее популярными среди которых являются конечно-разностные. Конечно-разностные методы отличаются надежностью, простотой при программировании, эффективностью и возможностью модификации и расширения при смене вычислительного алгоритма или при включении дополнительных физических процессов.

За последние полвека разработано большое количество разностных схем для численного решения задач газовой динамики. Одной из наиболее интересных и важных задач по моделированию течений сжимаемого газа является исследование течений с *ударными волнами и контактными разрывами*. В гравитационной газовой динамике и многих других прикладных задачах в общей структуре течения, как правило, присутствуют и сильные ударные волны, и контактные разрывы. Кроме того, в большинстве исследований общая точность численного моделирования задач гравитационной газовой динамики в первую оче-

редь зависит от того, насколько хорошо разрешены газодинамические разрывы.

Впервые разностная схема для расчета течений с разрывами предложена в 1950 г. фон Нейманом и Рихтмайером [259]. В процессе развития вычислительной математики для исследования разрывных течений были разработаны более точные и (или) более удобные схемы. Здесь мы рассмотрим ряд разностных схем сквозного счета, используемых для решения задач газовой динамики в *эйлеровых переменных*. Эти схемы исследованы как теоретически, так и с помощью одно- и двумерных тестов в работе [94].

Уравнения Эйлера представляют собой систему нелинейных гиперболических уравнений. С математической точки зрения наиболее интересной особенностью гиперболических систем является возможность появления *разрывных решений* даже из гладких начальных данных. Эти разрывы являются математической идеализацией резких градиентов, возникающих в гладких решениях полных уравнений Навье — Стокса в областях быстрого изменения параметров на расстояниях, много меньших, чем характерные размеры задачи. Именно использование математических особенностей гиперболических уравнений позволяет строить эффективные численные методы, несмотря на то, что получающиеся численные решения должны быть почти разрывными. Кроме того, для решения гиперболических систем можно применять явные методы, что часто дает существенный выигрыш во времени.

Исторически первые схемы, используемые для решения уравнений газовой динамики, либо давали очень размазанные разрывы (например, *схема Лакса — Фридрихса* [230, 250]), либо приводили к возникновению сильных осцилляций за ударной волной (например, *схема Лакса — Вендроффа* [251]). Очевидно, что общее направление развития численных методов для решения уравнений газовой динамики лежит на пути уменьшения численной вязкости при сохранении монотонности схемы. Для уменьшения численной вязкости схемы можно использовать характеристические свойства гиперболической системы. Впервые эти характеристические свойства были использованы в *схеме Куранта — Изаксона — Риса* [216], обладающей минимальной вязкостью из всех линейных монотонных схем первого порядка аппроксимации, для решения линейной системы гиперболических уравнений. Построение монотонных схем более высокого порядка аппроксимации входит в противоречие с *теоремой Годунова* [49], утверждающей, что из трех свойств разностной схемы для решения линейной системы гиперболического типа — линейности, монотонности, аппроксимации с поряд-

ком выше первого — одновременно могут иметь место только два. Для преодоления этого противоречия предложены схемы линейной гибридизации, представляющие собой суперпозицию схемы повышенного порядка аппроксимации, которая дает высокую точность в областях гладкости, и схемы первого порядка с достаточной вязкостью, обеспечивающей монотонность около скачков [178]. Дальнейшее развитие этой идеи привело к появлению схем с ограничителями антидиффузионных потоков, являющимися нелинейными функциями анализаторов гладкости [34, 35, 86, 209, 237, 284, 286, 287].

Важным свойством разностной схемы является консервативность, т. е. выполнение *законов сохранения* в разностном виде. Известно (см., например, [155, 171, 248, 252]), что схемы, не обладающие этим свойством, могут давать решения, весьма далекие от истинного, в частности, ударные волны, движущиеся с неправильными скоростями. Для обеспечения консервативности схемы естественно использовать запись схемы в потоковом виде. В этом случае искомые решения — функции — получаются в результате разностного дифференцирования функции соответствующего потока. Для вычисления потоков на границах разностных ячеек в 1959 г. С.К. Годуновым [49, 51] был предложен метод, основанный на решении *задачи о распаде произвольного разрыва* (или *задачи Римана*). Таким образом, был сделан важнейший шаг от схем, базирующихся на разложении в ряд Тейлора и основанных на предположении, что решение гладкое, к схемам, основанным на взаимодействии газодинамических разрывов. Дальнейшее развитие этого метода привело к схемам годуновского типа с приближенным решением *задачи о распаде разрыва* [238, 262, 274]. Далее будет рассмотрена одна из таких схем — *схема Рой с модификацией Эйнфельдта* [223] и с антидиффузионными потоками в форме Ошера [209]. Аналитически и численно будет проведено сравнение четырех разностных схем: *Лакса — Фридрихса*, *Роя*, *Роя — Эйнфельдта* и *Роя — Эйнфельдта — Ошера* повышенного порядка аппроксимации. Впервые такое сравнение выполнено в работе [94].

### 15.2.1. Основные уравнения

Рассмотрим следующие уравнения:

1) линейное *уравнение переноса*

$$\frac{\partial q}{\partial t} + a \frac{\partial q}{\partial x} = 0, \quad (15.19)$$

здесь значение  $a$  может быть как положительным, так и отрицательным;

2) произвольную систему линейных гиперболических уравнений

$$\frac{\partial \mathbf{q}}{\partial t} + \mathcal{A} \frac{\partial \mathbf{q}}{\partial x} = 0 \quad (15.20)$$

с гиперболической матрицей  $\mathcal{A}$ , т. е. имеющей действительные собственные числа  $\lambda^m$  и полный набор левых  $\mathbf{l}^m$  (являющихся строками матрицы  $\mathcal{L}$ ) и правых  $\mathbf{r}^m$  (являющихся столбцами матрицы  $\mathcal{R}$ ) собственных векторов:

$$\mathcal{A} = \mathcal{R} \Lambda \mathcal{L}, \quad \Lambda = \text{diag}(\lambda^1, \lambda^2, \dots, \lambda^N), \quad \mathcal{L} \mathcal{R} = \mathcal{I}$$

(в линейной алгебре такие матрицы называются матрицами простой структуры с действительными собственными числами [31]),  $\mathcal{I}$  — единичная матрица;

3) систему уравнений Эйлера для идеального газа в одномерном случае

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = 0, \quad (15.21)$$

где

$$\begin{aligned} \mathbf{q} &= (\rho, \rho u, \rho E)^T; \quad \mathbf{F} = (\rho u, \rho u^2 + P, \rho u h^*)^T; \\ E &= \varepsilon + \frac{u^2}{2}; \quad P = (\gamma - 1)\varepsilon\rho; \quad h^* = \varepsilon + \frac{P}{\rho} + \frac{u^2}{2}, \end{aligned}$$

и в двумерном случае

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = 0, \quad (15.22)$$

где

$$\begin{aligned} \mathbf{q} &= (\rho, \rho u, \rho v, \rho E)^T; \quad \mathbf{F} = (\rho u, \rho u^2 + P, \rho uv, \rho uh^*)^T; \\ \mathbf{G} &= (\rho v, \rho uv, \rho v^2 + P, \rho vh^*)^T; \\ \mathbf{v} &= (u, v)^T; \quad E = \varepsilon + \frac{\mathbf{v}^2}{2}; \quad h^* = \varepsilon + \frac{P}{\rho} + \frac{\mathbf{v}^2}{2}. \end{aligned}$$

Для системы (15.21) матрица

$$\mathcal{A} = \frac{\partial \mathbf{F}}{\partial \mathbf{q}} = \begin{pmatrix} 0 & 1 & 0 \\ -(3-\gamma) \frac{u^2}{2} & (3-\gamma)u & \gamma-1 \\ u \left( -h^* + (\gamma-1) \frac{u^2}{2} \right) & h^* - (\gamma-1)u^2 & \gamma u \end{pmatrix} \quad (15.23)$$

является гиперболической:  $\mathcal{A} = \mathcal{R}\Lambda\mathcal{L}$ , где

$$\mathcal{R} = \begin{pmatrix} 1 & 2 & 1 \\ u - c & 2u & u + c \\ h^* - uc & u^2 & h^* + uc \end{pmatrix}; \quad \Lambda = \text{diag}(u - c, u, u + c);$$

$$\mathcal{L} = \frac{\gamma - 1}{2c^2} \begin{pmatrix} \frac{u^2}{2} + \frac{uc}{\gamma - 1} & -u - \frac{c}{\gamma - 1} & 1 \\ \frac{c^2}{\gamma - 1} - \frac{u^2}{2} & u & -1 \\ \frac{u^2}{2} - \frac{uc}{\gamma - 1} & -u + \frac{c}{\gamma - 1} & 1 \end{pmatrix}; \quad c = \sqrt{\frac{P}{\rho}}.$$

Аналогично для системы (15.22) матрицы

$$\mathcal{A} = \frac{\partial \mathbf{F}}{\partial \mathbf{q}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -(3 - \gamma) \frac{\mathbf{v}^2}{2} + v^2 & (3 - \gamma)u & -(\gamma - 1)v & \gamma - 1 \\ -uv & v & u & 0 \\ u \left( -h^* + (\gamma - 1) \frac{\mathbf{v}^2}{2} \right) & h^* - (\gamma - 1)u^2 & -(\gamma - 1)uv & \gamma u \end{pmatrix}$$

и

$$\mathcal{B} = \frac{\partial \mathbf{G}}{\partial \mathbf{q}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -uv & v & u & 0 \\ -(3 - \gamma) \frac{\mathbf{v}^2}{2} + u^2 & -(\gamma - 1)u & (3 - \gamma)v & \gamma - 1 \\ v \left( -h^* + (\gamma - 1) \frac{\mathbf{v}^2}{2} \right) & -(\gamma - 1)uv & h^* - (\gamma - 1)v^2 & \gamma v \end{pmatrix}$$

также гиперболические. Матрица  $\mathcal{A}$  имеет разложение  $\mathcal{A} = \mathcal{R}_{\mathcal{A}}\Lambda_{\mathcal{A}}\mathcal{L}_{\mathcal{A}}$ , где

$$\mathcal{R}_{\mathcal{A}} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ u - c & u & u & u + c \\ v & v + c & v - c & v \\ h^* - uc & \frac{\mathbf{v}^2}{2} + vc & \frac{\mathbf{v}^2}{2} - vc & h^* + uc \end{pmatrix};$$

$$\Lambda_{\mathcal{A}} = \text{diag}(u - c, u, u, u + c);$$

$$\mathcal{L}_A = \frac{\gamma - 1}{2c^2} \begin{pmatrix} \frac{\mathbf{v}^2}{2} + \frac{uc}{\gamma - 1} & -u - \frac{c}{\gamma - 1} & -v & 1 \\ \frac{c^2}{\gamma - 1} - \frac{\mathbf{v}^2}{2} - \frac{vc}{\gamma - 1} & u & v + \frac{c}{\gamma - 1} & -1 \\ \frac{c^2}{\gamma - 1} - \frac{\mathbf{v}^2}{2} + \frac{vc}{\gamma - 1} & u & v - \frac{c}{\gamma - 1} & -1 \\ \frac{\mathbf{v}^2}{2} - \frac{uc}{\gamma - 1} & -u + \frac{c}{\gamma - 1} & -v & 1 \end{pmatrix}.$$

Матрица  $\mathcal{B}$  имеет разложение  $\mathcal{B} = \mathcal{R}_B \Lambda_B \mathcal{L}_B$ , где

$$\mathcal{R}_B = \begin{pmatrix} 1 & 1 & 1 & 1 \\ u & u+c & u-c & u \\ v-c & v & v & v+c \\ h^* - vc & \frac{\mathbf{v}^2}{2} + uc & \frac{\mathbf{v}^2}{2} - uc & h^* + vc \end{pmatrix};$$

$$\Lambda_B = \text{diag}(v - c, v, v, v + c);$$

$$\mathcal{L}_B = \frac{\gamma - 1}{2c^2} \begin{pmatrix} \frac{\mathbf{v}^2}{2} + \frac{vc}{\gamma - 1} & -u & -v - \frac{c}{\gamma - 1} & 1 \\ \frac{c^2}{\gamma - 1} - \frac{\mathbf{v}^2}{2} - \frac{uc}{\gamma - 1} & u + \frac{c}{\gamma - 1} & v & -1 \\ \frac{c^2}{\gamma - 1} - \frac{\mathbf{v}^2}{2} + \frac{uc}{\gamma - 1} & u - \frac{c}{\gamma - 1} & v & -1 \\ \frac{\mathbf{v}^2}{2} - \frac{vc}{\gamma - 1} & -u & -v + \frac{c}{\gamma - 1} & 1 \end{pmatrix},$$

а  $c$  — тот же параметр, что и в одномерном случае.

### 15.2.2. Схема Лакса — Фридрихса

Запишем скалярное уравнение переноса (15.19) в дивергентном виде:

$$\frac{\partial q}{\partial t} + \frac{\partial F}{\partial x} = 0,$$

где  $F = aq$ , и рассмотрим для него **схему Лакса — Фридрихса** [230, 250] на равномерной сетке с шагами  $\tau, h$  по времени и пространству соответственно:

$$\frac{\hat{q}_i - q_i}{\tau} + \frac{F_{i+1/2} - F_{i-1/2}}{h} = 0,$$

где

$$F_{i+1/2} = \frac{F_i + F_{i+1}}{2} - \frac{\nu}{2}(q_{i+1} - q_i), \quad \nu = |a|.$$

Эта схема представляет собой **схему с направленными разностями** (левой или правой) для обоих вариантов знака скорости. Записав данную схему в виде

$$\hat{q}_i = \frac{\tau(|a| + a)}{2h} q_{i-1} + \left(1 - \frac{\tau|a|}{h}\right) q_i + \frac{\tau(|a| - a)}{2h} q_{i+1},$$

можно легко увидеть, что при выполнении **критерия Куранта** [215]  $\frac{\tau|a|}{h} \leq 1$  она удовлетворяет принципу максимума [139] и, следовательно, не увеличивает **общую вариацию** TV решения:

$$\begin{aligned} \text{TV}(\hat{q}) &= \sum_i |\hat{q}_i - \hat{q}_{i-1}| \leqslant \\ &\leqslant \frac{\tau(|a| + a)}{2h} \text{TV}(q) + \left(1 - \frac{\tau|a|}{h}\right) \text{TV}(q) + \frac{\tau(|a| - a)}{2h} \text{TV}(q) = \text{TV}(q). \end{aligned}$$

Запишем линейную гиперболическую систему (15.20) в дивергентном виде:

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = 0,$$

где  $\mathbf{F} = \mathcal{A}\mathbf{q}$ , и затем запишем для нее такую же схему, что и для скалярного случая

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0,$$

где

$$\mathbf{F}_{i+1/2} = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{\nu}{2}(\mathbf{q}_{i+1} - \mathbf{q}_i), \quad \nu = \|\mathcal{A}\| = \max(|\lambda^1|, |\lambda^2|, \dots, |\lambda^N|).$$

Введем **инварианты Римана**  $\mathbf{s} = \mathcal{L}\mathbf{q}$  и от схемы в переменных  $\mathbf{q}$

$$\hat{\mathbf{q}}_i = \frac{\tau(\|\mathcal{A}\| + \mathcal{A})}{2h} \mathbf{q}_{i-1} + \left(1 - \frac{\tau\|\mathcal{A}\|}{h}\right) \mathbf{q}_i + \frac{\tau(\|\mathcal{A}\| - \mathcal{A})}{2h} \mathbf{q}_{i+1}$$

перейдем к схеме в переменных  $\mathbf{s}$

$$\hat{\mathbf{s}}_i = \frac{\tau(\|\mathcal{A}\| + \Lambda)}{2h} \mathbf{s}_{i-1} + \left(1 - \frac{\tau\|\mathcal{A}\|}{h}\right) \mathbf{s}_i + \frac{\tau(\|\mathcal{A}\| - \Lambda)}{2h} \mathbf{s}_{i+1}.$$

Эта схема расщепляется на независимые уравнения, каждое из которых дает невозрастание общей вариации  $k$ -го компонента инварианта Римана  $\mathbf{s}$  при выполнении критерия Куранта  $\frac{\tau\|\mathcal{A}\|}{h} \leq 1$ :

$$\begin{aligned} \text{TV}(\hat{\mathbf{s}}^k) &= \sum_i |\hat{s}_i^k - \hat{s}_{i-1}^k| \leqslant \frac{\tau(\|\mathcal{A}\| + \lambda^k)}{2h} \text{TV}(s^k) + \\ &\quad + \left(1 - \frac{\tau\|\mathcal{A}\|}{h}\right) \text{TV}(s^k) + \frac{\tau(\|\mathcal{A}\| - \lambda^k)}{2h} \text{TV}(s^k) = \text{TV}(s^k). \end{aligned}$$

И, наконец, построим схему для решения нелинейной системы (15.21) по аналогии со схемой для решения линейной системы:

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0, \quad (15.24)$$

где

$$\mathbf{F}_{i+1/2} = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{\nu}{2}(\mathbf{q}_{i+1} - \mathbf{q}_i), \quad \nu = \max(\|\mathcal{A}\|_i, \|\mathcal{A}\|_{i+1}).$$

Доказать монотонность схемы Лакса — Фридрихса для нелинейной системы аналитически не представляется возможным, поэтому исследуем ее численно на задаче о распаде разрыва со следующими параметрами:  $\{\rho, u, P\} = \{1, 0, 3\}$  при  $x < 1/2$  и  $\{\rho, u, P\} = \{1, 0, 1\}$  при  $x > 1/2$  (в дальнейшем — тест 1). Газ принимался идеальным с показателем адиабаты  $\gamma = 5/3$ . Решение этой задачи представляет собой простую волну разрежения и константные участки, разделенные контактным разрывом и ударной волной.

Рассмотрим результаты расчета задачи на сетке с 200 ячейками на каждой 4-й ячейке для момента времени  $t = 36$  (рис. 15.1: ВР — волна разрежения; КР — контактный разрыв; УВ — ударная волна; \* — решение по схеме Лакса — Фридрихса; • — решение по схеме Роя; — — точное решение).

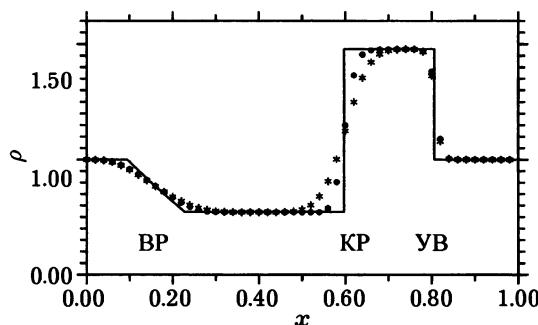


Рис. 15.1

Решение, полученное по схеме Лакса — Фридрихса, монотонно, но очень сильно размазано.

Диссипативные свойства схемы Лакса — Фридрихса хорошо видны на другом тесте — взаимодействии двух взрывных ударных волн [292]. В начальный момент времени распределение параметров среды следующее:  $\{\rho, u, P\} = \{1, 0, 1000\}$  при  $x < 1/10$ ;  $\{\rho, u, P\} = \{1, 0, 1/100\}$  при  $1/10 < x < 9/10$ ;  $\{\rho, u, P\} = \{1, 0, 100\}$  при  $x > 9/10$  (в дальнейшем — тест 2). Газ принимался идеальным с показателем адиабаты

$\gamma = 7/5$ . Как результат распада этого разрыва возникают две ударные волны УВ1 и УВ2 и два контактных разрыва КР1 и КР2, которые взаимодействуют между собой и порождают новые ударные волны УВ3 и УВ4 и контактные разрывы КР3 и КР4.

На рис. 15.2 представлены изолинии плотности на плоскости  $(x, t)$ . На рис. 15.2, б показано увеличенное изображение для прямоугольника, изображенного на рис. 15.2, а. Взаимодействие КР2 с УВ4 изменяет скорость распространения УВ4 и, кроме того, порождает КР4 и ВР, которая изменяет скорость распространения КР3, возникшего в результате взаимодействия УВ1 и УВ2. Анализ результатов показывает, что контактные разрывы передаются схемой Лакса — Фридрихса очень плохо, в частности, КР3 практически не виден так же, как и ВР. Забегая вперед, скажем, что ВР смогла передать только схема Рой — Эйнфельдта — Ошера (15.36).

### 15.2.3. Схемы годуновского типа (линейный случай)

Рассмотрим разностную схему в потоковой форме для гиперболической системы уравнений:

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0. \quad (15.25)$$

Будем считать решение постоянным по разностной ячейке. Предположим, что можно точно решить задачу о распаде разрыва (задачу Римана) между состояниями  $\mathbf{q}_L = \mathbf{q}_i$  и  $\mathbf{q}_R = \mathbf{q}_{i+1}$  и найти таким образом решение в центре разрыва  $\mathbf{q}^*(x_{i+1/2}, t)$ . Вычислить поток на границе между двумя ячейками можно исходя из решения задачи о распаде разрыва [49]:

$$\mathbf{F}_{i+1/2} = \frac{1}{\tau} \int_{t_0}^{t_0 + \tau} \mathbf{F}(\mathbf{q}^*(x_{i+1/2}, t)) dt. \quad (15.26)$$

В силу того, что волны, возникающие при распаде разрыва, центрированы, решение  $\mathbf{q}^*(x_{i+1/2}, t)$  не зависит от времени при малых  $\tau$ , следовательно,

$$\frac{1}{\tau} \int_{t_0}^{t_0 + \tau} \mathbf{F}(\mathbf{q}^*(x_{i+1/2}, t)) dt = \frac{1}{\tau} \int_{t_0}^{t_0 + \tau} \mathbf{F}(\mathbf{q}^*(x_{i+1/2}, t_0)) dt = \mathbf{F}(\mathbf{q}^*(x_{i+1/2}, t_0)).$$

Необходимо решить вопрос о том, как вычислять  $\mathbf{q}^*(x_{i+1/2}, t)$ . Рассмотрим сначала скалярное линейное уравнение (15.19). В этом случае

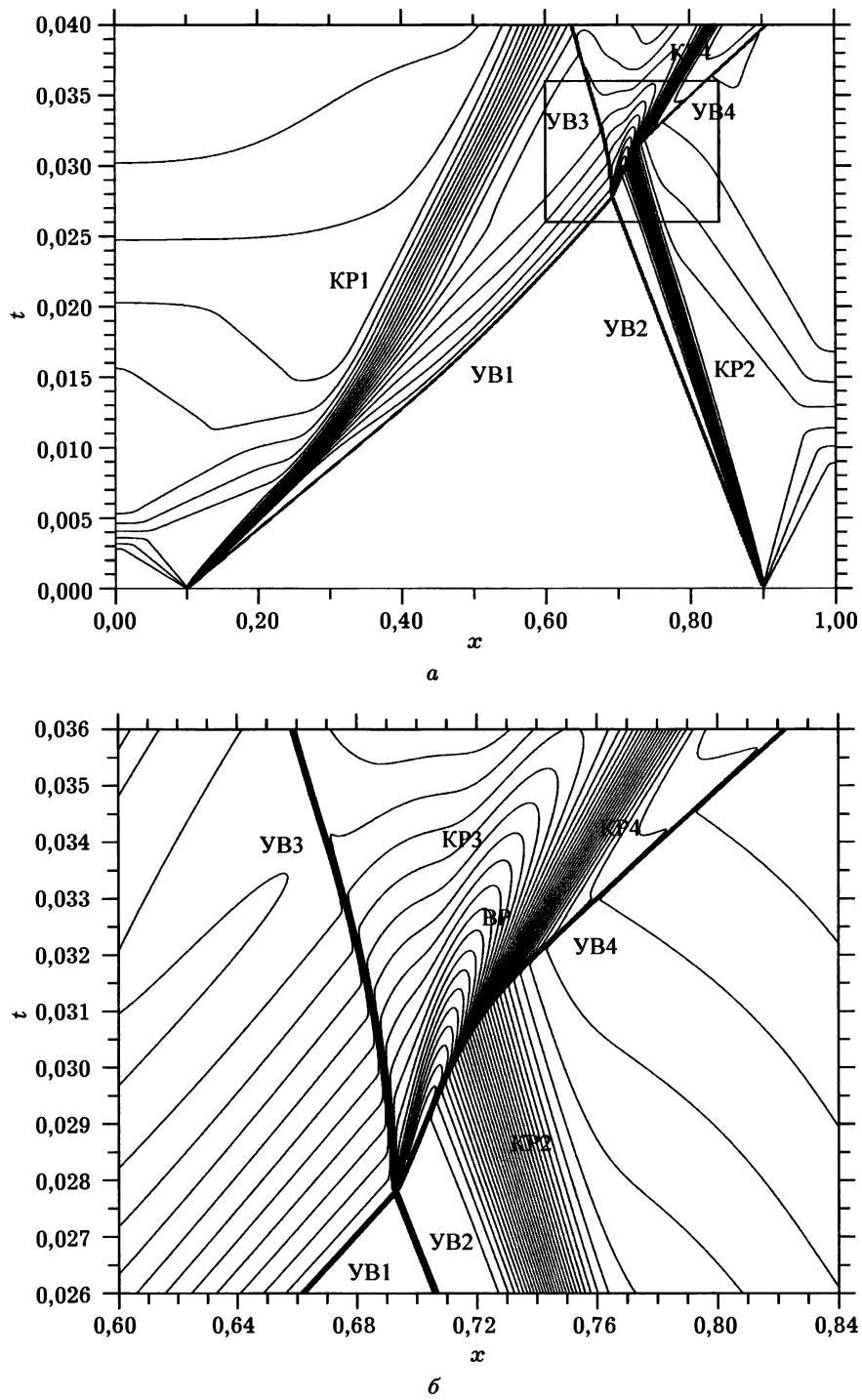


Рис. 15.2

$q^*(x_{i+1/2}, t)$  определяется просто:

$$q^*(x_{i+1/2}, t) = \begin{cases} q_L, & a > 0; \\ q_R, & a < 0, \end{cases} \quad \text{или} \quad q^*(x_{i+1/2}, t) = \vartheta(a)q_L + \vartheta(-a)q_R$$

и

$$F_{i+1/2} = aq^* = a\vartheta(a)q_L + a\vartheta(-a)q_R = a^+q_L + a^-q_R,$$

где  $\vartheta(x)$  — функция Хевисайда;  $a^+ = \max(a, 0) = (a + |a|)/2$ ;  $a^- = \min(a, 0) = (a - |a|)/2$ . Рассмотрим теперь линейную систему гиперболических уравнений (15.20). Переходя к **инвариантам Римана**  $s_L = \mathcal{L}\mathbf{q}_L$ ,  $s_R = \mathcal{L}\mathbf{q}_R$ , после несложных преобразований получаем величины

$$s^{*k}(x_{i+1/2}, t) = \begin{cases} s_L^k, & \lambda^k > 0; \\ s_R^k, & \lambda^k < 0, \end{cases} \quad \text{или} \quad s^{*k}(x_{i+1/2}, t) = \vartheta(\lambda^k)s_L^k + \vartheta(-\lambda^k)s_R^k,$$

$$\begin{aligned} q^{*k} &= \sum_m \mathcal{R}^{km} s^{*m} = \sum_m \mathcal{R}^{km} \vartheta(\lambda^m) s_L^m + \sum_m \mathcal{R}^{km} \vartheta(-\lambda^m) s_R^m = \\ &= \sum_m \mathcal{R}^{km} \vartheta(\lambda^m) \sum_l \mathcal{L}^{ml} q_L^l + \sum_m \mathcal{R}^{km} \vartheta(-\lambda^m) \sum_l \mathcal{L}^{ml} q_R^l, \end{aligned}$$

$$\begin{aligned} F_{i+1/2}^k &= \sum_n \mathcal{A}^{kn} q^{*n} = \\ &= \sum_n \mathcal{A}^{kn} \sum_m \mathcal{R}^{nm} \vartheta(\lambda^m) \sum_l \mathcal{L}^{ml} q_L^l + \sum_n \mathcal{A}^{kn} \sum_m \mathcal{R}^{nm} \vartheta(-\lambda^m) \sum_l \mathcal{L}^{ml} q_R^l = \\ &= \sum_l \sum_m \mathcal{L}^{ml} q_L^l \vartheta(\lambda^m) \sum_n \mathcal{A}^{kn} \mathcal{R}^{nm} + \sum_l \sum_m \mathcal{L}^{ml} q_R^l \vartheta(-\lambda^m) \sum_n \mathcal{A}^{kn} \mathcal{R}^{nm} = \\ &= \sum_l \sum_m \mathcal{L}^{ml} q_L^l \vartheta(\lambda^m) \lambda^m \mathcal{R}^{km} + \sum_l \sum_m \mathcal{L}^{ml} q_R^l \vartheta(-\lambda^m) \lambda^m \mathcal{R}^{km} = \\ &= \sum_l \left( \sum_m \lambda^{m+} \mathcal{R}^{km} \mathcal{L}^{ml} \right) q_L^l + \sum_l \left( \sum_m \lambda^{m-} \mathcal{R}^{km} \mathcal{L}^{ml} \right) q_R^l. \end{aligned}$$

Введем обозначения

$$\mathcal{A}^{kl+} = \sum_m \lambda^{m+} \mathcal{R}^{km} \mathcal{L}^{ml}, \quad \mathcal{A}^{kl-} = \sum_m \lambda^{m-} \mathcal{R}^{km} \mathcal{L}^{ml}.$$

Тогда окончательное выражение для потока  $\mathbf{F}_{i+1/2}$  можно записать следующим образом:

$$\mathbf{F}_{i+1/2} = \mathcal{A}^+ \mathbf{q}_L + \mathcal{A}^- \mathbf{q}_R.$$

Поскольку

$$\mathcal{A} = \mathcal{A}^+ + \mathcal{A}^-, \quad \mathcal{A}^+ - \mathcal{A}^- = \sum_m |\lambda^m| \mathcal{R}^{km} \mathcal{L}^{ml} = \mathcal{R} |\Lambda| \mathcal{L} = |\mathcal{A}|,$$

$$|\Lambda| = \text{diag}(|\lambda^1|, |\lambda^2|, \dots, |\lambda^N|), \quad \mathbf{F}_i = \mathcal{A} \mathbf{q}_i, \quad \mathbf{F}_{i+1} = \mathcal{A} \mathbf{q}_{i+1},$$

выражение для потока  $\mathbf{F}_{i+1/2}$  можно записать еще тремя эквивалентными способами:

$$\begin{aligned} \mathbf{F}_{i+1/2} &= \mathbf{F}_i + \mathcal{A}^- (\mathbf{q}_{i+1} - \mathbf{q}_i) = \\ &= \mathbf{F}_{i+1} - \mathcal{A}^+ (\mathbf{q}_{i+1} - \mathbf{q}_i) = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{|\mathcal{A}|}{2} (\mathbf{q}_{i+1} - \mathbf{q}_i). \end{aligned} \quad (15.27)$$

Отметим, что  $|\mathcal{A}|$  — это матрица, а  $\|\mathcal{A}\|$  — скаляр.

Схему (15.25) с потоками (15.27), т. е. схему Годунова для линейной системы, часто называют *схемой Куранта — Изаксона — Риса*. Она также может быть получена из соотношений на характеристиках [106, 216]. Эта схема замечательна тем, что обладает минимальной численной вязкостью из всех линейных монотонных схем. Действительно, уменьшим численную вязкость схемы (15.25)–(15.27) добавлением к (15.27) антидиффузионного члена  $\eta(\mathbf{q}_{i+1} - \mathbf{q}_i)$ ,  $\eta \geq 0$ . Снова переходя к инвариантам Римана, получаем

$$\begin{aligned} \hat{\mathbf{q}}_i &= \frac{\tau}{h} \left( \frac{|\mathcal{A}|}{2} + \frac{\mathcal{A}}{2} - \eta \right) \mathbf{q}_{i-1} + \left( 1 - \frac{\tau}{h} (|\mathcal{A}| - 2\eta) \right) \mathbf{q}_i + \frac{\tau}{h} \left( \frac{|\mathcal{A}|}{2} - \frac{\mathcal{A}}{2} - \eta \right) \mathbf{q}_{i+1}, \\ \hat{s}_i^k &= \frac{\tau}{h} \left( \frac{|\lambda^k|}{2} + \frac{\lambda^k}{2} - \eta \right) s_{i-1}^k + \left( 1 - \frac{\tau}{h} (|\lambda^k| - 2\eta) \right) s_i^k + \frac{\tau}{h} \left( \frac{|\lambda^k|}{2} - \frac{\lambda^k}{2} - \eta \right) s_{i+1}^k. \end{aligned}$$

Применение к последнему выражению принципа максимума [139] приводит к выводу, что при всех неотрицательных  $\eta$  схема монотонна только при  $\eta = 0$ . Заметим, что даже при  $\eta = 0$  для монотонности необходимо, чтобы был выполнен критерий Куранта  $\tau \|\mathcal{A}\| / h \leq 1$ .

#### 15.2.4. Схемы годуновского типа (нелинейный случай). Схема Роу

В методе, предложенном С.К. Годуновым, для общего случая системы нелинейных гиперболических уравнений информация о распространении и взаимодействии нелинейных волн включается в численную схему в форме решения задачи о распаде разрыва. Таким образом, в случае течений с газодинамическими разрывами метод Годунова дает более приемлемые результаты, чем методы, основанные на моделях гладкого

течения. Недостатком этого метода является его вычислительная сложность. Даже в случае простого уравнения состояния идеального газа для нахождения решения задачи о распаде разрыва требуется решить нелинейное алгебраическое уравнение (см., например, [135]). Задача еще больше усложняется для других уравнений состояния (см., например, [213, 214, 283]).

В то же время, поскольку входными данными, необходимыми для решения задачи о распаде разрыва, являются приближенные значения параметров, возникает вопрос: нельзя ли решать задачу о распаде разрыва также приближенно? Одним из методов, основанных на приближенном решении задачи Римана, является *схема Рой* [272, 274, 275], в которой решение нелинейной задачи о распаде разрыва заменяется на решение этой задачи для линейной системы, а она, как мы видели в 15.2.3, решается просто. Матрица  $\mathcal{A}(\mathbf{q}_L, \mathbf{q}_R)$  приближенной линейной системы должна быть, конечно, гиперболической и гладко переходить в матрицу Якоби  $\frac{\partial \mathbf{F}}{\partial \mathbf{q}}$  при  $\mathbf{q}_L \rightarrow \mathbf{q}_R$ . Кроме того, в работах [272, 274, 275] предложено брать матрицу  $\mathcal{A}(\mathbf{q}_L, \mathbf{q}_R)$  такой, чтобы удовлетворялось соотношение

$$\mathbf{F}_R - \mathbf{F}_L = \mathcal{A}(\mathbf{q}_L, \mathbf{q}_R)(\mathbf{q}_R - \mathbf{q}_L). \quad (15.28)$$

Соотношение (15.28) имеет глубокий смысл. Во-первых, при таком выборе матрицы приближенной линейной системы решение приближенной задачи о распаде разрыва, обозначенное ниже  $\bar{\mathbf{q}}$ , удовлетворяет тем же интегральным законам сохранения, что и нелинейной системы:

$$\int_{x_i}^{x_{i+1}} \bar{\mathbf{q}}(x, t_0 + \tau) dx = \int_{x_i}^{x_{i+1}} \mathbf{q}(x, t_0 + \tau) dx.$$

Это следует из соотношений

$$\begin{aligned} \int_{x_i}^{x_{i+1}} \mathbf{q}(x, t_0 + \tau) dx &= \int_{x_i}^{x_{i+1}} \mathbf{q}(x, t_0) dx - (\mathbf{F}_{i+1} - \mathbf{F}_i)\tau, \\ \int_{x_i}^{x_{i+1}} \bar{\mathbf{q}}(x, t_0 + \tau) dx &= \int_{x_i}^{x_{i+1}} \mathbf{q}(x, t_0) dx - \mathcal{A}(\mathbf{q}_{i+1} - \mathbf{q}_i)\tau. \end{aligned}$$

Во-вторых, в случае, когда  $\mathbf{q}_L$  и  $\mathbf{q}_R$  — состояния по разные стороны ударной волны или контактного разрыва, приближенное решение задачи о распаде разрыва совпадает с точным решением нелинейной задачи. Это следует из условий Гюгонио [242, 267]

$$\mathbf{F}_R - \mathbf{F}_L = D(\mathbf{q}_R - \mathbf{q}_L),$$

где  $D$  — скорость распространения ударной волны или контактного разрыва. В этом случае из соотношений (15.28) видно, что  $\mathbf{q}_R - \mathbf{q}_L$  — собственный вектор матрицы  $\mathcal{A}(\mathbf{q}_L, \mathbf{q}_R)$ , соответствующий собственному числу  $D$ , приближенное решение  $\bar{\mathbf{q}}$  также будет состоять из скачка  $\mathbf{q}_R - \mathbf{q}_L$ , движущегося со скоростью  $D$ .

Однако следует отметить, что решение линейной системы по схеме Рой, качественно повторяя решение газодинамической задачи, не содержит центрированных волн разрежения, а является системой скачков решения, распространяющихся со скоростями, соответствующими собственным числам матрицы  $\mathcal{A}(\mathbf{q}_L, \mathbf{q}_R)$ . В частности, некоторые из этих скачков могут не удовлетворять **энтропийному условию**. Впрочем, для большинства случаев (за исключением трансзвуковых волн разрежения) решение, полученное по схеме Рой, является удовлетворительным, даже если решение включает волны разрежения.

Вернемся к проблеме выбора матрицы  $\mathcal{A}(\mathbf{q}_L, \mathbf{q}_R)$ . Рой показал [272], что для системы уравнений газовой динамики (15.21) с уравнением состояния идеального газа существует промежуточное значение  $\mathbf{q}^*(\mathbf{q}_L, \mathbf{q}_R)$ , такое, что

$$\mathcal{A}(\mathbf{q}_L, \mathbf{q}_R) = \frac{\partial \mathbf{F}}{\partial \mathbf{q}}(\mathbf{q}^*)$$

удовлетворяет (15.28), а именно:

$$\rho^* = \sqrt{\rho_L \rho_R}, \quad u^* = \frac{\sqrt{\rho_L} u_L + \sqrt{\rho_R} u_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \quad h^{**} = \frac{\sqrt{\rho_L} h_L^* + \sqrt{\rho_R} h_R^*}{\sqrt{\rho_L} + \sqrt{\rho_R}}. \quad (15.29)$$

Чтобы вывести эти соотношения, нужно записать условие (15.28) для матрицы (15.23) и решить уравнения

$$\begin{aligned} u^{*2} \Delta \rho - 2u^* \Delta \rho u + \Delta \rho u^2 &= 0, \\ u^* h^{**} \Delta \rho + \Delta \rho u h^* - h^* \Delta \rho u - u^* \Delta \rho h^* &= 0, \end{aligned}$$

где  $\Delta y = y_R - y_L$ .

Промежуточное значение скорости звука может быть вычислено как  $c^* = \sqrt{(\gamma - 1)(h^{**} - u^{*2}/2)}$ . Это может привести к потере точности, если кинетическая энергия газа много больше тепловой. Поэтому для  $c^*$  предложена [223] эквивалентная формула, не содержащая разности больших величин:

$$c^* = \sqrt{\frac{\sqrt{\rho_L} c_L^2 + \sqrt{\rho_R} c_R^2}{\sqrt{\rho_L} + \sqrt{\rho_R}}} + \frac{\gamma - 1}{2} \frac{\sqrt{\rho_L \rho_R}}{(\sqrt{\rho_L} + \sqrt{\rho_R})^2} (u_R - u_L)^2.$$

В результате из (15.27) и (15.29) получаем схему Рой для решения уравнений одномерной газовой динамики (15.21):

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0, \quad (15.30)$$

где

$$\begin{aligned}\mathbf{F}_{i+1/2} &= \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2} \sum_m |\lambda^m(\mathbf{q}^*)| \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*), \\ \Delta s_{i+1/2}^m &= \mathbf{l}^m(\mathbf{q}^*)(\mathbf{q}_{i+1} - \mathbf{q}_i), \\ \Delta s_{i+1/2}^{1,3} &= \frac{1}{2c^{*2}} ((P_{i+1} - P_i) \mp \rho^* c^* (u_{i+1} - u_i)), \\ \Delta s_{i+1/2}^2 &= \frac{1}{2c^{*2}} (c^{*2} (\rho_{i+1} - \rho_i) - (P_{i+1} - P_i)).\end{aligned}$$

Результаты расчета задачи о распаде разрыва с параметрами  $\{\rho, u, P\} = \{1, 0, 3\}$  при  $x < 1/2$  и  $\{\rho, u, P\} = \{1, 0, 1\}$  при  $x > 1/2$  (тест 1) по схеме Рой (обозначено •) на сетке с 200 ячейками представлены на рис. 15.1. Видно, что получающееся решение монотонно и контактный разрыв передается лучше, чем при расчете по схеме Лакса — Фридрихса.

Результаты расчета задачи о взаимодействии двух взрывных ударных волн по схеме Рой (тест 2) показаны на рис. 15.3, аналогичном рис. 15.2.

Сравнение рис. 15.2 и 15.3 показывает, что численная диссипация схемы Рой меньше, чем схемы Лакса — Фридрихса, однако КРЗ и ВР воспроизводятся плохо.

