Министерство образования и науки Российской Федерации
Государственное образовательное учреждение высшего
профессионального образования
"Алтайский государственный технический
университет им. И.И. Ползунова"

С.А. КАНТОР

ОСНОВЫ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

Учебное пособие

Барнаул, 2010

УДК 519.6

Кантор С.А. Основы вычислительной математики: Учебное пособие. / Алт. госуд. технич. ун-т им. И.И.Ползунова. Барнаул, 2010. — 357с.

Учебное пособие является исправленным и дополненным вариантом вышедшего в 2005 году учебного пособия "Вычислительная математика". Оно предназначено для студентов вуза с повышенным объемом математической подготовки, в частности, для обучения специалистов по специальности 230105 - "Программное обеспечение вычислительной техники и автоматизированных систем", магистров по направлению 230100 "Информатика и вычислительная техника", бакалавров по направлению 231000 "Программная инженерия ". Кроме того, пособие может использоваться при получении дополнительного образования.

Первое издание учебного пособия "Вычислительная матемтика" было рекомендовано Научно-методическим Советом по проблемам подготовки преподавателей высшей школы в качестве учебного пособия для студентов, аспирантов и преподавателей высших учебных заведений при освоении дополнительной квалификации "Преподаватель" и "Преподаватель высшей школы" (протокол №18 от 7.04.2005 г.).

Рецензенты:
д.п.н., профессор Г.В. Лаврентьев (АлтГУ),
д.ф.-м.н., профессор Г.В. Пышнограй (АлтГТУ)

ВВЕДЕНИЕ

С момента своего рождения наиболее эффективное применение вычислительная техника нашла при проведении трудоемких расчетов в научных исследованиях. В настоящее время выработалась технология исследования сложных проблем с помощью ЭВМ, которая получила название **вычислительный эксперимент**. Эта технология все чаще вытесняет дорогостоящий натурный эксперимент. Можно указать такие крупные области применения вычислительного эксперимента, как энергетика, аэрокосмическая техника, обработка данных научного эксперимента, совершенствование технологических процессов.

Постановка вычислительного эксперимента делится на ряд этапов.

- 1. *Постановка задачи.* На этом этапе содержательно формулируется задача исследования, определяются конечные цели ее решения.

- 2. *Построение математической модели.* Здесь осуществляется математическая формулировка задачи в виде некоторых математических соотношений. При построении математической модели обычно пренебрегают целым рядом свойств, не оказывающих существенного влияния на ход исследуемого явления. При этом модель должна правильно (адекватно) описывать основные свойства изучаемого процесса. Примером математической модели, описывающей распространение тепла в некотором теле, служат краевые задачи для уравнения теплопроводности. Исследование математической модели зачастую провести очень сложно. Например, в явном виде решить многие типы уравнений, как правило, бывает трудно или вообще невозможно. Возникает проблема найти те или иные количественные характеристики решения. Они могут быть получены с помощью ЭВМ. Для этого требуется осуществление следующего этапа.

- 3. *Разработка численного метода.* Под численным методом понимается такая интерпретация математической модели, которая доступна для реализации на ЭВМ. Разработкой численных методов занимаются специалисты в области вычислительной математики. Исследователю же, занимающемуся прикладными вопросами, из имеющегося арсенала методов достаточно уметь выбрать тот, который наиболее пригоден в данном конкретном случае.

- 4. *Разработка алгоритма и построение программы для ЭВМ.* Вопрос о выборе алгоритма решения тесно связан с разработанным численным методом и часто их трудно разделить. Построение программы для ЭВМ, как правило, является одним из самых простых этапов.

- 5. *Проведение расчетов и анализ результатов.* Полученные результаты вычислений изучаются с точки зрения их соответствия исследуемому явлению, при необходимости вносятся коррективы в численный метод или математическую модель. В случае, когда полученные результаты расчетов имеют большой объем, возникают проблемы представления их в виде, удобном для анализа.