### 15.2.5. Энтропийное условие

Как уже отмечалось выше, в ряде случаев численное решение, полученное по схеме Рой, может соответствовать нефизическим разрывам. В качестве простого примера рассмотрим квазилинейное уравнение переноса

$$\frac{\partial q}{\partial t} + \frac{\partial F}{\partial x} = 0, \quad (15.31)$$

где  $F = \frac{q^2}{2}$ . Промежуточное значение  $q^*$  должно в этом случае определяться из соотношения  $F_R - F_L = \frac{\partial F}{\partial q}(q^*)(q_R - q_L)$ , из которого вытекает, что  $q^* = (q_L + q_R)/2$ . В результате получаем схему Рой для решения

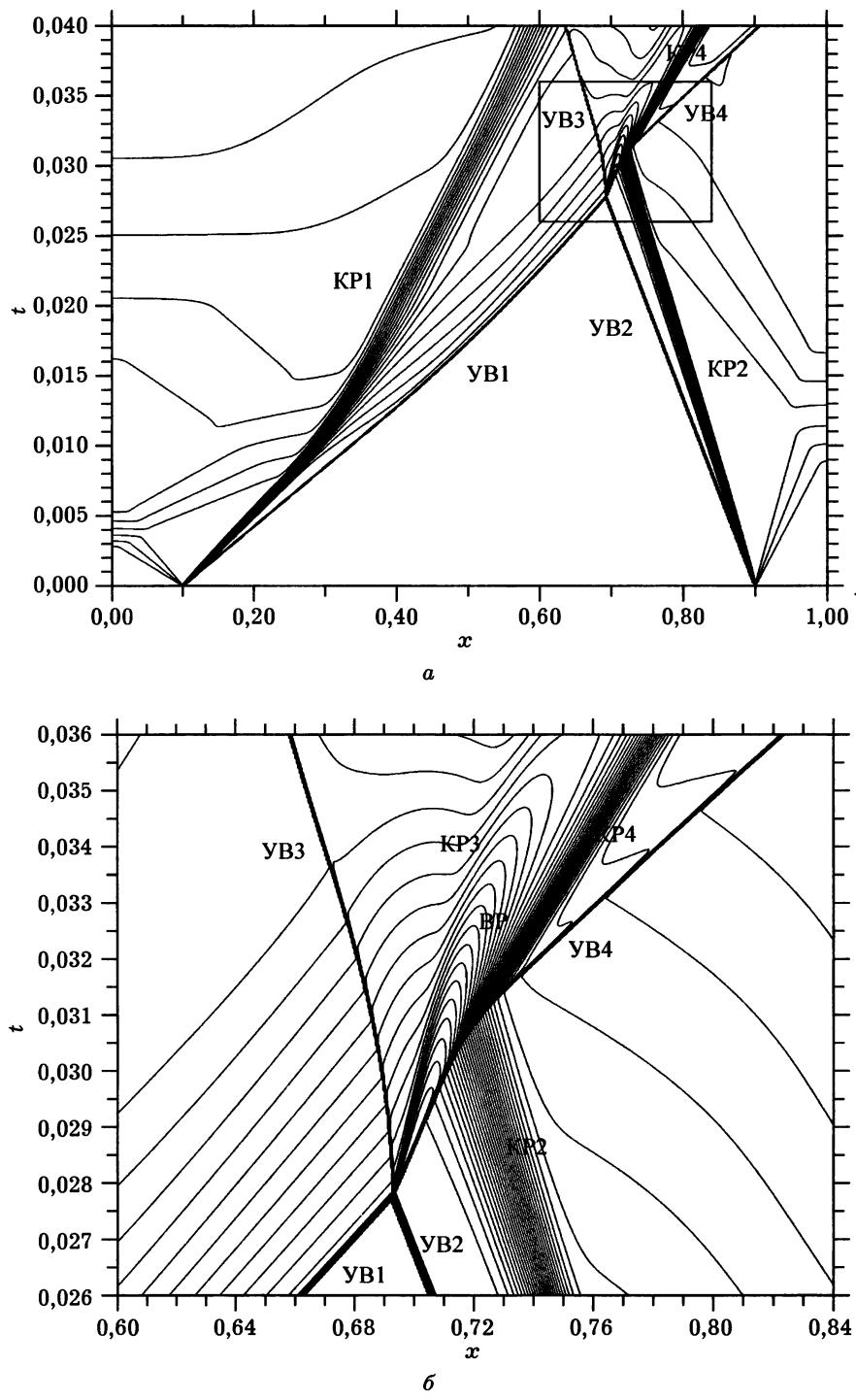


Рис. 15.3

уравнения (15.31):

$$\frac{\hat{q}_i - q_i}{\tau} + \frac{F_{i+1/2} - F_{i-1/2}}{h} = 0,$$

где

$$\begin{aligned} F_{i+1/2} &= \frac{F_i + F_{i+1}}{2} - \frac{1}{2} \left| \frac{\partial F}{\partial q}(q^*) \right| (q_{i+1} - q_i) = \\ &= \frac{1}{2} \left( \frac{q_i^2}{2} + \frac{q_{i+1}^2}{2} \right) - \frac{1}{4} |q_i + q_{i+1}| (q_{i+1} - q_i). \end{aligned}$$

Рассмотрим задачу со следующими начальными данными:

$$q(x, 0) = \begin{cases} -1, & x < 0; \\ 1, & x \geq 0. \end{cases}$$

Корректное решение этой задачи представляет собой волну разрежения:

$$q(x, t) = \begin{cases} -1, & x < -t; \\ x/t, & -t < x < t; \\ 1, & x > t. \end{cases}$$

Однако в результате непосредственной проверки убеждаемся, что численное решение по *схеме Рой* не соответствует аналитическому решению, так как всегда поток  $F_{i+1/2} = 1/2$ , следовательно, все время будет воспроизводиться начальный профиль решения. Фактически в данном случае численное решение по *схеме Рой* представляет собой неэволюционную ударную волну. Заметим, что данная проблема возникает, когда решением задачи является волна разрежения, содержащая звуковую точку, в которой  $\frac{\partial F}{\partial q} = 0$ .

Похожий результат можно получить и для уравнений газовой динамики. Рассмотрим задачу о распаде разрыва со следующими параметрами:  $\{\rho, u, P\} = \{8, 0, 480\}$  при  $x < 1/2$  и  $\{\rho, u, P\} = \{1, 0, 1\}$  при  $x > 1/2$  (в дальнейшем — тест 3). Газ принимался идеальным с показателем адиабаты  $\gamma = 5/3$ . Как и для теста 1, решение этой задачи представляет собой простую волну разрежения и константные участки, разделенные контактным разрывом и ударной волной. Однако в тесте 3 волна разрежения содержит звуковую точку (похожие результаты можно получить для теста 1 с помощью перехода в движущуюся систему координат).

Рассмотрим результаты расчета задачи на сетке с 200 ячейками на каждой 4-й ячейке на момент времени  $t = 7$  (рис. 15.4: ВР — волна разрежения; КР — контактный разрыв; УВ — ударная волна; ● — решение по схеме Рой (15.31);  $\Delta$  — решение по схеме Рой — Эйнфельдта (15.32); — — точное решение;  $---$  — скорость звука).

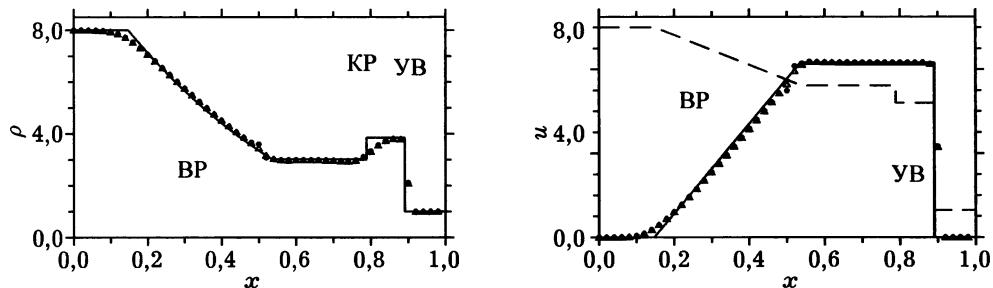


Рис. 15.4

Видно, что в звуковой точке образуется ударная волна разрежения [1].

Причина возникновения неэволюционных разрывов в схеме Роу достаточно очевидна — она связана с тем, что матрица вязкости  $|\mathcal{A}|$  имеет нулевое собственное число на стационарных ударных волнах, независимо от того, эволюционные они или нет. Это, безусловно, положительное свойство для расчета эволюционных ударных волн (приращение энтропии  $\Delta S > 0$ ), так как уменьшает численную вязкость, однако для предотвращения возникновения неэволюционных стационарных ударных волн ( $\Delta S < 0$ ) следовало бы увеличить численную вязкость схемы [212, 239, 252, 294]. Наиболее простой и эффективный способ предложен в [223].

Из диаграммы эволюционных и неэволюционных ударных волн при  $u > 0$  (рис. 15.5) следует, что замена в схеме Роу  $\lambda^1(\mathbf{q}^*)$  на  $\min\{\lambda^1(\mathbf{q}^*), \lambda_i^1\}$  увеличивает  $|\lambda^1|$  и, следовательно, увеличивает численную вязкость только для неэволюционных волн, в то время как на эволюционные волны эта операция не оказывает влияния. Аналогич-

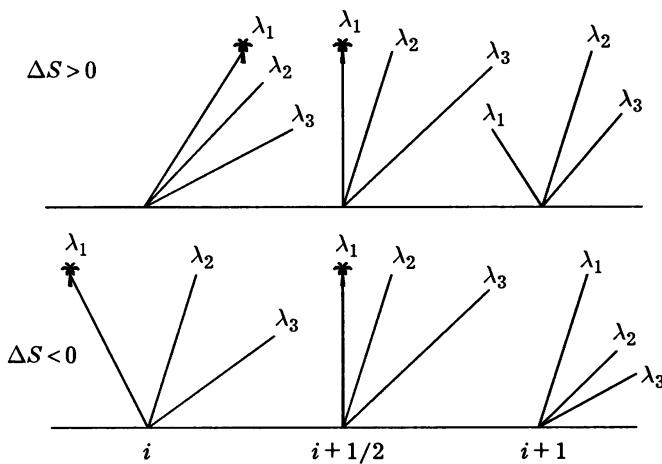


Рис. 15.5

ное действие оказывает замена  $\lambda^3(\mathbf{q}^*)$  на  $\max\{\lambda^3(\mathbf{q}^*), \lambda_{i+1}^3\}$  при  $u < 0$  (рис. 15.6).

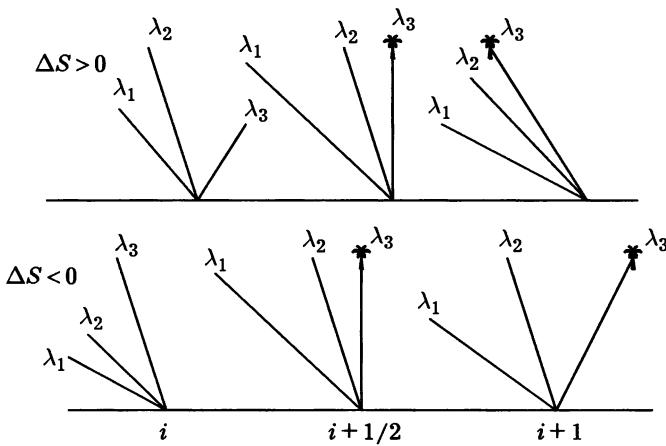


Рис. 15.6

В случаях, изображенных на рис. 15.5 и 15.6, для исключения нефизических скачков необходимо при  $u > 0$  взять  $\min\{\lambda_i^1, \lambda_{i+1/2}^1\}$ , а при  $u < 0$  —  $\max\{\lambda_{i+1}^3, \lambda_{i+1/2}^3\}$ .

В результате получаем *схему Рой — Эйнфельдта*

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0, \quad (15.32)$$

где

$$\mathbf{F}_{i+1/2} = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2} \sum_m |\tilde{\lambda}^m| \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*),$$

$$\tilde{\lambda}^1 = \min(\lambda^1(\mathbf{q}^*), \lambda_i^1), \quad \tilde{\lambda}^2 = \lambda^2(\mathbf{q}^*), \quad \tilde{\lambda}^3 = \max(\lambda^3(\mathbf{q}^*), \lambda_{i+1}^3).$$

Результаты расчета теста 3 по схеме Рой — Эйнфельдта (15.32) представлены на рис. 15.4 (изображен  $\Delta$ ). Видно, что ударная волна разрежения исчезла.

### 15.2.6. Схемы повышенного порядка аппроксимации

Приведенные выше примеры показывают, что схема Рой первого порядка аппроксимации, как, впрочем, и другие *схемы годуновского типа* первого порядка, дает очень размазанное численное решение около сильных и слабых разрывов. Далее мы покажем, как схема

Роу может быть преобразована к схеме второго или третьего порядка аппроксимации на гладких решениях, оставаясь, однако, монотонной, т. е. не дающей нефизических осцилляций около разрывов.

Снова рассмотрим скалярное уравнение переноса (15.19). Для простоты ограничимся случаем  $a > 0$ . Вместо кусочно-постоянной аппроксимации величин в ячейках введем кусочно-линейную аппроксимацию

$$q(x) = q_i + \sigma_i(x - x_i)$$

с некоторым подходящим значением наклона  $\sigma_i$ , которое будет определено ниже. Тогда в методе Годунова интеграл (15.26) будет равен

$$F_{i+1/2} = \frac{1}{\tau} \int_{t_0}^{t_0+\tau} aq^*(x_{i+1/2}, t) dt = aq_i + \frac{a}{2}(h - \tau a)\sigma_i.$$

Наклон  $\sigma_i$  можно выбирать многими способами, каждый из которых соответствует своей схеме. Если  $\sigma_i$  аппроксимирует  $\frac{\partial q}{\partial x}$ , то схема будет иметь второй порядок аппроксимации. Наиболее употребительными являются *схемы Лакса — Вендроффа* [251] с  $\sigma_i = (q_{i+1} - q_i)/h$ , *Бима — Ворминга* [197, 289] с  $\sigma_i = (q_i - q_{i-1})/h$  и *Фромма* [231] с  $\sigma_i = (q_{i+1} - q_{i-1})/(2h)$ . Потоки для этих схем могут быть вычислены по следующим выражениям:

схема Лакса — Вендроффа

$$F_{i+1/2} = aq_i + \frac{a}{2} \left(1 - \frac{\tau a}{h}\right) (q_{i+1} - q_i);$$

схема Бима — Ворминга

$$F_{i+1/2} = aq_i + \frac{a}{2} \left(1 - \frac{\tau a}{h}\right) (q_i - q_{i-1});$$

схема Фромма

$$F_{i+1/2} = aq_i + \frac{a}{2} \left(1 - \frac{\tau a}{h}\right) \frac{q_{i+1} - q_{i-1}}{2}.$$

Все три схемы, а также монотонную схему первого порядка аппроксимации можно записать в общем виде:

$$F_{i+1/2} = aq_i + \frac{a}{2} \left(1 - \frac{\tau a}{h}\right) \alpha(R) (q_{i+1} - q_i),$$

где

$$R = \frac{q_i - q_{i-1}}{q_{i+1} - q_i},$$

а функция  $\alpha(R)$  имеет вид

$\alpha(R) = 0$  для монотонной схемы первого порядка;

$\alpha(R) = 1$  для схемы Лакса — Вендроффа;

$\alpha(R) = R$  для схемы Бима — Ворминга;

$\alpha(R) = \frac{1+R}{2}$  для схемы Фромма.

Величина  $R$  обычно называется *анализатором гладкости*, а функция  $\alpha(R)$  — *ограничителем потока*. Для аппроксимации с порядком выше первого необходимо, чтобы  $\alpha(1) = 1$ . Отметим, что во всех указанных схемах  $\alpha(R)$  — линейная функция.

Хорошо известно, однако, что все три перечисленные схемы второго порядка аппроксимации являются немонотонными. Это легко показать, например, с помощью принципа максимума. Более того, среди линейных схем, т. е. схем вида

$$\hat{q}_i = \sum_k c_k q_{i+k}$$

с порядком аппроксимации выше первого нет монотонных схем [49].

Чтобы преодолеть запрет, накладываемый *теоремой Годунова*, предложено использовать нелинейные ограничители потока [237, 284]. Для обеспечения монотонности схемы необходимо, чтобы ограничитель потока удовлетворял условию  $0 \leq \alpha(R) \leq \text{minmod}(2, 2R)$ , где

$$\text{minmod}(x, y) = \frac{1}{2}(\text{sign}(x) + \text{sign}(y)) \min(|x|, |y|).$$

На рис. 15.7 приведена зависимость ограничителя антидиффузионного потока  $\alpha(R)$  от анализатора гладкости для различных схем. Область, в которой ограничитель потока удовлетворяет условию обеспечения монотонности, показана на плоскости  $(R, \alpha(R))$  штриховкой. На рис. 15.7 видно, что *схемы Лакса — Вендроффа, Бима — Ворминга и Фромма* немонотонны, так как их ограничители потока выходят за пределы области монотонности.

В дополнение к рассматриваемым схемам в [284] предложено использовать в качестве  $\alpha(R)$  выпуклую комбинацию ограничителя потока схем Лакса — Вендроффа ( $\alpha(R) = 1$ ) и Бима — Ворминга ( $\alpha(R) = R$ ) (симметричная разность областей монотонности показана на рис. 15.7 наклонной штриховкой). На рис. 15.7 изображены ограничители потока  $\alpha(R)$ , предложенные в различных работах:

minmod [284]

$$\alpha(R) = \text{minmod}(1, R);$$

superbee [273]

$$\alpha(R) = \max(0, \min(1, 2R), \min(2, R));$$

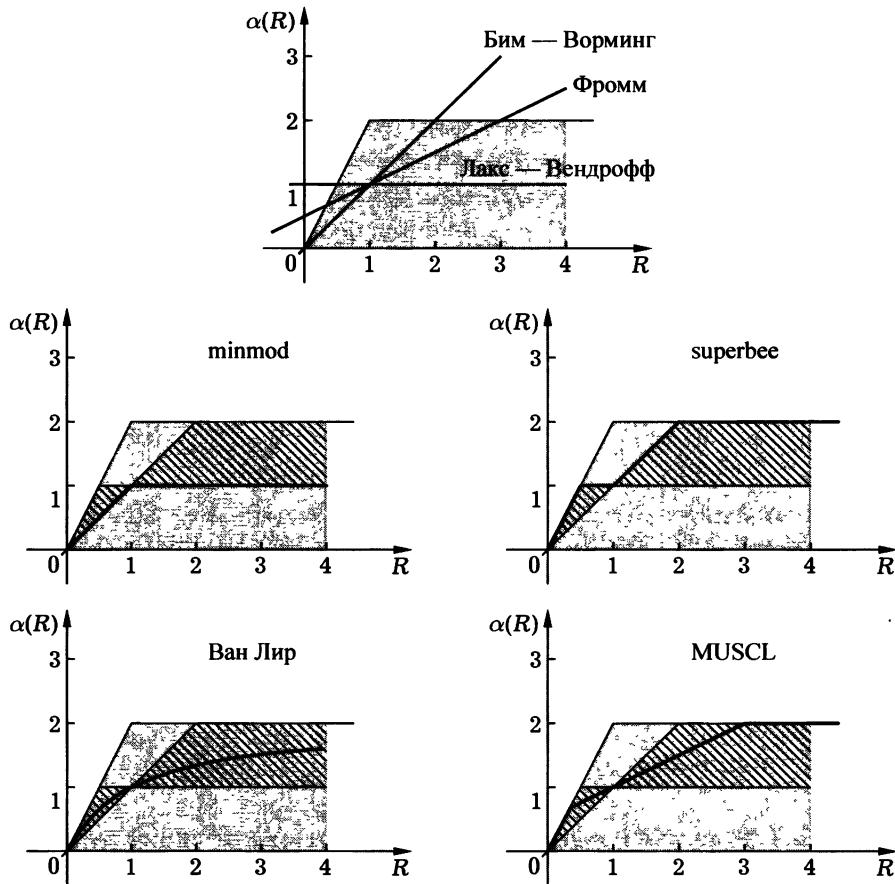


Рис. 15.7

MUSCL [287]

$$\alpha(R) = \max\left(0, \min\left(\frac{1+R}{2}, 2, 2R\right)\right);$$

ограничитель Ван Лира [286]

$$\alpha(R) = \frac{R + |R|}{1 + R}.$$

Несколько иной способ введения ограничителей потока предложен в работах [34, 35, 209]. Если рассмотренные выше схемы базировались на *схеме Лакса — Вендроффа* (ограничитель  $\alpha = 1$ ), то последние схемы при  $\alpha = 1$  будут превращаться в *схему с центральными разностями*. Рассмотрим общий случай уравнения (15.19), когда значение  $a$  может

быть как положительно, так и отрицательно. Будем обозначать поток в схеме первого порядка через  $F_{i+1/2}^I$ :

$$F_{i+1/2}^I = a^+ q_i + a^- q_{i+1}.$$

Добавим к  $F_{i+1/2}^I$  антидиффузионные члены  $\alpha^\pm$  с ограничителями потока:

$$F_{i+1/2} = F_{i+1/2}^I + \frac{\alpha_{i+1/2}^+}{2} F_{i+1/2}^+ - \frac{\alpha_{i+1/2}^-}{2} F_{i+1/2}^-,$$

$$F_{i+1/2}^+ = F_{i+1} - F_{i+1/2}^I = aq_{i+1} - a^+ q_i - a^- q_{i+1} = a^+ \Delta q_{i+1/2},$$

$$F_{i+1/2}^- = F_{i+1/2}^I - F_i = a^+ q_i + a^- q_{i+1} - aq_i = a^- \Delta q_{i+1/2},$$

где  $\Delta q_{i+1/2} = q_{i+1} - q_i$ . В результате получим схему

$$\begin{aligned} \frac{\hat{q}_i - q_i}{\tau} + a^+ \left( 1 + \frac{\alpha_{i+1/2}^+}{2} \frac{\Delta q_{i+1/2}}{\Delta q_{i-1/2}} - \frac{\alpha_{i-1/2}^+}{2} \right) \frac{q_i - q_{i-1}}{h} + \\ + a^- \left( 1 + \frac{\alpha_{i-1/2}^-}{2} \frac{\Delta q_{i-1/2}}{\Delta q_{i+1/2}} - \frac{\alpha_{i+1/2}^-}{2} \right) \frac{q_{i+1} - q_i}{h} = \\ = \frac{\hat{q}_i - q_i}{\tau} + a^+ \left( 1 + \frac{\alpha_{i+1/2}^+}{2} \frac{1}{R_{i+1/2}^+} - \frac{\alpha_{i-1/2}^+}{2} \right) \frac{q_i - q_{i-1}}{h} + \\ + a^- \left( 1 + \frac{\alpha_{i-1/2}^-}{2} \frac{1}{R_{i-1/2}^-} - \frac{\alpha_{i+1/2}^-}{2} \right) \frac{q_{i+1} - q_i}{h} = 0, \quad (15.33) \end{aligned}$$

где

$$\begin{aligned} \alpha_{i+1/2}^+ &= \alpha(R_{i+1/2}^+); \quad \alpha_{i+1/2}^- = \alpha(R_{i+1/2}^-); \\ R_{i+1/2}^+ &= \frac{\Delta q_{i-1/2}}{\Delta q_{i+1/2}}; \quad R_{i-1/2}^- = \frac{1}{R_{i+1/2}^+} = \frac{\Delta q_{i+1/2}}{\Delta q_{i-1/2}}. \end{aligned}$$

Легко показать, что данная схема является монотонной, если оба выражения (15.33) в круглых скобках удовлетворяют условиям

$$0 \leq (\dots) \leq \frac{h}{\tau |a|}. \quad (15.34)$$

Чтобы обеспечить первое из этих условий, достаточно взять  $\alpha = 0$  для  $R \leq 0$  и  $\alpha \leq 2$  для  $R > 0$ . Для аппроксимации с порядком выше первого

необходимо, чтобы  $\alpha(1) = 1$ . Следуя работе [209], выберем кусочно-линейную функцию  $\alpha(R)$  вида

$$\alpha(R) = \begin{cases} 0, & R \leq 0; \\ \left(\frac{1+\varphi}{2}\beta + \frac{1-\varphi}{2}\right)R, & 0 \leq R \leq \frac{1}{\beta}; \\ \frac{1+\varphi}{2} + \frac{1-\varphi}{2}R, & \frac{1}{\beta} \leq R \leq \beta; \\ \frac{1+\varphi}{2} + \frac{1-\varphi}{2}\beta, & R \geq \beta > 1, \end{cases}$$

или в другой форме

$$\alpha(R) = \frac{1+\varphi}{2} \minmod(1, \beta R) + \frac{1-\varphi}{2} \minmod(\beta, R).$$

Здесь  $\varphi$  и  $\beta$  — свободные параметры, последний должен удовлетворять соотношению  $1 < \beta \leq \beta_{\max}$ , где  $\beta_{\max}$  определяется из условия  $\alpha \leq 2$ :  $\beta_{\max} = (3 - \varphi)/(1 - \varphi)$ . Анализ данной схемы показывает, что она имеет третий порядок аппроксимации, если  $\varphi = 1/3$ , и второй в остальных случаях. Ограничение на временной шаг выводится из второго неравенства (15.34):

$$\frac{\tau|a|}{h} \leq \frac{4}{5 - \varphi + (1 + \varphi)\beta}.$$

Окончательное выражение для потока выглядит следующим образом (с учетом того, что  $\minmod(cx, cy) = c \minmod(x, y)$ ):

$$\begin{aligned} F_{i+1/2} &= a^+ q_i + a^- q_{i+1} + \\ &+ \frac{1+\varphi}{4} \minmod(a^+ \Delta q_{i+1/2}, \beta a^+ \Delta q_{i-1/2}) + \\ &+ \frac{1-\varphi}{4} \minmod(\beta a^+ \Delta q_{i+1/2}, a^+ \Delta q_{i-1/2}) - \\ &- \frac{1+\varphi}{4} \minmod(a^- \Delta q_{i+1/2}, \beta a^- \Delta q_{i+3/2}) - \\ &- \frac{1-\varphi}{4} \minmod(\beta a^- \Delta q_{i+1/2}, a^- \Delta q_{i+3/2}) = \\ &= F_{i+1/2}^I + \frac{1+\varphi}{4} \minmod(F_{i+1/2}^+, \beta F_{i-1/2}^+) + \\ &+ \frac{1-\varphi}{4} \minmod(\beta F_{i+1/2}^+, F_{i-1/2}^+) - \\ &- \frac{1+\varphi}{4} \minmod(F_{i+1/2}^-, \beta F_{i+3/2}^-) - \frac{1-\varphi}{4} \minmod(\beta F_{i+1/2}^-, F_{i+3/2}^-). \end{aligned}$$

Рассмотрим теперь линейную систему гиперболических уравнений (15.20). Монотонная схема для инвариантов Римана  $s = \mathcal{L}\mathbf{q}$  строится так же, как и для решения скалярного линейного уравнения переноса:

$$\begin{aligned} \frac{\hat{s}_i^k - s_i^k}{\tau} + \frac{(\mathcal{L}\mathbf{F})_{i+1/2}^k - (\mathcal{L}\mathbf{F})_{i-1/2}^k}{h} &= 0, \\ (\mathcal{L}\mathbf{F})_{i+1/2}^k &= \lambda^{k+} s_i^k + \lambda^{k-} s_{i+1}^k + \\ &+ \frac{1+\varphi}{4} \text{minmod}(\lambda^{k+} \Delta s_{i+1/2}^k, \beta \lambda^{k+} \Delta s_{i-1/2}^k) + \\ &+ \frac{1-\varphi}{4} \text{minmod}(\beta \lambda^{k+} \Delta s_{i+1/2}^k, \lambda^{k+} \Delta s_{i-1/2}^k) - \\ &- \frac{1+\varphi}{4} \text{minmod}(\lambda^{k-} \Delta s_{i+1/2}^k, \beta \lambda^{k-} \Delta s_{i+3/2}^k) - \\ &- \frac{1-\varphi}{4} \text{minmod}(\beta \lambda^{k-} \Delta s_{i+1/2}^k, \lambda^{k-} \Delta s_{i+3/2}^k). \end{aligned}$$

Переходя к переменным  $\mathbf{q} = \mathcal{R}s = \sum_m s^m \mathbf{r}^m$ , получим

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0,$$

$$\begin{aligned} \mathbf{F}_{i+1/2} &= \mathcal{A}^+ \mathbf{q}_i + \mathcal{A}^- \mathbf{q}_{i+1} + \\ &+ \frac{1+\varphi}{4} \sum_m \text{minmod}(\lambda^{m+} \Delta s_{i+1/2}^m \mathbf{r}^m, \beta \lambda^{m+} \Delta s_{i-1/2}^m \mathbf{r}^m) + \\ &+ \frac{1-\varphi}{4} \sum_m \text{minmod}(\beta \lambda^{m+} \Delta s_{i+1/2}^m \mathbf{r}^m, \lambda^{m+} \Delta s_{i-1/2}^m \mathbf{r}^m) - \\ &- \frac{1+\varphi}{4} \sum_m \text{minmod}(\lambda^{m-} \Delta s_{i+1/2}^m \mathbf{r}^m, \beta \lambda^{m-} \Delta s_{i+3/2}^m \mathbf{r}^m) - \\ &- \frac{1-\varphi}{4} \sum_m \text{minmod}(\beta \lambda^{m-} \Delta s_{i+1/2}^m \mathbf{r}^m, \lambda^{m-} \Delta s_{i+3/2}^m \mathbf{r}^m). \end{aligned}$$

Окончательно поток залишем в виде

$$\begin{aligned} \mathbf{F}_{i+1/2} &= \mathbf{F}_{i+1/2}^I + \frac{1+\varphi}{4} \sum_m \text{minmod}(\mathbf{F}_{i+1/2}^{m+}, \beta \mathbf{F}_{i-1/2}^{m+}) + \\ &+ \frac{1-\varphi}{4} \sum_m \text{minmod}(\beta \mathbf{F}_{i-1/2}^{m+}, \mathbf{F}_{i+1/2}^{m+}) - \\ &- \frac{1+\varphi}{4} \sum_m \text{minmod}(\mathbf{F}_{i+1/2}^{m-}, \beta \mathbf{F}_{i+3/2}^{m-}) - \\ &- \frac{1-\varphi}{4} \sum_m \text{minmod}(\beta \mathbf{F}_{i+3/2}^{m-}, \mathbf{F}_{i+1/2}^{m-}), \quad (15.35) \end{aligned}$$

где

$$\mathbf{F}_{i+1/2}^I = \mathcal{A}^+ \mathbf{q}_i + \mathcal{A}^- \mathbf{q}_{i+1} = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{|\mathcal{A}|}{2} (\mathbf{q}_{i+1} - \mathbf{q}_i);$$

$$\mathbf{F}_{i+1/2}^{m+} = \lambda^{m+} \Delta s_{i+1/2}^m \mathbf{r}^m = \lambda^{m+} (\mathbf{l}^m \cdot (\mathbf{q}_{i+1} - \mathbf{q}_i)) \mathbf{r}^m;$$

$$\mathbf{F}_{i+1/2}^{m-} = \lambda^{m-} \Delta s_{i+1/2}^m \mathbf{r}^m = \lambda^{m-} (\mathbf{l}^m \cdot (\mathbf{q}_{i+1} - \mathbf{q}_i)) \mathbf{r}^m.$$

И, наконец, схема для решения уравнений Эйлера строится по аналогии со схемой (15.35) для решения линейной системы гиперболических уравнений, причем потоки первого порядка аппроксимации вычисляются по формулам *схемы Ру — Эйнфельдта* (15.32). В результате получаем *схему Ру — Эйнфельдта — Ошера*

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0, \quad (15.36)$$

где

$$\begin{aligned} \mathbf{F}_{i+1/2} = & \mathbf{F}_{i+1/2}^I + \frac{1+\varphi}{4} \sum_m \text{minmod}(\mathbf{F}_{i+1/2}^{m+}, \beta \mathbf{F}_{i-1/2}^{m+}) + \\ & + \frac{1-\varphi}{4} \sum_m \text{minmod}(\beta \mathbf{F}_{i-1/2}^{m+}, \mathbf{F}_{i+1/2}^{m+}) - \\ & - \frac{1+\varphi}{4} \sum_m \text{minmod}(\mathbf{F}_{i+1/2}^{m-}, \beta \mathbf{F}_{i+3/2}^{m-}) - \\ & - \frac{1-\varphi}{4} \sum_m \text{minmod}(\beta \mathbf{F}_{i+3/2}^{m-}, \mathbf{F}_{i+1/2}^{m-}), \end{aligned}$$

$$\mathbf{F}_{i+1/2}^I = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2} \sum_m |\tilde{\lambda}^m(\mathbf{q}^*)| \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*),$$

$$\mathbf{F}_{i+1/2}^{m+} = \lambda^{m+}(\mathbf{q}^*) \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*) = \lambda^{m+}(\mathbf{q}^*) (\mathbf{l}^m(\mathbf{q}^*) \cdot (\mathbf{q}_{i+1} - \mathbf{q}_i)) \mathbf{r}^m(\mathbf{q}^*),$$

$$\mathbf{F}_{i+1/2}^{m-} = \lambda^{m-}(\mathbf{q}^*) \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*) = \lambda^{m-}(\mathbf{q}^*) (\mathbf{l}^m(\mathbf{q}^*) \cdot (\mathbf{q}_{i+1} - \mathbf{q}_i)) \mathbf{r}^m(\mathbf{q}^*).$$

Результаты расчета теста 1 и теста 3 по схеме Ру — Эйнфельдта — Ошера (15.36) представлены на рис. 15.8 и 15.9 ( $\star$  — расчет по схеме Ру — Эйнфельдта — Ошера; — — точное решение; ВР — волна разрежения; КР — контактный разрыв; УВ — ударная волна; — — — скорость звука). Как и ранее, расчет проводился на сетке с 200 ячейками, показана каждая 4-я ячейка, в окрестности КР и УВ показана каждая ячейка. Видно, что эта схема почти точно воспроизводит аналитическое решение.

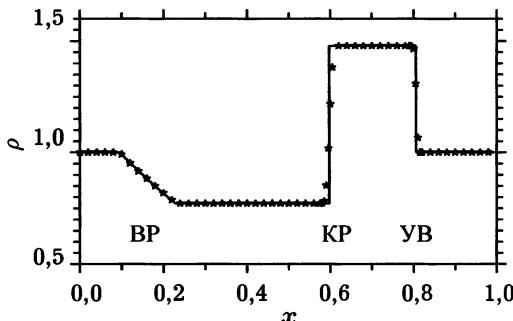


Рис. 15.8

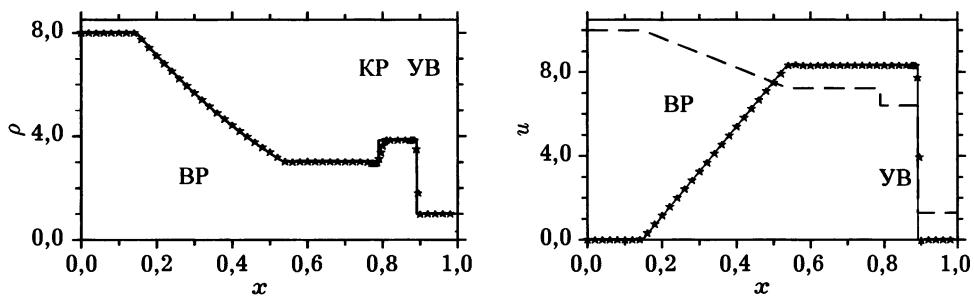


Рис. 15.9

Результаты расчета теста 2 по схеме Рой — Эйнфельдта — Ошера представлены на рис. 15.10. На рис. 15.10, *a* показаны изолинии плотности на плоскости  $(x, t)$ , а на рис. 15.10, *б* приведено увеличенное изображение прямоугольника на рис. 15.10, *a*. Видно, что все КР, включая КР3, передаются очень хорошо, равно как и ВР.

Следует отметить, что схема Рой — Эйнфельдта — Ошера повышенного порядка аппроксимации имеет расширенный шаблон в отличие от других рассмотренных схем, в которых поток на грани между двумя ячейками вычисляется по значениям параметров в этих ячейках. Расширенный шаблон затрудняет использование схемы Рой — Эйнфельдта — Ошера в алгоритмах сеточной адаптации [217, 246].

### 15.2.7. Схема Рой для двумерной газовой динамики

Схема Рой для решения уравнений двумерной газовой динамики строится аналогично одномерному случаю:

$$\frac{\hat{\mathbf{q}}_{i,j} - \mathbf{q}_{i,j}}{\tau} + \frac{\mathbf{F}_{i+1/2,j} - \mathbf{F}_{i-1/2,j}}{h} + \frac{\mathbf{G}_{i,j+1/2} - \mathbf{G}_{i,j-1/2}}{h} = 0,$$

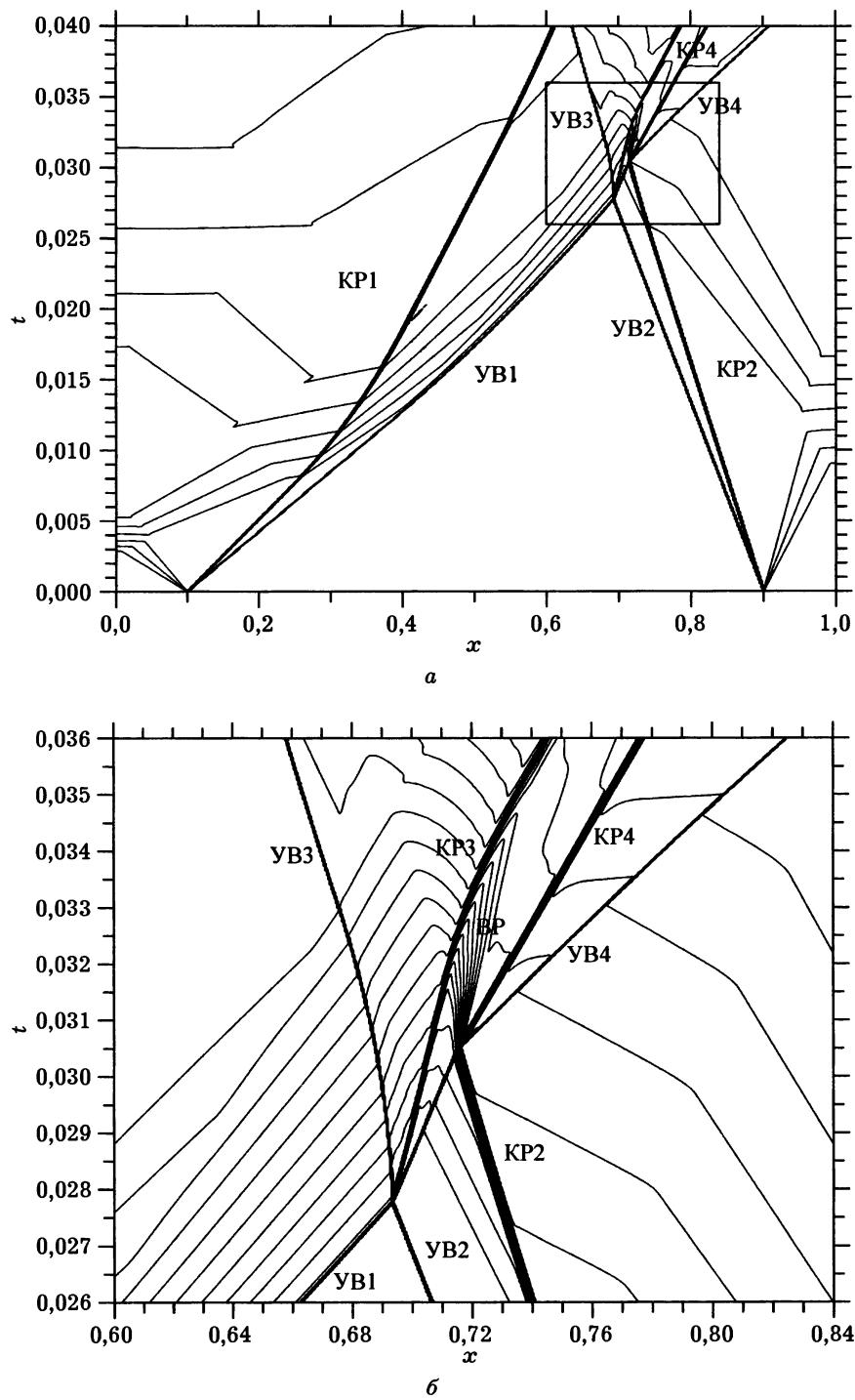


Рис. 15.10

где

$$\begin{aligned}
 \mathbf{F}_{i+1/2,j} &= \frac{\mathbf{F}_{i,j} + \mathbf{F}_{i+1,j}}{2} - \frac{1}{2} \sum_m |\lambda_{\mathcal{A}}^m(\mathbf{q}_{i+1/2,j}^*)| \Delta s_{i+1/2,j}^m \mathbf{r}_{\mathcal{A}}^m(\mathbf{q}_{i+1/2,j}^*); \\
 \Delta s_{i+1/2,j}^m &= l_{\mathcal{A}}^m(\mathbf{q}_{i+1/2,j}^*)(\mathbf{q}_{i+1,j} - \mathbf{q}_{i,j}); \\
 \Delta s_{i+1/2,j}^{1,4} &= \frac{(P_{i+1,j} - P_{i,j}) \mp \rho_{i+1/2,j}^* c_{i+1/2,j}^* (u_{i+1,j} - u_{i,j})}{2c_{i+1/2,j}^{*2}}; \\
 \Delta s_{i+1/2,j}^{2,3} &= \frac{c_{i+1/2,j}^{*2} (\rho_{i+1,j} - \rho_{i,j}) - (P_{i+1,j} - P_{i,j}) \pm \rho_{i+1/2,j}^* c_{i+1/2,j}^* (v_{i+1,j} - v_{i,j})}{2c_{i+1/2,j}^{*2}}; \\
 \rho_{i+1/2,j}^* &= \sqrt{\rho_{i,j} \rho_{i+1,j}}; \quad u_{i+1/2,j}^* = \frac{\sqrt{\rho_{i,j}} u_{i,j} + \sqrt{\rho_{i+1,j}} u_{i+1,j}}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i+1,j}}}; \\
 v_{i+1/2,j}^* &= \frac{\sqrt{\rho_{i,j}} v_{i,j} + \sqrt{\rho_{i+1,j}} v_{i+1,j}}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i+1,j}}}; \quad h_{i+1/2,j}^{**} = \frac{\sqrt{\rho_{i,j}} h_{i,j}^* + \sqrt{\rho_{i+1,j}} h_{i+1,j}^*}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i+1,j}}}; \\
 c_{i+1/2,j}^* &= \sqrt{\frac{\sqrt{\rho_{i,j}} c_{i,j}^2 + \sqrt{\rho_{i+1,j}} c_{i+1,j}^2}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i+1,j}}} + \frac{\gamma-1}{2} \frac{\sqrt{\rho_{i,j} \rho_{i+1,j}}}{(\sqrt{\rho_{i,j}} + \sqrt{\rho_{i+1,j}})^2} (\mathbf{v}_{i+1,j} - \mathbf{v}_{i,j})^2}; \\
 \mathbf{G}_{i,j+1/2} &= \frac{\mathbf{G}_{i,j} + \mathbf{G}_{i,j+1}}{2} - \frac{1}{2} \sum_m |\lambda_{\mathcal{B}}^m(\mathbf{q}_{i,j+1/2}^*)| \Delta s_{i,j+1/2}^m \mathbf{r}_{\mathcal{B}}^m(\mathbf{q}_{i,j+1/2}^*); \\
 \Delta s_{i,j+1/2}^m &= l_{\mathcal{B}}^m(\mathbf{q}_{i,j+1/2}^*)(\mathbf{q}_{i,j+1} - \mathbf{q}_{i,j}); \\
 \Delta s_{i,j+1/2}^{1,4} &= \frac{(P_{i,j+1} - P_{i,j}) \mp \rho_{i,j+1/2}^* c_{i,j+1/2}^* (v_{i,j+1} - v_{i,j})}{2c_{i,j+1/2}^{*2}}; \\
 \Delta s_{i,j+1/2}^{2,3} &= \frac{c_{i,j+1/2}^{*2} (\rho_{i,j+1} - \rho_{i,j}) - (P_{i,j+1} - P_{i,j}) \pm \rho_{i,j+1/2}^* c_{i,j+1/2}^* (u_{i,j+1} - u_{i,j})}{2c_{i,j+1/2}^{*2}}; \\
 \rho_{i,j+1/2}^* &= \sqrt{\rho_{i,j} \rho_{i,j+1}}; \quad u_{i,j+1/2}^* = \frac{\sqrt{\rho_{i,j}} u_{i,j} + \sqrt{\rho_{i,j+1}} u_{i,j+1}}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i,j+1}}}; \\
 v_{i,j+1/2}^* &= \frac{\sqrt{\rho_{i,j}} v_{i,j} + \sqrt{\rho_{i,j+1}} v_{i,j+1}}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i,j+1}}}; \quad h_{i,j+1/2}^{**} = \frac{\sqrt{\rho_{i,j}} h_{i,j}^* + \sqrt{\rho_{i,j+1}} h_{i,j+1}^*}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i,j+1}}}; \\
 c_{i,j+1/2}^* &= \sqrt{\frac{\sqrt{\rho_{i,j}} c_{i,j}^2 + \sqrt{\rho_{i,j+1}} c_{i,j+1}^2}{\sqrt{\rho_{i,j}} + \sqrt{\rho_{i,j+1}}} + \frac{\gamma-1}{2} \frac{\sqrt{\rho_{i,j} \rho_{i,j+1}}}{(\sqrt{\rho_{i,j}} + \sqrt{\rho_{i,j+1}})^2} (\mathbf{v}_{i,j+1} - \mathbf{v}_{i,j})^2}.
 \end{aligned}$$

Заменяя  $\lambda^m$  на  $\tilde{\lambda}^m$ , согласно (15.32), получаем схему Рой — Эйнфельдта, а добавляя антидиффузионные члены, согласно (15.36), — схему Рой — Эйнфельдта — Ошера.

Для сравнения разностных схем Лакса — Фридрихса, Роу — Эйнфельдта и Роу — Эйнфельдта — Ошера повышенного порядка аппроксимации, используемых для решения уравнений двумерной газовой динамики, рассмотрена задача о течении сверхзвукового потока с числом Маха  $M = 3$  в канале со ступенькой [224, 292]. Ширина канала принималась равной 1, длина — 3, высота ступеньки — 0,2, ее левый край располагался на расстоянии 0,6 от левой границы канала. Газ считался идеальным с показателем адиабаты  $\gamma = 7/5$ . На левой границе задавались граничные значения  $\rho = 1$ ,  $u = 3$ ,  $v = 0$  и  $P = 1/\gamma$ , на правой границе течение предполагалось также сверхзвуковым, поэтому граничные условия на правой границе не задавались. Верхняя и нижняя границы считались твердыми. Работы по численному решению этой задачи [224, 282, 288, 292] показали, что после касания отошедшей ударной волны с верхней стенкой образуются «ножка Маха» и контактный разрыв, в котором развивается неустойчивость Кельвина — Гельмгольца.

На рис. 15.11–15.14 представлены результаты расчета этой задачи (в дальнейшем — тест 4) по различным схемам и на различных сетках. Показаны изолинии плотности и векторы скорости для расчета задачи о течении сверхзвукового потока в канале со ступенькой.

На рис. 15.11 представлены результаты расчета теста 4 на момент времени  $t = 4$  по схеме Лакса — Фридрихса (двумерный вариант (15.24)). Видно, что ударные волны сильно размазаны. «Ножка Маха» образуется при расчете на сетке  $320 \times 960$ , но контактный разрыв за ней не разрешается.

На рис. 15.12 представлены результаты расчета теста 4 на момент времени  $t = 4$  по схеме Роу (двумерный вариант (15.30)). Ударные волны передаются лучше, однако на углу ступеньки образуется волна разрежения.

На рис. 15.13 представлены результаты расчета теста 4 на момент времени  $t = 4$  по схеме Роу — Эйнфельдта (двумерный вариант (15.32)). Здесь течение на углу ступеньки передается корректно — образуется простая волна разрежения.

И, наконец, на рис. 15.14 представлены результаты расчета теста 4 на момент времени  $t = 4$  по схеме Роу — Эйнфельдта — Ошера (двумерный вариант (15.36)) повышенного порядка аппроксимации. Отчетливо виден контактный разрыв за «ножкой Маха» и развивающаяся в нем неустойчивость Кельвина — Гельмгольца. Эта неустойчивость особенно хорошо видна на рис. 15.15, где показаны изолинии энтропии  $P/\rho^\gamma$  в области контактного разрыва для расчета на сетке  $320 \times 960$ .

Помимо качества приближенного решения, полученного по той или иной схеме, важной характеристикой являются вычислительные затра-

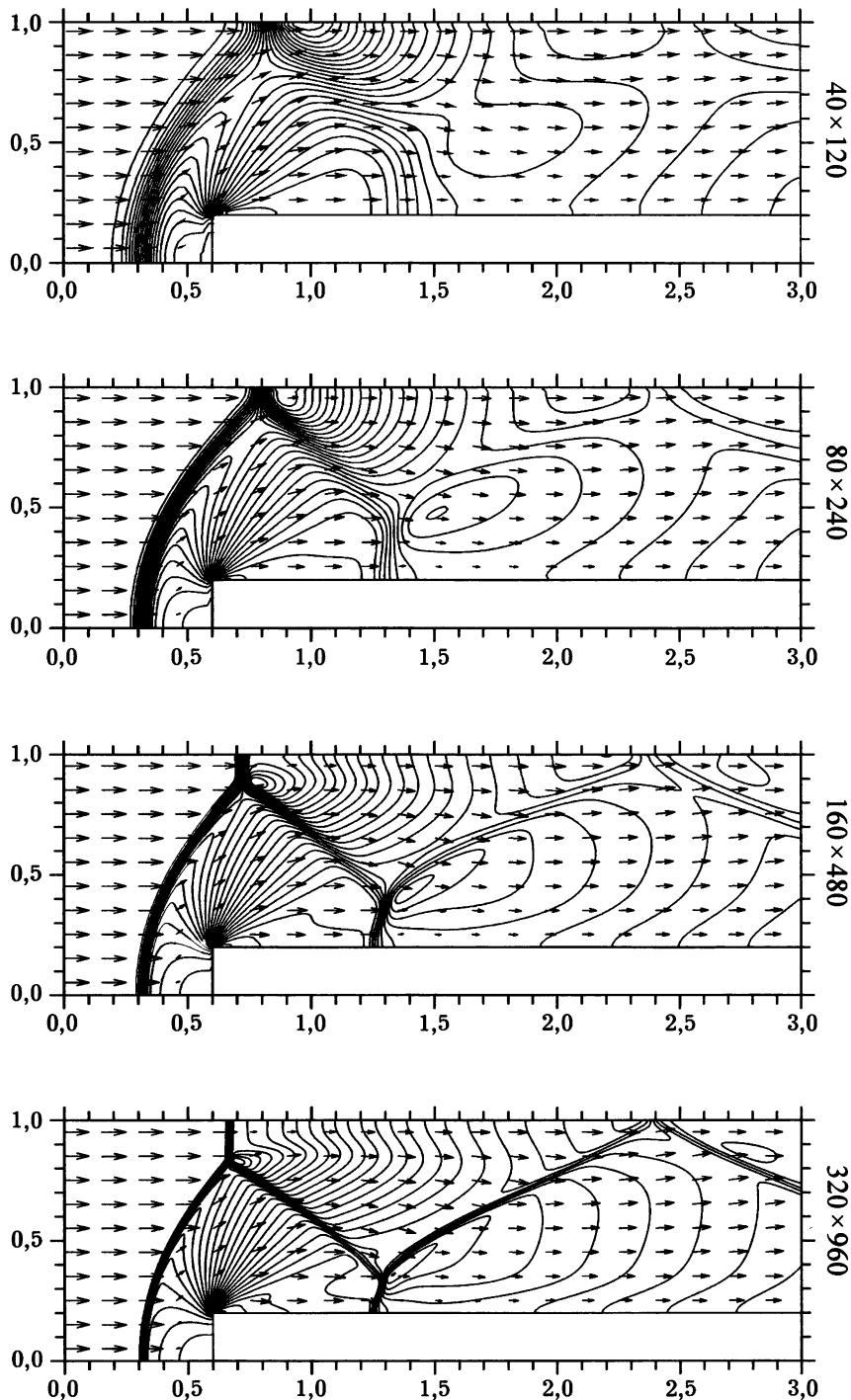


Рис. 15.11

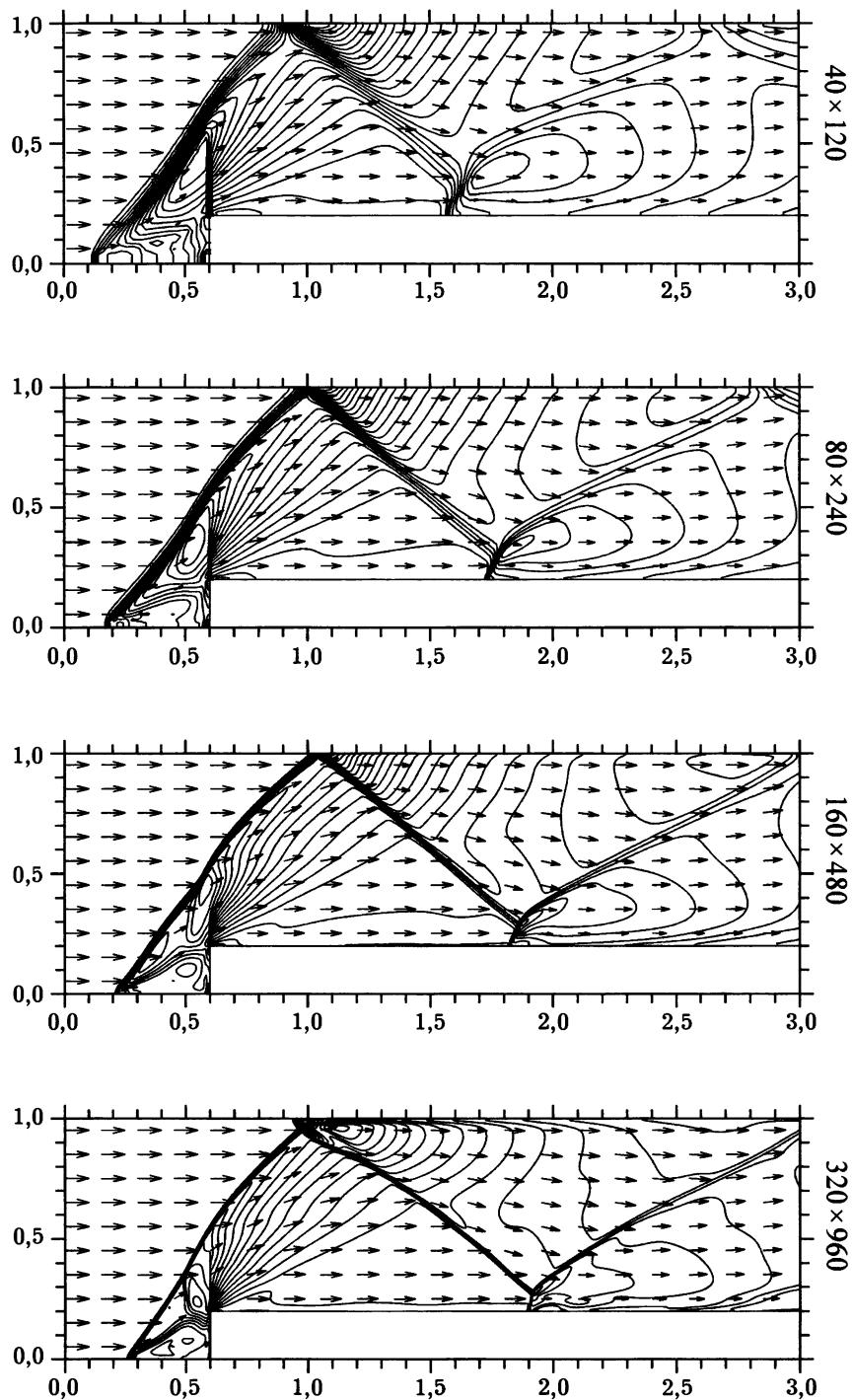


Рис. 15.12

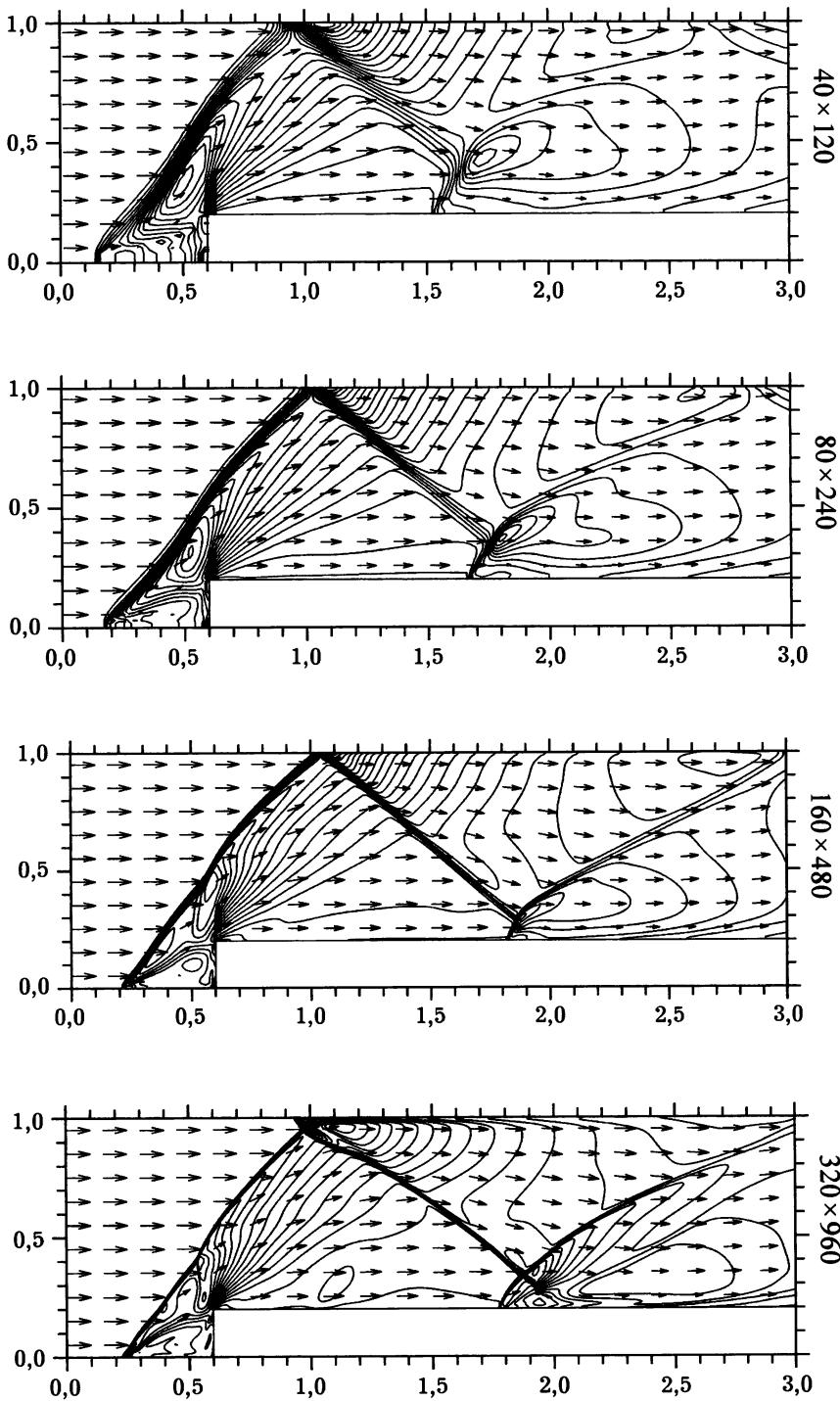


Рис. 15.13

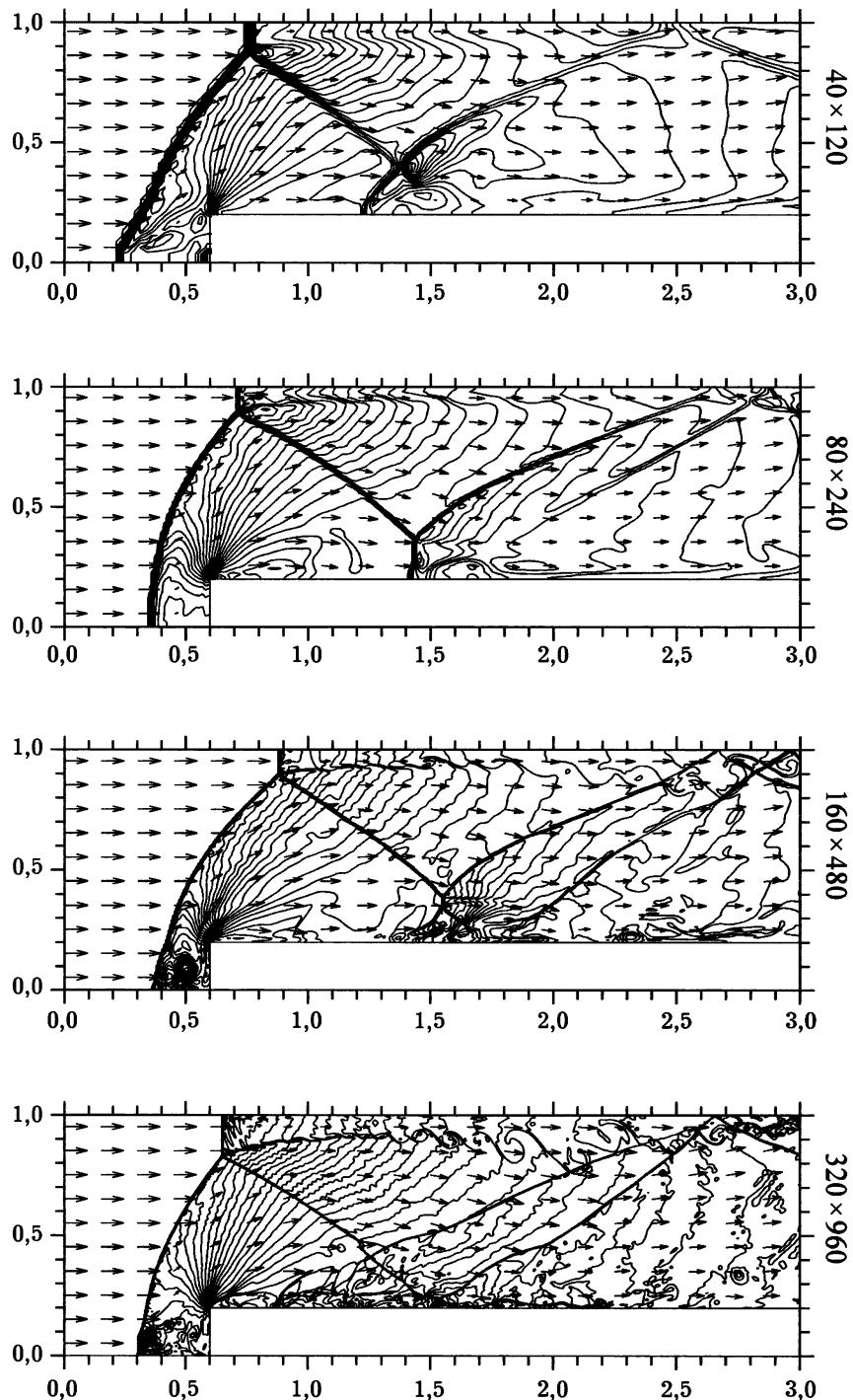


Рис. 15.14

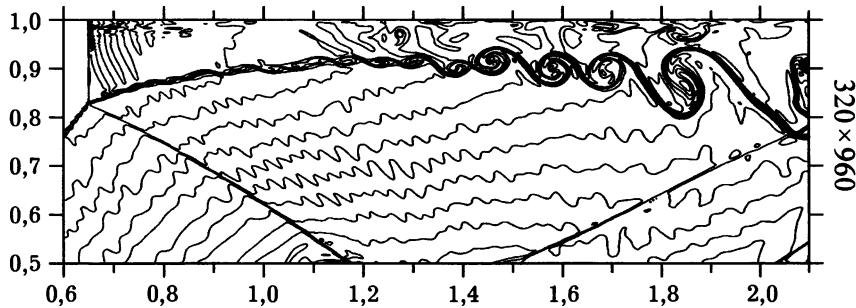


Рис. 15.15

ты. В табл. 15.1 представлено отношение расчетного времени в секундах (на компьютере Pentium/200) к модельному для различных схем и сеток. Расчеты проводились с числом Куранта 0,6 для первых трех схем и с числом 0,1 для схемы Рой — Эйнфельдта — Ошера. Отметим резкое возрастание (в 18 раз) вычислительной трудоемкости расчета задачи по схеме Рой — Эйнфельдта — Ошера по сравнению с расчетом по другим схемам, обусловленное в основном ограничением числа Куранта (в 6 раз) и в меньшей степени (в 3 раза) усложнением алгоритма.

Таблица 15.1

Схема	Сетка			
	40 × 120	80 × 240	160 × 480	320 × 960
Лакса — Фридрихса	30	240	1900	15000
Рой	50	455	3600	29000
Рой — Эйнфельдта	60	460	3700	29500
Рой — Эйнфельдта — Ошера	1100	9000	72000	575000

### 15.2.8. Упрощенная схема Рой — Эйнфельдта — Ошера и схема Лакса — Фридрихса — Ошера

Схема Рой — Эйнфельдта — Ошера, обладая высокой точностью расчета, имеет существенный недостаток — большое время расчета, которое прежде всего определяется многократными обращениями к функции *minmod*. Для уменьшения времени расчетов следует использовать *упрощенную схему Рой — Эйнфельдта — Ошера*, в которой дополнительные потоки  $F_{i+1/2}^{m+}$  и  $F_{i+1/2}^{m-}$  просуммированы для положительных и отрицательных значений  $\lambda$  и обращение к функции *minmod* проводится меньшее количество раз (для простоты приведен одномерный вариант):

$$\frac{\hat{q}_i - q_i}{\tau} + \frac{F_{i+1/2} - F_{i-1/2}}{h} = 0, \quad (15.37)$$

где

$$\begin{aligned}\mathbf{F}_{i+1/2} = & \mathbf{F}_{i+1/2}^I + \frac{1+\varphi}{4} \text{minmod}(\mathbf{F}_{i+1/2}^+, \beta \mathbf{F}_{i-1/2}^+) + \\ & + \frac{1-\varphi}{4} \text{minmod}(\beta \mathbf{F}_{i-1/2}^+, \mathbf{F}_{i+1/2}^+) - \\ & - \frac{1+\varphi}{4} \text{minmod}(\mathbf{F}_{i+1/2}^-, \beta \mathbf{F}_{i+3/2}^-) - \\ & - \frac{1-\varphi}{4} \text{minmod}(\beta \mathbf{F}_{i+3/2}^-, \mathbf{F}_{i+1/2}^-);\\ \mathbf{F}_{i+1/2}^I = & \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2} \sum_m |\tilde{\lambda}^m(\mathbf{q}^*)| \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*);\end{aligned}$$

$$\mathbf{F}_{i+1/2}^+ = \sum_{\lambda^m \geq 0} \lambda^m(\mathbf{q}^*) \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*) = \sum_{\lambda^m \geq 0} \lambda^m(\mathbf{q}^*) (\mathbf{l}^m(\mathbf{q}^*) \cdot (\mathbf{q}_{i+1} - \mathbf{q}_i)) \mathbf{r}^m(\mathbf{q}^*);$$

$$\mathbf{F}_{i+1/2}^- = \sum_{\lambda^m \leq 0} \lambda^m(\mathbf{q}^*) \Delta s_{i+1/2}^m \mathbf{r}^m(\mathbf{q}^*) = \sum_{\lambda^m \leq 0} \lambda^m(\mathbf{q}^*) (\mathbf{l}^m(\mathbf{q}^*) \cdot (\mathbf{q}_{i+1} - \mathbf{q}_i)) \mathbf{r}^m(\mathbf{q}^*).$$

Результаты расчета теста 4 по упрощенной схеме Рой — Эйнфельдта — Ошера (15.37) представлены на рис. 15.16 (изображены изолинии плотности и векторы скорости), из которого видно, что решение по упрощенной схеме Рой — Эйнфельдта — Ошера является немонотонным и обладает очень сильными осцилляциями.

Логическим продолжением описанного упрощения является **схема Лакса — Фридрихса — Ошера** [33], в которой поток первого порядка вычисляется по схеме Лакса — Фридрихса (15.24), а поправки высокого порядка строятся по аналогии с упрощенной схемой Рой — Эйнфельдта — Ошера (15.37):

$$\frac{\hat{\mathbf{q}}_i - \mathbf{q}_i}{\tau} + \frac{\mathbf{F}_{i+1/2} - \mathbf{F}_{i-1/2}}{h} = 0, \quad (15.38)$$

где

$$\begin{aligned}\mathbf{F}_{i+1/2} = & \mathbf{F}_{i+1/2}^I + \frac{1+\varphi}{4} \text{minmod}(\mathbf{F}_{i+1/2}^+, \beta \mathbf{F}_{i-1/2}^+) + \\ & + \frac{1-\varphi}{4} \text{minmod}(\beta \mathbf{F}_{i-1/2}^+, \mathbf{F}_{i+1/2}^+) - \\ & - \frac{1+\varphi}{4} \text{minmod}(\mathbf{F}_{i+1/2}^-, \beta \mathbf{F}_{i+3/2}^-) - \\ & - \frac{1-\varphi}{4} \text{minmod}(\beta \mathbf{F}_{i+3/2}^-, \mathbf{F}_{i+1/2}^-);\\ \mathbf{F}_{i+1/2}^I = & \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{\nu}{2} (\mathbf{q}_{i+1} - \mathbf{q}_i);\end{aligned}$$

$$\nu = \max(|\lambda_i^1|, |\lambda_i^2|, \dots, |\lambda_i^N|, |\lambda_{i+1}^1|, |\lambda_{i+1}^2|, \dots, |\lambda_{i+1}^N|);$$

$$\mathbf{F}_{i+1/2}^+ = \mathbf{F}_{i+1} - \mathbf{F}_{i+1/2}^I; \quad \mathbf{F}_{i+1/2}^- = \mathbf{F}_{i+1/2}^I - \mathbf{F}_i.$$

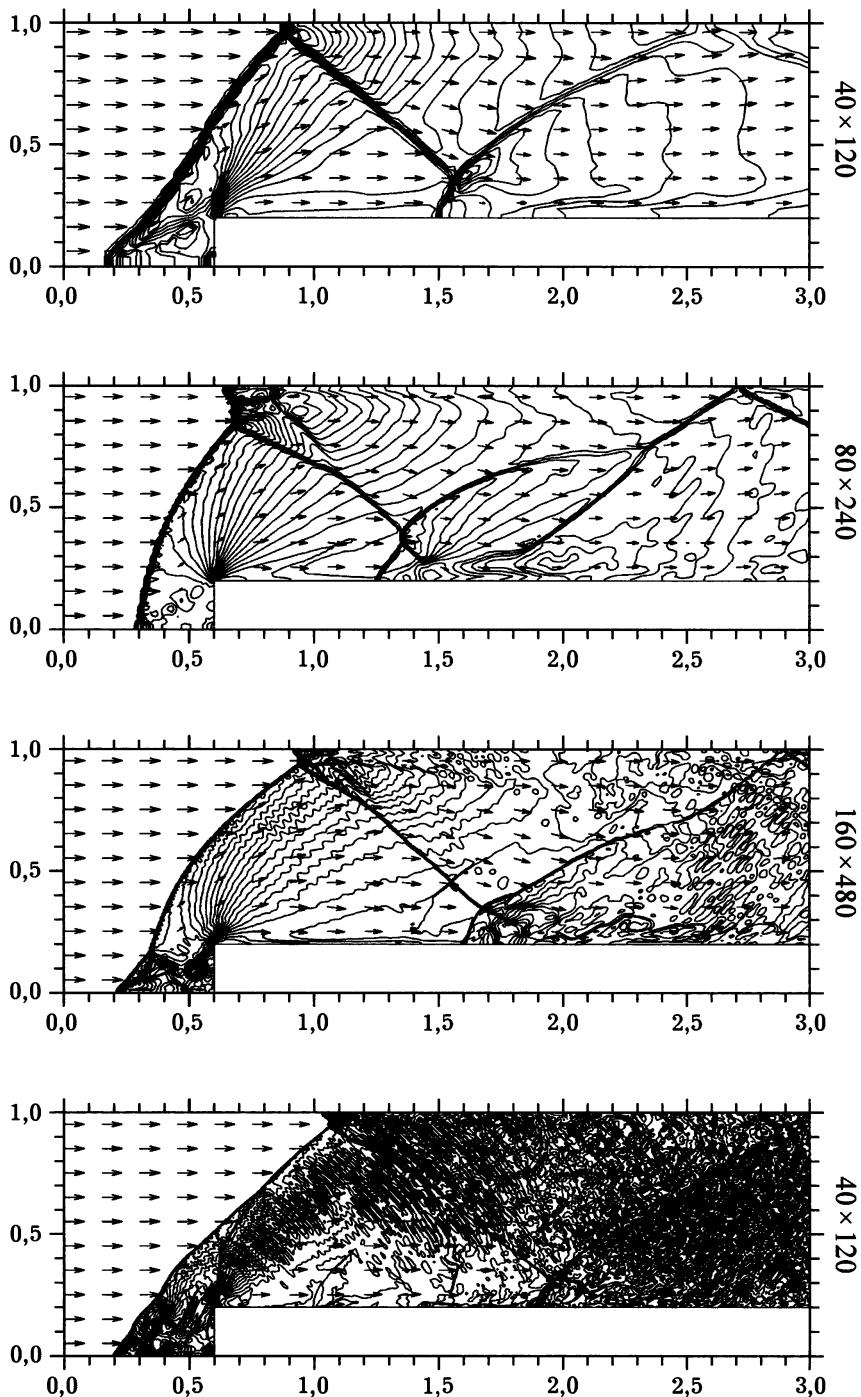


Рис. 15.16

Результаты расчета теста 4 по схеме Лакса — Фридрихса — Ошера (15.38) представлены на рис. 15.17 (изображены изолинии плотности и векторы скорости). Видно, что при расчете по *схеме Лакса — Фридрихса — Ошера* получаются вполне удовлетворительные результаты, хотя отсутствие неустойчивости Кельвина — Гельмгольца показывает, что ее аппроксимационная вязкость выше, чем при расчете по *схеме Роу — Эйнфельдта — Ошера*. Однако расчет по *схеме Лакса — Фридрихса — Ошера* выполняется в 2,5 раза быстрее, чем по *схеме Роу — Эйнфельдта — Ошера*.

### 15.2.9. Схема Роу для решения уравнений трехмерной газовой динамики

Представим *схему Роу для решения уравнений трехмерной газовой динамики*, построенную аналогично одномерному и двумерному случаям:

$$\frac{\hat{\mathbf{q}}_{i,j,k} - \mathbf{q}_{i,j,k}}{\tau} + \frac{\mathbf{F}_{i+1/2,j,k} - \mathbf{F}_{i-1/2,j,k}}{h} + \\ + \frac{\mathbf{G}_{i,j+1/2,k} - \mathbf{G}_{i,j-1/2,k}}{h} + \frac{\mathbf{H}_{i,j,k+1/2} - \mathbf{H}_{i,j,k-1/2}}{h} = 0,$$

где

$$\begin{aligned}\mathbf{q} &= (\rho, \rho u, \rho v, \rho w, \rho E)^T; \\ \mathbf{F} &= (\rho u, \rho u^2 + P, \rho uv, \rho uw, \rho uh^*)^T; \\ \mathbf{G} &= (\rho v, \rho uv, \rho v^2 + P, \rho vw, \rho vh^*)^T; \\ \mathbf{H} &= (\rho w, \rho uw, \rho vw, \rho w^2 + P, \rho wh^*)^T; \\ \mathbf{v} &= (u, v, w)^T; \quad E = \varepsilon + \frac{\mathbf{v}^2}{2}; \quad h^* = \varepsilon + \frac{P}{\rho} + \frac{\mathbf{v}^2}{2}.\end{aligned}$$

Запишем расчетные формулы для трехмерного случая. Слагаемые, связанные с координатой  $x$ , имеют следующий вид:

$$\mathcal{A} = \frac{\partial \mathbf{F}}{\partial \mathbf{q}},$$

$$\begin{aligned}\mathbf{F}_{i+1/2,j,k} &= \frac{\mathbf{F}_{i,j,k} + \mathbf{F}_{i+1,j,k}}{2} - \\ &- \frac{1}{2} \sum_m |\lambda_{\mathcal{A}}^m(\mathbf{q}_{i+1/2,j,k}^*)| \Delta s_{i+1/2,j,k}^m \mathbf{r}_{\mathcal{A}}^m(\mathbf{q}_{i+1/2,j,k}^*),\end{aligned}$$

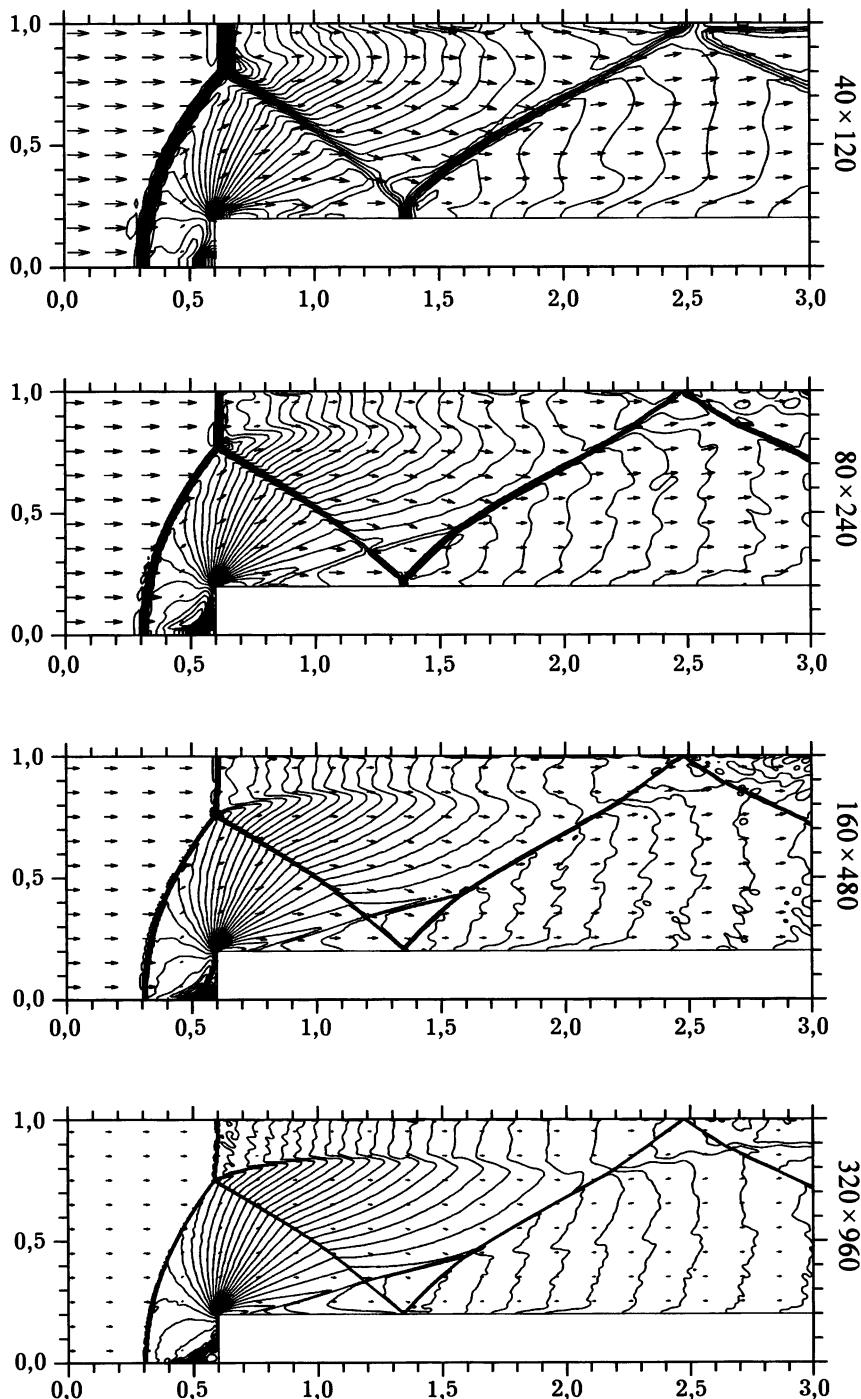


Рис. 15.17

где

$$\Delta s_{i+1/2,j,k}^{1,5} = \frac{(P_{i+1,j,k} - P_{i,j,k}) \mp \rho_{i+1/2,j,k}^* c_{i+1/2,j,k}^* (u_{i+1,j,k} - u_{i,j,k})}{2c_{i+1/2,j,k}^{*2}};$$

$$\Delta s_{i+1/2,j,k}^2 = \frac{\rho_{i+1/2,j,k}^* c_{i+1/2,j,k}^* (v_{i+1,j,k} - v_{i,j,k})}{2c_{i+1/2,j,k}^{*2}};$$

$$\Delta s_{i+1/2,j,k}^3 = \frac{\rho_{i+1/2,j,k}^* c_{i+1/2,j,k}^* (w_{i+1,j,k} - w_{i,j,k})}{2c_{i+1/2,j,k}^{*2}};$$

$$\Delta s_{i+1/2,j,k}^4 = \frac{c_{i+1/2,j,k}^{*2} (\rho_{i+1,j,k} - \rho_{i,j,k}) - (P_{i+1,j,k} - P_{i,j,k})}{2c_{i+1/2,j,k}^{*2}};$$

$$\rho_{i+1/2,j,k}^* = \sqrt{\rho_{i,j,k} \rho_{i+1,j,k}}, \quad \mathbf{v}_{i+1/2,j,k}^* = \frac{\sqrt{\rho_{i,j,k}} \mathbf{v}_{i,j,k} + \sqrt{\rho_{i+1,j,k}} \mathbf{v}_{i+1,j,k}}{\sqrt{\rho_{i,j,k}} + \sqrt{\rho_{i+1,j,k}}};$$

$$h_{i+1/2,j,k}^{**} = \frac{\sqrt{\rho_{i,j,k}} h_{i,j,k}^* + \sqrt{\rho_{i+1,j,k}} h_{i+1,j,k}^*}{\sqrt{\rho_{i,j,k}} + \sqrt{\rho_{i+1,j,k}}};$$

$$\mathcal{R}_A = \begin{pmatrix} 1 & 0 & 0 & 2 & 1 \\ u-c & 0 & 0 & 2u & u+c \\ v & 2c & 0 & 2v & v \\ w & 0 & 2c & 2w & w \\ h^* - uc & 2vc & 2wc & \mathbf{v}^2 & h^* + uc \end{pmatrix}, \quad \Lambda_A = \text{diag}(u-c, u, u, u, u+c).$$

Слагаемые, связанные с координатой  $y$ , имеют следующий вид:

$$\mathcal{B} = \frac{\partial \mathbf{G}}{\partial \mathbf{q}},$$

$$\mathbf{G}_{i,j+1/2,k} = \frac{\mathbf{G}_{i,j,k} + \mathbf{G}_{i,j+1,k}}{2} - \frac{1}{2} \sum_m |\lambda_B^m(\mathbf{q}_{i,j+1/2,k}^*)| \Delta s_{i,j+1/2,k}^m \mathbf{r}_B^m(\mathbf{q}_{i,j+1/2,k}^*),$$

где

$$\Delta s_{i,j+1/2,k}^{1,5} = \frac{(P_{i,j+1,k} - P_{i,j,k}) \mp \rho_{i,j+1/2,k}^* c_{i,j+1/2,k}^* (v_{i,j+1,k} - v_{i,j,k})}{2c_{i,j+1/2,k}^{*2}};$$

$$\Delta s_{i,j+1/2,k}^2 = \frac{1}{2c_{i,j+1/2,k}^{*2}} [\rho_{i,j+1/2,k}^* c_{i,j+1/2,k}^* (u_{i,j+1,k} - u_{i,j,k})];$$

$$\Delta s_{i,j+1/2,k}^3 = \frac{1}{2c_{i,j+1/2,k}^{*2}} [\rho_{i,j+1/2,k}^* c_{i,j+1/2,k}^* (w_{i,j+1,k} - w_{i,j,k})];$$

$$\Delta s_{i,j+1/2,k}^4 = \frac{1}{2c_{i,j+1/2,k}^{*2}} [c_{i,j+1/2,k}^{*2} (\rho_{i,j+1,k} - \rho_{i,j,k}) - (P_{i,j+1,k} - P_{i,j,k})];$$

$$\rho_{i,j+1/2,k}^* = \sqrt{\rho_{i,j,k}\rho_{i,j+1,k}}; \quad \mathbf{v}_{i,j+1/2,k}^* = \frac{\sqrt{\rho_{i,j,k}}\mathbf{v}_{i,j,k} + \sqrt{\rho_{i,j+1,k}}\mathbf{v}_{i,j+1,k}}{\sqrt{\rho_{i,j,k}} + \sqrt{\rho_{i,j+1,k}}};$$

$$h_{i,j+1/2,k}^{**} = \frac{\sqrt{\rho_{i,j,k}}h_{i,j,k}^* + \sqrt{\rho_{i,j+1,k}}h_{i,j+1,k}^*}{\sqrt{\rho_{i,j,k}} + \sqrt{\rho_{i,j+1,k}}};$$

$$\mathcal{R}_B = \begin{pmatrix} 1 & 0 & 0 & 2 & 1 \\ u & 2c & 0 & 2u & u \\ v-c & 0 & 0 & 2v & v+c \\ w & 0 & 2c & 2w & w \\ h^*-vc & 2uc & 2wc & \mathbf{v}^2 & h^*+vc \end{pmatrix}, \quad \Lambda_B = \text{diag}(v-c, v, v, v, v+c).$$

Слагаемые, связанные с координатой  $z$ , имеют следующий вид:

$$\mathcal{C} = \frac{\partial \mathbf{H}}{\partial \mathbf{q}},$$

$$\mathbf{H}_{i,j,k+1/2} = \frac{\mathbf{H}_{i,j,k} + \mathbf{H}_{i,j,k+1}}{2} - \frac{1}{2} \sum_m |\lambda_C^m(\mathbf{q}_{i,j,k+1/2}^*)| \Delta s_{i,j,k+1/2}^m \mathbf{r}_C^m(\mathbf{q}_{i,j,k+1/2}^*),$$

где

$$\Delta s_{i,j,k+1/2}^{1,5} = \frac{(P_{i,j,k+1} - P_{i,j,k}) \mp \rho_{i,j,k+1/2}^* c_{i,j,k+1/2}^* (w_{i,j,k+1} - w_{i,j,k})}{2c_{i,j,k+1/2}^{*2}};$$

$$\Delta s_{i,j,k+1/2}^2 = \frac{\rho_{i,j,k+1/2}^* c_{i,j,k+1/2}^* (u_{i,j,k+1} - u_{i,j,k})}{2c_{i,j,k+1/2}^{*2}};$$

$$\Delta s_{i,j,k+1/2}^3 = \frac{\rho_{i,j,k+1/2}^* c_{i,j,k+1/2}^* (v_{i,j,k+1} - v_{i,j,k})}{2c_{i,j,k+1/2}^{*2}};$$

$$\Delta s_{i,j,k+1/2}^4 = \frac{c_{i,j,k+1/2}^{*2} (\rho_{i,j,k+1} - \rho_{i,j,k}) - (P_{i,j,k+1} - P_{i,j,k})}{2c_{i,j,k+1/2}^{*2}};$$

$$\rho_{i,j,k+1/2}^* = \sqrt{\rho_{i,j,k}\rho_{i,j,k+1}}, \quad \mathbf{v}_{i,j,k+1/2}^* = \frac{\sqrt{\rho_{i,j,k}}\mathbf{v}_{i,j,k} + \sqrt{\rho_{i,j,k+1}}\mathbf{v}_{i,j,k+1}}{\sqrt{\rho_{i,j,k}} + \sqrt{\rho_{i,j,k+1}}};$$

$$h_{i,j,k+1/2}^{**} = \frac{\sqrt{\rho_{i,j,k}}h_{i,j,k}^* + \sqrt{\rho_{i,j,k+1}}h_{i,j,k+1}^*}{\sqrt{\rho_{i,j,k}} + \sqrt{\rho_{i,j,k+1}}};$$

$$\mathcal{R}_C = \begin{pmatrix} 1 & 0 & 0 & 2 & 1 \\ u & 2c & 0 & 2u & u \\ v & 0 & 2c & 2v & v \\ w-c & 0 & 0 & 2w & w+c \\ h-wc & 2uc & 2vc & \mathbf{v}^2 & h+wc \end{pmatrix}, \quad \Lambda_C = \text{diag}(w-c, w, w, w, w+c).$$

Заменив  $\lambda^m$  на  $\tilde{\lambda}^m$ , согласно (15.32), получим *схему Roy — Эйнфельдта*, а добавив антидиффузионные члены, согласно (15.36), — *схему Roy — Эйнфельдта — Ошера*.

### 15.2.10. Другие схемы газовой динамики

Рассмотренные в 14 разностные схемы, используемые для решения простейшего уравнения гиперболического типа, свидетельствуют о значительном многообразии схем для решения уравнений газовой динамики, поскольку фактически в соответствии с каждой новой схемой для простейшего уравнения может быть создана новая схема и для системы уравнений гиперболического типа. Основанием для этого является наличие полного набора собственных векторов у матрицы системы и фактическое расщепление задачи на совокупность отдельных подзадач. При интегрировании по разностной ячейке возникает балансовое соотношение, которое является суммой одномерных выражений для каждой из граней. Для аппроксимации этих выражений можно применять соотношения, используемые для решения одномерного *уравнения переноса*. Приведенное описание является только наброском метода построения схем для решения уравнений газовой динамики. При конкретной реализации возможны различные нюансы и трудности.