Основным содержанием данного курса является вопрос построения и исследования численных методов, поэтому курс часто называют "Численные методы".

С развитием вычислительной техники развивались библиотеки научных программ и пакеты прикладных программ, которые являются важнейшим инструментом выполнения типовых расчетов во многих областях знаний. Каждый такой пакет является комплексом взаимосвязанных прикладных программ, специальных и общих средств системного обеспечения. Многие пакеты позволяют проделать, не выходя из них, всю нужную работу или весьма значительную часть: провести моделирование, расчеты, оформить результаты, подготовить презентацию. В последние десятилетия наметился существенный прогресс в разработке интегрированных математических систем, резко уменьшивший затраты времени на их освоение и программирование (Eureka фирмы Borland, Mathcad фирмы MathSoft, MatLab фирмы MathWorks, Scilab - фирмы Scilab, Maxima - группа независимых разработчиков и другие).

Среды общения с ЭВМ на естественном для математика языке позволяют больше внимания уделять постановке проблемы, математическому моделированию реальных ситуаций, анализу ответа. Для записи команд и отдельных выражений в математических пакетах используется входной язык, напоминающий Бейсик с примесью Фортрана и Паскаля. Он может рассматриваться как функциональный язык высокого уровня и легко осваивается пользователями с опытом процедурного программирования. Его версии в отдельных пакетах различаются в некоторых деталях. Применяется он либо непосредственно (MatLab, Derive, Scilab), либо в комбинации с "кнопочно-шаблонным" интерфейсом (Maple, Mathematica, Mathcad, Scientific Workplace).

Последние версии современных пакетов реализуют основные концепции объектно ориентированного программирования: производные типы объектов, иерархическое наследование свойств, возможность определения над объектами новых операций и переопределения стандартных, модульное программирование.

Значительное сходство в возможностях и технологии применения пакетов сочетается с множеством заметных различий во входном языке. Самыми богатыми возможностями обладают профессиональные пакеты Mathematica и Maple. Подмножества символьной математики из Maple входят в последние версии Mathcad, MatLab, Scientific Workplace.

Остановимся кратко на характеристиках наиболее распространенных пакетов.

Система Mathematica разработана фирмой Wolfram Research и является мощным средством выполнения математически исследований как в символьной, так и в численной форме. Система справедливо считается мировым лидером среди компьютерных систем символьной математики. Она используется во многих ведущих университетах мира и получила широкое распространение в образовательных учреждениях всех континентов. Список монографий, посвященных применениям Mathematica, содержит сотни наименований; издаются журналы, проводятся ежегодные конференции. Разработчик поддерживает в сети Internet свободный доступ к громадному числу научных, методических и учебных продуктов, созданных сотрудниками фирмы и пользователями, число которых превысило миллион. На базе Mathematica создано около 100 специализированных коммерческих пакетов. Из вычислительных возможностей пакета отметим

- проведение различных вычислений с высокой степенью точности;
- алгебраические и численные вычисления производных и интегралов;
- решение систем алгебраических, дифференциальных и разностных уравнений;

- поддержка целого ряда функций матричных и векторных вычислений;
- поддержка интервальной арифметики;
- широкий набор встроенных математических функций;
- вычисления вычетов и преобразований на их основе.

Пакет обеспечивает вычисления как в действительной, так и в комплексной области. Он имеет стандартный оконный интерфейс работы с файлами и редактирования набора, выбора размера и начертания шрифтов, управления графическим выводом (двумерным и трехмерным), звукового сопровождения и анимации. Представляя в среде пакета Mathematica вычислительный алгоритм и поясняющее его описание, пользователь формирует так называемый блокнот (notepad), который можно затем сохранять для последующего использования, выполнять полностью или по частям (ячейкам). В дальнейшем документ можно модифицировать как в вычислительной части, так и в части имеющейся в нем текстовой, графической и формульной информации. Допускается включение в документ чертеже рисунков, полученных в других вычислительных и графических пакетах. Выходной документ может быть подготовлен совместимым с MS Word, MS Excel , PowerPoint и т.д. Выход системы можно преобразовать для вывода на печать в форматы упомянутых систем, а также TEXa и построенных на его базе издательских систем LATEX[1] и $\mathcal{A}MS$-TEX.

Своим названием (MATrix LABoratory) система MatLab обязана ориентации на матричные и векторные вычисления. Пакет MatLab прошел многолетний путь развития от системы для больших ЭВМ (конец 70-х годов), опиравшейся на знаменитые пакеты Фотран программ для линейной алгебры, до интегрированной среды, ориентированной на массовые персональные компьютеры (с середины 80-х). MatLab - хорошо апробированная и надежная система, рассчитанная на широкий круг задач и представление данных в универсальной (матрично-векторной) форме. Считается, что эта система фактически стала международным стандартом учебного и научного программного обеспечения для решения задач технических и научных вычислений. На сегодняшний день MatLab работает на большинстве современных операционных систем, включая GNU/Linux, Mac OS, Solaris и Microsoft Windows.

Для MatLab имеется возможность создавать специальные наборы инструментов (toolbox), расширяющие его функциональность. Наборы инструментов представляют собой коллекции функций, написанных на языке MatLab для решения определённого класса задач. Компания Mathworks за дополнительную плату поставляет наборы инструментов, которые используются во многих областях, включая следующие:

- *Цифровая обработка сигналов, изображений и данных*: DSP Toolbox, Image Processing Toolbox, Wavelet Toolbox, Communication Toolbox, Filter Design Toolbox — наборы функций, позволяющих решать широкий спектр задач обработки сигналов, изображений, проектирования цифровых фильтров и систем связи.
- *Системы управления*: Control Systems Toolbox, μ-Analysis and Synthesis Toolbox, Robust Control Toolbox, System Identification Toolbox, LMI Control Toolbox, Model Predictive Control Toolbox, Model-Based Calibration Toolbox — наборы функций, облегчающих анализ и синтез динамических систем, проектирование,

[1]Компьютерный набор данного учебного пособия выполнен автором в издательской системе LATEX

моделирование и идентификацию систем управления, включая современные алгоритмы управления, такие как робастное управление, ЛМН-синтез, μ-синтез и другие.

- *Финансовый анализ*: GARCH Toolbox, Fixed-Income Toolbox, Financial Time Series Toolbox, Financial Derivatives Toolbox, Financial Toolbox, Datafeed Toolbox — наборы функций, позволяющие быстро и эффективно собирать, обрабатывать и передавать различную финансовую информацию.

- *Анализ и синтез географических карт, включая трёхмерные*: Mapping Toolbox.

- *Сбор и анализ экспериментальных данных*: Data Acquisition Toolbox, Image Acquisition Toolbox, Instrument Control Toolbox, Link for Code Composer Studio — наборы функций, позволяющих сохранять и обрабатывать данные, полученные в ходе экспериментов, в том числе в реальном времени. Поддерживается широкий спектр научного и инженерного измерительного оборудования.

- *Визуализация и представление данных*: Virtual Reality Toolbox — позволяет создавать интерактивные миры и визуализировать научную информацию с помощью технологий виртуальной реальности и языка VRML.

- *Средства разработки*: MatLab Builder for COM, MatLab Builder for Excel, MatLab Builder for NET, MatLab Compiler, Filter Design HDL Coder — наборы функций, позволяющих создавать независимые приложения из среды MatLab.

- *Взаимодействие с внешними программными продуктами*: MatLab Report Generator, Excel Link, Database Toolbox, MatLab Web Server, Link for ModelSim — наборы функций, позволяющие сохранять данные различных видов таким образом, чтобы другие программы могли с ними работать.

- *Базы данных*: Database Toolbox — инструменты работы с базами данных.