В данной главе подробно рассмотрены лишь схемы типа *схемы Roy* и *схемы Roy — Эйнфельдта — Ошера*. Однако основные соотношения (матрицы, собственные вектора и собственные значения и т. п.) позволяют сравнительно легко преобразовать эти схемы в другие. Подробные обзоры современных схем для решения систем уравнений гиперболического типа можно найти в работах [98] и [285].

Отметим, что в [37] изложены результаты применения разрывного метода Галеркина (RKDG) при решении уравнений газовой динамики на той же системе тестов. Данный метод является весьма многообещающим, и его использование дает более высокое качество решения. Однако метод RKDG является более сложным по сравнению с разностными схемами.

## 16. ТЕОРЕТИЧЕСКИЕ И АЛГОРИТМИЧЕСКИЕ ОСНОВЫ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

Представлены основы *метода конечных элементов*, описана его связь с *методами взвешенных невязок*, *методом Бубнова — Галеркина*, *вариационно-разностными методами*. На примерах двумерного *уравнения Пуассона* и уравнений линейной теории упругости показана технология применения конечных элементов, включая сборку *матрицы жесткости* и вычисление необходимых интегралов. На примере конкретной задачи кратко представлен *метод граничных элементов*.

### 16.1. Метод конечных элементов и его варианты

Рассмотрим проекционно-сеточные методы, и, в частности, метод конечных элементов (МКЭ). Представленный материал имеет обзорный характер; наша основная цель — попытаться дать читателю представление об основных теоретических и алгоритмических идеях, лежащих в основе МКЭ.

В настоящее время МКЭ является одним из наиболее широко применяемых методов для решения задач, встречающихся в научной и инженерной практике. Примерами таких задач являются линейные и нелинейные задачи механики деформируемого твердого тела, теплопроводности, гидродинамики, электродинамики и многих других.

Популярность МКЭ обусловлена разными факторами, основными среди которых являются его универсальность и технологичность, т. е. возможность достаточно простого способа построения *аппроксимаций* метода для широкого класса задач в многомерных областях сложной формы и структуры, возможность повышения точности без принципиального качественного усложнения вычислительного алгоритма в целом.

Литература по МКЭ и его применению для решения тех или иных задач очень обширна. Для ознакомления с общей теорией МКЭ можно рекомендовать [110, 112, 115, 116, 123, 133, 166, 207]. Теоретические вопросы, связанные со *слабыми и вариационными постановками задач*, рассмотрены, например, в [67, 93, 104]. К работам прикладной направленности можно отнести [45, 60, 72, 73, 114, 122, 125, 157, 158, 165, 296–298]. Использование МКЭ для решения задач теории упругости рассмотрено

в [71, 296–298 и др.]. Исследованию и численному решению вариационных неравенств, описывающих, например, контактное взаимодействие деформируемых упругих тел, течения пластических сред, задачи со свободной поверхностью в гидродинамике, посвящены книги [48, 85, 241, 247, 249, 293]. Задачи гидродинамики рассмотрены в [167, 298]. Вопросы совместного использования МКЭ и *многосеточных методов* рассмотрены в [189, 207].

Отметим, что одним из наиболее полных классических трудов по приложениям МКЭ является трехтомник [296–298].

Математической основой МКЭ является теория проекционных (в частности, проекционно-сеточных) методов. Общие вопросы обоснования этих методов для решения абстрактных операторных уравнений рассмотрены в [93].

С теоретической точки зрения МКЭ входит в класс *проекционно-сеточных* или *вариационно-сеточных* методов. От других методов этого класса его отличает прежде всего использование базисных функций с *финитным носителем*, обычно заданных на треугольных или четырехугольных *сетках* — в пространственно-двумерном случае, либо на сетках из тетраэдров или восьмигранников — в пространственно-трехмерном случае.

Рассматриваемые далее варианты МКЭ основаны на методе Бубнова — Галеркина или на методе Ритца. Существуют и другие варианты, которые здесь не рассматриваются.

Цель данной главы — изложение главных идей, лежащих в основе теории и практики применения проекционно-сеточных методов, с минимальным углублением в детали теории.

Теория проекционно-сеточных методов изложена в основном в соответствии с [112]. Алгоритмическим аспектам МКЭ также посвящены многочисленные статьи и книги, например [193, 194, 296–298], которым мы и следуем.

## 16.2. Метод взвешенных невязок

Одной из наиболее общих форм *проекционных методов* решения краевых задач, в которые входят методы Бубнова — Галеркина и, в частности, метод конечных элементов (МКЭ), является *метод взвешенных невязок*.

Предположим, что необходимо найти решение уравнения

$$Au = f, \quad (16.1)$$

рассматриваемого в области  $\Omega$  с граничными условиями

$$lu = g, \quad (16.2)$$

заданными на границе  $\partial\Omega$  области  $\Omega$ . Здесь  $A$  — некоторый линейный дифференциальный оператор, а оператор  $l$  задает граничные условия того или иного типа на границе  $\partial\Omega$ . Разыскиваемое решение  $u$  должно входить в область определения оператора  $A$ , принадлежащую некоторому пространству  $U$ , и удовлетворять поставленным граничным условиям.

Будем также считать, что правая часть  $f$  уравнения (16.1) принадлежит некоторому гильбертову пространству  $H$ , включающему область значений оператора  $A$ . Обычно  $H = L_2(\Omega)$ . Пространство  $U$  также будем считать гильбертовым.

Запишем задачу (16.1)–(16.2) в несколько ином виде. Требуется определить функцию  $u$ , такую, что

$$u \in D(A) \subset U : \quad Au = f, \quad (16.3)$$

где  $D(A)$  — область определения оператора  $A$ ,

$$D(A) = \{u : Au \in H, (lu - g)|_{\partial\Omega} = 0\}.$$

Таким образом, область определения  $D(A)$  оператора задачи, в которой разыскивается решение, состоит из функций, имеющих требуемую гладкость и удовлетворяющих граничным условиям (16.2).

В дальнейшем всегда будем считать, что решение этой задачи существует, единственно и непрерывно зависит от коэффициентов, правой части и граничных условий, т. е. задача поставлена корректно.

Если функция  $u$  является решением задачи (16.3), то она, очевидно, удовлетворяет и следующему уравнению:

$$u \in D(A) : \quad (Au - f, v) = 0, \quad (16.4)$$

справедливому для всех функций  $v \in H$ , называемых часто *пробными*. Здесь  $(f, g)$  представляет собой скалярное произведение в пространстве  $H$ . Если  $H = L_2(\Omega)$ , то  $(f, g)$  определяется как

$$(f, g) = \int_{\Omega} fg d\Omega.$$

Таким образом, выражение (16.4) представляет собой условие ортогональности невязки

$$Ru = Au - f$$

произвольному элементу пространства пробных функций. Если при подходящем выборе пространства функций это утверждение обратимо, т. е. из равенства (16.4), верного при любом  $v \in H$ , следует равенство нулю первого сомножителя скалярного произведения, то решение  $u$ , определяемое из (16.4), будет являться решением задачи (16.1).

В соответствии с методом взвешенных невязок приближенное решение задачи (16.1) разыскивается в виде

$$u_h = \varphi_0 + \sum_{i=1}^N u_i \varphi_i, \quad (16.5)$$

где  $\varphi_0$  — некоторая заданная функция (учитывающая, например, неоднородные граничные условия задачи и/или особенности ее решения);  $u_i$  — некоторые коэффициенты, подлежащие определению;  $\varphi_i$  — некоторые наперед заданные **базисные функции**, образующие полную систему в  $U$ . При этом необходимо, чтобы приближенное решение удовлетворяло граничным условиям и имело требуемую гладкость, т. е.

$$u_h \in D(A).$$

При описании проекционно-сеточных методов мы используем обозначения, традиционные для литературы по данным методам. Поэтому здесь и далее приближенное решение обозначим  $u_h$ .

Неизвестные коэффициенты  $u_i$  определяются из условий, которые являются следствием уравнения (16.4):

$$(Au_h - f, \psi_j) = 0, \quad j = 1, 2, \dots, N, \quad (16.6)$$

где  $\psi_j \in H$  — набор некоторых пробных функций из пространства  $H$ , образующих в  $H$  полную систему.

Уравнение (16.6) представляет собой систему уравнений для определения коэффициентов  $u_i$ . Если оператор  $A$  является линейным, то (16.6) можно записать в виде системы линейных алгебраических уравнений (СЛАУ) вида

$$\sum_{i=1}^N u_i (A\varphi_i, \psi_j) = (f - A\varphi_0, \psi_j), \quad j = 1, 2, \dots, N, \quad (16.7)$$

относительно коэффициентов  $u_i$ .

Тот или иной выбор базисных и пробных функций приводит к тем или иным вариантам метода взвешенных невязок. Рассмотрим примеры этих вариантов.

**Пример 16.1. 1.** Выбор пробных функций в виде  $\psi_j = \delta(x - x_j)$ , где точки  $x_j$  попарно не совпадают и принадлежат области  $\Omega$ , в которой разыскивается решение задачи, приводит к так называемому **методу коллокаций**. В этом случае неизвестные коэффициенты определяются из условия равенства нулю **невязки**  $Ru$  решения на некотором заданном наборе точек.

Выбор координат точек  $x_j$  является параметром метода и, в частности, влияет на точность решения при неизменном числе точек.

Очевидно, что выбор таких пробных функций требует существенной коррекции теории. Мы этого делать не будем, ограничившись лишь данным описанием.

2. **Метод наименьших квадратов** соответствует случаю  $\psi_j = A\varphi_j$ . Этот метод подробно рассмотрен в 16.4.2.

3. **Метод Бубнова — Галеркина** соответствует случаю, когда наборы базисных и пробных функций совпадают, т. е.  $\psi_j = \varphi_j$ . Этот метод подробно рассмотрен в 16.3.

4. **Обобщенный метод Галеркина** соответствует случаю  $\psi_j = B\varphi_j$ , где оператор  $B$  выбирается из тех или иных дополнительных условий. Например, такие варианты метода взвешенных невязок используют для построения схем МКЭ для решения задач конвекции-диффузии. В этом случае помимо точности в первую очередь рассматриваются вопросы, связанные с устойчивостью метода.

5. Вариант метода Бубнова — Галеркина, в котором базисные и пробные функции являются собственными функциями той или иной вспомогательной задачи, называется **спектральным методом**. Такой метод особенно эффективен при исследовании вопросов устойчивости различных процессов, например, в задачах гидродинамики. #

Существуют и другие варианты метода взвешенных невязок.

Таким образом, метод взвешенных невязок включает широкий набор методов, пригодных для решения самых разнообразных задач. При этом возникает много вопросов, связанных с выбором базисных и пробных функций, пространств, которым они принадлежат, со способом учета граничных условий того или иного рода. Некоторые из этих вопросов будут рассмотрены ниже.

### 16.3. Метод Бубнова — Галеркина

Как уже отмечалось, метод Бубнова — Галеркина, включающий в себя и метод конечных элементов (МКЭ), соответствует выбору совпадающих *базисных и пробных функций*.

При этом обычно метод Бубнова — Галеркина применяется в несколько ином виде, чем это рассматривалось ранее: для построения его аппроксимаций обычно используют не уравнение (16.4), а так называемую слабую постановку задачи, *решение* которой называют **слабым** или **обобщенным**. В данном случае задача записывается в некотором виде, который накладывает минимальные естественные ограничения на гладкость искомого решения. Это особенно актуально в приложениях, содержащих задачи с разрывными коэффициентами или какими-либо другими неоднородностями.

С точки зрения теории наиболее простыми являются задачи с симметричным и положительно определенным оператором, часто встречающиеся в приложениях.

### 16.3.1. Обобщенные решения и слабая постановка задачи

Напомним, что оператор  $A^*$  называется *сопряженным* оператору  $A$ , если для любых функций  $u$  и  $v$  из области  $D(A)$  определения оператора  $A$  справедливо  $(Au, v) = (u, A^*v)$ . Если при этом  $A = A^*$ , то оператор  $A$  называют *самосопряженным*.

Напомним, что оператор называется *положительно определенным*, если существует такое  $\alpha > 0$ , что  $(Au, u) \geq \alpha \|u\|^2$ ,  $u \in D(A)$ , где  $\|\cdot\|$  — норма в гильбертовом пространстве  $U$ ,  $D(A) \subset U$ .

Для таких операторов оказывается, что при выполнении условия *самосопряженности* выражение

$$[u, v] = (Au, v) \quad (16.8)$$

является само по себе скалярным произведением, которое называют *энергетическим скалярным произведением*. Соответствующая *энергетическая норма* имеет вид  $\|u\| = \sqrt{[u, u]}$ . *Пространство* функций с такой нормой будем называть *энергетическим*.

Переход от классической постановки задачи в виде операторного уравнения (16.4) к слабой постановке обычно осуществляется с помощью той или иной *формулы Грина*. Как правило, она имеет вид

$$(Au, v) = a(u, v) + \langle \delta u, \gamma v \rangle, \quad (16.9)$$

где  $a(\cdot, \cdot)$  является некоторой билинейной симметричной положительно определенной формой, соответствующей оператору  $A$ , а  $\langle \mu, \nu \rangle$  определяется как

$$\langle \mu, \nu \rangle = \int_{\partial\Omega} \mu \nu dS$$

и представляет собой скалярное произведение в пространстве функций, заданных на границе  $\partial\Omega$  области  $\Omega$ . Операторы  $\delta$  и  $\gamma$  также определены на функциях, заданных на границе области, и определяются оператором  $A$ . При этом обычно (но не всегда) оператор  $\gamma$  взятия следа функции задает просто ограничение функции, определенной в области  $\Omega$ , на ее границу  $\partial\Omega$ , а оператор  $\delta$  является некоторым дифференциальным оператором более низкого порядка, чем порядок оператора  $A$ , и действует на границе рассматриваемой области.

Приведем некоторые примеры, поясняющие формулу Грина (16.9). Для построения формулы Грина необходим только оператор  $A$  задачи, а не ее полная постановка, поэтому граничные условия задавать не будем.

**Пример 16.2.** Рассмотрим уравнение Лапласа в двумерной области  $\Omega$ :

$$-\Delta u = 0.$$

Оператор этой задачи (*оператор Лапласа*) в декартовой ортогональной системе координат  $Oxy$  имеет вид

$$-\Delta = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}.$$

Умножим выражение  $-\Delta u$  на произвольную достаточно гладкую функцию  $v$ , заданную в области  $\Omega$ , и проинтегрируем получившееся выражение по всей области  $\Omega$ . В результате

$$\int_{\Omega} -\Delta u v d\Omega = \int_{\Omega} \nabla u \nabla v d\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial \vec{n}} v dS.$$

Таким образом, в данном случае в формуле Грина (16.9)

$$\begin{aligned} Au &= -\Delta u, \quad a(u, v) = \int_{\Omega} \nabla u \nabla v d\Omega, \\ \delta u &= -\nabla u \vec{n} \Big|_{\partial\Omega}, \quad \gamma u = u \Big|_{\partial\Omega}, \end{aligned}$$

где  $\vec{n}$  — единичная внешняя нормаль к границе  $\partial\Omega$  области  $\Omega$ .

Отсюда также следует, что в пространстве функций, имеющих вторые непрерывные производные и обращающихся в нуль на границе области, оператор Лапласа является симметричным, т. е.

$$(-\Delta u, v) = (u, -\Delta v).$$

**Пример 16.3.** Рассмотрим теперь задачу линейной теории упругости. Система соотношений, описывающих напряженно-деформированное состояние упругого тела, включает в себя уравнения равновесия

$$\sigma_{ij,j} + f_i = 0,$$

материальные соотношения, выражающие обобщенный *закон Гука*,

$$\sigma_{ij} = 2\mu\varepsilon_{ij} + \lambda\varepsilon_{kk}\delta_{ij}$$

и кинематические соотношения

$$\varepsilon_{ij}(u) = \frac{u_{i,j} + u_{j,i}}{2}.$$

Уравнение, выражающее условие равновесия тела, также удобно записывать в безындексной форме:

$$\operatorname{div} \sigma + f = 0.$$

Здесь  $\sigma$  — тензор напряжений;  $\varepsilon$  — тензор деформации;  $u$  — вектор перемещения;  $\sigma_{ij}$ ,  $\varepsilon_{ij}$  и  $u_i$  — их компоненты в выбранной системе координат;  $f$  — вектор объемной плотности внешних сил;  $\lambda = \lambda(x)$ ,  $\mu = \mu(x)$  — коэффициенты Ламе, в общем случае являющиеся заданными функциями точки пространства;  $\delta_{ij}$  — компоненты единичного тензора;  $\sigma_{ij,j} = \frac{\partial \sigma_{ij}}{\partial x_j}$ ,  $u_{i,j} = \frac{\partial u_i}{\partial x_j}$ . Мы используем записи  $\varepsilon_{ij}(u)$  и  $\varepsilon(u)$  для того, чтобы указать тот вектор перемещения, по которому в соответствии с кинематическими соотношениями вычисляется тензор деформации. Аналогичным является смысл обозначений  $\sigma_{ij}(u)$  и  $\sigma(u)$ .

Действуя так же, как и в примере 16.2, т. е. умножив левую часть уравнения равновесия на произвольную достаточно гладкую функцию и проинтегрировав получившееся выражение по частям, получим формулу Грина, для которой операторы  $A$ ,  $\delta$ ,  $\gamma$  и билинейная форма  $a(\cdot, \cdot)$  имеют следующий вид:

$$Au = -\operatorname{div} \sigma(u), \quad a(u, v) = \int_{\Omega} \sigma(u) \varepsilon(v) d\Omega,$$

$$\delta u = -\sigma(u) \cdot \vec{n}|_{\partial\Omega}, \quad \gamma u = u|_{\partial\Omega}.$$

Полученные выражения верны для произвольной зависимости  $\sigma$  от  $\varepsilon$  и приведенной выше зависимости  $\varepsilon$  от  $u$ .

Оператор рассматриваемой задачи также является самосопряженным и положительно определенным при подходящем задании области определения оператора задачи, например при нулевых перемещениях на границе. #

Можно привести другие примеры формулы Грина для тех или иных операторов.

Покажем теперь, как именно строится слабая постановка задачи.

Рассмотрим, как и ранее, уравнение (16.1). Умножим его на пока произвольную достаточно гладкую функцию  $v$  и проинтегрируем получившееся выражение по области  $\Omega$  с использованием формулы Грина, соответствующей оператору задачи.

В результате получим

$$a(u, v) + \langle \delta u, \gamma v \rangle = (f, v). \quad (16.10)$$

Дальнейшее преобразование этого выражения связано с учетом граничных условий того или иного рода. Будем считать, что граница области  $\Gamma = \partial\Omega$  разбита на две непересекающиеся части  $\Gamma_D$  и  $\Gamma_N$ :

$$\partial\Omega = \Gamma_D \cup \Gamma_N,$$

причем на  $\Gamma_D$  заданы граничные условия первого рода

$$u|_{\Gamma_D} = g, \quad (16.11)$$

а на оставшейся части  $\Gamma_N$  граничные условия заданы граничные условия

$$\delta u|_{\Gamma_N} = \mu, \quad (16.12)$$

где  $g$  и  $\mu$  — некоторые известные функции, определяющие граничные условия.

Разбив интегрирование по границе области в (16.10) на две части:

$$\langle \delta u, \gamma v \rangle = \int_{\partial\Omega} \delta u \gamma v dS = \int_{\Gamma_D \cup \Gamma_N} \delta u \gamma v dS = \langle \delta u, \gamma v \rangle_{\Gamma_D} + \langle \delta u, \gamma v \rangle_{\Gamma_N},$$

где

$$\langle \varphi, \psi \rangle_{\Gamma_D} = \int_{\Gamma_D} \varphi \psi dS, \quad \langle \varphi, \psi \rangle_{\Gamma_N} = \int_{\Gamma_N} \varphi \psi dS,$$

с учетом граничного условия (16.12) получим

$$a(u, v) + \langle \delta u, \gamma v \rangle_{\Gamma_D} + \langle \mu, \gamma v \rangle_{\Gamma_N} = (f, v).$$

При этом можно показать, что *граничные условия* (16.11) и (16.12) являются соответственно *главными* и *естественными*. Формальная разница между ними заключается в том, что естественные граничные условия исключаются из рассмотрения в слабой постановке задачи, так как вводятся непосредственно в уравнение указанным выше способом.

Границное условие первого рода при этом никуда не «исчезает» и должно учитываться в дальнейшем. Можно показать, что полная корректная *обобщенная* (или *слабая*) *постановка задачи* имеет

вид: определить функцию  $u$ , такую, что

$$a(u, v) = (f, v) - \langle \mu, \gamma v \rangle_{\Gamma_N}, \quad u|_{\Gamma_D} = g, \quad (16.13)$$

для произвольной достаточно гладкой функции  $v$ , равной нулю на части  $\Gamma_D$  расчетной области.

В качестве частного случая рассмотрим ситуацию, когда на всей границе области задано граничное условие первого рода. При этом формально  $\Gamma_N = \emptyset$ ,  $\Gamma = \Gamma_D$  и слабая постановка задачи имеет вид: определить функцию  $u$ , такую, что

$$a(u, v) = (f, v), \quad u|_{\Gamma_D} = g, \quad (16.14)$$

для произвольной достаточно гладкой функции  $v$ , обращающейся в нуль на всей границе  $\Gamma$  расчетной области.

Отметим, что задачи (16.13) или (16.14) можно записать в виде: определить функцию  $u$ , такую, что

$$u \in V : \quad a(u, v) = F(v), \quad v \in V_0, \quad (16.15)$$

где  $V$  и  $V_0$  являются некоторыми подходящими пространствами. При этом в определении пространства  $V$  учитываются граничные условия первого рода, накладываемые на решение (например,  $u|_{\Gamma_D} = g$ ), а в определении пространства  $V_0$  учитываются граничные условия первого рода, накладываемые на пробные функции (например,  $v|_{\Gamma_D} = 0$ ).

Задачи вида (16.15) называют *вариационными уравнениями*. Для их исследования построена хорошо разработанная и подробная теория, позволяющая исследовать как сами задачи, так и их аппроксимации.

При этом важно отметить, что существенным является не конкретный вид билинейной формы  $a(\cdot, \cdot)$  и правой части  $F(\cdot)$ , а такие их свойства, как симметричность, положительная определенность и непрерывность.

Далее в этой главе будем говорить именно о задаче (16.15).

Рассмотрим подробнее понятие главных и естественных граничных условий, необходимое с точки зрения применения и исследования как слабых постановок краевых задач, так и построения расчетных схем методом Бубнова — Галеркина.

Для этого вернемся к задаче (16.3). Напомним, что выбор области определения  $D(A)$  оператора задачи связан с двумя основными критериями:

- 1) необходимой гладкостью решения (т. е. наличием у произвольной функции из этого пространства необходимого количества производных);

2) учетом граничных условий, которым должно удовлетворять решение рассматриваемой задачи.

При этом, как уже отмечалось, соотношение (16.8) представляет собой энергетическое скалярное произведение в пространстве  $D(A)$ .

Функция  $u$ , удовлетворяющая соотношению (16.13), может обладать меньшей гладкостью, чем функции из пространства  $D(A)$ , так как в выражение для билинейной формы  $a(\cdot, \cdot)$  формально входят производные меньших порядков (см. пример 16.2).

Корректное определение того пространства, которому принадлежит решение слабой постановки задачи, выглядит следующим образом. Это решение разыскивается в пространстве  $V = H_A$ , которое является пополнением пространства  $D(A)$  по норме, порожденной энергетическим скалярным произведением (16.8). В результате такого пополнения функции из пространства  $V$  обладают, вообще говоря, меньшей по сравнению с функциями из пространства  $D(A)$  гладкостью.

В определение пространства  $D(A)$  помимо гладкости входят также граничные условия задачи. При этом в результате пополнения в пространстве  $V$  могут появиться функции, не удовлетворяющие граничным условиям, участвующим в определении пространства  $D(A)$ .

Если функции из пространства  $D(A)$  удовлетворяли некоторому граничному условию  $lu = 0$  и в результате пополнения функции из пространства  $V$  также удовлетворяют этому же граничному условию, то такое *граничное условие* называется *главным*.

Если же функции из пространства  $D(A)$  удовлетворяли некоторому граничному условию  $lu = 0$ , а функции из пространства  $V$  ему уже не удовлетворяют, то такое *граничное условие* называют *естественным*.

Практическая разница между главными и естественными граничными условиями заключается в том, что вариационное уравнение для определения слабого решения необходимо дополнять лишь главными граничными условиями. При этом решение такой задачи будет автоматически удовлетворять естественным граничным условиям (которые входят непосредственно в уравнение).

Приведем простой признак определения того, каким является граничное условие — главным или естественным. Пусть оператор  $A$  задачи является дифференциальным оператором порядка  $2m$ . Тогда те граничные условия, в которые входят производные решения порядка  $m$  и выше, будут естественными граничными условиями, а все остальные — главными [112].

### 16.3.2. Аппроксимация методом Бубнова — Галеркина

Будем считать, что слабая постановка рассматриваемой задачи уже сформулирована и имеет вид (16.15).

В соответствии с методом Бубнова — Галеркина ее решение определяется следующим образом.

Необходимо выбрать систему базисных (и пробных) функций  $\varphi_i$ , принадлежащих энергетическому пространству  $V$ . Это означает, что указанные базисные функции должны удовлетворять главным граничным условиям задачи, но могут не удовлетворять естественным граничным условиям.

Для простоты будем рассматривать случай, когда на решение наложены только главные граничные условия: пространства  $V$  и  $V_0$  совпадают, т. е.  $V = V_0$ . В качестве примера можно взять задачу Дирихле для уравнения Пуассона с нулевыми граничными условиями. В этом случае  $V = V_0$  состоит из достаточно гладких функций, обращающихся в нуль на границе области.

Всевозможные линейные комбинации конечного числа базисных функций образуют некоторое конечномерное подпространство  $V_h$  в пространстве  $V$ .

Приближенное решение задачи разыскивается в виде

$$u_h = \sum_{i=1}^N u_i \varphi_i.$$

Его коэффициенты определяются из условий

$$a(u_h, v_h) = F(v_h), \quad v_h \in V_h.$$

Чтобы это соотношение выполнялось, достаточно потребовать его выполнения для каждой из базисных функций  $\varphi_i$ , т. е.

$$a(u_h, \varphi_i) = F(\varphi_i), \quad i = 1, 2, \dots, N, \tag{16.16}$$

или в развернутом виде

$$\sum_{j=1}^N u_j a(\varphi_j, \varphi_i) = F(\varphi_i), \quad i = 1, 2, \dots, N.$$

Последнее соотношение представляет собой систему линейных алгебраических уравнений (СЛАУ) для определения коэффициентов  $u_i$ :

$$A_h u = b,$$

где матрица  $A_h$  имеет размеры  $N \times N$ , и ее элементы равны

$$(A_h)_{ij} = a(\varphi_j, \varphi_i);$$

вектор  $b$  имеет размерность  $N$ , и его элементы равны  $b_i = F(\varphi_i)$ . Отметим, что матрица  $A_h$  наследует свойства билинейной формы  $a(\cdot, \cdot)$ , т. е. является симметричной и положительно определенной, если исходная билинейная форма удовлетворяет этим же условиям.

### 16.3.3. Сходимость метода Бубнова — Галеркина

Отметим основные утверждения, лежащие в обосновании сходимости методов рассматриваемого класса.

Основными теоретическими результатами, лежащими в основе рассмотренных выше методов, являются **лемма Лакса — Мильграма** и **лемма Сеа**.

Первая из них утверждает, что если билинейная форма  $a(\cdot, \cdot)$ , заданная в пространстве  $V \times V$ , является непрерывной и положительно определенной, т. е. для произвольных  $u, v \in V$  справедливы неравенства

$$|a(u, v)| \leq \alpha \|u\|_V \|v\|_V \quad \text{и} \quad |a(u, u)| \geq \beta \|u\|_V^2,$$

то задача, состоящая в определении  $u \in V$ , такого, что

$$a(u, v) = F(v), \quad v \in V,$$

где  $F \in V'$ , имеет единственное решение, непрерывно зависящее от  $F$ , т. е.

$$\|u\|_V \leq \frac{1}{\beta} \|F\|_{V'}$$

Здесь  $V'$  — пространство, двойственное к  $V$ , условие  $F \in V'$  означает, что  $F$  — непрерывный линейный функционал на  $V$ . Пространство  $V$  является гильбертовым.

Отметим, что в данном случае не требуется симметричности билинейной формы  $a(\cdot, \cdot)$ .

Это утверждение гарантирует корректную постановку задачи как для дифференциального, так и для конечномерного случая, потому что, как можно заметить, задачи для обобщенного (16.15) и приближенного (16.16) решений исходного уравнения отличаются лишь тем, что в первом случае решение разыскивается в функциональном пространстве  $V$ , а во втором — в некотором конечномерном пространстве  $V_h$ .

Таким образом, решение как исходного вариационного уравнения, так и конечномерной задачи, существует и единствено.

В свою очередь, лемма Сеа является основным результатом, гарантирующим сходимость *метода Бубнова — Галеркина*. Она утверждает, что в условиях леммы Лакса — Мильграма в случае  $V_h \subset V$  ошибка

приближенного решения удовлетворяет неравенству

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Здесь  $C$  — постоянная, не зависящая от  $V_h$ . Следовательно, достаточное условие сходимости состоит в существовании такого семейства подпространств  $\{V_h\}$  пространства  $V$ , что для всякого  $u \in V$

$$\lim_{h \rightarrow 0} \inf_{v_h \in V_h} \|u - v_h\|_V = 0.$$

Величина в правой части неравенства леммы Сеа легко оценивается сверху ошибкой аппроксимации, в частности, *интерполяции* решения по выбранной системе базисных функций:

$$\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - \tilde{u}_h\|_V.$$

Правая часть этого неравенства не зависит от оператора задачи и определяется только способом выбора конечномерного пространства, в котором разыскивается решение.

Таким образом, задача оценки ошибки приближенного решения сводится к задаче построения интерполяции произвольной функции. Эта задача гораздо проще исходной. Для ее решения могут использоваться методы теории интерполяции функций (см. 3).

Оба результата можно найти практически в любой книге по теории МКЭ, например [166, 207], см. также [78, 93, 113].

Построенная процедура получения ошибки является простейшей. Но и оценка, которая получается, дается в слабой, энергетической, норме. Тем не менее можно получить оценки и в более сильных нормах, например, в равномерной. Они требуют гораздо более тонкого анализа и здесь не рассматриваются.

#### 16.4. Вариационно-сеточные методы

*Вариационно-сеточные методы* являются другой разновидностью методов, которые в итоге приводят к аппроксимациям задачи методом конечных элементов (МКЭ).

Так же как и в случае *проекционно-сеточных методов*, решение разыскивается в виде линейной комбинации тех или иных *базисных функций*. Существенное их отличие от проекционно-сеточных методов заключается в том, что источником для построения аппроксимаций

является не слабая постановка задачи в виде того или иного вариационного уравнения, а формулировка исходной задачи в виде задачи минимизации некоторого функционала. При решении прикладных задач такой функционал часто имеет явное физическое значение, например функционал энергии, который достигает своего минимума на решении задачи. В этом смысле постановки задач в виде задач минимизации некоторого функционала являются более естественными, так как фактически соответствуют реализации тех или иных вариационных принципов механики сплошной среды.

Перейдем к описанию двух распространенных вариационно-сеточных методов.

#### 16.4.1. Метод Ритца

**Метод Ритца** является одним из распространенных способов построения расчетных схем. Он применяется для решения задач с симметричным положительно определенным оператором. В этом случае метод Ритца приводит к той же конечномерной задаче (системе алгебраических уравнений для коэффициентов линейной комбинации, в виде которой разыскивается решение), что и рассмотренный ранее *метод Бубнова — Галеркина*.

Рассмотрим, как и ранее, задачу (16.3), считая *оператор  $A$  симметричным и положительно определенным* на своей области определения  $D(A)$ . Напомним, что область определения  $D(A)$  в такой постановке задает не только требуемую гладкость решения, но и накладываемые на него граничные условия. Для простоты будем считать их однородными.

Можно показать [112, 166], что в этом случае задача (16.3) эквивалентна следующей вариационной задаче: найти функцию  $u \in D(u)$ , которая обеспечивает минимум функционала

$$\mathcal{F}(v) = \frac{1}{2}(Av, v) - (f, v), \quad (16.17)$$

т. е.

$$\mathcal{F}(u) = \min_{v \in D(A)} \mathcal{F}(v),$$

или, в других обозначениях,

$$u = \operatorname{argmin}_{v \in D(A)} \mathcal{F}(v).$$

Если оператор исходной задачи является симметричным и положительно определенным на  $D(A)$ , то указанная задача имеет единственное решение, совпадающее с решением задачи (16.3).

Для определения приближенного решения снова будем искать его в виде (16.5), где базисные функции  $\varphi_i \in D(A)$ .

Приближенное решение при этом определяется из условия минимума функционала

$$\mathcal{F}(u_h) = \min \mathcal{F}(v_h), \quad (16.18)$$

где минимум разыскивается среди всевозможных линейных комбинаций вида (16.5).

Поскольку при сделанных ранее предположениях рассматриваемый функционал имеет единственный минимум, то неизвестные значения  $u_i$  определяются из необходимых условий минимума функционала (16.18)

$$\frac{\partial \mathcal{F}(u_h)}{\partial u_i} = 0, \quad i = 1, 2, \dots, N,$$

что приводит, в свою очередь, к системе линейных алгебраических уравнений (СЛАУ) вида

$$\sum_{i=1}^N u_i (A\varphi_i, \varphi_j) = (f, \varphi_j), \quad j = 1, 2, \dots, N,$$

т. е. к тем же самим уравнениям, что и обычный метод Бубнова — Галеркина (уравнение (16.7) при  $\varphi_0 = 0$ ,  $\psi_i = \varphi_i$ ).

Как и для метода Бубнова — Галеркина, постановку задачи (16.17) можно «ослабить» тем же способом, что и ранее, т. е. используя формулу Грина и новое пространство  $V = H_A$ , которое строится как пополнение пространства  $D(A)$  по энергетической норме, соответствующей оператору задачи.

В этом случае постановка задачи будет следующей: определить функцию  $u \in V$ , которая обеспечивает минимум функционала

$$\mathcal{F}(v) = \frac{1}{2}a(v, v) - F(v),$$

т. е.

$$\mathcal{F}(u) = \min_{v \in V} \mathcal{F}(v).$$

Здесь билинейная форма  $a(\cdot, \cdot)$  и линейный функционал  $F(\cdot)$  являются такими же, как и в уравнении (16.15). Главные и естественные граничные условия определяются так же, как и ранее.

Построенная задача эквивалентна слабой постановке (16.15).

В результате аппроксимации этой задачи возникает СЛАУ, совпадающая с (16.16).

Несмотря на то что в рассмотренном случае методы Ритца и Бубнова — Галеркина приводят к одинаковым конечномерным задачам, их разделение имеет смысл. Это связано с дальнейшим обобщением описываемых методов и слабых постановок на задачи с более сложными, например несимметричными и не положительно определенными операторами, а также с более сложными граничными условиями.

Так, использование метода Бубнова — Галеркина и слабых постановок задач позволяет рассматривать аппроксимации для задач с несимметричным и не положительно определенным оператором, что недопустимо в методе Ритца, так как при этом нарушается условие единственности минимума соответствующего функционала. При этом общая схема метода и способ получения слабой постановки и ее аппроксимаций практически не меняются. Использованные выше свойства оператора задачи облегчают теоретическое исследование метода, но с формальной точки зрения для построения ее слабой постановки и дальнейшей аппроксимации необходима лишь формула Грина.

В то же время формулировка исходной задачи в виде задачи поиска минимума функционала позволяет рассматривать задачи с более сложными граничными условиями, например с граничными условиями, заданными внутри расчетной области, а не на ее границе. При этом часть граничных условий может рассматриваться как некие ограничения, которые снимаются не путем выбора подходящих пространств базисных и пробных функций, а, например, с помощью метода множителей Лагранжа. То же относится и к некоторым уравнениям, которые входят в формулировку полной задачи. Типичным примером подобных постановок являются задачи типа задачи Стокса, в которые помимо основного уравнения входят условия соленоидальности решения (см., например, [167]). Такого рода условия также могут трактоваться как дополнительные ограничения, которые снимаются методом множителей Лагранжа.

### 16.4.2. Метод наименьших квадратов

Рассмотрим задачу (16.3), но не будем накладывать теперь никаких ограничений на оператор  $A$  помимо его невырожденности. При этом будем считать, что правая часть  $f$  рассматриваемой задачи принадлежит пространству  $D(A^*)$ , где  $A^*$  — оператор, сопряженный оператору  $A$ .

Подействуем этим оператором на уравнение (16.3), в результате получим следующую задачу: определить  $u \in D(A)$ , такое, что

$$A^*Au = A^*f. \quad (16.19)$$

При этом оператор  $A^*A$  последней задачи уже является симметричным вне зависимости от того, был ли симметричным оператор  $A$ .

**Метод наименьших квадратов** представляет собой метод Ритца для задачи (16.19).

Минимизируемый в данном случае функционал имеет вид

$$\mathcal{F}(u) = \frac{1}{2} (A^* Au, u) - (A^* f, u). \quad (16.20)$$

Используя метод Ритца как в предыдущем случае, получаем систему уравнений вида (16.7) с пробными функциями  $\psi_i = A\varphi_i$ .

Минимизируемый функционал (16.20) можно записать в виде

$$\mathcal{F}(u) = \frac{1}{2} \|Au - f\|^2 - \frac{1}{2} \|f\|^2.$$

Поэтому задача его минимизации эквивалентна задаче минимизации функционала

$$\mathcal{J}(u) = \|Au - f\|^2,$$

откуда и появилось название метода.

Метод наименьших квадратов имеет как достоинства, так и недостатки по сравнению с методом Ритца. К достоинствам метода относится прежде всего то, что в отличие от метода Ритца на оператор  $A$  не накладываются никакие дополнительные требования типа симметричности и положительной определенности. К недостаткам можно отнести то, что базисные функции МНК должны принадлежать пространству  $D(A)$ , т. е. иметь требуемую («высокую») гладкость, а также удовлетворять как главным, так и естественным граничным условиям исходной задачи (16.3) (точнее говоря, главным условиям задачи (16.19), порядок которой в два раза выше, чем порядок исходного уравнения).

## 16.5. Метод конечных элементов

**Метод конечных элементов** (МКЭ) представляет собой один из рассмотренных ранее методов с использованием *финитных базисных функций*. Использование таких функций получило распространение в основном по следующим причинам:

возможность решения задач в областях сложной формы, для которых трудно построить нефинитные базисные функции;

относительная легкость построения *аппроксимаций* высокого порядка;

*разреженность матрицы* получаемой системы линейных алгебраических уравнений (СЛАУ), что особенно актуально при решении задач большой размерности.

Разреженность матрицы связана с финитностью базисных функций и тем фактом, что дифференциальные операторы являются локальными, т. е. их значение в какой-либо точке зависит только от значений их аргументов в бесконечно малой окрестности этой точки. При нарушении хотя бы одного из этих свойств получаемая система уравнений не будет иметь, вообще говоря, разреженную матрицу, например при использовании финитных базисных функций для аппроксимации задачи с нелокальным, в частности интегральным, оператором, либо нефинитных базисных функций для аппроксимации задачи с локальным оператором. Такая ситуация возникает, например, при рассмотрении *спектральных методов*.

Ранее кратко были рассмотрены некоторые вопросы, связанные с основными идеями проекционно-сеточных методов. В этом смысле, как уже отмечалось, математическая теория МКЭ является следствием общей теории метода Бубнова — Галеркина или метода Ритца.

Поэтому далее на примере ряда модельных задач более детально будут рассмотрены некоторые вопросы, связанные с использованием и реализацией метода конечных элементов.

### 16.5.1. Двумерное уравнение Пуассона

В качестве первого примера рассмотрим *уравнение Пуассона* в двумерной области  $\Omega \subset \mathbb{R}^2$ :

$$\begin{aligned} -\Delta u &= f \text{ в } \Omega, \\ u &= g \text{ на } \Gamma_D, \\ -\frac{\partial u}{\partial \vec{n}} &= \mu \text{ на } \Gamma_N, \end{aligned}$$

где  $\Gamma_D$  — часть границы области, на которой заданы граничные условия первого рода;  $\Gamma_N$  — часть границы области, на которой заданы граничные условия второго рода,  $\Gamma_D \cup \Gamma_N = \partial\Omega$ ,  $\Gamma_D \cap \Gamma_N = \emptyset$ ,  $\Gamma_D \neq \emptyset$ .

Будем считать, что в дальнейшем все рассматриваемые функции имеют нужное количество производных, поэтому не будем накладывать формальные условия на гладкость.

Представим решение задачи в виде  $u = u_0 + u_g$ , где функция  $u_0$  обращается в нуль на части  $\Gamma_D$  границы области, а  $u_g$  — некоторая произвольная, но наперед заданная функция, значения которой совпадают с  $g$  на границе области,  $u_g|_{\Gamma_D} = g$ .

Тогда можно перейти к следующей задаче с однородными граничными условиями первого рода на  $\Gamma_D$  относительно функции  $u_0$ :

$$\begin{aligned} -\Delta u_0 &= f + \Delta u_g \quad \text{в } \Omega, \\ u_0 &= 0 \quad \text{на } \Gamma_D, \\ -\frac{\partial u_0}{\partial \vec{n}} &= \mu + \frac{\partial u_g}{\partial \vec{n}} \quad \text{на } \Gamma_N. \end{aligned}$$

*Слабая постановка задачи* для определения  $u_0$  может быть получена способом, описанным в 16.3.1: определить  $u_0 \in V_D$ , такое, что

$$\int_{\Omega} \nabla u_0 \cdot \nabla v \, d\Omega = \int_{\Omega} fv \, d\Omega - \int_{\Gamma_N} \mu v \, dS - \int_{\Omega} \nabla u_g \cdot \nabla v \, d\Omega, \quad v \in V_D,$$

где пространство  $V_D$  состоит из функций, имеющих суммируемые с квадратом первые производные и обращающихся в нуль на части  $\Gamma_D$  границы расчетной области:

$$V_D = \{v \in V: v|_{\Gamma_D} = 0\},$$

а пространство  $V$  состоит из произвольных заданных в  $\Omega$  функций, имеющих суммируемые с квадратом первые производные.

Для аппроксимации задачи с помощью МКЭ рассмотрим конечно-мерное пространство  $V_h$ , аппроксимирующее пространство  $V$  и пространство  $V_{D,h} = V_h \cap V_D(\Omega)$ , элементы которого приближают элементы пространства  $V_D$ .

Пусть функция  $u_{g,h} \in V_h$  представляет собой аппроксимацию функции  $u_g$ , задающей граничное условие первого рода. В качестве функции  $u_{g,h}$  может быть взят интерполянт функции  $u_g$  по системе базисных функций пространства  $V_h$ .

Тогда конечномерная задача примет вид: определить  $u_{0,h} \in V_{D,h}$ , такую, что

$$\int_{\Omega} \nabla u_{0,h} \cdot \nabla v_h \, d\Omega = \int_{\Omega} fv_h \, d\Omega - \int_{\Gamma_N} \mu v_h \, d\gamma - \int_{\Omega} \nabla u_{g,h} \cdot \nabla v_h \, d\Omega, \quad v_h \in V_{D,h}.$$

Пусть  $\varphi_i$ ,  $i = \overline{1, N}$ , — базис в пространстве  $V_h$ , причем часть функций  $\varphi_i$  с номерами  $i \in I$  образуют базис в пространстве  $V_{D,h}$ . Другими словами, множество индексов  $I$  соответствует тем базисным функциям из  $V_h$ , которые обращаются в нуль на  $\Gamma_D$ . Количество индексов в множестве  $I$  будем считать равным  $M = |I| < N$ ,  $|I| \geq 1$ .

Тогда последнее уравнение эквивалентно следующему:

$$\int_{\Omega} \nabla u_{0,h} \cdot \nabla \varphi_i d\Omega = \int_{\Omega} f \varphi_i d\Omega - \int_{\Gamma_N} \mu \varphi_i d\gamma - \int_{\Omega} \nabla u_{g,h} \cdot \nabla \varphi_i d\Omega, \quad i \in I,$$

или, что то же самое, для  $\forall \varphi_i \in V_{D,h}$ .

Представляя неизвестное решение в виде линейной комбинации базисных функций:

$$u_{0,h} = \sum_{i \in I} u_{0,h,i} \varphi_i, \quad u_{g,h} = \sum_{i=1}^N u_{g,h,i} \varphi_i,$$

окончательно получим СЛАУ для определения неизвестных коэффициентов  $U_h = \{u_{0,h,i}\}$ :

$$Au_{0,h} = b,$$

где  $A = A_{M \times M}$ ,  $b = b_{M \times 1}$ ,

$$A_{ij} = \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j d\Omega, \quad i, j \in I,$$

$$b_i = \int_{\Omega} f \varphi_i d\Omega - \int_{\Gamma_N} \mu \varphi_i d\gamma - \sum_{j=1}^N u_{g,h,j} \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j d\Omega, \quad i \in I. \quad (16.21)$$

Матрицу  $A$  обычно называют **матрицей жесткости** задачи, даже если рассматривается не аппроксимация задачи теории упругости, а какая-либо другая.

Приближенное решение исходной задачи будет иметь вид

$$u_h = u_{0,h} + u_{g,h} = \sum_{i \in I} u_{0,h,i} \varphi_i + \sum_{i=1}^N u_{0,g,h,i} \varphi_i.$$

Описанные построения в принципе верны для произвольных аппроксимаций задачи, т. е. не зависят от способа построения пространств  $V_h$  и других, а следовательно, и от выбора базисных функций.

В случае МКЭ базисные функции являются финитными и обычно задаются тем или иным образом на треугольной либо четырехугольной сетке, построенной в области  $\Omega$ .

Поэтому будем считать, что в области  $\Omega$  задана **правильная триангуляция  $T$** , т. е. построено такое разбиение области  $\Omega$  на треугольные ячейки (конечные элементы), что любые два треугольника имеют

либо общее ребро, либо общую вершину, либо пустое пересечение. Таким образом,

$$\Omega = \bigcup_{T \in \mathcal{T}} T.$$

Эта триангуляция, в свою очередь, порождает разбиение границы  $\Gamma$  области  $\Omega$  на граничные элементы  $E$ , представляющие собой стороны треугольников, которые образуют *триангуляцию*  $\mathcal{T}$ . Также будем считать, что какое-либо граничное ребро  $E$  целиком принадлежит либо  $\Gamma_N$ , либо  $\Gamma_D$ . При этом объединение всех таких ребер совпадает с полной границей  $\Gamma$  области  $\Omega$ .

Каждый треугольник  $T$  при этом задается набором трех своих узлов  $P_k$  с координатами  $P_k = (x_k, y_k)$ . В дальнейшем будем считать, что узлы треугольника обходятся в положительном направлении, т. е. против хода часовой стрелки.

В простейшем случае кусочно-линейных базисных функций  $\varphi_k$  принимает значение единица в узле  $P_k$  и нуль во всех остальных узлах. В пределах одного треугольника она продолжена линейно.

Базис пространства  $V_h$  состоит из базисных функций, соответствующих всем узлам триангуляции, а базис пространства  $V_{D,h}$  — из базисных функций, соответствующих всем узлам, не лежащим на части  $\Gamma_D$  границы.

Теперь в силу свойства аддитивности интеграла относительно области интегрирования соотношения (16.21) могут быть записаны в виде

$$A_{ij} = \sum_{T \in \mathcal{T}} \int_T \nabla \varphi_i \cdot \nabla \varphi_j d\Omega, \quad i, j \in I,$$

$$b_i = \sum_{T \in \mathcal{T}} \int_T f \varphi_i d\Omega - \sum_{E \in \Gamma_N} \int_E \mu \varphi_i d\gamma - \sum_{j=1}^N u_{g,h,j} \sum_{T \in \mathcal{T}} \int_T \nabla \varphi_i \cdot \nabla \varphi_j d\Omega, \quad i \in I.$$

Таким образом, задача вычисления интегралов в выражениях для коэффициентов матрицы жесткости задачи и ее правой части сводится к задаче вычисления тех же интегралов по отдельным треугольникам или их сторонам.

**Сборка матрицы жесткости.** Рассмотрим *алгоритм сборки матрицы жесткости* задачи. По сути, он представляет собой удобный алгоритмический прием для вычисления значений коэффициентов матрицы жесткости, в целом не привязанный ни к конкретному виду

базисных функций, ни к способу разбиения области на конечные элементы.

В соответствии с ним вычисление коэффициентов матрицы жесткости сводится к вычислению матриц жесткости отдельных конечных элементов. При проведении вычислений в пределах одного конечного элемента, конечно же, проявляются особенности и в его форме, и в способе выбора базисных функций. Однако эти особенности соответствуют не каждому отдельному конечному элементу, а классу элементов заданного типа. При эффективной реализации алгоритм сборки очень удобен и позволяет разрабатывать качественные программы, позволяющие использовать самые разные базисные функции и конечные элементы.

Рассмотрим один из треугольников  $T$  триангуляции. Будем считать, что его вершины имеют координаты  $P_1 = (x_1, y_1)$ ,  $P_2 = (x_2, y_2)$ ,  $P_3 = (x_3, y_3)$ . Пусть  $\varphi_1, \varphi_2, \varphi_3$  — базисные функции, соответствующие этим вершинам и данному треугольнику. Таким образом,

$$\varphi_j(x_i, y_i) = \delta_{ij}, \quad i, j = 1, 2, 3,$$

и функции  $\varphi_j$  являются линейными в пределах  $T$ .

Можно показать, что в этом случае внутри рассматриваемого конечного элемента базисные функции задаются соотношениями

$$\varphi_i(x, y) = \frac{\det \begin{pmatrix} 1 & x & y \\ 1 & x_{i+1} & y_{i+1} \\ 1 & x_{i+2} & y_{i+2} \end{pmatrix}}{\det \begin{pmatrix} 1 & x_i & y_i \\ 1 & x_{i+1} & y_{i+1} \\ 1 & x_{i+2} & y_{i+2} \end{pmatrix}}, \quad i = 1, 2, 3,$$

где для удобства обозначения формально считается, что  $P_4 = P_1$ ,  $x_4 = x_1$ ,  $y_4 = y_1$ , аналогично индекс 5 идентичен индексу 2, т. е. значения индексов вычисляются по модулю числа 3.

Непосредственным дифференцированием можно убедиться, что

$$\nabla \varphi_i(x, y) = \frac{1}{2|T|} \begin{pmatrix} y_{i+1} - y_{i+2} \\ x_{i+2} - x_{i+1} \end{pmatrix},$$

где  $|T|$  — площадь треугольника  $T$ ,

$$|T| = \frac{1}{2} \det \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}.$$

В результате получается следующее выражение для **матрицы жесткости конечного элемента**  $T$ :

$$\begin{aligned} A_{T,ij} &= \int_T \nabla \varphi_i \nabla \varphi_j d\Omega = \\ &= \frac{|T|}{(2|T|)^2} \begin{pmatrix} y_{i+1} - y_{i+2} \\ x_{i+2} - x_{i+1} \end{pmatrix}^T \begin{pmatrix} y_{j+1} - y_{j+2} \\ x_{j+2} - x_{j+1} \end{pmatrix}, \quad i, j = 1, 2, 3. \end{aligned}$$

В матричном виде это соотношение можно записать так:

$$A_T = \frac{1}{2}|T|GG^T, \quad G = \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

В силу линейности базисных функций приведенные выше выражения являются точными. При использовании базисных функций более высокого порядка или при решении задачи с переменными коэффициентами для вычисления элементов матрицы жесткости может потребоваться использование тех или иных квадратурных формул.

Аналогично может быть рассмотрена правая часть  $b$  полученной СЛАУ. Вклад в правую часть дают плотность объемных сил  $f$ , граничные условия второго рода, заданные на части границы  $\Gamma_N$ , и член, содержащий функцию  $u_g$ , задающую граничное условие первого рода.

В этом случае можно прийти к следующим выражениям:

$$F_{T,i} = \int_T f \varphi_i dx \approx \frac{1}{6} \det \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} f(x_s, y_s), \quad i = 1, 2,$$

где  $(x_s, y_s)$  — координаты **барицентра треугольника**, т. е. точки, все барицентрические координаты которой одинаковы ( $1/3$  в двумерном случае,  $1/4$  в трехмерном, см. 12.2.2, 16.5.4). Такое выражение соответствует использованию одноточечной квадратурной формулы для вычисления соответствующего интеграла.

Аналогично можно получить

$$G_{T,i} = \int_E g \varphi_i dx \approx \frac{|E|}{2} g(x_M, y_M), \quad i = 1, 2,$$

где  $|E|$  — длина ребра  $E$ , а  $(x_M, y_M)$  — координаты его центра.

Выражение для слагаемого, содержащего функцию  $u_{g,h}$ , которая задает граничное условие первого рода, может быть реализовано численно аналогичным способом. Подробно оно не будет рассматриваться, потому что существуют более удобные методы учета граничных условий, которые будут описаны ниже.

Таким образом можно вычислить все необходимые величины, отнесенные к одному треугольнику или ребру триангуляции.

Для вычисления коэффициентов матрицы жесткости полной задачи необходимо использовать соответствие между локальной нумерацией узлов одного треугольника и глобальной нумерацией узлов в области. Например, узлы конечного элемента  $T$  имеют глобальные номера  $\pi(T, 1), \pi(T, 2), \pi(T, 3)$ , где  $\pi(T, k)$  — глобальный номер узла конечного элемента  $T$  с локальным номером  $k = 1, 2, 3$ .

При программной реализации метода функция  $\pi$  обычно задается в виде двумерного массива  $S$ , представляющего собой список вершин треугольников, образующих триангуляцию. Так, если элемент  $T$  имеет в этом списке номер  $l$ , то  $\pi(T, k) = S(l, k)$ .

Пусть  $u_T = (u_{T,1}, u_{T,2}, u_{T,3})^T$  — вектор значений решения в узлах данного конечного элемента, а  $u = (u_1, u_2, \dots, u_N)^T$  — вектор решения во всей области.

Тогда связь между локальным вектором решения  $u_T$  и глобальным вектором решения  $u$  может быть записана в виде

$$u = N_T u_T, \quad u_T = N_T^T u,$$

где  $N_T$  — матрица размерами  $N \times 3$ , все элементы которой, за исключением трех, равны нулю. Оставшиеся элементы с индексами  $(\pi(T, 1), 1), (\pi(T, 2), 2), (\pi(T, 3), 3)$  равняются единице.

Непосредственной проверкой можно убедиться, что матрица жесткости и вектор правой части могут быть записаны в виде

$$A = \sum_{T \in \mathcal{T}} N_T A_T N_T^T, \quad f = \sum_{T \in \mathcal{T}} N_T F_T, \quad g = \sum_{E \in \Gamma} N_E G_E,$$

где индекс  $T$  или  $E$  обозначает, что соответствующая матрица относится к конечному элементу  $T$  либо к граничному ребру  $E$ .

Матрица  $N_E$  для какого-либо ребра  $E$  имеет два столбца, соответствующих двум узлам, образующим ребро, и  $M$  строк. Она устроена аналогично матрице для конечного элемента, т. е. все ее элементы равны нулю, за исключением двух, равных единице — с индексами  $(1, k_1)$  и  $(2, k_2)$ , где  $k_1, k_2$  равняются глобальным номерам узлов, образующих данное ребро.

Приведенные выражения дают удобный способ вычисления коэффициентов матрицы жесткости. При реализации МКЭ матрицы  $N$  никогда не формируются явным образом. Данный способ можно представить в виде следующего алгоритма:

- 1) положить величины  $A, F, G$  равными нулю;

2) в цикле по конечным элементам выполнить следующие действия:

- вычислить матрицу жесткости  $A_T$  и вектор правой части  $F_T$  (соответствующий приложенной объемной силе) конечного элемента  $T$ ;
- присвоить

$$I := (\pi(T, 1), \pi(T, 2), \pi(T, 3)),$$

$$A(I, I) := A(I, I) + A_T, \quad b(I, I) = b(I) + F_T;$$

3) в цикле по граничным элементам выполнить следующие действия:

- вычислить вектор правой части  $G_E$ , соответствующий ребру  $E$ ;
- присвоить

$$I := (\pi(T, k_1), \pi(T, k_2)),$$

$$b(I) := b(I) + G_E,$$

где  $k_1, k_2$  — локальные номера узлов, образующих ребро в локальной нумерации узлов того конечного элемента, которому это ребро принадлежит.

Здесь для произвольных целочисленных векторов  $I = (i_1, i_2, \dots, i_n)$ ,  $J = (j_1, j_2, \dots, j_m)$  матрица  $A(I, J)$  представляет собой подматрицу размером  $n \times m$  матрицы  $A$ , стоящую на пересечении строк с номерами  $i_1, i_2, \dots, i_n$  и столбцов с номерами  $j_1, j_2, \dots, j_m$ . Аналогичные обозначения используются и для векторов.

Граничные условия первого рода могут учитываться непосредственно с использованием приведенных выше выражений для вычисления матриц жесткости отдельных конечных элементов. Однако этот способ не всегда является удобным с точки зрения реализации МКЭ и используется в основном при его теоретическом исследовании.

Рассмотрим другой способ. Сначала формально рассмотрим вектор неизвестных, соответствующих всем узлам, заданным в области  $\Omega$ . При этом упорядочим узлы так, чтобы сначала шли номера, не попадающие на часть границы  $\Gamma_D$ , т. е. перенумеруем узлы (и базисные функции) так, чтобы множество  $I = \{1, 2, \dots, M\}$ .

Сформируем для этого случая матрицу жесткости  $A$  и вектор правой части  $b$ .

Тогда полная система уравнений  $Au = b$  может быть записана в следующем блочном виде:

$$\begin{pmatrix} A_{II} & A_{I\Gamma_D} \\ A_{I\Gamma_D}^T & A_{\Gamma_D\Gamma_D} \end{pmatrix} \begin{pmatrix} u_{h,I} \\ u_{h,\Gamma_D} \end{pmatrix} = \begin{pmatrix} b_I \\ b_{\Gamma_D} \end{pmatrix}.$$

При этом вектор  $u_{h,I}$  соответствует узлам, не попадающим на  $\Gamma_D$ , а значение элементов вектора  $u_{h,\Gamma_D}$  уже задано из граничных условий,

$$u_{h,\Gamma_D} = g_h.$$

Это позволяет исключить  $u_{h,\Gamma_D}$  из первой блочной строки рассмотренной выше системы уравнений, в результате чего можно записать

$$A_{II}u_{h,I} = b_I - A_{I\Gamma_D}g_h.$$

Матрицы  $A_{II}$ ,  $A_{I\Gamma_D}$  легко могут быть получены из полной матрицы жесткости  $A$  удалением строк и столбцов с номерами  $M+1$ ,  $M+2, \dots, N$ .

Вектор  $b_{\Gamma_D}$ , вообще говоря, неизвестен до решения задачи и представляет собой аппроксимацию выражения

$$\int_{\Gamma_D} \frac{\partial u}{\partial n} u_g dS,$$

зависящего от неизвестного решения  $u$ . Его значения могут быть вычислены после решения задачи.

В рассматриваемом способе эти значения не требуются и не используются.

С использованием введенных в 16.5.1 обозначений можно показать, что такой подход соответствует выбору функции  $u_{g,h}$ , равной нулю во всех узлах, кроме тех, которые принадлежат части  $\Gamma_D$  границы области.

Наконец, третий способ заключается в непосредственной модификации матрицы жесткости и вектора правой части в соответствии с заданными граничными условиями первого рода, а именно: если в узле с номером  $i$  задано граничное условие  $u_{h,i} = g_i$ , то необходимо положить  $A_{ij} = 0$ ,  $j = 1, 2, \dots, i-1, i+1, \dots, N$ ,  $A_{ii} = 1$ ,  $F_i = g_i$ .

Довольно часто для учета граничных условий используется так называемый **метод штрафа**. В этом случае исходная задача, например вида

$$-\Delta u = f \text{ в } \Omega; \quad u = g \text{ на } \partial\Omega,$$

с граничными условиями первого рода заменяется на задачу с граничным условием третьего рода вида

$$-\Delta u_\varepsilon = f \text{ в } \Omega; \quad u_\varepsilon + \varepsilon \frac{\partial u_\varepsilon}{\partial \vec{n}} = g \text{ на } \partial\Omega. \quad (16.22)$$

В последней задаче граничное условие является естественным и, следовательно, не накладывает каких-либо ограничений на пространства базисных и пробных функций.

Можно показать, что при достаточно малом значении параметра  $\varepsilon$  решение  $u_\varepsilon$  близко к  $u$  [112].

Качественные соображения, подтверждающие справедливость этого утверждения, следующие. Запишем граничное условие для  $u_\varepsilon$  в виде

$$\frac{\partial u}{\partial \vec{n}} = \frac{1}{\varepsilon} (g - u_\varepsilon).$$

Обе части этого равенства являются ограниченными величинами. При этом величина  $1/\varepsilon$  велика, если параметр  $\varepsilon$  мал. Тогда в силу ограниченности производной величина  $g - u_\varepsilon$  в правой части равенства мала.

Указанный способ учета граничных условий имеет название «метод штрафа» по следующим причинам. Замена исходной задачи на задачу минимизации выпуклого функционала проводится при рассмотрении граничных условий первого рода как ограничений, накладываемых на аргумент целевой функции (в данном случае целевой функцией является функционал Ритца). Одним из классических способов сведения этой задачи к задаче безусловной оптимизации является метод штрафа, описанный во многих источниках. Условие стационарности такого модифицированного функционала и приводит к *слабой постановке задачи*, которая эквивалентна слабой постановке задачи (16.22).

Недостатком этого метода является появление в задаче малого параметра, что негативно сказывается на свойствах результирующей системы уравнений, а именно на ее числе *обусловленности*.

### 16.5.2. Линейная задача теории упругости

Рассмотрим математическую постановку задачи: требуется определить поле перемещения  $u = (u_i) = (u_1, u_2, \dots, u_d)$  в области  $\Omega \in \mathbb{R}^d$ ,  $d = 2$  или  $d = 3$ , с границей  $\Gamma = \partial\Omega$ , удовлетворяющее уравнению

$$\operatorname{div} \sigma(u) = -f$$

и граничным условиям на  $\Gamma$ . Здесь  $\sigma$  — тензор напряжений, компоненты которого в декартовой ортогональной системе координат  $Ox_1x_2\dots x_d$  будем обозначать  $\sigma_{ij}$ ;  $f = (f_1, f_2, \dots, f_d)$  — вектор объемной плотности сил, действующих на тело.

В координатном представлении задача может быть записана как

$$\frac{\partial \sigma_{ij}}{\partial x_j} = -f_i.$$

При этом используется правило суммирования по повторяющимся индексам от 1 до  $d$ .

Отметим, что в случае декартовой ортогональной системы координат с базисными векторами  $e_i$  справедливо следующее соотношение для компонент  $u_i$  вектора перемещения:

$$u_i = (u, e_i).$$

Будем считать, что, как и раньше, границу  $\Gamma$  можно представить в виде объединения двух частей  $\Gamma = \Gamma_D \cup \Gamma_N$ . В отличие от рассмотренного в 16.5.1 уравнения Пуассона, где  $\Gamma_N \cap \Gamma_D = \emptyset$ , для рассматриваемой задачи части границы  $\Gamma_N$  и  $\Gamma_D$  могут иметь непустое пересечение.

Границное условие Дирихле на  $\Gamma_D$  (кинематические граничные условия) в общем случае может быть записано в виде

$$Mu = w, \quad (16.23)$$

где  $M$  — матрица размером  $d \times d$ . Значения ее элементов в данной точке зависят от граничного условия. Например, если в точке задан полный вектор перемещения  $u_D$ , то нужно формально положить

$$M = \begin{pmatrix} e_1 \\ \vdots \\ e_d \end{pmatrix} = E_{d \times d},$$

где  $e_i$ ,  $i = \overline{1, d}$ , — базисные векторы в пространстве  $\mathbb{R}^d$ ;  $E_{d \times d}$  — единичная матрица размером  $d \times d$ ;  $w = u_D$ . В этом случае формально имеем

$$(e_i, u) = u_i = w_i, \quad i = \overline{1, d},$$

т. е. компоненты поля перемещения принимают заданные значения.

Например, если в точке задано равенство нулю нормальной компоненты поля перемещения, т. е. тело может «скользить» по заданной поверхности, то необходимо положить

$$M = \begin{pmatrix} n \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} n_1 & n_2 & \cdots & n_d \\ 0 & 0 & \cdots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

и  $w = 0$ , где  $n = (n_1, n_2, \dots, n_d)$  — вектор единичной внешней нормали к границе области в заданной точке.

Отсутствие каких-либо граничных условий первого рода формально соответствует нулевым матрице  $M$  и вектору  $w$ .

Границные условия первого рода другого вида, например, в случае задания только некоторых компонент поля перемещения могут быть рассмотрены аналогично.

Зависимость тензора напряжений  $\sigma$  от перемещения  $u$  в рассматриваемой модели имеет вид  $\sigma(u) = \mathbb{C} : \varepsilon(u)$ , где тензор деформации

$$\varepsilon(u) = \frac{1}{2}(\nabla(u) + \nabla(u)^T),$$

а тензор  $\mathbb{C}$  четвертого ранга называют тензором коэффициентов упругости. Его вид определяется моделью рассматриваемой среды. В простейшем случае линейной однородной изотропной среды его компоненты имеют вид

$$C_{ijkl} = \frac{2\mu\nu}{1-2\nu} \delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \quad (16.24)$$

где

$$\mu = \frac{E}{2(1+\nu)}.$$

Здесь  $\mu$  — модуль сдвига;  $E$  и  $\nu$  — модуль Юнга и коэффициент Пуассона соответственно. Тензор  $\mathbb{C}$  симметричен по первой паре индексов, по второй паре индексов и по парам индексов. Последнее означает, что тензор не меняется при замене первой пары индексов на вторую, и наоборот. Знак двоеточия в записи тензора напряжений означает внутреннее произведение (или свертку [82]) тензоров.

В этом случае связь между напряжениями и деформациями может быть непосредственно записана в виде

$$\sigma_{ij} = 2\mu\varepsilon_{ij} + \lambda\varepsilon_{kk}\delta_{ij}, \quad (16.25)$$

где

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)} =$$

модуль упругости.

Пару коэффициентов  $\lambda$  и  $\mu$  называют константами Ламе.

В дальнейшем мы будем рассматривать только случай (16.24). Обобщение на более сложный случай тензора коэффициентов упругости не представляет труда.

Границное условие второго рода (динамическое граничное условие) может быть записано в виде

$$\sigma(u)n = p,$$

где  $p$  — заданный вектор нормальных напряжений;  $n$  — вектор единичной внешней нормали к границе области.

Отметим, что в классической постановке приведенная выше форма задания граничных условий верна с той оговоркой, что их количество

в одном узле всегда должно равняться трем. В этом случае на части границы  $\Gamma_N \cap \Gamma_D$ , где рассматриваются граничные условия обоих типов, задание граничного условия второго рода в полном виде, вообще говоря, неверно, следует рассматривать лишь необходимое количество проекций соотношения  $\sigma(u)n = p$  на соответствующее количество направлений.

С учетом приведенной выше зависимости  $\sigma$  от  $\varepsilon$  и произвольной зависимости  $\varepsilon$  от перемещения  $u$  слабая постановка задачи примет вид: определить  $u \in V$ , такое, что

$$\int_{\Omega} \varepsilon(v) : \mathbb{C} : \varepsilon(u) d\Omega = \int_{\Omega} fv d\Omega + \int_{\Gamma_N} pv dS, \quad v \in V_D,$$

$$Mu = w \text{ на } \Gamma_D.$$

Здесь  $V$  — пространство достаточно гладких векторных полей, а пространство  $V_D$  определено как

$$V_D = \{v \in V: Mv = 0 \text{ на } \Gamma_D\}.$$

Отметим, что слабая постановка задачи построена корректно с точки зрения задания граничных условий. Например, если в точке границы заданы нулевые нормальные перемещения (одно условие первого рода), то слабое решение задачи будет автоматически удовлетворять не условию  $\sigma(u)n = p$ , а граничному условию  $\tau\sigma(u)n = \tau p$ , где  $\tau$  — произвольный вектор в касательной плоскости в данной точке.

При аппроксимации приведенной задачи методом Бубнова — Галеркина пространство  $V$ , в котором разыскивается решение, заменяется на конечномерное пространство  $V_h$ . Тогда приближенное решение определяется как решение следующей задачи: определить  $u_h \in V_h$ , такое, что  $M_h u_h = w_h$  на  $\Gamma_D$  и для любого  $v_h \in V_{D,h}$  справедливо

$$\int_{\Omega} \varepsilon(v_h) : \mathbb{C} : \varepsilon(u_h) d\Omega = \int_{\Omega} fv_h d\Omega + \int_{\Gamma_N} pv_h dS.$$

Пространство  $V_h$  может быть построено разными способами.

Пока лишь отметим, что при заданном пространстве  $V_h$  пространство пробных функций  $V_{D,h}$  формально определяется как  $V_{D,h} = V_h \cap V_D$ , т. е.

$$V_{D,h} = \{v_h \in V_h: Mv_h = 0 \text{ на } \Gamma_D\}.$$

В отличие от уравнения Пуассона на данном этапе не будем определять базисные функции пространства  $V_{D,h}$ , так как конкретный вид

его базиса в данном случае указать гораздо сложнее, хотя и возможно. Эта сложность связана прежде всего с наличием в постановке задачи непустого пересечения частей границы  $\Gamma_N$  и  $\Gamma_D$  и наличия достаточно общего граничного условия  $Mu = w$ . Ниже будет указан способ учета граничных условий, формально не требующий построения соответствующих базисных функций в  $V_{D,h}$ . Пока лишь отметим, что  $V_{D,h}$  является подпространством пространства  $V_h$ .

Перейдем к рассмотрению пространства  $V_h$ .

В случае использования МКЭ оно строится следующим образом.

Пусть  $T$  — триангуляция области  $\Omega$ , т. е.  $\Omega = \bigcup_{T \in T} T$ , где  $T$  — тре-

угольники для случая двумерной задачи ( $d = 2$ ) или тетраэдры для случая трехмерной задачи ( $d = 3$ ). Границы и ребра элементов дискретизации образуют разбиение границы  $\Gamma$  расчетной области. Будем считать, что любая грань  $E \subset \Gamma$  какого-либо элемента  $T$  принадлежит полностью части границы  $\Gamma_N$ ,  $\Gamma_D$  либо им обеим.

Пусть  $C$  — множество узлов триангуляции  $T$ . Будем считать, что общее количество узлов в области равняется  $N$ .

Тогда набор векторных базисных функций  $\eta_i$ ,  $i = \overline{1, dN}$ , представляющих базис в пространстве  $V_h$ , для случая  $d = 3$  будет иметь вид

$$(\eta_1, \eta_2, \dots, \eta_{dN}) = (\varphi_1 e_1, \varphi_1 e_2, \varphi_1 e_3, \dots, \varphi_d e_d, \dots, \varphi_N e_1, \varphi_N e_2, \varphi_N e_3),$$

где  $\varphi_i$  — обычная скалярная кусочно-линейная базисная функция, соответствующая узлу  $P_i$ , т. е.  $\varphi_j(P_i) = \delta_{ij}$ , а на ребра, грани и вовнутрь конечных элементов  $T$  функция  $\varphi_i$  продолжена линейно. Отметим также, что с каждым узлом триангуляции связано такое количество базисных функций, которое соответствует пространственной размерности задачи, и для интерполяции каждой компоненты вектора перемещения (рассматриваемого как скалярная функция) используется обычная кусочно-линейная интерполяция на сетке из треугольников или тетраэдров.

Таким образом, с каждым узлом связаны  $d = 2$  или  $d = 3$  базисные функции (степени свободы), соответствующие перемещениям вдоль двух или трех независимых направлений, заданных базисными векторами используемой системы координат  $e_1, e_2$  или  $e_1, e_2, e_3$ .

Следовательно, построенная выше конечномерная задача (без учета граничных условий первого рода) может быть записана в виде

$$\int_{\Omega} \varepsilon(\eta_j) : \mathbb{C} : \varepsilon(u_h) d\Omega = \int_{\Omega} f \eta_j d\Omega + \int_{\Gamma_N} p \eta_j dS, \quad j = \overline{1, dN},$$

или, что то же самое, для любого  $\eta_i \in V_h$ .

Представив решение задачи в виде линейной комбинации приведенных выше базисных функций:

$$u_h = \sum_{i=1}^{dN} u_i \eta_i,$$

получим следующую систему уравнений относительно вектора неизвестных  $u_h$ :

$$\sum_{i=1}^{dN} \left( \int_{\Omega} \varepsilon(\eta_j) : \mathbb{C} : \varepsilon(\eta_i) d\Omega \right) u_i = \int_{\Omega} f \eta_j d\Omega + \int_{\Gamma_N} p \eta_j dS, \quad j = \overline{1, dN},$$

или в матричном виде  $Au_h = F$ , где

$$A_{ij} = \int_{\Omega} \varepsilon(\eta_j) : \mathbb{C} : \varepsilon(\eta_i) d\Omega, \quad F_j = \int_{\Omega} f \eta_j d\Omega + \int_{\Gamma_N} p \eta_j dS, \quad i, j = \overline{1, dN}.$$

Для вычисления коэффициентов матрицы жесткости  $A$  и вектора правой части  $F$ , как и в 16.5.1, удобно перейти к рассмотрению отдельных конечных элементов: представим интегралы в матрице жесткости и векторе правой части в виде суммы интегралов по отдельным элементам, т. е.

$$F_j = \sum_{T \in \mathcal{T}} \int_T f \eta_j \, d\Omega + \sum_{E \in \Gamma_N} \int_E p \eta_j \, dS, \quad i, j = \overline{1, dN}.$$

Рассмотрим теперь один конечный элемент  $T$  с узлами  $P_1, P_2, \dots, P_K$ . Для задачи пространственной размерности  $d$  с конечным элементом связано  $dK$  базисных функций (степеней свободы), которые имеют вид

$$\eta_{\pi(T,1)} = \varphi_{P_1} e_1, \quad \dots, \quad \eta_{\pi(T,d)} = \varphi_{P_1} e_d,$$

• • • • • • • • • • • • • • •

$$\eta_{\pi(TdK-(d-1))} = \varphi_{P_K} e_1, \dots, \eta_{\pi(TdK-1)} = \varphi_{P_K} e_d.$$

$$\eta_{\pi(T,dK-(d-1))} = \varphi_{P_K} e_1, \quad \dots, \quad \eta_{\pi(T,dK)} = \varphi_{P_K} e_d,$$

где  $e_i$  —  $i$ -й единичный вектор (орт). Базисные функции  $\varphi_{P_j}$  представляют собой скалярные базисные функции элемента  $T$ , соответствующие узлу  $P_j$ . Функция  $\pi(T, i)$  определяет по локальным номерам 1, 2, ...,  $dK$  — степеням свободы элемента — их глобальные номера.

Тогда выражение для матрицы жесткости элемента  $T$  примет вид

$$A_{T,ij} = \int_T \varepsilon(\eta_{\pi(T,j)}) : \mathbb{C} : \varepsilon(\eta_{\pi(T,i)}) d\Omega, \quad i, j = \overline{1, dK}.$$

При этом элемент  $A_{pq}$  глобальной матрицы жесткости будет представлять собой сумму величин  $A_{T,ij}$  по всем элементам  $T$ , таким, что  $p = \pi(T, i)$  и  $q = \pi(T, j)$ .

Приведенные выше соотношения верны как в двумерном, так и в трехмерном случаях.

Рассмотрим теперь отдельно построение матриц жесткости конечных элементов для двумерного и трехмерного случаев.

**Двумерный случай.** Здесь конечный элемент представляет собой треугольник  $T$  с вершинами  $P_k = (x_k, y_k)$ ,  $k = \overline{1, 3}$ .

При вычислении локальных матриц жесткости элементов обычно используют матричные, а не тензорные представления определяющих соотношений теории упругости.

Для двумерного случая связь между перемещениями и деформацией можно записать в виде

$$\gamma^*(u) = \begin{pmatrix} \frac{\partial u_1}{\partial x} \\ \frac{\partial u_2}{\partial y} \\ \frac{\partial u_2}{\partial x} + \frac{\partial u_1}{\partial y} \end{pmatrix} = \begin{pmatrix} \varepsilon_{11}(u) \\ \varepsilon_{22}(u) \\ 2\varepsilon_{12}(u) \end{pmatrix}.$$

Отметим, что последний элемент вектора деформации  $\gamma^*$  равен удвоенному элементу  $\varepsilon_{12}$  тензора деформации. Это связано с тем, что тензоры напряжений и деформации являются симметричными, а в векторных обозначениях используются только различные (с учетом симметрии) элементы тензора деформации.

Тогда для рассматриваемой однородной и изотропной среды, в которой связь между напряжениями и деформациями определяется выражением (16.25), тензорное соотношение  $\sigma(u) = \mathbb{C} : \varepsilon(u)$  (см. выражение (16.24)) в матричных обозначениях примет вид

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{pmatrix} = \begin{pmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{pmatrix} \begin{pmatrix} \varepsilon_{11}(u) \\ \varepsilon_{22}(u) \\ 2\varepsilon_{12}(u) \end{pmatrix} = C\gamma^*(u).$$

Следовательно,

$$\begin{aligned} \varepsilon(v) : \mathbb{C} : \varepsilon(u) &= \sigma_{11}\varepsilon_{11} + \sigma_{22}\varepsilon_{22} + \sigma_{12}\varepsilon_{12} + \sigma_{21}\varepsilon_{21} = \\ &= \sigma_{11}\varepsilon_{11} + \sigma_{22}\varepsilon_{22} + 2\sigma_{12}\varepsilon_{12} = \gamma^{*T}(v)C\gamma^*(u). \end{aligned}$$

С учетом этих соотношений получим следующее выражение для элементов матрицы жесткости конечного элемента:

$$A_{T,ij} = \int_T \gamma^{*T}(\eta_{\pi(T,j)}) C \gamma^*(\eta_{\pi(T,i)}) d\Omega = |T| \gamma^{*T}(\eta_{\pi(T,j)}) C \gamma^*(\eta_{\pi(T,i)}),$$

где  $|T|$  — площадь элемента  $T$ .

Отметим, что последнее соотношение в силу линейности базисных функций и того факта, что стороны конечного элемента — отрезки прямых (т. е. конечный элемент — обычный треугольник), является точным. При рассмотрении более сложных конечных элементов для вычисления коэффициентов матрицы жесткости обычно используют ту или иную *квадратурную формулу*. Если она точна на многочленах соответствующей степени, а коэффициенты уравнения являются многочленами порядка не выше, чем используемые базисные функции, то полученные выше соотношения также будут точными и для этого случая.

Путем непосредственных вычислений нетрудно проверить, что

$$\begin{aligned} & \left. \gamma^* \left( \sum_{j=1}^6 u_j \eta_{\pi(T,j)} \right) \right|_T = \\ & = \begin{pmatrix} \varphi_{P_1,x} & 0 & \varphi_{P_2,x} & 0 & \varphi_{P_3,x} & 0 \\ 0 & \varphi_{P_1,y} & 0 & \varphi_{P_2,y} & 0 & \varphi_{P_3,y} \\ \varphi_{P_1,y} & \varphi_{P_1,x} & \varphi_{P_2,y} & \varphi_{P_2,x} & \varphi_{P_3,y} & \varphi_{P_3,x} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_6 \end{pmatrix} = R \begin{pmatrix} u_1 \\ \vdots \\ u_6 \end{pmatrix}. \end{aligned}$$

В записи матрицы индексы  $x$  и  $y$  (например,  $x$  в выражении  $\varphi_{P_2,x}$ ) означают дифференцирование по переменным  $x$  и  $y$ .

Тогда выражение для всех коэффициентов матрицы жесткости одного элемента в матричном виде может быть записано как

$$A_T = |T| R^T C R. \quad (16.26)$$

Для удобного вычисления значений градиента базисных функций используем соотношение

$$\begin{pmatrix} \nabla \varphi_{P_1} \\ \nabla \varphi_{P_2} \\ \nabla \varphi_{P_3} \end{pmatrix} = \frac{1}{2|T|} \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

**Трехмерный случай.** Здесь вектор деформации  $\gamma^*$  имеет следующий вид:

$$\gamma^*(u) = \begin{pmatrix} \frac{\partial u_1}{\partial x} \\ \frac{\partial u_2}{\partial y} \\ \frac{\partial u_3}{\partial z} \\ \frac{\partial u_3}{\partial z} \\ \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \\ \frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x} \\ \frac{\partial u_2}{\partial z} + \frac{\partial u_3}{\partial y} \end{pmatrix} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{13} \\ 2\varepsilon_{23} \end{pmatrix}.$$

Используя аналогично двумерному случаю соотношение

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{23} \end{pmatrix} = \begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{13} \\ 2\varepsilon_{23} \end{pmatrix} = C\gamma^*(u),$$

получим соотношение  $\varepsilon(v) : \mathbb{C} : \varepsilon(u) = \gamma^{*T}(v)C\gamma^*(u)$ .

Наконец, выражение для значения вектора деформации при произвольно заданном кусочно-линейном поле перемещения внутри тетраэдра имеет вид

$$\gamma^*\left(\sum_{j=1}^{12} u_j \eta_{\pi(T,j)}\right) \Big|_T = (\Phi_{P_1}, \Phi_{P_2}, \Phi_{P_3}, \Phi_{P_4}) \begin{pmatrix} u_1 \\ \vdots \\ u_{12} \end{pmatrix},$$

где матрица  $\Phi_{P_i}$ ,  $i = \overline{1, 4}$ , задается следующим образом:

$$\Phi_{P_i} = \begin{pmatrix} \varphi_{P_i,x} & 0 & 0 \\ 0 & \varphi_{P_i,y} & 0 \\ 0 & 0 & \varphi_{P_i,z} \\ \varphi_{P_i,y} & \varphi_{P_i,x} & 0 \\ \varphi_{P_i,z} & 0 & \varphi_{P_i,x} \\ 0 & \varphi_{P_i,z} & \varphi_{P_i,y} \end{pmatrix},$$

и снова получается выражение (16.26) для вычисления матрицы жесткости элемента с матрицей  $R$  вида

$$R = (\Phi_{P_1}, \Phi_{P_2}, \Phi_{P_3}, \Phi_{P_4}).$$

Выражение для вычисления компонент градиентов базисных функций компактно можно записать в виде

$$\begin{pmatrix} \nabla \varphi_{P_1} \\ \nabla \varphi_{P_2} \\ \nabla \varphi_{P_3} \\ \nabla \varphi_{P_4} \end{pmatrix} = \frac{1}{2|T|} \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Для вычисления компонент вектора правой части конечного элемента, как в двумерном, так и в трехмерном случае, воспользуемся одноточечной квадратурной формулой с узлом в центре тяжести конечного элемента.

Тогда для двумерного случая ( $d = 2, K = 3$ ) вклад объемных сил можно представить в виде

$$F_{T,i} = \int_T f \cdot \eta_{\pi(T,i)} d\Omega \approx \frac{1}{6} |T| f_i(P_s), \quad i = 1, 2, \dots, 6,$$

для трехмерного случая ( $d = 3, K = 4$ ) — в виде

$$F_{T,i} = \int_T f \cdot \eta_{\pi(T,i)} d\Omega \approx \frac{1}{24} |T| f_i(P_s), \quad i = 1, 2, \dots, 12.$$

Для вычисления вклада поверхностных сил можно использовать тот же подход — одноточечную квадратурную формулу для интегрирования по ребру границы двумерного треугольного конечного элемента или по треугольной грани тетраэдра. В этом случае соответствующие выражения примут вид

$$\int_E g \cdot \eta_k dS \approx \frac{1}{\varkappa} |E| g_i(P_s), \quad i = 1 + \text{mod}(k - 1, d),$$

где индекс  $i$  пробегает значения от 1 до  $d(K - 1)$ , а  $\varkappa = 2$  для двумерного случая ( $d = 2$ ) и  $\varkappa = 3$  для трехмерного случая ( $d = 3$ ).

**Границные условия первого рода.** Для учета граничных условий первого рода можно использовать различные подходы. Все они начинаются с аппроксимации соотношения (16.23).

Один из них заключается в сборке полной матрицы жесткости описанным выше способом с учетом только граничных условий второго рода.

Далее каждая строка матрицы жесткости, соответствующая степени свободы, на которую наложены граничные условия первого рода,

заменяется на строку, состоящую из нулей, за исключением диагонального элемента, который полагается равным единице. Соответствующая компонента вектора правой части полагается равной значению граничного условия.

При другом подходе непосредственно используют способ задания граничных условий, описанный при постановке задачи.

Как было показано, в произвольной точке границы они могут быть записаны в виде

$$Mu = \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix} u = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1d} \\ m_{21} & m_{22} & \cdots & m_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ m_{d1} & m_{d2} & \cdots & m_{dd} \end{pmatrix} u = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}.$$

Здесь  $i = \overline{1, d}$  и  $m_i = (m_{i1}, m_{i2}, \dots, m_{id})$  — некоторые вектор-функции точки границы, а  $w_i$  — некоторые скалярные функции. Примеры были рассмотрены выше.

При аппроксимации задачи система соотношений в точках границы заменяется на систему линейных уравнений вида

$$Bu_h = W, \quad (16.27)$$

где  $B$  — матрица размером  $n \times dN$ , каждая строка которой представляет собой вектор размерности  $dN$  и содержит значения вектор-функции  $m_i$  в граничных узлах сетки. Аналогично вектор  $W$  размерности  $n$  содержит значение функций  $w_j$  в соответствующем узле. При этом значение  $n$  равняется количеству граничных условий, которые необходимо поставить в конечномерной задаче. Считается, что ранг матрицы  $B$  равняется  $n$ . В частности, при ее построении не учитывается вырожденный случай (отсутствие граничных условий первого рода), т. е. те узлы, в которых все  $m_j$  равняются нулю.

Одним из распространенных способов учета граничных условий является метод множителей Лагранжа. В этом случае получается следующая система уравнений для определения решения:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u_h \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ W \end{pmatrix},$$

где  $\lambda$  — вектор множителей Лагранжа.

Матрица построенной системы является невырожденной, симметричной, но не является положительно определенной. Для ее решения могут быть использованы как общие методы решения СЛАУ, так и специальные методы решения задач о седловой точке. Последние пред-

ставляют собой, по существу, методы минимизации конечномерного квадратичного функционала при наличии ограничений-равенств.

Существуют и другие способы учета граничных условий. В основном они отличаются вариантом включения граничных условий в полную задачу.

Так, например, граничные условия можно учитывать следующим образом. Сначала строится система уравнений, в которой не учитываются граничные условия первого рода. Далее каждая строка этой системы уравнений и соответствующая компонента вектора правой части заменяются на соответствующее уравнение системы (16.27). Полученная система уравнений решается тем или иным способом. Отметим, что при таком подходе результирующая система уравнений, вообще говоря, не будет симметричной.

Еще одним распространенным способом учета граничных условий является *метод штрафа*, рассмотренный выше для случая уравнения Пуассона.

### 16.5.3. Численное интегрирование

Как уже отмечалось выше, при вычислении коэффициентов матрицы жесткости необходимо уметь интегрировать соответствующие выражения.

Обычно на практике для этого используют те или иные *квадратурные формулы*. В силу того, что базисные функции являются многочленами того или иного порядка в пределах одного конечного элемента, наиболее распространены многоточечные квадратурные формулы, которые точны на многочленах нужного порядка. Применение таких квадратурных формул дает точный результат, если подынтегральное выражение является многочленом соответствующего порядка. Например, если коэффициенты задачи постоянны, то достаточно выбрать квадратурные формулы, которые точны на многочленах, используемых при построении конечно-элементных базисных функций.

При этом точность метода (порядок используемых базисных функций) и точность квадратурной формулы должны быть согласованы естественным образом: если метод имеет порядок  $p$  при условии, что коэффициенты конечномерной задачи вычисляются точно, то, чтобы при использовании квадратурной формулы порядок не ухудшился, необходимо применять квадратурные формулы порядка не ниже, чем точность  $p$  метода.

Большое количество подходящих квадратурных формул приведено, например, в [113, 296].

### 16.5.4. Конечные элементы высокого порядка

Ранее были рассмотрены простейшие конечные элементы и базисные функции, являющиеся линейными в пределах одного конечного элемента-треугольника.

Представленные теория и алгоритмические подходы позволяют использовать базисные функции более высокого порядка, обеспечивающие соответственно и большую точность решения.

В общем случае [166] конечный элемент состоит из следующих компонентов:

- 1) собственно конечного элемента — множества точек пространства, которое задает элементарную область, участвующую в разбиении исходной расчетной области (треугольник в приведенных выше примерах);
- 2) некоторого набора функций, линейной оболочкой которых приближается решение в пределах заданного конечного элемента;
- 3) некоторого набора функционалов, которые для заданной в пределах конечного элемента функции позволяют определить неизвестные параметры линейной комбинации.

Например, для рассмотренного ранее линейного конечного элемента на треугольнике указанные компоненты таковы:

- 1) множество точек — треугольник;
- 2) базисные функции — соответствующий набор из трех линейных в пределах заданного треугольника функций;
- 3) функционалы, представляющие собой  $\delta$ -функции Дирака, связанные с вершинами треугольника, и сопоставляющие произвольной заданной в треугольнике функции три ее значения в вершинах треугольника.

Отметим, что разновидностей конечных элементов и соответствующих базисных функций очень много. Существуют лагранжевы и эрмитовы конечные элементы, а также некоторые другие их разновидности (см., например, [166, 207, 296, 297]).

В качестве более сложного примера выберем лагранжев двумерный конечный элемент второго порядка и на его примере рассмотрим основные вопросы, связанные с использованием базисных функций высокого порядка, а также покажем основные приемы работы с ними.

Отметим, что с точки зрения теории метода никаких принципиально новых проблем при использовании таких функций не возникает.

Как и ранее, сам конечный элемент  $T$  является треугольником. Его узлы  $P_1, P_2, \dots, P_6$  образуют три вершины треугольника и три середины его сторон (рис. 16.1). Базис состоит из шести функций

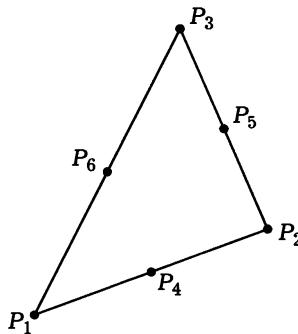


Рис. 16.1

$\varphi_i(x, y)$ ,  $i = \overline{1, 6}$ , представляющих собой квадратичные многочлены двух переменных в пределах элемента  $T$ :

$$\varphi_i(x, y) = a_i x^2 + b_i x y + c_i y^2 + d_i x + e_i y + f_i.$$

Для определения неизвестных коэффициентов  $a_i, b_i, \dots, f_i$  потребуем, чтобы многочлен  $\varphi_i$ , соответствующий узлу  $P_i$ , принимал значение 1 в этом узле и 0 во всех остальных, т. е.

$$\varphi_i(P_j) = \delta_{ij}.$$

Эти условия приводят к следующей системе уравнений относительно неизвестных коэффициентов  $a_i, b_i, \dots, f_i$ , соответствующих базисной функции  $\varphi_i$ :

$$a_i x_j^2 + b_i x_j y_j + c_i y_j^2 + d_i x_j + e_i y_j + f_i = \delta_{ij},$$

где индекс  $j$  пробегает все значения от 1 до 6. Таким образом, получаем систему из шести уравнений для шести неизвестных.

В матричном виде ее можно записать как

$$C \cdot N = E_{6 \times 6},$$

где матрицы

$$C = \begin{pmatrix} x_1^2 & x_1 y_1 & y_1^2 & x_1 & y_1 & 1 \\ x_2^2 & x_2 y_2 & y_2^2 & x_2 & y_2 & 1 \\ x_3^2 & x_3 y_3 & y_3^2 & x_3 & y_3 & 1 \\ x_4^2 & x_4 y_4 & y_4^2 & x_4 & y_4 & 1 \\ x_5^2 & x_5 y_5 & y_5^2 & x_5 & y_5 & 1 \\ x_6^2 & x_6 y_6 & y_6^2 & x_6 & y_6 & 1 \end{pmatrix} \quad \text{и} \quad N = \begin{pmatrix} a_1 & a_2 & \dots & a_6 \\ b_1 & b_2 & \dots & b_6 \\ \dots & \dots & \dots & \dots \\ f_1 & f_2 & \dots & f_6 \end{pmatrix},$$

а  $E_{6 \times 6}$  — единичная матрица размером 6.

Таким образом, матрица неизвестных коэффициентов базисных функций может быть вычислена как

$$N = C^{-1}.$$

Далее, в матричных обозначениях базисная функция  $\varphi_i$  может быть записана в виде

$$\varphi_i(x, y) = P(x, y) \cdot (a_i, b_i, \dots, f_i)^T,$$

где вектор-строка  $P(x, y) = (x^2, xy, y^2, x, y, 1)$ . Тогда для полного вектора базисных функций  $\varphi(x, y) = (\varphi_1(x, y), \varphi_2(x, y), \dots, \varphi_6(x, y))^T$

$$\varphi(x, y) = PN = PC^{-1}.$$

Теперь, имея вектор  $u_T = (u_1, u_2, \dots, u_6)$  узловых значений конечного элемента, можно получить значение интерполянта в любой точке конечного элемента  $T$ :

$$u(x, y) = P(x, y)C^{-1}u_T.$$

Аналогичным образом могут быть получены выражения для векторов градиентов базисных функций и аналогичных величин.

Тем не менее такой способ не очень удобен. Прежде всего проблема возникает в связи с необходимостью обращать матрицу  $C$  для каждого конечного элемента.

Иногда коэффициенты  $a_i, b_i, \dots, f_i$  могут быть легко получены в явном виде, например когда  $T$  является прямоугольным треугольником. Этот факт можно использовать и для получения коэффициентов в более общем случае — если удается построить отображение, переводящее прямоугольный треугольник в заданный (см., например, [112]).

Рассмотрим другой способ описания конечных элементов высокого порядка, более удобный на практике на примере лагранжевых конечных элементов на треугольнике.

При дальнейших построениях будем использовать понятие **барицентрических координат**, которые обозначим через  $\lambda_1, \lambda_2, \lambda_3$ .

Для треугольника  $T$  с вершинами  $(a_{11}, a_{12}), (a_{21}, a_{22}), (a_{31}, a_{32})$  взаимно однозначная связь между барицентрическими  $\lambda_1, \lambda_2, \lambda_3$  и декартовыми  $(x, y)$  координатами произвольной точки  $P \in T$  выражается следующими соотношениями:

$$\begin{aligned} x &= \lambda_1 a_{11} + \lambda_2 a_{21} + \lambda_3 a_{31}, \\ y &= \lambda_1 a_{12} + \lambda_2 a_{22} + \lambda_3 a_{32}, \\ 1 &= \lambda_1 + \lambda_2 + \lambda_3. \end{aligned}$$

Вершине треугольника с координатами  $P_i = (a_{i1}, a_{i2})$  при этом соответствуют значения барицентрических координат  $\lambda_i = 1$ . Следовательно, оставшиеся две барицентрические координаты равны нулю.

Отметим, что  $\lambda_i$  является линейной функцией декартовых координат, причем ее линии уровня параллельны стороне треугольника, противолежащей вершине  $P_i = (a_{i1}, a_{i2})$ .

Явные выражения для  $\lambda_i$  имеют вид

$$\lambda_i = \frac{a_i + b_i x + c_i y}{2\Delta},$$

где

$$a_1 = a_{21}a_{32} - a_{31}a_{22}; \quad b_1 = a_{22} - a_{32}; \quad c_1 = a_{31} - a_{21},$$

выражения для  $a_2, a_3, b_2, b_3$  и  $c_2, c_3$  получаются из приведенных выше путем циклической перестановки первых индексов;

$$\Delta = \frac{1}{2} \det \begin{pmatrix} 1 & a_{11} & a_{12} \\ 1 & a_{21} & a_{22} \\ 1 & a_{31} & a_{32} \end{pmatrix}.$$

Запишем простейшие линейные базисные функции в барицентрических координатах:

$$\varphi_1 = \lambda_1, \quad \varphi_2 = \lambda_2, \quad \varphi_3 = \lambda_3.$$

Приведенные выше формулы и соотношения для этих функций получены именно из такого представления.

Можно показать [296], что для элементов более высокого порядка (второго, третьего и выше) базисные функции могут быть записаны следующим образом:

$$\varphi_i = l_I^I(\lambda_1) l_J^J(\lambda_2) l_K^K(\lambda_3),$$

где функции  $l_k^n$  — базисные интерполяционные многочлены Лагранжа вида

$$l_k^n(\xi) = \frac{(\xi - \xi_0)(\xi - \xi_1) \cdots (\xi - \xi_{k-1})(\xi - \xi_{k+1}) \cdots (\xi - \xi_n)}{(\xi_k - \xi_0)(\xi_k - \xi_1) \cdots (\xi_k - \xi_{k-1})(\xi_k - \xi_{k+1}) \cdots (\xi_k - \xi_n)},$$

принимающие значение 1 при  $\xi = \xi_k$  и 0 при  $\xi = \xi_i, i = \overline{1, n}, i \neq k$ .

Использованные индексы  $I, J, K$  определяются барицентрическими координатами  $\lambda_1(P_i), \lambda_2(P_i)$  и  $\lambda_3(P_i)$  узла с номером  $i$  (рис. 16.2). На рис. 16.2 и далее целое число  $M$  задает степень (порядок) базисных функций, оно на единицу меньше количества узлов, расположенных

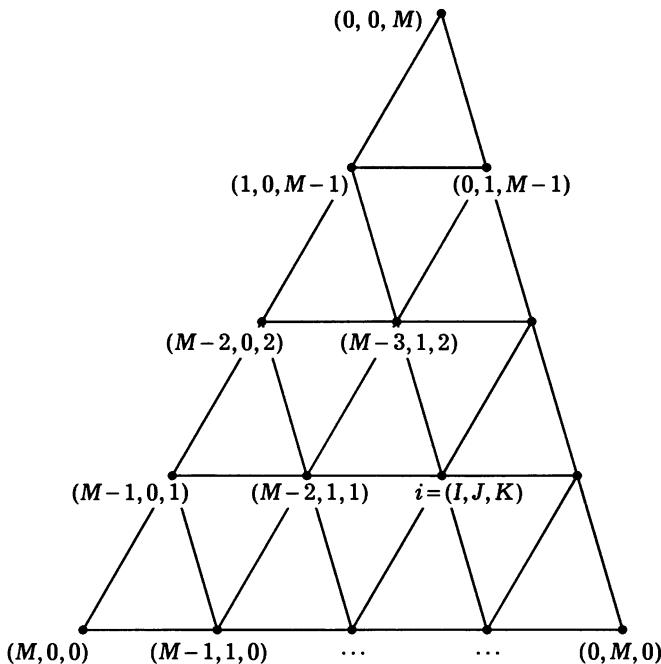


Рис. 16.2

на одной стороне треугольного конечного элемента. Индексы  $I, J, K$  задают номера узлов некоторой треугольной сетки, построенной внутри данного конечного элемента  $T$ .

Узлы интерполяции, которые необходимо задать для построения базисных функций, определяются следующим образом: если номеру  $i$  узла  $P_i$  соответствуют индексы  $I, J, K$ , то, по определению,

$$\lambda_{1I} = \lambda_1(P_i), \quad \lambda_{2J} = \lambda_2(P_i), \quad \lambda_{3K} = \lambda_3(P_i).$$

Непосредственной проверкой можно убедиться, что базисная функция  $\varphi_i$  принимает значение 1 в узле с барицентрическими координатами  $\lambda_1 = \lambda_{1I}$ ,  $\lambda_2 = \lambda_{2J}$ ,  $\lambda_3 = \lambda_{3K}$  (т. е. в узле  $P_i$ ) и значение 0 во всех остальных узлах.

Другими словами, процесс построения базисной функции выглядит следующим образом.

Сначала задается набор узлов конечного элемента, соответствующий треугольной сетке (см. рис. 16.2). При этом через каждый такой узел проходят три отрезка, параллельные сторонам треугольника и соответствующие фиксированному значению одного из индексов  $i, j$  или  $k$ . Положим значение базисной функции равным единице в рассматриваемом узле и нулю — во всех остальных. По этим значениям построим три одномерных интерполяционных полинома, соответствующих трем

отрезкам, проходящим через рассматриваемую точку. Произведение этих полиномов и будет искомой базисной функцией. Таким образом, речь идет о последовательной интерполяции на треугольнике с использованием барицентрических координат.

В приведенных выражениях для базисных функций слагаемое максимального порядка имеет вид

$$\lambda_1^I \lambda_2^J \lambda_3^K.$$

Поскольку в силу способа введения индексов  $I, J, K$  всегда справедливо равенство  $I + J + K = M$ , то и порядок базисных функций равняется  $M$ . Величина  $M$ , в свою очередь, на единицу меньше количества узлов на одной стороне треугольника (см. рис. 16.2).

Отметим, что приведенные выше выражения верны не только для случая равномерной треугольной сетки в треугольнике (равномерного расположения узлов на сторонах конечного элемента).

Теперь для базисных функций второго порядка формально получим (см. рис. 16.1)

$$\begin{aligned}\varphi_i &= (2\lambda_i - 1)\lambda_i, \quad i = 1, 2, 3, \\ \varphi_4 &= 4\lambda_1\lambda_2, \quad \varphi_5 = 4\lambda_2\lambda_3, \quad \varphi_6 = 4\lambda_1\lambda_3.\end{aligned}$$

Для примера рассмотрим подробно построение базисной функции  $\varphi_1$ .

Для элемента второго порядка имеем  $M = 2$ . Узлу с номером  $i = 1$  соответствуют  $I = 2$  и  $J = K = 0$ . Таким образом, получаем

$$\varphi_1 = \frac{(\lambda_1 - \lambda_{10})(\lambda_1 - \lambda_{11})}{(\lambda_{12} - \lambda_{10})(\lambda_{12} - \lambda_{11})}.$$

Согласно данному выше определению,  $\lambda_{10}$  равно значению координаты  $\lambda_1$  всех узлов, соответствующих слою треугольной сетки при  $I = 0$ . Этому сечению принадлежит всего один узел —  $P_2$ , следовательно,  $\lambda_{10} = 0$ .

Значение  $\lambda_{11}$  равно значению координаты  $\lambda_1$  всех узлов, у которых  $I = 1$ . Таких узлов всего два —  $P_4$  и  $P_6$  (см. рис. 16.1 и 16.2). Этим узлам соответствует координата  $\lambda_1 = 1/2$ , и, следовательно,  $\lambda_{11} = 1/2$ .

Аналогично  $\lambda_{12}$  равно значению координаты  $\lambda_1$  всех узлов, у которых  $I = 2$ . Такой узел также всего один —  $P_1$ . Следовательно,  $\lambda_{12} = 1$ .

Отсюда получаем

$$\varphi_1 = \frac{(\lambda_1 - 0)(\lambda_1 - 1/2)}{(1 - 0)(1 - 1/2)} = (2\lambda_1 - 1)\lambda_1.$$

И так далее для всех оставшихся базисных функций.

Аналогично могут быть построены и функции более высоких порядков. По существу, описанный выше способ является реализацией метода последовательной двумерной интерполяции при специальном выборе системы координат.

Приведем значения интеграла, часто встречающегося при вычислениях с использованием барицентрических координат:

$$\int\limits_T \lambda_1^a \lambda_2^b \lambda_3^c dS = \frac{a!b!c!}{(a+b+c+2)!} 2\Delta.$$

При вычислении производных от базисных функций можно воспользоваться связью между декартовыми и барицентрическими координатами и правилом дифференцирования сложной функции.

Следовательно, использование барицентрических координат, с одной стороны, существенно упрощает построение и использование базисных функций и конечных элементов высокого порядка, а с другой, дает возможность применять практически без изменений алгоритм построения матрицы жесткости (алгоритм сборки).

Таким же образом с привлечением барицентрических координат может быть рассмотрен и трехмерный случай, в котором конечные элементы представляют собой тетраэдры.

То же касается большинства используемых на практике конечных элементов, например в форме тетраэдров, параллелепипедов, призм и т. д. Этот материал подробно изложен в [296–298].

## 16.6. О применении МКЭ к решению других задач

Как уже отмечалось выше, *проекционно-сеточные методы* являются очень мощным и широко распространенным аппаратом для решения задач самых разных классов.

Основной идеей методов является поиск *слабого решения задачи*, т. е. использование специальных интегральных постановок задач, в которых ослабляется условие на гладкость решения. Чаще всего такие постановки имеют смысл *интегральных законов сохранения* и могут быть получены для большинства прикладных задач.

Математический аппарат проекционных методов основан на теории абстрактных *вариационных уравнений*. При теоретическом исследовании используется не конкретный вид вариационного уравнения, а такие его свойства, как, например, линейность, *положительная определенность* или монотонность (для нелинейных задач) соответствующей билинейной формы и т. д. Конкретный вид вариационного уравнения

не важен. Именно этот факт является основой хорошо разработанной в настоящее время общей теории проекционно-сеточных методов.

Другой важной частью этой теории является то, что оценка ошибки МКЭ сводится к оценке ошибки *интерполяции* произвольной функции по системе базисных функций в пределах одного конечного элемента. Таким образом, задача построения оценки МКЭ сводится к задаче построения ошибки интерполяции, которая никак не связана с рассматриваемым уравнением и обычно является гораздо более простой задачей.

Ранее мы не рассматривали всю совокупность задач, для решения которых также возможно использование МКЭ.

К таким задачам можно отнести, в частности, нелинейные задачи самого разного происхождения, например нелинейные задачи теории упругости и теплопроводности. При их аппроксимации обычно сначала получают систему нелинейных алгебраических уравнений, которую решают тем или иным итерационным методом, например *методом Ньютона*. Другой способ заключается в построении линеаризации задачи и ее последующем решении подходящим итерационным методом. Получаемая на каждой итерации система линейных алгебраических уравнений (СЛАУ) затем аппроксимируется некоторым образом.

Другими задачами, которые решаются МКЭ, являются так называемые задачи со свободными границами. Их постановка имеет вид не вариационных уравнений, а вариационных неравенств. К ним относятся задачи гидродинамики со свободной поверхностью, контактные задачи теории упругости и другие.

Метод конечных элементов также используется для решения нестационарных задач. При этом после пространственной аппроксимации задачи с помощью МКЭ получается система обыкновенных дифференциальных уравнений (ОДУ), которая затем может решаться подходящим методом. Другой способ решения нестационарных задач заключается в построении временных аппроксимаций для дифференциальной задачи и сведении ее к последовательности стационарных задач, которые необходимо решать на каждом временном слое. Для построения пространственной аппроксимации таких задач может использоваться МКЭ.

С алгоритмической точки зрения основой для эффективной реализации МКЭ является использование матричных соотношений и алгоритма сборки матрицы жесткости. При этом важную роль играет применение квадратурных формул, точных на многочленах заданной степени, которые используют при вычислении матриц жесткости отдельных элементов.

## 16.7. Основы метода граничных элементов

**Метод граничных элементов**, как и метод конечных элементов, относится к классу *проекционно-сеточных методов*.

Его отличительной особенностью является то, что для построения *аппроксимаций* исходной задачи, определенной в пространственной области  $\Omega$ , используется другая, эквивалентная исходной, постановка задачи. Новая постановка состоит в нахождении функций, определенных только на границе  $\partial\Omega$  расчетной области.

Указанная постановка имеет вид того или иного *граничного интегрального уравнения* относительно значений решения и его производных на границе области. Метод граничных элементов относится к проекционно-сеточным методам решения этого уравнения. Метод обладает как достоинствами, так и недостатками. Достоинством метода граничных элементов является прежде всего то, что неизвестное решение задачи определено лишь на границе расчетной области. Поэтому размерность конечномерной задачи существенно меньше, чем размерность задачи со сходным шагом расчетной сетки в случае МКЭ. При этом не возникает проблем, связанных с построением пространственной сетки, требуется аппроксимировать лишь одномерную (для случая двумерной задачи) или двумерную (для случая трехмерной задачи) границу области.

Еще одним достоинством является зачастую более точная передача решения внутри расчетной области, особенно в случае, когда решение имеет особенности типа пограничных слоев, концентраторов напряжений и т. д. Это связано с тем, что в методе граничных элементов изначально решение определяется только на границе области и далее восстанавливается в любой наперед заданной внутренней точке области некоторым прямым методом с любой наперед заданной точностью. Также метод граничных элементов позволяет численно определять решение в неограниченных областях. Это одна из причин его популярности при расчете внешних задач акустики, электродинамики, сейсмологии и других, связанных с изучением распространения волн в тех или иных средах в неограниченном пространстве.

Другая распространенная область применения этого метода связана с его использованием в виде отдельной части более «крупного» алгоритма для решения сложных связанных задач, например контактных задач теории упругости или связанных задач о взаимодействии упругого твердого тела и потока жидкости.

Метод граничных элементов, как и МКЭ, позволяет проводить расчеты в областях сложной формы, а не только простой. При этом его

использование гораздо легче с алгоритмической точки зрения, так как нет необходимости выполнять триангуляцию полной пространственной области, а достаточно рассматривать только ее границу, что существенно проще.

К недостаткам можно отнести следующее. Обычно метод граничных элементов приводит к необходимости решать системы линейных алгебраических уравнений (СЛАУ) с заполненной и несимметричной матрицей. Первое связано с нелокальностью интегральных операторов, с помощью которых записываются граничные интегральные уравнения, и с характерным видом *фундаментального решения* оператора задачи, которое обычно не является функцией с конечным носителем и отлично от нуля во всей области.

Другим недостатком является то, что для построения граничных интегральных уравнений и уравнений метода граничных элементов требуется знать аналитическое выражение фундаментального решения оператора задачи. Для широкого класса задач фундаментальное решение может быть построено, например, для *уравнения Лапласа*, линейного стационарного *уравнения теплопроводности* с постоянными коэффициентами, уравнений линейной теории упругости для однородной изотропной и анизотропной сред, некоторых задач акустики (например, связанных с уравнением Гельмгольца), различных нестационарных задач и многих других.

Однако чаще всего фундаментальное решение неизвестно. Например, это касается большинства задач с переменными коэффициентами или нелинейных задач. Отметим также, что в простейших случаях, когда коэффициенты задачи являются кусочно-постоянными функциями, применение метода граничных элементов все же возможно.

В связи с этим в инженерной практике метод граничных элементов используется гораздо реже, чем МКЭ. В основном он применяется для решения относительно простых задач с постоянными или кусочно-постоянными коэффициентами.

Первые работы, посвященные методу граничных элементов, относятся к началу 1960-х годов. Несмотря на то что с тех пор метод много исследовался и интенсивно развивался, он редко (по сравнению с МКЭ) используется для решения крупномасштабных задач, возникающих в приложениях. Так, например, при использовании МКЭ число неизвестных степеней свободы доходит до нескольких миллионов, а при использовании метода граничных элементов оно, в связи с указанными выше проблемами, составляет всего несколько тысяч. При этом для формирования заполненной матрицы, возникающей при использовании метода, необходимо  $O(N^2)$  операций. Дополнительно  $O(N^3)$  операций

необходимо для обращения получившейся системы уравнений, например, прямыми методами.

Однако в середине 1980-х годов был предложен так называемый Fast Multipole Boundary Element Method, эффективность которого на несколько порядков выше, чем у обычного метода граничных элементов, и близка к  $O(N)$  [255, 260]. В настоящий момент значительная часть современных работ по методу граничных элементов посвящена именно этой его разновидности. В качестве примера задачи, успешно решенной с его помощью, можно привести расчет волокнистого композита с непосредственным учетом десятков тысяч волокон за время порядка часов.

Математической основой метода граничных элементов являются теория проекционно-сеточных методов и теория граничных интегральных уравнений. Последняя, в свою очередь, основана на математической теории потенциала (см., например, [27, 59, 170]).

Отметим, что теория потенциала является довольно «тонкой» и требует отдельного очень детального изложения. Поэтому далее в тексте при использовании результатов теории потенциала они просто констатируются в каждом конкретном случае, когда это требуется. В целом же изложение ведется на нестрогом, но довольно общем уровне, без явного указания на гладкость рассматриваемых функций, областей, границ и т. д.

Подчеркнем, что наша цель — дать общее представление о методе граничных элементов и идеях, лежащих в его основе.

Также отметим, что существуют различные варианты метода граничных элементов, отличающиеся способом построения аппроксимаций или способом построения граничных интегральных уравнений (см., например, [21]).

Рассмотрим простейшие из них, в частности, так называемый прямой метод граничных элементов, и основные идеи, лежащие в основе метода [21, 206]. Основным примером является уравнение Лапласа.

### 16.7.1. Постановка задачи.

#### Границные интегральные уравнения

Рассмотрим следующую задачу: определить функцию  $u$ , удовлетворяющую уравнению

$$Au = 0 \tag{16.28}$$

в области  $\Omega$  и граничным условиям

$$u|_{\Gamma_D} = g, \quad \delta u|_{\Gamma_N} = \mu, \tag{16.29}$$

на ее границе  $\Gamma = \partial\Omega$ . Здесь  $\Gamma_D$  — часть границы, на которой заданы граничные условия Дирихле, а на  $\Gamma_N$  заданы граничные условия Неймана,  $\Gamma_D \cup \Gamma_N = \Gamma$ ,  $\Gamma_D \cap \Gamma_N = \emptyset$ .

Оператор  $\delta$  является обобщенным оператором нормальной производной (см. 16.3.1). Обычно  $\delta u$  имеет смысл того или иного физического потока величины, распределение которой в пространстве описывается решением задачи.

В соответствии с устоявшимися обозначениями, используемыми при рассмотрении метода граничных элементов, в дальнейшем будем использовать обозначение

$$q = q(u) = \delta u.$$

Будем считать, что для оператора  $A$  задачи справедлива *формула Грина* (см. 16.3.1):

$$(Au, v)_\Omega = a(u, v) + \langle \delta u, \gamma v \rangle_\Gamma,$$

где билинейная форма  $a(\cdot, \cdot)$  является симметричной, а  $(\cdot, \cdot)_\Omega$  и  $\langle \cdot, \cdot \rangle_\Gamma$  представляют собой обычные скалярные произведения в пространствах  $L_2(\Omega)$  и  $L_2(\Gamma)$ .

Применяя повторно формулу Грина, получим следующее соотношение:

$$(Au, v)_\Omega - (u, Av)_\Omega = \langle \delta u, \gamma v \rangle_\Gamma - \langle \gamma u, \delta v \rangle_\Gamma. \quad (16.30)$$

Умножив уравнение (16.28) на произвольную функцию  $v$  и проинтегрировав результат по всей области, с использованием формулы Грина (16.30) получим следующее соотношение, верное для произвольных  $u$  и  $v$ :

$$(u, Av)_\Omega = \langle \gamma u, \delta v \rangle_\Gamma - \langle \delta u, \gamma v \rangle_\Gamma. \quad (16.31)$$

Полученное уравнение является основой для получения граничного интегрального уравнения, соответствующего задаче (16.28)–(16.29), и его аппроксимации методом граничных элементов.

С учетом граничных условий (16.29) последнее уравнение можно преобразовать к виду

$$(u, Av)_\Omega = \langle g, \delta v \rangle_{\Gamma_D} + \langle \gamma u, \delta v \rangle_{\Gamma_N} - \langle \delta u, \gamma v \rangle_{\Gamma_D} - \langle \mu, \gamma v \rangle_{\Gamma_N}. \quad (16.32)$$

Это соотношение верно для функции  $u$ , являющейся решением задачи, и произвольной функции  $v$ . Иногда это соотношение называют *обратной слабой постановкой задачи* [21].

Граничное интегральное уравнение получается как следствие обратной слабой постановки при специальном выборе пространства *пробных функций*  $v$ .

Дальнейшее рассмотрение требует введения понятия фундаментального решения.

**Фундаментальным решением** оператора  $A$  называют функцию  $u^*$ , удовлетворяющую уравнению [27]

$$Au^* = \delta_\xi(x), \quad (16.33)$$

где  $\delta_\xi(x)$  —  $\delta$ -функция Дирака, связанная с точкой  $\xi \in \Omega$ . Далее всюду точку  $\xi$ , являющуюся носителем  $\delta$ -функции, будем считать параметром, т. е. при вычислении тех или иных производных либо интегралов, вычисления всегда проводятся относительно переменной  $x$ .

В соответствии с определением фундаментальное решение не единственно, а определено с точностью до произвольной функции, являющейся решением уравнения (не краевой задачи) (16.28).

Для получения граничного интегрального уравнения в качестве функции  $v$  в соотношении (16.32) обычно используют фундаментальное решение оператора уравнения, т. е.  $v = u^*(\xi, x)$ .

В соответствии со свойствами  $\delta$ -функции из (16.31) получим

$$u(\xi) = \langle \gamma u, \delta u^* \rangle_\Gamma - \langle \delta u, \gamma u^* \rangle_\Gamma, \quad (16.34)$$

где  $\xi$  лежит строго внутри области  $\Omega$ , а не на ее границе  $\Gamma$ .

Если точка  $\xi \in \Gamma$ , то предыдущее выражение обычно имеет вид

$$c(\xi)u(\xi) = \langle \gamma u, \delta u^* \rangle_\Gamma - \langle \delta u, \gamma u^* \rangle_\Gamma, \quad (16.35)$$

где коэффициент  $c(\xi)$  появляется в силу того, что первое слагаемое в правой части уравнения (16.34) обычно не является непрерывной функцией параметра  $\xi$  в области и на ее границе. Вид этой функции определяется конкретной задачей и пока не рассматривается.

Отметим также, что формально интегралы в правой части уравнения (16.35) записаны как скалярные произведения в пространстве  $L_2(\Gamma)$ . Однако подынтегральные выражения и множители в этом скалярном произведении не принадлежат пространству  $L_2(\Gamma)$ . Это связано с тем, что фундаментальное решение задачи обычно имеет особенность в точке — носителе  $\delta$ -функции в правой части уравнения (16.33).

Более того, значение первого интеграла в (16.35)

$$\langle \gamma u, \delta u^* \rangle_\Gamma = - \int_{\Gamma} u \frac{\partial u^*(\xi, x)}{\partial \vec{n}_x} dS_x$$

понимается в смысле главного значения по Коши.

Укажем также, что в теории потенциала первое слагаемое правой части (16.35) называют потенциалом двойного слоя с плотностью  $u$ , а второе — потенциалом простого слоя с плотностью  $\delta u$  [27, 59, 170].

Для получения граничного интегрального уравнения необходимо учесть граничные условия (16.29), накладываемые на решение задачи. Для этого разобьем в уравнениях (16.34) и (16.35) интегрирование по полной границе  $\Gamma$  на интегрирование по ее частям  $\Gamma_D$  и  $\Gamma_N$ .

Тогда с учетом граничных условий (16.29) получим для случая  $\xi \in \Omega$ :

$$u(\xi) = \langle g, \delta u^*(\xi, x) \rangle_{\Gamma_D} + \langle \gamma u, \delta u^*(\xi, x) \rangle_{\Gamma_N} - \\ - \langle \delta u, \gamma u^*(\xi, x) \rangle_{\Gamma_D} - \langle \mu, \gamma u^*(\xi, x) \rangle_{\Gamma_N},$$

для случая  $\xi \in \Gamma$ :

$$c(\xi)u(\xi) = \langle g, \delta u^*(\xi, x) \rangle_{\Gamma_D} + \langle \gamma u, \delta u^*(\xi, x) \rangle_{\Gamma_N} - \\ - \langle \delta u, \gamma u^*(\xi, x) \rangle_{\Gamma_D} - \langle \mu, \gamma u^*(\xi, x) \rangle_{\Gamma_N}. \quad (16.36)$$

Уравнение (16.36) является *граничным интегральным уравнением* и связывает значения неизвестной функции и ее производных на границе.

При этом неизвестными в уравнении являются значения функции на той части границы, где заданы граничные условия второго рода, и значения потока решения на той части границы, где заданы граничные условия первого рода.

Уравнение (16.36) является основой так называемого *прямого метода граничных элементов*.

Если решается задача Дирихле и на всей границе заданы граничные условия первого рода, то уравнение (16.36) является *уравнением Фредгольма первого рода* относительно потока решения на границе; если на всей границе заданы граничные условия второго рода, то уравнение (16.36) является *уравнением Фредгольма второго рода* относительно значения решения задачи на границе области.

Если граничные значения решения и его потока известны, то решение во внутренних точках может быть восстановлено с помощью соотношения (16.34).

Отметим, что возможны и иные формы (см., например, [21]) написания граничного интегрального уравнения, соответствующего задаче (16.28), (16.29).

**Пример 16.4.** Рассмотрим краевую задачу для уравнения Лапласа в области  $\Omega$  следующего вида:

$$\begin{aligned} -\Delta u &= 0 \text{ в } \Omega; \\ u|_{\Gamma_D} &= g, \quad \delta u|_{\Gamma_N} = \mu. \end{aligned}$$

Обычная формула Грина для оператора данной задачи рассмотрена в примере 16.2. Формула Грина (16.30) для этого случая имеет вид

$$\int_{\Omega} (-\Delta uv + u\Delta v) d\Omega = - \int_{\Gamma} \left( \frac{\partial u}{\partial \vec{n}} v - u \frac{\partial v}{\partial \vec{n}} \right) dS,$$

где  $\vec{n}$  — вектор единичной внешней нормали к границе  $\Gamma$  области  $\Omega$ ,  $\delta u \equiv -\frac{\partial u}{\partial \vec{n}}$ .

Фундаментальное решение оператора Лапласа имеет вид (см., например, [27])

$$u^*(\xi, x) = -\frac{1}{2\pi} \ln \frac{1}{|x - \xi|}$$

в двумерном случае ( $n = 2$ ) и

$$u^*(\xi, x) = \frac{1}{4\pi} \frac{1}{|x - \xi|}$$

в трехмерном случае ( $n = 3$ ). Здесь  $|x| = \|x\|$  — длина вектора  $x$ .

Для простоты далее будем рассматривать только двумерный случай.

В этом случае уравнение (16.34) при  $\xi \in \Omega$  примет вид

$$u(\xi) + \int_{\Gamma} u(x) \frac{\partial u^*(\xi, x)}{\partial \vec{n}_x} dS_x = \int_{\Gamma} \frac{\partial u}{\partial \vec{n}_x} u^*(\xi, x) dS_x$$

или, если  $\xi \in \Gamma$ ,

$$c(\xi)u(\xi) + \int_{\Gamma} u(x) \frac{\partial u^*(\xi, x)}{\partial \vec{n}_x} dS_x = \int_{\Gamma} \frac{\partial u}{\partial \vec{n}_x} u^*(\xi, x) dS_x, \quad (16.37)$$

где функция  $c(\xi) = \frac{\pi + \alpha(\xi)}{2\pi}$ ,  $\alpha(\xi)$  — внутренний угол границы в точке  $\xi$ .

Если граница является гладкой кривой с непрерывным вектором внешней нормали, т. е. не содержит угловых точек, то

$$\alpha(\xi) = 0, \quad c(\xi) = 1/2.$$

Для уравнения Лапласа значение  $c(\xi)$  определяется выражением

$$c(\xi) = - \int_{\Gamma} \frac{\partial u^*(\xi, x)}{\partial \vec{n}_x} dS_x.$$

Подставляя известные значения краевых условий, из (16.36) получим граничное интегральное уравнение

$$\begin{aligned} c(\xi)u(\xi) + \int_{\Gamma_D} g(x) \frac{\partial u^*(\xi, x)}{\partial \vec{n}_x} dS_x + \int_{\Gamma_N} u(x) \frac{\partial u^*(\xi, x)}{\partial \vec{n}_x} dS_x = \\ = \int_{\Gamma_D} \frac{\partial u}{\partial \vec{n}_x} u^*(\xi, x) dS_x - \int_{\Gamma_N} \mu(x) u^*(\xi, x) dS_x. \quad \# \end{aligned}$$

При построении далее аппроксимаций задачи с алгоритмической точки зрения удобно использовать не непосредственно граничное интегральное уравнение (16.36), а соотношение (16.35) или (16.37), которое будем записывать в следующем виде:

$$c(\xi)u(\xi) = - \int_{\Gamma} G(\xi, x)q(x) dS_x + \int_{\Gamma} F(\xi, x)u(x) dS_x, \quad (16.38)$$

где

$$G(\xi, x) = u^*(\xi, x); \quad q(x) = \delta u; \quad F(\xi, x) = \delta u^*(\xi, x) = \delta G(x, \xi).$$

**Пример 16.5.** Для частного случая двумерного уравнения Лапласа, рассмотренного ранее,

$$G(\xi, y) = -\frac{1}{2\pi} \ln \frac{1}{|\xi - x|} = -\frac{1}{2\pi} \ln \frac{1}{r}, \quad F(\xi, x) = -\frac{\partial G(\xi, x)}{\partial \vec{n}_x} = -\frac{1}{2\pi r} \frac{\partial r}{\partial \vec{n}},$$

где  $r = |x - y|$ .  $\#$

Вид уравнения (16.38) характерен для граничного интегрального уравнения, используемого в прямом методе граничных элементов. Аналогичный вид имеет, например, соответствующее уравнение для трехмерного уравнения Лапласа, для уравнений двумерной и трехмерной упругости и т. д.

В дальнейшем мы будем рассматривать это уравнение без учета конкретного вида функций  $F$  и  $G$ .

### 16.7.2. Аппроксимации метода граничных элементов

Метод граничных элементов относится к проекционно-сеточным методам решения соответствующего граничного интегрального уравнения.

Традиционно (см., например, [21]) с алгоритмической точки зрения он состоит из следующих основных этапов.

1. Задается разбиение границы  $\Gamma$  расчетной области на *границные элементы*. В пределах каждого такого элемента искомое решение и его нормальная производная (поток) аппроксимируются по той или иной системе базисных функций, связанных с граничной сеткой.

2. Для получения СЛАУ и определения решения чаще всего применяется метод коллокаций. Обычно на этом этапе используют уравнение (16.38), т. е. граничные условия на данном этапе не задаются.

3. Путем наложения граничных условий получается СЛАУ, решение которой дает значения неизвестного решения. Затем при необходимости восстанавливаются значения решения и его производных во внутренних точках расчетной области.

Рассмотрим эти этапы подробнее на простом примере.

Возьмем простейший случай: будем считать, что неизвестные значения решения и потока являются постоянными в пределах одного граничного элемента, т. е. для приближения решения будем использовать кусочно-постоянные базисные функции.

Они представляют собой функции минимальной гладкости, которые можно использовать для решения задачи. В рассмотренные выше граничные интегральные уравнения в качестве неизвестных входят одновременно сама функция и поток на границе, которые являются независимыми переменными.

Пусть граница области  $\Gamma$  представлена в виде объединения граничных элементов:

$$\Gamma = \bigcup_{i=1}^N \Gamma_i.$$

Граничная сетка представляет собой разбиение границы на элементарные ячейки  $\Gamma_i$ , например на отрезки (в двумерном случае) или треугольники (в трехмерном случае) (рис. 16.3).

В общем случае для аппроксимации границы и значений решения и потоков на ней могут использоваться произвольные одномерные или двумерные конечные элементы, как в обычном МКЭ.

Будем считать, что общее количество граничных элементов равно  $N$ , причем  $N = N_D + N_N$ , где  $N_D$  — число граничных элементов, образующих  $\Gamma_D$ ;  $N_N$  — число граничных элементов, образующих  $\Gamma_N$ .

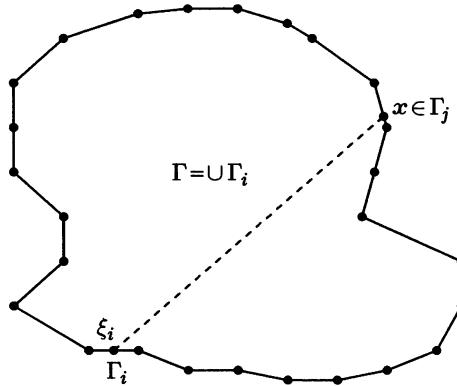


Рис. 16.3

В силу того, что используются кусочно-постоянные *базисные функции*, количество узловых значений решения, связанных с границей  $\Gamma$ , также равняется  $N$ .

Из соотношения (16.38) получим

$$c(\xi_i)u(\xi_i) = - \int_{\Gamma} G(\xi_i, x)q(x) dS_x + \int_{\Gamma} F(\xi_i, x)u(x) dS_x.$$

Отсюда, разбив интегрирование по всей границе в правой части последнего равенства на интегрирование по отдельным граничным элементам, получим

$$c(\xi_i)u(\xi_i) = - \sum_{j=1}^N \int_{\Gamma_j} G(\xi_i, x)q(x) dS_x + \sum_{j=1}^N \int_{\Gamma_j} F(\xi_i, x)u(x) dS_x.$$

Это соотношение является точным (при условии, что граница расчетной области аппроксимируется точно), и оно справедливо, если  $u$  является решением задачи.

Пусть  $u_i = u(\xi_i)$  — значение приближенного решения в центре  $\xi_i$  граничного элемента  $\Gamma_i$ ,  $i = \overline{1, N}$ . Аналогично, пусть  $q_i = q(\xi_i)$ ,  $i = \overline{1, N}$ , является значением плотности потока решения в точке  $\xi_i$  границы.

Заменив решение и его поток на кусочно-постоянный интерполант, с учетом того, что в пределах одного граничного элемента неизвестное решение и его плотность потока постоянны, получим

$$c_i u_i - \sum_{j=1}^N \left( \int_{\Gamma_j} F(\xi_i, x) dS_x \right) u_j = - \sum_{j=1}^N \left( \int_{\Gamma_j} G(\xi_i, x) dS_x \right) q_j,$$

где  $c_i = c(\xi_i)$ .

Эти соотношения можно записать в матричном виде следующим образом:

$$H_h u_h = -G_h q_h, \quad (16.39)$$

где

$$u_h = (u_1, u_2, \dots, u_N)^T, \quad q_h = (q_1, q_2, \dots, q_N)^T \quad —$$

векторы узловых значений решения и потоков, а матрицы  $H_h$  и  $G_h$  имеют размер  $N \times N$ ; их компоненты можно записать в виде

$$H_h = \text{diag}(c_1, c_2, \dots, c_N) - F_h,$$

$$F_{h,ij} = \int_{\Gamma_j} F(\xi_i, x) dS_x,$$

$$G_{h,ij} = \int_{\Gamma_j} G(\xi_i, x) dS_x.$$

Здесь  $\text{diag}(c_1, c_2, \dots, c_N)$  представляет собой диагональную матрицу с указанными коэффициентами на диагонали.

Отметим, что для случая кусочно-постоянных базисных и пробных функций значения коэффициентов  $c_i = c(\xi_i)$  соответствуют случаю гладкой границы [21].

Например, для уравнения Лапласа необходимо положить  $c_i = 1/2$  для всех значений  $i = \overline{1, N}$ .

Уравнение (16.39) связывает между собой значения решения и его потока во всех точках границы. При этом  $N_D$  компонент вектора решения  $u_h$  и  $N_N$  компонент вектора потока  $q_h$  известны из граничных условий. Поэтому матричное уравнение (16.39) представляет собой СЛАУ для определения  $N_N$  неизвестных компонент вектора  $u_h$ , относящихся к части  $\Gamma_N$  границы  $\Gamma$ , и  $N_D$  компонент вектора  $q_h$ , относящихся к части  $\Gamma_D$  границы  $\Gamma$ .

Перенеся все неизвестные в левую часть уравнения, придем к СЛАУ

$$Ax = b, \quad (16.40)$$

где вектор  $x$  размерности  $N$  состоит из неизвестных компонент векторов  $u_h$  и  $q_h$ .

Опишем алгоритм более подробно: будем считать, что граничные элементы (и соответствующие узлы) пронумерованы последовательно, причем сначала идут  $N_D$  узлов с номерами 1, 2, ...,  $N_D$ , соответствующих части  $\Gamma_D$  границы, а потом  $N_N$  узлов с номерами  $N_D + 1, N_D + 2, \dots, N$ , соответствующих части  $\Gamma_N$  границы.

Тогда система уравнений (16.39) может быть записана в следующем блочном виде:

$$(H_{h,D}, H_{h,N}) \begin{pmatrix} u_{h,D} \\ u_{h,N} \end{pmatrix} = -(G_{h,D}, G_{h,N}) \begin{pmatrix} q_{h,D} \\ q_{h,N} \end{pmatrix},$$

где матрицы  $H_{h,D}$  и  $G_{h,D}$  имеют размер  $N \times N_D$ , а матрицы  $H_{h,N}$  и  $G_{h,N}$  имеют размер  $N \times N_N$ .

Здесь векторы  $u_h$  и  $q_h$  также записаны в следующем блочном виде:

$$u_h = (u_{h,D}, u_{h,N})^T, \quad q_h = (q_{h,D}, q_{h,N})^T,$$

где блоки  $u_{h,D}$  и  $q_{h,D}$  длиной  $N_D$  соответствуют узлам на части границы  $\Gamma_D$ , а блоки  $u_{h,N}$  и  $q_{h,N}$  длиной  $N_N$  соответствуют узлам на части  $\Gamma_N$ .

Выполняя матричное умножение в последнем уравнении, получим

$$H_{h,D}u_{h,D} + H_{h,N}u_{h,N} = -G_{h,D}q_{h,D} - G_{h,N}q_{h,N}.$$

Теперь с учетом того, что на части  $\Gamma_D$  границы заданы граничные условия первого рода, т. е. известны значения элементов вектора  $u_{h,D}$ , а на части  $\Gamma_N$  границы известны значения потока (известны значения элементов вектора  $q_{h,N}$ ), получаем следующее уравнение:

$$H_{h,N}u_{h,N} + G_{h,D}q_{h,D} = -G_{h,N}\mu_h - H_{h,D}g_h,$$

где  $\mu_h$  и  $g_h$  — подходящие аппроксимации функций  $\mu$  и  $g$  из (16.29).

Записывая полученные уравнения снова в блочном виде, получим СЛАУ (16.40), в которой

$$A = (H_{h,N}, G_{h,D}), \quad x = (u_{h,N}, q_{h,D})^T, \quad b = -G_{h,N}\mu_h - H_{h,D}g_h.$$

Отсюда видно, что матрица  $A$  в (16.40) не является, вообще говоря, симметричной.

Решение этой системы дает приближенные значения решения задачи и его потока на границе области.

В дальнейшем решение задачи во внутренних точках области может быть восстановлено в соответствии с уравнением (16.34).

Рассмотренная схема построения аппроксимаций является достаточно общей и справедлива для очень широкого класса задач, граничное интегральное уравнение для которых формально имеет вид (16.38).

Заметим, что решение интегрального уравнения Фредгольма первого рода может потребовать применения *регуляризации*.

### 16.7.3. Алгоритмические аспекты

В целом технология метода граничных элементов совпадает с такой для обычного МКЭ. Возможным также является использование алгоритма сборки *матрицы жесткости*. Вычисление ее коэффициентов тоже сводится к интегрированию соответствующих соотношений по отдельным граничным элементам. Поэтому здесь мы рассмотрим подробнее только процедуру вычисления коэффициентов матрицы жесткости, т. е. величин  $H_{h,ij}$  и  $G_{h,ij}$ . Сложность заключается в том, что в рассматриваемом случае подынтегральное выражение имеет особенность, а соответствующий интеграл понимается в смысле главного значения по Коши.

Подынтегральные выражения для вычисления коэффициентов  $H_{ij}$ ,  $G_{ij}$  при  $i \neq j$  не содержат особенность. Для их вычисления можно использовать те же самые квадратурные формулы, например гауссовские, что и в случае МКЭ.

При вычислении коэффициентов  $H_{h,ii}$  и  $G_{h,ii}$ , у которых подынтегральные выражения имеют особенность, необходимо использовать специальные *квадратурные формулы*, учитывающие эту особенность (см., например, [21]).

Основные принципы построения таких формул рассматриваются в **4.5**.

Отметим также, что в простейших случаях, например двумерных и трехмерных задач для уравнения Лапласа, указанные интегралы могут быть вычислены аналитически.

## ЛИТЕРАТУРА

1. Абакумов М.В., Мухин С.И., Попов Ю.П., Попов С.Б. Особенности схемы Roe при расчёте задач обтекания // Препр. ИПМ им. М.В. Келдыша РАН. 1996. № 46. 30 с.
2. Абрамович М., Стиган И. Справочник по специальным функциям с формулами, графиками и математическими таблицами. М.: Наука, 1979. 832 с.
3. Агафонов С.А., Герман А.Д., Муратова Т.В. Дифференциальные уравнения. М.: Изд-во МГТУ им. Н.Э. Баумана, 1997. 336 с.
4. Альберт Д., Нильсон Э., Уолли Д. Теория сплайнов и ее приложения. М.: Мир, 1972. 318 с.
5. Александрикова Т.А., Галанин М.П. Нелинейная монотонизация схемы К.И. Бабенко для численного решения квазилинейного уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 2003. № 62. 35 с.
6. Амосов А.А., Дубинский Ю.А., Копченова Н.В. Вычислительные методы для инженеров. М.: Высш. шк., 1994. 544 с.
7. Астраганцев Г.П. Об одном итерационном методе решения сеточных эллиптических задач // ЖВМ и МФ. 1971. Т. 11, № 2. С. 439–448.
8. Бабенко К.И. Основы численного анализа. Москва; Ижевск: НИЦ «Регулярная и хаотическая динамика», 2002. 848 с.
9. Бабенко К.И., Воскресенский Г.П. Численный метод расчета пространственного обтекания тел сверхзвуковым потоком газа // ЖВМ и МФ. 1961. Т. 1, № 6. С. 1051–1060.
10. Бабенко К.И., Воскресенский Г.П., Любимов А.Н., Русанов В.В. Пространственное обтекание гладких тел идеальным газом. М: Наука, 1964. 505 с.
11. Бабушка И., Витасек Э., Прагер М. Численные процессы решения дифференциальных уравнений. М.: Мир, 1969. 368 с.
12. Бахвалов Н.С. Численные методы. М.: Наука, 1973. 632 с.
13. Бахвалов Н.С. О сходимости одного релаксационного метода при естественных ограничениях на эллиптический оператор // ЖВМ и МФ. 1966. Т. 6, № 5. С. 861–883.
14. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. М.: Бином. Лаборатория знаний, 2007. 636 с.
15. Бахвалов Н.С., Лапин А.В., Чижонков Е.В. Численные методы в задачах и упражнениях. М.: Высш. шк., 2000. 190 с.
16. Белоцерковский О.М. Численное моделирование в механике сплошных сред. М.: Наука: Физматлит, 1984. 520 с.
17. Березин И.С., Жидков Н.П. Методы вычислений. Т. 1. М.: Наука, 1966. 632 с.
18. Березин И.С., Жидков Н.П. Методы вычислений. Т. 2. М.: Физматгиз, 1962. 640 с.

19. Борис Дж.П., Бук Д.Л. Решение уравнений непрерывности методом коррекции потоков // Сб.: Вычислительные методы в физике. Управляемый термоядерный синтез. М: Мир, 1980. С. 92–141.
20. Боровин Г.К., Комаров М.М., Ярошевский В.С. Ошибки-ловушки при программировании на Фортране. М.: Наука: Физматлит, 1987. 144 с.
21. Бреббия К., Теллес Ж., Броубел Л. Методы граничных элементов. М.: Мир, 1987. 524 с.
22. Вазов В., Форсайт Дж. Разностные методы решения дифференциальных уравнений. М.: Иностр. литература, 1963. 487 с.
23. Варга Р. Функциональный анализ и теория аппроксимации в численном анализе. М.: Мир, 1974. 126 с.
24. Васильев Ф.П. Численные методы решения экстремальных задач. М.: Наука, 1988. 552 с.
25. Вержбицкий В.М. Основы численных методов. М.: Вышш. шк., 2002. 840 с.
26. Верлань А.Ф., Сизиков В.С. Методы решения интегральных уравнений с программами для ЭВМ. Киев: Наук. думка, 1978. 292 с.
27. Владимиров В.С. Уравнения математической физики. М.: Наука, 1981. 512 с.
28. Власова Е.А., Зарубин В.С., Кувыркин Г.Н. Приближенные методы математической физики. М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. 700 с.
29. Воеводин В.В. Вычислительные основы линейной алгебры. М.: Наука, 1977. 304 с.
30. Воеводин В.В. Математические модели и методы в параллельных процессах. М.: Наука: Физматлит, 1986. 296 с.
31. Воеводин В.В., Кузнецов Ю.А. Матрицы и вычисления. М.: Наука, 1984. 320 с.
32. Волков Е.А. Численные методы. М.: Наука, 1982. 254 с.
33. Вязников К.В., Жорняк Н.С., Тишкун В.Ф., Фаворский А.П. О методе построения квазимонотонных разностных схем повышенного порядка аппроксимации // Препр. ИПМ им. М.В. Келдыша АН СССР. 1989. № 141. 19 с.
34. Вязников К.В., Тишкун В.Ф., Фаворский А.П. Квазимонотонные разностные схемы для уравнений газодинамики // Препр. ИПМ им. М.В. Келдыша АН СССР. 1987. № 175. 24 с.
35. Вязников К.В., Тишкун В.Ф., Фаворский А.П. Построение монотонных разностных схем повышенного порядка аппроксимации для систем уравнений гиперболического типа // Математическое моделирование. 1989. Т. 1, № 5. С. 95–120.
36. Вязников К.В., Тишкун В.Ф., Фаворский А.П., Шашков М.Ю. Квазимонотонные разностные схемы повышенного порядка точности // Препр. ИПМ им. М.В. Келдыша АН СССР. 1987. № 36. 27 с.
37. Галанин М.П., Грищенко Е.В., Савенков Е.Б., Токарева С.А. Применение RKDG метода для численного решения задач газовой динамики // Препр. ИПМ им. М.В. Келдыша РАН. 2006. № 52. 31 с.

38. Галанин М.П., Еленина Т.Г. Нелинейная монотонизация разностных схем для линейного уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 1999. № 44. 30 с.
39. Галанин М.П., Еленина Т.Г. Сравнительный анализ разностных схем для линейного уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 1998. № 52. 33 с.
40. Галанин М.П., Еленина Т.Г. Тестирование разностных схем для линейного уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 1999. № 40. 42 с.
41. Галанин М.П., Еленина Т.Г. Нелинейная монотонизация схемы К.И. Бабенко («квадрат») для уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 2002. № 4. 26 с.
42. Галанин М.П., Савенков Е.Б., Токарева С.А. Применение разрывного метода Галеркина для численного решения квазилинейного уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 2005. № 105. 35 с.
43. Галанин М.П., Щеглов И.А. Разработка и реализация алгоритмов трехмерной триангуляции сложных пространственных областей: прямые методы // Препр. ИПМ им. М.В. Келдыша РАН. 2006. № 10. 32 с.
44. Галанин М.П., Щеглов И.А. Разработка и реализация алгоритмов трехмерной триангуляции сложных пространственных областей: итерационные методы // Препр. ИПМ им. М.В. Келдыша РАН. 2006. № 9. 32 с.
45. Галлager R. Метод конечных элементов. Основы. М.: Мир, 1984. 428 с.
46. Гантмахер Ф.Р. Теория матриц. М.: Наука: Физматлит, 2004. 264 с.
47. Гласко В.Б. Обратные задачи математической физики. М.: Изд-во МГУ. 1984. 112 с.
48. Гловински Р., Лионс Ж.Л., Тремольер Р. Численное исследование вариационных неравенств. М.: Мир, 1979. 576 с.
49. Годунов С.К. Разностный метод численного расчета разрывных решений гидродинамики // Математический сборник. 1959. 47 (89). С. 271–306.
50. Годунов С.К. Уравнения математической физики. М: Наука: Физматлит, 1979. 392 с.
51. Годунов С.К., Забродин А.В., Прокопов Г.П. Разностная схема для двумерных нестационарных задач газовой динамики и расчет обтекания с отошедшей ударной волной // ЖВМ и МФ. 1961. Т. 1, № 6. С. 1020–1050.
52. Годунов С.К., Забродин А.В., Иванов М.Я., Крайко А.Н. и др. Численные решения многомерных задач газовой динамики. М: Наука, 1976. 400 с.
53. Годунов С.К., Рябенький В.С. Разностные схемы. М.: Наука, 1977. 440 с.
54. Головизнин В.М., Карабасов С.А. Метод прыжкового переноса для численного решения гиперболических уравнений. Точный алгоритм для моделирования конвекции на эйлеровых сетках // Препр. ИБРАЭ, 2002. № IBRAE-2000-04. 40 с.
55. Головизнин В.М., Карабасов С.А. Нелинейная коррекция схемы «Кабаре» // Математическое моделирование. 1998. Т. 10, № 12. С. 107–123.

56. Головизнин В.М., Самарский А.А. Разностная аппроксимация конвективного переноса с пространственным расщеплением временной производной // Математическое моделирование. 1998. Т. 10, № 1. С. 86–100.
57. Голуб Дж., Ван Лоун Ч. Матричные вычисления. М.: Мир, 1999. 548 с.
58. Гольдин В.Я., Калиткин Н.Н., Тишова Т.В. Нелинейные разностные схемы для гиперболических уравнений // ЖВМ и МФ. 1965. Т. 5, № 5. С. 938–944.
59. Гюнтер Н. Теория потенциала и ее применение к основным задачам математической физики. М.: ГИТТЛ, 1953. 416 с.
60. Деклу Ж. Метод конечных элементов. М.: Мир, 1976. 96 с.
61. Демидович Б.П., Марон И.А. Основы вычислительной математики. СПб.: Лань, 2007. 672 с.
62. Демидович Б.П., Марон И.А., Шувалова Э.З. Численные методы анализа. СПб.: Лань, 2008. 400 с.
63. Джордж А., Лю Дж. Численное решение больших разреженных систем уравнений. М.: Мир, 1984. 333 с.
64. Днестровский Ю.Н., Костомаров Д.П. Математическое моделирование плазмы. М: Наука, 1982. 319 с.
65. Дробышевич В.И., Дымников В.П., Ривин Г.С. Задачи по вычислительной математике. М.: Наука: Физматлит, 1980. 144 с.
66. Дьяконов Е.Г. Минимизация вычислительной работы. Асимптотики оптимальные алгоритмы для эллиптических задач. М.: Наука: Физматлит, 1989. 272 с.
67. Дюво Г., Лионс Ж.-Л. Неравенства в механике и физике. М.: Мир, 1980. 384 с.
68. Еленина Т.Г. Решение нелинейной монотонизированной разностной схемы К.И. Бабенко («квадрат») // Препр. ИПМ им. М.В. Келдыша РАН. 2002. № 75. 34 с.
69. Зарубин В.С. Математическое моделирование в технике. М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. 496 с.
70. Зарубин В.С., Кувыркин Г.Н. Математические модели механики и электродинамики сплошной среды. М.: Изд-во МГТУ им. Н.Э. Баумана, 2008. 512 с.
71. Зарубин В.С., Селиванов В.В. Вариационные и численные методы механики сплошной среды. М.: Изд-во МГТУ им. Н.Э. Баумана, 1993. 360 с.
72. Зенкевич О. Метод конечных элементов в технике. М.: Мир, 1975. 271 с.
73. Зенкевич О., Морган К. Конечные элементы и аппроксимация. М.: Мир, 1986. 318 с.
74. Иванов А.А., Тишкин В.Ф., Фаворский А.П., Яцук А.Н. Построение квазимонотонной схемы повышенного порядка аппроксимации для уравнения переноса // Препр. ИПМ им. М.В. Келдыша РАН. 1993. № 69. 25 с.
75. Иванов В.В. Методы вычислений на ЭВМ: Справ. пособие. Киев: Наук. думка, 1986. 584 с.
76. Иванов В.К., Васин В.В., Танана В.П. Теория линейных некорректных задач и ее приложения. М.: Наука. 1978. 206 с.

77. Икрамов Х.Д. Численные методы для симметричных линейных систем. М.: Наука, 1988. 159 с.
78. Ильин В.П. Методы и технологии конечных элементов. Новосибирск: Изд. ИВМ и МГ СО РАН, 2007. 371 с.
79. Ильин В.П., Кузнецов Ю.И. Трехдиагональные матрицы и их приложения. М.: Наука, 1985. 208 с.
80. Калиткин Н.Н. Численные методы. М.: Наука, 1978. 512 с.
81. Калиткин Н.Н., Альшин А.Б., Альшина Е.А., Рогов Б.В. Вычисления на квазивременных сетках. М.: Физматлит, 2005. 224 с.
82. Канатников А.Н., Крищенко А.П. Линейная алгебра. М.: Изд-во МГТУ им. Н.Э. Баумана, 1999. 336 с.
83. Канторович Л.В., Крылов В.И. Приближенные методы высшего анализа. М.: Физматгиз, 1962. 708 с.
84. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. М.: Мир, 1998. 575 с.
85. Киндерлерер Д., Стампакъя Г. Введение в вариационные неравенства и их приложения. М.: Мир, 1983. 256 с.
86. Колган В.П. Применение принципа минимаксных значений производных к построению конечно-разностных схем для расчета разрывных решений газовой динамики // Уч. записки ЦАГИ. 1972. 3, 68. С. 68–77.
87. Коллатц Л. Функциональный анализ и вычислительная математика. М.: Мир, 1969. 448 с.
88. Коллатц Л. Задачи на собственные значения. М.: Наука, 1968. 503 с.
89. Коллатц Л., Альбрехт Ю. Задачи по прикладной математике. М.: Мир, 1978. 168 с.
90. Копченова Н.В., Марон Н.А. Вычислительная математика в примерах и задачах. СПб.: Лань, 2008. 368 с.
91. Коробов Н.М. Теоретико-числовые методы в приближенном анализе. М.: МЦНМО, 2004. 295 с.
92. Костомаров Д.П., Фаворский А.П. Вводные лекции по численным методам. М.: Логос, 2004. 184 с.
93. Красносельский М.А., Вайникко Г.М., Забрейко П.П., Рутицкий Я.Б., Степенко В.Я. Приближенное решение операторных уравнений. М.: Наука, 1969. 456 с.
94. Кузнецов О.А. Численное исследование схемы Роу с модификацией Эйнфельдта для уравнений газовой динамики // Препр. ИПМ им. М.В. Келдыша РАН. 1998. № 43. 44 с.
95. Крылов В.И., Бобков В.В., Монастырный П.И. Вычислительные методы. Т. 1. М.: Физматгиз, 1976. 304 с.
96. Крылов В.И., Бобков В.В., Монастырный П.И. Вычислительные методы. Т. 2. М.: Физматгиз, 1977. 399 с.
97. Куликовский А.Г., Любимов Г.А. Магнитная гидродинамика. М.: Логос, 2005. 328 с.
98. Куликовский А.Г., Погорелов Н.В., Семенов А.Ю. Математические вопросы численного решения гиперболических систем уравнений. М.: Физматлит, 2001. 608 с.

99. *Курант Р.* Уравнения с частными производными. М.: Мир, 1964. 830 с.
100. *Курант Р., Гильберт Д.* Методы математической физики. Т. 1. М.-Л.: Гостехиздат, 1951. 476 с.
101. *Лаврентьев М.М., Романов В.Г., Шишатский С.П.* Некорректные задачи математической физики и анализа. М.: Наука, 1980. 286 с.
102. *Ладыженская О.А.* Краевые задачи математической физики. М.: Наука, 1988. 286 с.
103. *Лебедев В.И.* Функциональный анализ и вычислительная математика. М.: Физматлит, 2000. 296 с.
104. *Лионс Ж.-Л.* Некоторые методы решения нелинейных краевых задач. М.: Мир, 1972. 588 с.
105. *Локуциевский О.М., Гавриков М.Б.* Начала численного анализа. М.: ТОО «Янус», 1995. 581 с.
106. *Магомедов К.М., Холодов А.С.* Сеточно-характеристические численные методы. М.: Наука, 1988. 288 с.
107. *Мак-Кракен Д., Дорн У.* Численные методы и программирование на ФОРТРАНЕ. М.: Мир, 1977. 534 с.
108. *Мартинсон Л.К., Малов Ю.И.* Дифференциальные уравнения математической физики. М.: Изд-во МГТУ им. Н.Э. Баумана, 1996. 368 с.
109. *Мартиценко С.И.* Универсальная многосеточная технология для численного решения краевых задач на структурированных сетках // Вычислительные методы и программирование. 2000. Т. 1, № 1. С. 85–104.
110. *Марчук Г.И.* Методы вычислительной математики. М.: Наука, 1989. 608 с.
111. *Марчук Г.И.* Методы расщепления. М.: Наука, 1988. 263 с.
112. *Марчук Г.И., Агошков В.И.* Введение в проекционно-сеточные методы. М.: Наука, 1981. 416 с.
113. *Марчук Г.И., Шайдуров В.В.* Повышение точности решений разностных схем. М.: Наука, 1979. 319 с.
114. *Митчелл Э., Уайт Р.* Метод конечных элементов для уравнений с частными производными. М.: Мир, 1981. 216 с.
115. *Михлин С.Г.* Численная реализация вариационных методов. М.: Наука, 1966. 432 с.
116. *Михлин С.Г.* Вариационные методы в математической физике. М.: Наука, 1970. 512 с.
117. *Михлин С.Г., Смолицкий Х.Л.* Приближенные методы решения дифференциальных и интегральных уравнений. М.: Наука, 1983. 383 с.
118. *Морозов В.А.* Регулярные методы решения некорректно поставленных задач. М.: Наука, 1987. 240 с.
119. *Мухин С.И., Попов С.Б., Попов Ю.П.* Исследование свойств разностных схем высокого порядка точности для гиперболических уравнений // Препр. ИПМ АН СССР. 1984. № 36. 18 с.
120. *Никифоров А.Ф., Уваров В.Б.* Специальные функции математической физики. М.: Наука: Физматлит, 1984. 344 с.
121. *Никольский С.М.* Квадратурные формулы. М.: Наука, 1988. 255 с.

122. Норри Д., де Фриз Ж. Введение в метод конечных элементов. М.: Мир, 1981. 304 с.
123. Обэн Ж.-П. Приближенное решение эллиптических краевых задач. М.: Мир, 1977. 384 с.
124. Овсянников Л.В. Лекции по основам газовой динамики. М.: Наука, 1981. 368 с.
125. Оден Дж. Конечные элементы в нелинейной механике сплошных сред. М.: Мир, 1976. 464 с.
126. Ольшанский М.А. Лекции и упражнения по многосеточным методам. М.: Физматлит, 2005. 168 с.
127. Ортега Дж. Введение в параллельные и векторные методы решения линейных систем. М.: Мир, 1991. 367 с.
128. Ортега Дж., Пул У. Введение в численные методы решения дифференциальных уравнений. М.: Наука, 1986. 288 с.
129. Ортега Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. М.: Мир, 1975. 560 с.
130. Патанкар С. Численные методы решения задач теплообмена и динамики жидкости. М.: Энергоатомиздат, 1984. 152 с.
131. Писанецки С. Технология разреженных матриц. М.: Мир, 1988. 411 с.
132. Поттер Д. Вычислительные методы в физике. М.: Мир, 1975. 392 с.
133. Ректорис К. Вариационные методы в математической физике и технике. М.: Мир, 1985. 590 с.
134. Рихтмайер Р., Мортон К. Разностные методы решения краевых задач. М.: Мир, 1972. 418 с.
135. Рождественский Б.Л., Яненко Н.Н. Системы квазилинейных уравнений и их приложение к газовой динамике. М.: Наука, 1978. 688 с.
136. Русанов В.В. Разностная схема 3-го порядка точности для сквозного расчета разрывных решений // ДАН СССР. 1968. Т. 180, № 6. С. 1303–1305.
137. Рябенький В.С. Введение в вычислительную математику. М.: Наука, 1994. 336 с.
138. Рябенький В.С. Метод разностных потенциалов и его приложения. М.: Физматлит, 2002. 496 с.
139. Самарский А.А. Теория разностных схем. М.: Наука, 1989. 616 с.
140. Самарский А.А. Введение в численные методы. СПб.: Лань, 2005. 288 с.
141. Самарский А.А., Андреев В.Б. Разностные методы для эллиптических уравнений. М.: Наука, 1976. 352 с.
142. Самарский А.А., Вабищевич П.Н. Нелинейные монотонные схемы для уравнения переноса // ДАН. 1998. Т. 361, № 1. С. 21–23.
143. Самарский А.А., Вабищевич П.Н., Самарская Е.А. Задачи и упражнения по численным методам. М.: Едиториал УРСС, 2003. 208 с.
144. Самарский А.А., Вабищевич П.Н. Численные методы решения задач конвекции-диффузии. М.: Едиториал УРСС, 2003. 246 с.
145. Самарский А.А., Вабищевич П.Н. Численные методы решения обратных задач математической физики. М.: Едиториал УРСС, 2004. 480 с.

146. Самарский А.А., Вабищевич П.Н. Вычислительная теплопередача. М.: Едиториал УРСС, 2003. 782 с.
147. Самарский А.А., Вабищевич П.Н. Аддитивные схемы для задач математической физики. М.: Едиториал УРСС, 2001. 312 с.
148. Самарский А.А., Галактионов В.А., Курдюмов С.П., Михайлов А.П. Режимы с обострением в задачах для квазилинейных параболических уравнений. М.: Наука: Физматлит, 1987. 480 с.
149. Самарский А.А., Гулин А.В. Численные методы. М.: Наука: Физматлит, 1989. 416 с.
150. Самарский А.А., Гулин А.В. Устойчивость разностных схем. М.: URSS, 2005. 384 с.
151. Самарский А.А., Гулин А.В. Численные методы математической физики. М.: Научный мир, 2000. 316 с.
152. Самарский А.А., Лазаров Р.Д., Макаров В.Л. Разностные схемы для дифференциальных уравнений с обобщенными решениями. М.: Выш. шк., 1987. 296 с.
153. Самарский А.А., Михайлов А.П. Математическое моделирование. Идеи. Методы. Примеры. М.: Наука: Физматлит, 1997. 320 с.
154. Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. М.: Наука, 1978. 592 с.
155. Самарский А.А., Попов Ю.П. Разностные методы решения задач газовой динамики. М.: Наука, 1980. 352 с.
156. Сборник задач по методам вычислений / Под ред. П.И. Монастырного. М.: Наука: Физматлит, 1994. 320 с.
157. Сегерлинд Л. Применение метода конечных элементов. М.: Мир, 1979. 392 с.
158. Сильвестер П., Феррари Р. Метод конечных элементов для радиоинженеров и инженеров-электриков. М.: Мир, 1986. 229 с.
159. Скворцов А.В. Обзор алгоритмов построения триангуляции Делоне // Вычислительные методы и программирование. 2002. № 3. С. 14–39.
160. Скворцов А.В. Алгоритмы построения триангуляции с ограничениями // Вычислительные методы и программирование. 2002. № 3. С. 82–92.
161. Соболев С.Л. Введение в теорию кубатурных формул. М.: Наука, 1974. 808 с.
162. Соболь И.М. Численные методы Монте-Карло. М.: Наука, 1973. 311 с.
163. Станкова Е.Н., Затевахин М.А. Многосеточные методы: введение в стандартные методы ([http://www.csa.ru/skif/kurs\\_3/multigrid](http://www.csa.ru/skif/kurs_3/multigrid)).
164. Степочкин С.Б., Субботин Ю.Н. Сплайны в вычислительной математике. М.: Наука, 1976. 248 с.
165. Стрэнг Г., Фикс Дж. Теория метода конечных элементов. М.: Мир, 1977. 349 с.
166. Съярле Ф. Метод конечных элементов для эллиптических задач. М.: Мир, 1980. 512 с.
167. Темам Р. Уравнения Навье — Стокса: Теория и численный анализ. М.: Мир, 1981. 408 с.

168. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1986. 288 с.
169. Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г. Численные методы решения некорректных задач. М.: Наука, 1990. 232 с.
170. Тихонов А.Н., Самарский А.А. Уравнения математической физики. М.: Изд-во МГУ: Наука, 2004. 798 с.
171. Тихонов А.Н., Самарский А.А. О сходимости разностных схем в классе разрывных коэффициентов // ДАН СССР. 1959. Т. 124, № 3. С. 529–532.
172. Турчак Л.И. Основы численных методов. М.: Наука: Физматлит, 1987. 320 с.
173. Тьюарсон Р. Разреженные матрицы. М.: Мир, 1977. 189 с.
174. Уилкинсон Дж.Х. Алгебраическая проблема собственных значений. М.: Наука, 1970. 564 с.
175. Фаддеев Д.К., Фаддеева В.Н. Вычислительные методы линейной алгебры. СПб.: Лань, 2002. 736 с.
176. Федоренко Р.П. Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 528 с.
177. Федоренко Р.П. Приближенное решение задач оптимального управления. М.: Наука, 1978. 488 с.
178. Федоренко Р.П. Применение разностных схем высокой точности для численного решения гиперболических уравнений // ЖВМ и МФ. 1962. Т. 2, № 6. С. 1122–1128.
179. Федоренко Р.П. Релаксационный метод решения разностных эллиптических уравнений // ЖВМ и МФ. 1961. Т. 1, № 5. С. 922–927.
180. Федоренко Р.П. Скорость сходимости одного итерационного метода // ЖВМ и МФ. 1964. Т. 4, № 3. С. 227–235.
181. Флетчер К. Вычислительные методы в динамике жидкости. Т. 1. Основные положения и общие методы. М.: Мир, 1991. 502 с.
182. Флетчер К. Вычислительные методы в динамике жидкости. Т. 2. Методы расчета различных течений. М.: Мир, 1991. 552 с.
183. Формалев В.Ф., Ревизников Д.Л. Численные методы. М.: Физматлит, 2004. 400 с.
184. Хайрер Э., Ваннер Г., Нерсетт С.П. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990. 512 с.
185. Хайрер Э., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. М.: Мир, 1999. 685 с.
186. Хейгеман Л., Янг Д. Прикладные итерационные методы. М.: Мир, 1986. 446 с.
187. Хемминг Р.В. Численные методы. М.: Наука, 1972. 400 с.
188. Хорн Р., Джонсон Ч. Матричный анализ. М.: Мир, 1989. 655 с.
189. Шайдуров В.В. Многосеточные методы конечных элементов. М.: Наука, 1989. 288 с.
190. Шокин Ю.И., Яненко Н.Н. Метод дифференциального приближения. Новосибирск: Наука, 1985. 364 с.

191. Эстербю О., Златев З. Прямые методы для разреженных матриц. М.: Мир, 1987. 118 с.
192. Яненко Н.Н. Метод дробных шагов решения многомерных задач математической физики. Новосибирск: Наука, 1967. 196 с.
193. Alberty J., Carstensen C., Funken S.A., Klose R. Matlab implementation of the finite element method in elasticity // Computing. 2000. 69, 3. P. 239–263.
194. Alberty J., Carstensen C., Funken S. Remarks around 50 lines of Matlab: Short finite element implementation // Numerical Algorithms. 1999. 20, 2–3. P. 117–137.
195. Babushka I., Rheinboldt W.C. A-posteriori Error Estimates for Finite Element Method // Int. J. Numer. Meth. Eng. 1978. V. 12. P. 1597–1615.
196. Baker T.J. Automatic Mesh Generation for Complex Three-Dimensional Regions Using a Constrained Delaunay Triangulation // Engineering With Computers. Springer-Verlag. 1989. № 5. P. 161–175.
197. Beam R.M., Warming R.F. An implicit finite-difference algorithm for hyperbolic system in conservation-law form // J. Comp. Phys. 1976. V. 22. P. 87–110.
198. Bern M., Eppstein D. Mesh Generation and Optimal Triangulation // Computing in Euclidean Geometry, World Scientific Publishing Co. 1995. P. 23–90.
199. Blandford D.K., Bleloch G., Cardoze D., Kadow C. Compact Representations of Simplicial Meshes In Two and Three Dimensions // Proceedings of 12th International Meshing Roundtable, Sandia National Laboratories. 2003. P. 135–146.
200. Boris J.P., Book D.L. Flux corrected transport, I, SHASTA // J. Comp. Phys. 1973. V. 11. P. 33–69.
201. Boris J.P., Book D.L., Hain K. Flux-corrected transport: generalization of method // J. Comp. Physics. 1975. V. 18. P. 248–283.
202. Bornemann F.A., Deuflhard P. The cascadic multigrid method for elliptic problems // Numer. Math. 1996. V. 75. P. 135–152.
203. Borouchaki H., Lo S.H. Fast Delaunay Triangulation In Three Dimensions // Computer Methods In Applied Mechanics And Engineering, Elsevier. 1995. V. 128. P. 153–167.
204. Brandt A. Multi-level adaptive solutions to boundary value problems // Math. Comput. 1977. V. 31. P. 333–390.
205. Brandt A. Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems // Proc. 3rd Int. Conf. on Numerical Methods in Fluid Mechanics. / H. Cabannes and R. Temam (eds) / Lecture Notes in Physics. 1973. V. 18. P. 82–89.
206. Brebbia C.A., Dominguez J. Boundary Elements. An Introductory Course. WIT Press, 1997. 313 p.
207. Brenner S.C., Scott L.R. The Mathematical Theory of Finite Element Methods. Series: Texts in Applied Mathematics, Vol. 15. 2nd ed. Springer, 2002. 361 p.
208. Buratynski E.K. A Three-Dimensional Unstructured Mesh Generator for Arbitrary Internal Boundaries // Numerical Grid Generation in Computational Fluid Mechanics, Pineridge Press. 1988. P. 621–631.

209. *Chakravarthy S.R., Osher S.* A new class of high accuracy TVD schemes for hyperbolic conservation laws // AIAA Pap. 1985. N 85-0363. 11 p.
210. *Cavalcanti P.R., Mello U.T.* Three-Dimensional Constrained Delaunay Triangulation: A Minimalist Approach // Proceedings of the 8th International Meshing Roundtable. 1999. P. 119–129.
211. *Cockburn B.* An introduction to the discontinuous Galerkin method for convection-dominated problems // SIAM J. Sci. Comput. 2001. V. 16. P. 173–261.
212. *Colella P.* Glimm's method for gas dynamics // SIAM J. Sci. Stat. Comput. 1982. V. 3. P. 76–110.
213. *Colella P., Glaz H.* Efficient algorithms for the solution of the Riemann problem for real gases // Lawrence Berkeley Laboratory Report LBL-15776, 1983.
214. *Colella P., Glaz H.M.* Efficient solution algorithms for the Riemann problem for real gases // J. Comp. Phys. 1985. V. 59, № 2. P. 264–289.
215. *Courant R., Friedrichs K.O., Lewy H.* Über die partiellen Differenzengleichungen der mathematischen Physik // Math. Ann. 1928. V. 100, № 1–2. P. 32–74.
216. *Courant R., Isaacson E., Rees M.* On the solution of nonlinear hyperbolic differential equations by finite differences // Commun. Pure Appl. Math. 1952. V. 5, № 3. P. 243–255.
217. *De Zeeuw D.L.* A quadtree-based adaptively-refined Cartesian-grid algorithm for solution of Euler equations // PhD thesis, University of Michigan, 1993.
218. *Tamal Dey K., Sugihara K., Bajaj C.L.* Delaunay Triangulations In Three Dimensions With Finite Precision Arithmetic // Computer Aided Geometric Design, North-Holland. 1992. № 9. P. 457–470.
219. *Dendy Jr. J. E.* Black box multigrid // J. Comput. Phys. 1982. V. 48. P. 366–386.
220. *Dendy Jr. J.E.* Black box multigrid for systems // Appl. Math. Comp. 1986. V. 19. P. 57–74.
221. *Djidjev H.N.* Force-Directed Methods For Smoothing Unstructured Triangular And Tetrahedral Meshes // Proceedings of 9th International Meshing Roundtable, Sandia National Laboratories. October 2000. P. 395–406.
222. *Durbeck L.* Evaporation: A Technique For Visualizing Mesh Quality // Proceedings of 8th International Meshing Roundtable, South Lake Tahoe, CA, U.S.A. October 1999. P. 259–265.
223. *Einfeldt B.* On Godunov-type methods for gas dynamics // SIAM J. Numer. Anal. 1988. V. 25, № 2. P. 294–318.
224. *Emery A.E.* An evaluation of several differencing methods for inviscid fluid flow problems // J. Comp. Phys. 1968. V. 2. P. 306–331.
225. *Field D.A.* Laplacian Smoothing And Delaunay Triangulations // Communications in Applied Numerical Methods. 1988. V. 4. P. 709–712.
226. *Frey P.J., Borouchaki H., George P.-L.* Delaunay Tetrahedralization Using an Advancing-Front Approach // Proceedings of 5th International Meshing Roundtable, Sandia National Laboratories. October 1996. P. 31–46.
227. *Freitag L.A., Ollivier-Gooch C.* A Comparison of Tetrahedral Mesh Improvement Techniques // Proceedings of 5th International Meshing Roundtable, Sandia National Laboratories. October. 1996. P. 87–106.

228. *Freitag L.A., Ollivier-Gooch C.* Tetrahedral Mesh Improvement Using Swapping and Smoothing // International Journal for Numerical Methods in Engineering. 1995. V. 40. P. 3979–4002.
229. *Freitag L.A., Ollivier-Gooch C.* The Effect Of Mesh Quality On Solution Efficiency // Proceedings of 6th International Meshing Roundtable, Sandia National Laboratories. October. 1997. P. 249.
230. *Friedrichs R.O.* Symmetric hyperbolic linear differential equations // Commun. Pure Appl. Math. 1954. V. 7, № 2. P. 345–392.
231. *Fromm J.E.* A method for reducing dispersion in convective difference schemes // J. Comp. Phys. 1968. V. 3. P. 176–189.
232. *George P.L.* TET MESHING: Construction, Optimization and Adaptation // Proceedings of 8th International Meshing Roundtable. 1999. P. 133–141.
233. *Golias N.A., Tsiboukis T.D.* An Approach to Refining Three-Dimensional Tetrahedral Meshes Based on Delaunay Transformations // International Journal for Numerical Methods in Engineering, John Wiley. 1994. № 37. P. 793–812.
234. *Goloviznin V.M., Karabasov S.A.* Non-linear correction of “Cabaret” scheme // Second International Conference “Finite Difference Methods: Theory and Application”, Proceedings. V. 2. 1998. Minsk. P. 7–18.
235. *Hackbusch W.* Multi-grid methods and applications. Berlin: Springer, 1985.
236. *Hackbusch W.* Robust multi-grid methods, the frequency decomposition multi-grid algorithm // Proc. 4th GAMM-seminar, Kiel, 1988, W. Hackbusch (ed.) (Notes on Numerical Fluid Mechanics 123) Vieweg, Braunschweig. 1989. P. 96–104.
237. *Harten A.* High resolution schemes for hyperbolic conservation laws // J. Comp. Phys. 1983. V. 49, № 3. P. 357–393.
238. *Harten A., Lax P.D., Van Leer B.* On upstream differencing and Godunov-type schemes for hyperbolic conservation laws // SIAM Rev. 1983. V. 25, № 1. P. 35–61.
239. *Harten A., Hyman J.M.* Self-adjusting grid methods for one-dimensional hyperbolic conservation laws // J. Comp. Phys. 1983. V. 50. P. 235–269.
240. *Hazlewood C.* Approximating Constrained Tetrahedralizations // Computer Aided Geometric Design. 1993. V. 10. P. 67–87.
241. *Hlavaváček I., Haslinger J., Nečas J., Lovíšek J.*, Solution of variational inequalities in mechanics // Series: Applied Mathematical Sciences. V. 66. Springer, 1988. 275 p.
242. *Hugoniot H.* Sur la propagation du mouvement dans les corps et spécialement dans les gaz parfaits // Journal de l'école polytechnique. 1889. V. 58, P. 1–125.
243. *Joe B.* Delaunay Triangular Meshes in Convex polygons // SIAM J. Sci. Stat. Comput. 1986, V. 7. P. 514–539.
244. *Joe B.* Construction Of Three-Dimensional Delaunay Triangulations Using Local Transformations // Computer Aided Geometric Design. 1991. V. 8. P. 123–142.
245. *Joe B.* Construction of Three-Dimensional Improved-Quality Triangulations Using Local Transformations // Siam J. Sci. Comput. 1995. V. 16. P. 1292–1307.

246. *Khokhlov A.M.* Fully threaded tree for adaptive refinement fluid dynamics simulation // J. Comp. Phys. 1998. V. 143, № 2. P. 519–543.
247. *Kikuchi N., Oden J.T.* Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods. SIAM: Reissue edition, 1995. 495 p.
248. *Kimoto P.A., Chernoff D.F.* Convergence properties of finite-difference hydrodynamics schemes in the presence of shocks // Astrophys. J. Suppl. 1995. V. 96. P. 627.
249. *Laursen, T.* Computational Contact and Impact Mechanics: Fundamentals of Modelling of Interfacial Phenomena in Nonlinear Finite Element Analysis. Springer, 2006. 454 p.
250. *Lax P.D.* Weak solutions of nonlinear hyperbolic equations and their numerical computations // Commun. Pure Appl. Math. 1954. V. 7, № 1. P. 159–193.
251. *Lax P.D., Wendroff B.* Difference schemes for hyperbolic equations with high order of accuracy // Commun. Pure Appl. Math. 1964. V. 17, № 3. P. 381–398.
252. *LeVeque R.J.* Numerical Methods for conservation laws. Basel: Birkhäuser–Verlag, 1990. 214 p.
253. *Lewis R.W., Zheng Yao, Gethin D.T.* Three-Dimensional Unstructured Mesh Generation: Part 3. Volume Meshes // Computer Methods In Applied Mechanics And Engineering, Elsevier. 1996. V. 134. P. 285–310.
254. *Liu A., Joe B.* On The Shape Of Tetrahedra From Bisection // Mathematics of Computation. 1994. V. 63, № 207. P. 141–154.
255. *Liu Y.J., Nishimura N.* The fast multiple boundary element method for potential problem: A tutorial // Engineering Analysis with Boundary Elements. 2006. V. 30. P. 371–381.
256. *Lo S.H.* Volume Discretization into Tetrahedra I. Verification and Orientation of Boundary Surfaces // Computers and Structures, Pergamon Press, 1991. V. 39, № 5. P. 493–500.
257. *Lo S.H.* Volume Discretization into Tetrahedra II. 3D Triangulation by Advancing Front Approach // Computers and Structures, Pergamon. 1991. V. 39, № 5. P. 501–511.
258. *Lohner R.* Generation Of Three-Dimensional Unstructured Grids By The Advancing Front Method // Proceedings of the 26th AIAA Aerospace Sciences Meeting, Reno, Nevada, 1988.
259. *von Neumann J., Richtmyer R.D.* A method for numerical calculation of hydrodynamic shocks // J. Appl. Phys. 1950. V. 21, № 3. P. 232–237.
260. *Nishimura N.* Fast multiple accelerated boundary integral equation methods // Appl. Mech. Rev. 2002. V. 55. P. 299–324.
261. *Osher S., Chakravarthy S.* High resolution schemes and the entropy conditions // SIAM J. Num. Analysis. 1984. V. 21. P. 955–984.
262. *Osher S., Solomon F.* Upwind difference schemes for hyperbolic systems of conservation laws // Math. Comput. 1982. V. 38, № 158. P. 339–374.
263. *Owen S.J.* A Survey of Unstructured Mesh Generation Technology // Proceedings of 7th International Meshing Roundtable, Dearborn, MI. 1998. P. 239–269.

264. *Parthasarathy V.N., Graichen C.M., Hathaway A.F.* A Comparison of Tetrahedron Quality Measures // Finite Elements in Analysis and Design, Elsevier. 1993. № 15. P. 255–261.
265. *Pirzadeh S.* Unstructured Viscous Grid Generation by Advancing-Layers Method // AIAA-93-3453-CP, AIAA. 1993. P. 420–434.
266. *Rajan V.T.* Optimality of Delaunay Triangulation in  $R^d$  // Proc. 7th ACM Symp. Comp. Geometry. 1991. P. 357–363.
267. *Rankine W.J.M.* On the thermodynamic theory of waves of finite longitudinal disturbances // Trans. Roy. Soc. of London. 1870. V. 160. P. 277–286.
268. *Rassineux A.* Generation and Optimization of Tetrahedral Meshes by Advancing Front Technique // International Journal for Numerical Methods in Engineering, Wiley. 1998. V. 41. P. 651–674.,
269. *Rebay S.* Efficient Unstructured Mesh Generation by Means of Delaunay Triangulation and Bowyer-Watson Algorithm // Journal Of Computational Physics. 1993. V. 106. P. 125–138.
270. *Rivara M.-C.* Selective Refinement/Derefinement Algorithms For Sequences Of Nested Triangulations // International Journal for Numerical Methods in Engineering. 1998. № 28. P. 2889–2906.
271. *Rivara M.-C., Levin C.* A 3D Refinement Algorithm Suitable For Adaptive And Multigrid Techniques // Communications in Applied Numerical Methods. 1998. № 8. P. 281–290.
272. *Roe P.L.* Approximate Riemann solvers, parameter vectors, and difference schemes // J. Comp. Phys. 1981. V. 34, № 2. P. 357–372.
273. *Roe P.L.* Some contribution to the modelling of discontinuous flows // in Proc. 1983 AMS-SIAM Summer Seminar on Large Scale Computing in Fluid Mechanics, Lectures in Applied Math., Philadelphia: SIAM. 1983. V. 22. P. 163–193.
274. *Roe P.L.* Characteristic-based schemes for the Euler equations // Ann. Rev. Fluid Mech. 1986. V. 18. P. 337–365.
275. *Roe P.L.* The use of the Riemann problem in finite-difference schemes // Lect. Notes Phys. 1980. V. 141. P. 354–359.
276. *Ruppert J.* A Delaunay refinement algorithm for quality 2-dimensional mesh generation // Journal of Algorithms. 1995. № 18. P. 548–585.
277. *Sevono E.* Towards an adaptive advancing front method // Proceedings, 6th International Meshing Roundtable, Sandia National Laboratories. October 1997. P. 349–360.
278. *Shephard M.S., Georges M.K.* Three-Dimensional Mesh Generation by Finite Octree Technique // International Journal for Numerical Methods in Engineering. 1991. V. 32. P. 709–749.
279. *Shephard M.S., Guerinoni F., Flaherty J.E., Ludwig R.A., Baehmann P.L.* Finite octree mesh generation for automated adaptive 3D Flow Analysis // Numerical grid generation in computational fluid mechanics. Miami, 1988. P. 709–718.
280. *Shimada K., Gossard D.C.* Bubble Mesh: Automated Triangular Meshing of Non-manifold Geometry by Sphere Packing // Proceedings of 3rd Symposium on Solid Modeling and Applications. 1995. P. 409–419.

281. *Shimada K., Yamada A., Itoh T.* Anisotropic Triangular Meshing of Parametric Surfaces via Close Packing of Ellipsoidal Bubbles // Proceedings of 6th International Meshing Roundtable. 1997. P. 375–390.
282. *Shu Chi-Wang.* Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws // ICASE Report N 97–65, NASA/CR-97-206253, 1997.
283. *Suresh A., Liou M.S.* Osher's scheme for real gases // AIAA J. 1991. V. 29. P. 920–926.
284. *Sweby P.K.* High resolution schemes using flux limiters for hyperbolic conservation laws // SIAM J. Numer. Anal. 1984. V. 21, № 5. P. 995–1011.
285. *Toro E.F.* Riemann Solvers and Numerical Methods for Fluid Dynamics. A Practical Introduction. Berlin: Springer. 1997. 624 p.
286. *Van Leer B.* Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme // J. Comp. Phys. 1974. V. 14, № 4. P. 361–370.
287. *Van Leer B.* Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection // J. Comp. Phys. 1977. V. 23, № 3. P. 276–239.
288. *Van Leer B.* Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's Method // J. Comp. Phys. 1979. V. 32, № 1. P. 101–136.
289. *Warming R.F., Beam R.M.* Upwind second order difference schemes and applications in aerodynamic flows // AIAA J. 1976. V. 14. P. 1241–1249.
290. *Watson D.F.* Computing the Delaunay Tessellation with Application to Voronoi Polytopes // The Computer Journal. 1981. V. 24 (2). P. 167–172.
291. *Wesseling P.* An Introduction to Multigrid Methods. Chichester: Wiley, 1991.
292. *Woodward P., Colella P.* The numerical simulations of two-dimensional fluid flow with strong shocks // J. Comp. Phys. 1984. V. 54, № 1. P. 115–173.
293. *Wriggers, P.* Computational Contact Mechanics. Springer, 2006. 518 p.
294. *Yee H.C.* Upwind and symmetric shock-capturing schemes // NASA Ames Technical Memorandum 89464, 1987. V. 31.
295. *Yerry M.A., Shephard M.S.* Three-Dimensional Mesh Generation by Modified Octree Technique // International Journal for Numerical Methods in Engineering. 1984. V. 20. P. 1965–1990.
296. *Zienkiewicz O.C., Taylor R.L.* The Finite Element Method, Vol. 1: The Basis. London: Butterworth Heinemann. 736 p.
297. *Zienkiewicz O.C., Taylor R.L.* The Finite Element Method, Vol. 2: Solid and Structural Mechanics. London: Butterworth Heinemann. 480 p.
298. *Zienkiewicz O.C., Taylor, R.L.* The Finite Element Method, Vol. 3: Fluid Dynamics. London: Butterworth Heinemann. 320 p.

# ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

## **А**лгоритм насыщающий 127

- ненасыщающий 127
  - регуляризующий 331
  - сборки матрицы жесткости 522
  - устойчивый 24
- Анализатор гладкости 479
- Аппроксимация абсолютная 222
- безусловная 222
  - локальная 222
  - разностная оператора Лапласа в криволинейных координатах 311
  - разностного метода,  $p$ -го порядка 167
  - уравнения 167
  - суммарная 304
  - условная 222

## **Б**арицентр треугольника 524

## **В**ариация общая 465

- Волна бегущая 259
- плоская 282
  - разрежения 442
  - ударная 416
- Выделение особенностей аддитивное 154
- мультипликативное 154
- Вязкость аппроксимационная (численная) 283

## **Г**армоника 244

## **Д**ефект сплайна 131

- Дискретизация 339
- Дисперсия 283
- Дифференцирование численное 162, 334
- Дополнение ортогональное 34

## **З**адача корректно поставленная 202

- по Тихонову 328
- Коши 165, 202
- для уравнения Лапласа 335
- переноса 264
- с финитными начальными данными 267
- краевая 202
- начально-краевая для уравнения переноса 264
- (смешанная) 202
- некорректная (по Адамару) 326
- некорректно поставленная 202

## **З**адача о распаде разрыва (задача Римана) 467

- условно-корректная 328
  - Штурма — Лиувилля 319
  - разностная 230
  - в двумерном случае 293
- Закон Гука 507
- сохранения 456
  - импульса 456
  - массы 456
  - энергии 456
- Значение собственное 38, 39

## **И**нварианты Римана 465, 469

- Интерполянт 107
- Интерполяция 103
- кусочно-линейная 103
  - полиномиальная 106
  - последовательная 133
  - тригонометрическая 127
- Итерации внешние 96
- внутренние 96
  - сглаживающие 389

## **К**вазирешение 330

- Константа Лебега 120
- Координаты барицентрические 359, 542
- Критерий Адамара 46, 58
- Делоне 342
  - Куранта 266, 465
- Круги Гершгорина 47

## **Л**емма Гершгорина 47

- Лакса — Мильграма 513
- Сея 513

## **М**ажоранта 237

- Мантисса 19
- Матрица вырожденная 46
- Гильберта 55
  - Грама 125, 126
  - жесткости 521
  - конечного элемента 524
  - невырожденная 46
  - разреженная 83
- Метод Адамса 176
- баланса 215
  - безусловно устойчивый 180
  - Бубнова — Галеркина 505
  - вариационно-сеточный 514
  - взвешенных невязок 502

Метод Галеркина 198, 309, 310  
 -- для решения интегральных уравнений 325  
 -- обобщенный 505  
 -- разрывный 444  
 - гармоник 245  
 - Гаусса 47  
 -- с выбором главного элемента 51  
 - гибридный 100  
 - Гира 185  
 - граничной коррекции 342, 361  
 - граничных элементов 548  
 -- прямой 553  
 - деления отрезка пополам («вилки») 87  
 - дифференциального приближения 280  
 - дробления 355  
 - замены ядра вырожденным 323  
 - Зейделя 67  
 -- нелинейный 99  
 - интегро-интерполяционный 215  
 - интерполяционный 94  
 - исчерпывания 342, 378  
 - итерационный вариационного типа 71  
 -- двухслойный 64  
 -- нестационарный 65  
 -- неявный 65  
 -- односторонний 64  
 -- стационарный 65  
 -- трехслойный 77  
 -- триангуляции области 340  
 -- явный 65  
 - квадратного корня 62  
 - коллокаций 504  
 - конечных элементов 518  
 - линейный  $m$ -шаговый 175  
 - минимальных невязок 75  
 -- погрешностей 76  
 -- поправок 76  
 - многосеточный каскадный 406  
 -- классический 390, 392  
 - наименьших квадратов 105, 196, 505, 518  
 - на основе шаблонов 342  
 - неопределенных коэффициентов 217  
 - неявный, решения задачи Коши 175  
 - Ньютона 91, 94, 99  
 -- модифицированный 99  
 -- с параметром 99  
 - отображения (изопараметрический) 342  
 - парабол 95  
 - Пикара 98  
 -- модифицированный 99  
 - полностью неявный, решения задачи Коши 175  
 - полунонеявный решения задачи Коши 184

Метод последовательного интегрирования 156  
 - последовательных приближений решения интегральных уравнений 321  
 - предиктор-корректор 169  
 - прогонки 56  
 -- встречной 58  
 -- левой 58  
 -- матричной 58, 60  
 -- потоковой 58  
 -- правой 58  
 -- пятидиагональной 58  
 -- циклической 58  
 - проекционно-сеточный 218  
 - простой итерации 65  
 - прямой 47  
 - разделения переменных 230  
 - разностной аппроксимации 214  
 - расщепления 450  
 - регуляризации 82, 332  
 -- А.Н. Тихонова 82  
 -- СЛАУ 81  
 - релаксации 98  
 -- верхней 68  
 -- нелинейный 90  
 - Ритца 197, 304, 305, 515  
 - Ричардсона с чебышевскими параметрами 65  
 - Рунге — Кутты 169  
 -- двухшаговый 169  
 --  $m$ -шаговый 170  
 -- Ромберга численного дифференцирования 160  
 - скорейшего спуска 76  
 - сопряженных градиентов 78  
 -- направлений 77  
 -- невязок 78  
 -- погрешностей 78  
 -- поправок 78  
 - спектральный 505  
 - стрельбы 188  
 - типа «простой итерации» 88  
 - триангуляции области прямой 340  
 - условно устойчивый 180  
 - установления 297  
 - Холецкого 62  
 - хорд (секущих) 93, 94  
 - штрафа 527  
 - Эйлера 166, 168  
 -- неявный 180  
 - экстраполяции Ричардсона 142  
 - энергетических неравенств 256  
 - явный, решения задачи Коши 175  
 - Якоби 67  
 -- нелинейный 99  
 -- ячеек 155  
 -  $A(\alpha)$ -устойчивый 184  
 -  $A$ -устойчивый 184

- Метод  $p$ -го порядка точности 167  
 Многообразие линейное 32  
 Множество корректности 328  
 – практической эквивалентности 329  
 Модель математическая I, 16
- Н**евязка 73, 389  
 Неравенство Коши — Буняковского 34  
 – типа вложения 297  
 Норма 33  
 – вектора евклидова (шаровая, сферическая) 42  
 – кубическая 40  
 – октаэдрическая 41  
 – евклидова 34  
 – интерполяционного полинома 120  
 – матрицы 44  
 – максимальная 44  
 – согласованная 43  
 – спектральная 45  
 – оператора 35  
 – подчиненная 35  
 – эквивалентная 33  
 – энергетическая 506  
 Нормы согласованные 208
- О**бласть значений оператора 35  
 – определения оператора 35  
 – устойчивости 182  
 Объект технический I, 16  
 Ограничитель потока 479  
 Оператор кососимметричный 37  
 – линейный 35  
 – неотрицательный 37  
 – непрерывный 35  
 – нормальный 37  
 – обратный 36  
 – ограниченный 35  
 – положительно определенный 37  
 – положительный 37  
 – проектирования 207  
 – пролонгации 392, 399  
 – регуляризующий 329  
 – самосопряженный 36  
 – сжимающий 97  
 – сопряженный 36  
 – сужения 392, 396  
 Операция ограничения функции 102  
 Определитель Вандермонда 107  
 Основание системы счисления 19
- П**акет волновой 283  
 Параметр регуляризации 82  
 Переменные лагранжевые 457  
 – эйлеровы 457  
 Погрешность абсолютная 21  
 – алгоритма 24
- Погрешность аппроксимации (невязка)  
 – правой части 212  
 – разностного оператора 212  
 – разностной задачи 212  
 – вычислений 19  
 – вычисления функции 26  
 – линейная абсолютная 26  
 – предельная абсолютная 27  
 – неустранимая 19  
 – относительная 21  
 – разностной схемы 212  
 – численного метода 19, 167  
 Подпространство 32  
 Полином базисный 108  
 – тригонометрический 128  
 – интерполяционный 107  
 – в форме Лагранжа 108  
 – – – Ньютона 111, 112  
 – остаточный член 108  
 – Эрмита 114  
 – (многочлен) Фурье 126  
 – наименее уклоняющийся от нуля 117  
 – обобщенный 124  
 – Тейлора 117  
 – Чебышева 66, 118  
 Поправка 73, 389  
 Последовательность фундаментальная 33  
 Постановка задачи обобщенная 509  
 – слабая 509  
 – – – обратная 551  
 Поток тепловой 216  
 – физический 445  
 – численный 445  
 Правило Рунге для квадратурных формул 142  
 – – – оценки ошибки метода решения ОДУ 174  
 Предобуславливатель 64  
 Преобразование изопараметрическое 359  
 – Лапласа 335  
 Приближение дифференциальное разностной схемы первое 280  
 – наилучшее равномерное 121  
 Признак Неймана 244  
 – спектральный 244  
 Принцип замороженных коэффициентов 258  
 – максимума 228  
 – сжимающих отображений 97  
 Проблема собственных значений полная 79  
 – – – частичная (ограниченная) 80  
 Произведение скалярное 34, 211  
 – – – энергетическое 506  
 Производная разностная вторая 162  
 – – левая 160

- Производная разностная правая 160  
 -- центральная 160  
 Пролонгация 398  
 Пространство гильбертово 34  
 - евклидово 34  
 - линейное 31  
 -- нормированное 33  
 -- подпространство 32  
 -- полное 33  
 --  $n$ -мерное 32  
 - Соболева 331  
 - унитарное 34  
 - энергетическое 37, 506  
 Процедура сглаживающая 389  
 Процесс Эйткена 162, 174
- P**адиус спектральный оператора 36  
 - числовой оператора 37  
 Разность разделенная 111  
 Разрыв 441  
 - контактный 459  
 Решение автомодельное 259  
 - обобщенное 305, 310  
 - разрывное 460  
 - слабое 505  
 - фундаментальное 552
- C**глаживание 389  
 - последующее 398  
 - предварительное 398  
 - функции 334  
 Сглаживатель 389  
 Сетка 203  
 - грубая 391, 393  
 - мелкая 391, 393  
 - прямоугольная 133  
 - равномерная 203  
 - связная 235  
 - структурированная 342  
 - узлы 203  
 - шаги 203  
 Система ОДУ жесткая 181  
 - ортонормированная 34  
 - полная 34  
 - функций чебышевская 107  
 Скорость групповая 283  
 Слой временной 210  
 Соотношение дисперсионное 283  
 Спектр оператора 38  
 Сплайн интерполяционный 131  
 -- степени  $m$  131  
 -  $B$ -сплайн 132  
 Стабилизатор Тихонова 331  
 Сумма квадратурная 138  
 -- узлы 138  
 Суммирование ряда Фурье 335  
 Схема Бима — Ворминга 478  
 - В.В. Русанова 422
- Схема в потоковой форме 467  
 - годуновского типа 467  
 - двухслойная асимметричная 424  
 - Дюфорта — Франкела («ромб») 262  
 - К.И. Бабенко («квадрат») 427  
 ---- с коррекцией типа «лимитера» 427  
 - коррекции потоков 421  
 - Кранка — Николсона 423  
 - Куранта — Изаксона — Риса 470  
 - Лакса 421  
 - Лакса — Вендроффа 277, 420, 478  
 -- Фридрихса для уравнений газовой динамики 464  
 -- Фридрихса — Ошера 494  
 - неявная 206  
 -- с левой разностью 423  
 -- типа Лакса — Вендроффа с весом 423  
 - переменных направлений 300  
 - полностью неявная 206  
 - прыжкового переноса 442  
 - разностная 203  
 -- бездиссипативная 283  
 -- двухслойная 225  
 -- диссипативная 283  
 -- для интегрального уравнения 319  
 -- консервативная 227  
 -- корректная 226  
 -- локально-одномерная 302  
 -- монотонная 229  
 -- неконсервативная 227  
 -- однородная 228  
 -- повышенного порядка для уравнения Пуассона 292  
 -- продольно-поперечная 300  
 -- равномерно устойчивая 225  
 -- суммарной аппроксимации 304  
 -- трехслойная 262  
 -- устойчивая 224  
 -- экономичная 300  
 - решения ОДУ симметричные 168  
 - Ричардсона («крест») 262  
 - Роу 471  
 -- для трехмерной газовой динамики 496  
 - Роу — Эйнфельдта 477  
 -- Эйнфельдта — Ошера 484  
 ---- упрощенная 493  
 - Р.П. Федоренко 426  
 - с весами 248  
 -- аппроксимация 249  
 -- устойчивость 251  
 -- асимптотическая 254  
 -- в  $C$  254  
 --  $L_2$  251  
 -- по правой части 253  
 - симметричные решения ОДУ 168

- Схема с монотонизацией по области зависимости** 428  
 – направленными разностями 465  
 – со «сглаживанием» 426  
 – с центральной разностью 421  
 ----- для уравнения переноса 270, 272  
 -- экспоненциальной подгонкой 193  
 -- «лимитерами» 422, 425  
 – Фромма 478  
 – явная 206  
 -- с левой разностью 420  
 ----- 2-го порядка аппроксимации по  $x$  423  
 – «бегущего» счета для уравнения переноса 266  
 ----- теплопроводности 260  
 – «кабаре» 424  
 -- с монотонизаторами 424  
 – «левый уголок» 266  
 -- неявная 270  
 – «парабола» 427  
 -- с «лимитерами» 427  
 – «правый уголок» 268  
 -- неявная 271  
**Сходимость линейная со скоростью геометрической прогрессии** 90  
 – метода простой итерации 88  
 – разностной схемы 225  
 – стационарного метода 96  
 – с  $p$ -м порядком 92  
 – численного метода 166  
**Счет на установление** 297
- Теорема Вейерштрасса** 106  
 – Годунова 279  
 – Марцинкевича 121  
 – Тихонова 328, 331  
 – Фабера 120  
**Технология многосеточная универсальная** 409  
**Тождество интегральное** 309  
**Трейд** 368  
**Триангуляция** 339  
 – Делоне 367  
 -- с ограничениями 373  
 – правильная 521  
 – сложных областей 345
- Узлы нерегулярные** 205  
 – регулярные 205  
**Уравнение вариационное** 510  
 – волновое 284  
 – граничное интегральное 553  
 – интегральное 317  
 -- Вольтерра 1-го рода 326  
 ----- 2-го рода 319  
 -- Фредгольма 1-го рода 326  
 ----- 2-го рода 318
- Уравнение колебаний** 201  
 – конвекции-диффузии 281  
 – Лапласа 201  
 – переноса 201, 264, 415  
 -- двумерное 415  
 -- квазилинейное 415, 436  
 -- линейное 415  
 – Пуассона 201  
 – состояния 455  
 – теплопроводности 201  
 – характеристическое  $m$ -шагового разностного метода 177
- Уравнения газовой динамики** 455  
 -- в дивергентной форме 456  
 ----- интегральной форме 456  
 ----- лагранжевых переменных 457  
 ----- недивергентной форме 456  
 ----- гиперболичность 457  
 ----- изотермические 458  
 – Эйлера 455  
**Уровень сеточный нулевой** 391  
 -- первый 391  
**Условие граничное главное** 509, 511  
 -- естественное 509, 511  
 – диагонального преобладания 57  
 – корней 178  
 – начальное финитное 419  
 – положительности коэффициентов 235  
 – энтропийное 472, 473  
**Условия Гюгонио** 471  
**Устойчивость** 224  
 – абсолютная 224  
 – асимптотическая 225, 254  
 – безусловная 224  
 – по граничным условиям 225  
 -- начальным данным 225  
 -- правой части 225  
 – условная 224  
**Уточнение итерационное** 79
- Флип** 368  
**Форма записи каноническая двухслойного итерационного метода** 64  
**Формула Гаусса** 150  
 – Грина 506  
 – квадратурная 138  
 -- интерполяционного типа 141, 143  
 -- левых прямоугольников 140  
 -- наивысшей алгебраической степени (Гаусса) 150  
 -- Ньютона — Котеса 146  
 -- правых прямоугольников 140  
 -- Симпсона 142  
 -- трапеций 140, 141  
 -- Филона 151  
 -- центральных прямоугольников 139  
 --  $l$ -го порядка точности 140

- Ф**ункция базисная 104, 504  
– одномерная 104  
– пробная 503  
– сеточная 203  
– формы 104
- Х**арактеристика аппроксимационная 345  
– уравнения переноса 264
- Ц**икл многосеточный адаптивный 405
- Ц**ифра числа верная 21  
– значащая 21
- Ч**исло жесткости 181  
– на временном интервале 182
- Ч**исло Куранта 266, 420  
– обусловленности 53
- Ш**аблон 205, 348  
Шаг временной курантовский 266
- Э**ксперимент вычислительный I, 16  
Экстраполяция 103  
Элемент граничный 556  
– конечный 104  
– двумерный 136  
– одномерный 104  
– наилучшего приближения 124  
– собственный 38, 39  
Элементы линейно независимые 32
- Я**дро вырожденное 323  
– оператора 36  
– приближенное 324

# ОГЛАВЛЕНИЕ

<b>Предисловие . . . . .</b>	<b>5</b>
<b>Основные обозначения . . . . .</b>	<b>7</b>
<b>Введение . . . . .</b>	<b>11</b>
B.1. Предмет и содержание книги . . . . .	11
B.2. История вычислений . . . . .	15
B.2.1. Исторические сведения . . . . .	15
B.2.2. Вычислительный эксперимент . . . . .	16
B.3. Ошибки при вычислениях . . . . .	18
B.3.1. Хранение чисел на ЭВМ и ошибки округления . . . . .	19
B.3.2. Ошибки арифметических операций . . . . .	22
B.3.3. Погрешность алгоритма . . . . .	24
B.4. Библиографические комментарии . . . . .	27
<b>ЧАСТЬ I. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЧИСЛЕННЫХ МЕТОДОВ . . . . .</b>	<b>29</b>
<b>1. Задачи линейной алгебры. Решение систем линейных алгебраических уравнений . . . . .</b>	<b>31</b>
1.1. Элементы функционального анализа и линейной алгебры . . . . .	31
1.1.1. Линейные пространства . . . . .	31
1.1.2. Операторы в линейных нормированных пространствах . . . . .	34
1.1.3. Операторы в гильбертовом пространстве . . . . .	36
1.1.4. Операторы в конечномерном пространстве . . . . .	38
1.1.5. Нормы векторов и матриц . . . . .	39
1.1.6. Другие нормированные пространства . . . . .	45
1.1.7. Критерий Адамара и лемма Гершгорина . . . . .	46
1.2. Прямые методы решения СЛАУ . . . . .	47
1.2.1. Схема метода Гаусса . . . . .	47
1.2.2. Расчетные формулы метода Гаусса . . . . .	49
1.2.3. Число действий в методе Гаусса . . . . .	49
1.2.4. Выбор главного элемента . . . . .	50
1.3. Обусловленность СЛАУ . . . . .	52
1.4. Метод прогонки решения СЛАУ с трехдиагональной матрицей . . . . .	55
1.4.1. Метод правой прогонки . . . . .	55
1.4.2. Методы левой и встречной прогонок . . . . .	59
1.4.3. Метод матричной прогонки . . . . .	60

---

1.5. Метод квадратного корня . . . . .	61
1.6. Итерационные методы решения СЛАУ . . . . .	64
1.6.1. Каноническая форма одношаговых итерационных методов . . . . .	64
1.6.2. Примеры одношаговых итерационных методов . .	65
1.6.3. Условия сходимости стационарных итерационных методов . . . . .	68
1.7. Итерационные методы решения СЛАУ вариационного типа . . . . .	71
1.7.1. Расчетные формулы . . . . .	72
1.7.2. Оценка скорости сходимости . . . . .	73
1.7.3. Частные случаи методов . . . . .	75
1.8. Методы сопряженных направлений . . . . .	77
1.9. Итерационное уточнение решения . . . . .	79
1.10. Решение проблемы собственных значений . . . . .	79
1.11. О регуляризации плохо обусловленных СЛАУ . . . . .	81
1.12. Хранение больших разреженных матриц . . . . .	83
1.13. Библиографические комментарии . . . . .	84
<b>2. Решение нелинейных уравнений . . . . .</b>	<b>86</b>
2.1. Решение скалярных уравнений . . . . .	86
2.1.1. Метод деления отрезка пополам (метод «вилки») . . . . .	87
2.1.2. Итерационные методы решения типа простой итерации . . . . .	88
2.1.3. Варианты метода простой итерации . . . . .	90
2.2. Решение систем нелинейных уравнений . . . . .	96
2.2.1. Сходимость стационарного метода . . . . .	96
2.2.2. Примеры итерационных методов . . . . .	98
2.3. Библиографические комментарии . . . . .	101
<b>3. Методы интерполяирования функций . . . . .</b>	<b>102</b>
3.1. Постановка задачи интерполяции. Простейшие варианты интерполяирования . . . . .	102
3.1.1. Кусочно-линейная интерполяция . . . . .	103
3.1.2. Варианты интерполяции . . . . .	106
3.2. Полиномиальная интерполяция . . . . .	106
3.2.1. Интерполяционный полином в форме Лагранжа .	108
3.2.2. Интерполяционный полином в форме Ньютона .	111
3.2.3. Интерполяционный полином Эрмита . . . . .	114
3.3. Сходимость и устойчивость полиномиальной интерполяции . . . . .	117
3.3.1. Оптимизация узлов сетки . . . . .	117

3.3.2. Устойчивость интерполяционного полинома относительно погрешностей функции . . . . .	119
3.3.3. Устойчивость интерполяционного полинома относительно априорной информации . . . . .	120
3.3.4. Наилучшие приближения в гильбертовом пространстве . . . . .	124
3.3.5. Насыщаемость алгоритма интерполяции. Тригонометрическая интерполяция . . . . .	127
3.4. Сплайн-интерполяция . . . . .	129
3.5. Двумерная интерполяция . . . . .	133
3.5.1. Прямоугольная сетка . . . . .	133
3.5.2. Треугольная сетка . . . . .	135
3.6. Библиографические комментарии . . . . .	137
<b>4. Методы численного интегрирования и дифференцирования . . . . .</b>	<b>138</b>
4.1. Простейшие квадратурные формулы . . . . .	138
4.1.1. Формула прямоугольников . . . . .	139
4.1.2. Формула трапеций . . . . .	140
4.1.3. Формула Симпсона . . . . .	141
4.2. Квадратурные формулы интерполяционного типа . . . . .	142
4.3. Квадратурные формулы Гаусса . . . . .	147
4.4. Интегрирование быстроосциллирующих функций . . . . .	151
4.5. Вычисление несобственных интегралов I и II рода . . . . .	152
4.6. Вычисление кратных интегралов . . . . .	155
4.7. Численное дифференцирование . . . . .	157
4.8. Библиографические комментарии . . . . .	164
<b>5. Численное решение задачи Коши для обыкновенных дифференциальных уравнений . . . . .</b>	<b>165</b>
5.1. Постановка задачи и простейшие методы . . . . .	165
5.1.1. Симметричная схема . . . . .	168
5.1.2. Метод Рунге — Кутты второго порядка . . . . .	169
5.2. Методы Рунге — Кутты . . . . .	170
5.3. Многошаговые разностные методы . . . . .	175
5.3.1. Погрешность аппроксимации многошаговых методов . . . . .	176
5.3.2. Устойчивость и сходимость разностных методов . . . . .	177
5.3.3. Примеры методов Адамса . . . . .	179
5.4. Понятие о методах решения жестких систем . . . . .	180
5.4.1. Условно устойчивые и безусловно устойчивые разностные методы . . . . .	180

---

5.4.2. Понятие жесткой системы ОДУ . . . . .	181
5.4.3. Решение жестких систем . . . . .	182
5.5. Библиографические комментарии . . . . .	186
<b>6. Решение краевых задач для систем обыкновенных дифференциальных уравнений . . . . .</b>	<b>188</b>
6.1. Постановка задачи. Метод стрельбы . . . . .	188
6.2. Разностные методы . . . . .	191
6.2.1. Линейная краевая задача второго порядка . . . . .	192
6.2.2. Нелинейные задачи . . . . .	194
6.3. Методы Ритца и Галеркина . . . . .	195
6.3.1. Метод Ритца . . . . .	196
6.3.2. Метод Галеркина . . . . .	197
6.3.3. Выбор системы функций . . . . .	198
6.4. Библиографические комментарии . . . . .	198
<b>7. Элементы теории разностных схем . . . . .</b>	<b>200</b>
7.1. Постановка задачи и основные понятия . . . . .	200
7.1.1. Постановка задачи . . . . .	200
7.1.2. Сетка и сеточные функции . . . . .	202
7.2. Обозначения и некоторые разностные соотношения . .	210
7.3. Методы и приемы конструирования разностных схем .	214
7.3.1. Метод разностной аппроксимации . . . . .	214
7.3.2. Интегро-интерполяционный метод . . . . .	215
7.3.3. Метод неопределенных коэффициентов . . . . .	217
7.3.4. Другие методы получения алгебраических уравнений . . . . .	218
7.3.5. Аппроксимации в нерегулярных точках . . . . .	219
7.4. Основные качественно-количественные характеристики разностных схем и их виды . . . . .	221
7.4.1. Аппроксимация . . . . .	221
7.4.2. Устойчивость . . . . .	224
7.4.3. Сходимость . . . . .	225
7.4.4. Качественно-количественные виды схем . . . . .	227
7.5. Разделение переменных в дискретном случае . . . . .	229
7.6. Принцип максимума для разностных схем . . . . .	234
7.7. Устойчивость разностных схем . . . . .	237
7.7.1. Применение принципа максимума к исследованию устойчивости по граничным условиям первого рода и начальным данным . . . . .	238
7.7.2. Признаки равномерной устойчивости . . . . .	238
7.7.3. Использование метода разделения переменных .	243

7.7.4. Необходимый «спектральный» признак устойчивости схемы по начальным данным . . . . .	244
7.7.5. Метод энергетических неравенств . . . . .	245
7.8. Библиографические комментарии . . . . .	246
<b>8. Численное решение параболических уравнений . . . . .</b>	<b>248</b>
8.1. Линейное одномерное уравнение теплопроводности с постоянными коэффициентами. Схема с весами . . . . .	248
8.1.1. Аппроксимация схемы с весами . . . . .	249
8.1.2. Устойчивость схемы с весами . . . . .	251
8.1.3. Сходимость и точность схемы с весами . . . . .	257
8.2. Некоторые другие задачи и схемы . . . . .	257
8.2.1. Задача с переменными коэффициентами . . . . .	258
8.2.2. Схема «бегущего» счета для решения уравнения теплопроводности . . . . .	260
8.2.3. Трехслойные схемы . . . . .	262
8.3. Библиографические комментарии . . . . .	263
<b>9. Численное решение гиперболических уравнений . . . . .</b>	<b>264</b>
9.1. Линейное одномерное уравнение переноса . . . . .	264
9.1.1. Явная схема с левой разностью (схема 1) . . . . .	266
9.1.2. Явная схема с правой разностью (схема 2) . . . . .	268
9.1.3. Явная схема с центральной разностью (схема 3)	270
9.1.4. Неявная схема с левой разностью (схема 4) . . . . .	270
9.1.5. Неявная схема с правой разностью (схема 5) . . . . .	271
9.1.6. Неявная схема с центральной разностью (схема 6)	272
9.1.7. Уравнение переноса с отрицательной или переменной скоростью . . . . .	274
9.1.8. Интерполяционный метод построения некоторых других схем для уравнения переноса . . . . .	275
9.2. Монотонность схем для уравнения переноса . . . . .	278
9.3. Дифференциальное приближение . . . . .	280
9.4. Волновое уравнение . . . . .	284
9.5. Библиографические комментарии . . . . .	288
<b>10. Численное решение эллиптических уравнений . . . . .</b>	<b>289</b>
10.1. Решение задачи Дирихле для уравнения Пуассона . . . . .	289
10.2. Разностная схема для уравнения Пуассона повышенного порядка точности . . . . .	292
10.3. Собственные функции разностного оператора Лапласа и их применение . . . . .	293
10.3.1. Разностная задача Штурма — Лиувилля в двумерном случае . . . . .	293

---

10.3.2. Численное нахождение решения разностной задачи . . . . .	295
10.4. Экономичные разностные схемы для решения уравнения теплопроводности в многомерном случае . . . . .	299
10.4.1. Продольно-поперечная схема . . . . .	300
10.4.2. Локально-одномерная схема . . . . .	302
10.5. Проекционные методы решения эллиптических уравнений . . . . .	304
10.5.1. Метод Ритца . . . . .	304
10.5.2. Метод Галеркина . . . . .	309
10.6. Оператор Лапласа в криволинейных координатах и его разностная аппроксимация . . . . .	311
10.6.1. Цилиндрические координаты . . . . .	312
10.6.2. Сферические координаты . . . . .	314
10.7. Библиографические комментарии . . . . .	315
<b>11. Численное решение интегральных уравнений . . . . .</b>	<b>317</b>
11.1. Корректно поставленные задачи . . . . .	317
11.1.1. Разностный метод численного решения . . . . .	319
11.1.2. Метод последовательных приближений . . . . .	321
11.1.3. Замена ядра вырожденным . . . . .	323
11.1.4. Метод Галеркина (метод моментов) . . . . .	325
11.2. Некорректные задачи . . . . .	326
11.2.1. Предпосылки метода регуляризации . . . . .	327
11.2.2. Понятие регуляризирующего оператора и пример регуляризации операторного уравнения первого рода . . . . .	329
11.2.3. Примеры некорректно поставленных задач . . . . .	334
11.3. Библиографические комментарии . . . . .	336
<b>ЧАСТЬ II. ИЗБРАННЫЕ ВОПРОСЫ ТЕОРИИ И ПРАКТИКИ ЧИСЛЕННЫХ МЕТОДОВ . . . . .</b>	<b>337</b>
<b>12. Методы триангуляции пространственных областей . . . . .</b>	<b>339</b>
12.1. Методы триангуляции и оценка качества сетки . . . . .	339
12.1.1. Классификация методов . . . . .	341
12.1.2. Оценка качества сетки . . . . .	344
12.1.3. Особенности построения сеток в сложных областях . . . . .	345
12.2. Прямые методы . . . . .	347
12.2.1. Методы на основе шаблонов . . . . .	348
12.2.2. Методы отображения . . . . .	357

<b>12.3. Методы граничной коррекции . . . . .</b>	<b>361</b>
12.3.1. Построение первичной сетки . . . . .	362
12.3.2. Коррекция первичной сетки . . . . .	364
<b>12.4. Методы на основе критерия Делоне . . . . .</b>	<b>367</b>
12.4.1. Построение триангуляции Делоне на заданном наборе точек . . . . .	369
12.4.2. Триангуляции Делоне с ограничениями . . . . .	373
12.4.3. Особенности технической реализации алгоритмов на основе критерия Делоне . . . . .	377
<b>12.5. Метод исчерпывания . . . . .</b>	<b>378</b>
<b>12.6. Оптимизация сеток . . . . .</b>	<b>381</b>
12.6.1. Оптимизация расположения узлов, или сглаживание сетки . . . . .	382
12.6.2. Оптимизация связей . . . . .	383
12.6.3. Сгущение сетки . . . . .	384
<b>13. Многосеточные методы . . . . .</b>	<b>386</b>
13.1. Проблема решения больших сеточных задач . . . . .	386
13.2. Основы многосеточных методов . . . . .	387
13.3. Классические многосеточные методы . . . . .	392
13.3.1. Пример одномерной задачи . . . . .	393
13.3.2. Основные направления развития КММ . . . . .	402
13.4. Универсальная многосеточная технология . . . . .	408
<b>14. Численное решение уравнения переноса . . . . .</b>	<b>415</b>
14.1. Уравнение переноса: постановка задачи . . . . .	415
14.2. Линейное одномерное уравнение переноса . . . . .	419
14.2.1. Постановка задачи для линейного одномерного уравнения переноса. Тестовые задачи . . . . .	419
14.2.2. Разностные схемы для линейного одномерного уравнения переноса . . . . .	420
14.2.3. Метод нелинейной монотонизации разностных схем для линейного одномерного уравнения переноса . . . . .	430
14.2.4. Результаты расчетов для одномерного линейного уравнения переноса . . . . .	433
14.3. Одномерное квазилинейное уравнение . . . . .	436
14.3.1. Постановка задачи для квазилинейного одномерного уравнения переноса. Тестовые задачи . . . . .	436
14.3.2. Нелинейная монотонизация схемы К.И. Бабенко для квазилинейного одномерного уравнения переноса . . . . .	438

---

14.3.3. Результаты расчетов для одномерного квазилинейного уравнения переноса . . . . .	440
14.3.4. Решение квазилинейного уравнения переноса с помощью разрывного метода Галеркина . . . . .	444
14.4. Двумерное линейное уравнение переноса . . . . .	448
14.4.1. Постановка задачи для линейного двумерного уравнения переноса. Тестовые задачи . . . . .	448
14.4.2. Разностные схемы для численного решения линейного двумерного уравнения . . . . .	450
14.4.3. Результаты расчетов для линейного двумерного уравнения переноса . . . . .	451
<b>15. Численное решение уравнений газовой динамики . . . . .</b>	<b>455</b>
15.1. Уравнения газовой динамики . . . . .	455
15.2. Разностная схема Роу — Эйнфельдта — Ошера . . . . .	459
15.2.1. Основные уравнения . . . . .	461
15.2.2. Схема Лакса — Фридрихса . . . . .	464
15.2.3. Схемы годуновского типа (линейный случай) . . . . .	467
15.2.4. Схемы годуновского типа (нелинейный случай). Схема Роу . . . . .	470
15.2.5. Энтропийное условие . . . . .	473
15.2.6. Схемы повышенного порядка аппроксимации . . . . .	477
15.2.7. Схема Роу для двумерной газовой динамики . . . . .	485
15.2.8. Упрощенная схема Роу — Эйнфельдта — Ошера и схема Лакса — Фридрихса — Ошера . . . . .	493
15.2.9. Схема Роу для решения уравнений трехмерной газовой динамики . . . . .	496
15.2.10. Другие схемы газовой динамики . . . . .	500
<b>16. Теоретические и алгоритмические основы метода конечных элементов . . . . .</b>	<b>501</b>
16.1. Метод конечных элементов и его варианты . . . . .	501
16.2. Метод взвешенных невязок . . . . .	502
16.3. Метод Бубнова — Галеркина . . . . .	505
16.3.1. Обобщенные решения и слабая постановка задачи	506
16.3.2. Аппроксимация методом Бубнова — Галеркина	512
16.3.3. Сходимость метода Бубнова — Галеркина . . . . .	513
16.4. Вариационно-сеточные методы . . . . .	514
16.4.1. Метод Ритца . . . . .	515
16.4.2. Метод наименьших квадратов . . . . .	517
16.5. Метод конечных элементов . . . . .	518
16.5.1. Двумерное уравнение Пуассона . . . . .	519

16.5.2. Линейная задача теории упругости . . . . .	528
16.5.3. Численное интегрирование . . . . .	539
16.5.4. Конечные элементы высокого порядка . . . . .	540
16.6. О применении МКЭ к решению других задач . . . . .	546
16.7. Основы метода граничных элементов . . . . .	548
16.7.1. Постановка задачи. Граничные интегральные уравнения . . . . .	550
16.7.2. Аппроксимации метода граничных элементов .	556
16.7.3. Алгоритмические аспекты . . . . .	560
<b>Литература . . . . .</b>	<b>561</b>
<b>Предметный указатель . . . . .</b>	<b>576</b>

*Научное издание*

**Математическое моделирование  
в технике и в технологиях**

**Галанин Михаил Павлович  
Савенков Евгений Борисович**

**МЕТОДЫ ЧИСЛЕННОГО АНАЛИЗА  
МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ**

Редактор *Н.Г. Ковалевская*

Технический редактор *Э.А. Кулакова*

Художники *С.С. Водчиз, Н.Г. Столярова*

Корректор *О.В. Калашникова*

Компьютерная верстка *А.Н. Канатникова*

Оригинал-макет подготовлен  
в Издательстве МГТУ им. Н.Э. Баумана

Санитарно-эпидемиологическое заключение  
№ 77.99.60.953.Д.003961.04.08 от 22.04.2008 г.

Подписано в печать 25.10.2010. Формат 70×100 1/16.  
Усл. печ. л. 48,1. Тираж 1000 экз. (1-й з-д 1–500). Заказ 2043

1166 - 00

Издательство МГТУ им. Н.Э. Баумана

105005, Москва, 2-я Бауманская, 5.

E-mail: [press@bmstu.ru](mailto:press@bmstu.ru)

<http://www.press.bmstu.ru>

Отпечатано  
в ГУП ППП «Типография «Наука».  
121099, Москва, Шубинский пер., 6.

ISBN 978-5-7038-3252-3



9 785703 832523