- *Научные и математические пакеты*: Bioinformatics Toolbox, Curve Fitting Toolbox, Fixed-Point Toolbox, Fuzzy Logic Toolbox, Genetic Algorithm and Direct Search Toolbox, OPC Toolbox, Optimization Toolbox, Partial Differential Equation Toolbox, Spline Toolbox, Statistic Toolbox, RF Toolbox — наборы специализированных математических функций, позволяющие решать широкий спектр научных и инженерных задач, включая разработку генетических алгоритмов, решение дифференциальных уравнений в частных производных, целочисленные проблемы, оптимизацию систем и другие.

- *Нейронные сети*: Neural Network Toolbox — инструменты для синтеза и анализ нейронных сетей.

- *Нечёткая логика*: Fuzzy Logic Toolbox — инструменты для построения и анализа нечётких множеств.

- *Символьные вычисления*: Symbolic Math Toolbox — инструменты для символьных вычислений с возможностью взаимодействия с символьным процессором программы Maple.

Помимо вышеперечисленных, существуют тысячи других наборов инструментов для MatLab, написанных другими компаниями и энтузиастами.

В целом MatLab можно охарактеризовать как мощную и хорошо сбалансированную математическую систему, посредством подпакетов ориентированную преимущественно на инженерные приложения. Однако, для своей работы этот пакет требует существенных ресурсов. Так, например, версия 7.7 требует более 2-х Гб дискового пространства.

Система Scilab как по интерфейсу, так и по своим возможностям похожа на MatLab. Основное отличие состоит в том, что Scilab является свободно распространяемым программным продуктом.

Пакет Mathcad (MathSoft Inc.) является одним из наиболее удобных для несложных расчетов на персональных компьютерах. Появившись в 1986 году для платформы MS-DOS, Mathcad впервые среди программ подобного рода использовал наборную математическую нотацию, совмещённую с автоматической системой вычислений. Пакет имеет естественный входной язык представления математических зависимостей и инструменты их набора типа предлагаемых в Microsoft Equation. В частности, войдя в окно работы с матрицами, можно получить шаблоны для матрицы или вектора, при необходимости их скорректировав. Пакет физико-математический: он позволяет вводить размерности переменных задачи и автоматически контролирует соответствие размерностей операндов результата. Mathcad оборудован текстовым процессором, позволяющим, например, оформить статью без помощи специализированных средств Word.

Вообще Mathcad - это полноценное Windows приложение со средствами обмена в статике и в динамике (Clipboard, OLE). В частности, очень прост экспорт графиков Mathcad. Возможность только собственными средствами сформулировать задачу в привычных обозначениях, исследовать задачу, обработать исходные данные, выбрать метод решения, получить его, задокументировать и передать по сети разработчики считают основным достоинством пакета.

Maple это среда для выполнения символьных, численных и графических вычислений профессиональными математиками, разработанная фирмой Waterloo Maple Software (University of Waterloo, Канада) и Высшей технической школой в Цюрихе. Она воплотила колоссальный математический потенциал, включает широчайший арсенал средств ("от элементарной арифметики до общей теории относительности") и активно используется научной общественностью, о чем свидетельствуют многочисленные ссылки в научных статьях. Выпускается журнал "Maple Technical Newsletter" и даже газета. Подмножества символьного инструментария Maple являются существенными элементами последних версий популярных пакетов Mathcad и MatLab.

Maxima как и Scilab относится к числу свободно распространяемого программного обеспечения. Они имеет широчайший набор средств для проведения аналитических вычислений, численных вычислений и построения графиков. По набору возможностей система близка к таким коммерческим системам как Maple и Mathematica. В то же время она обладает высочайшей степенью переносимости. Это единственная из существующих систем аналитических вычислений, которая может работать на всех основных современных операционных системах на компьютерах, начиная от самых мощных вплоть до наладонных компьютеров.

В связи с широким распространением математических пакетов, возникли сайты, журналы ориентированные, в первую очередь, на пользователей пакетов MatLab, Mathcad, Mathematica и Maple. К ним относятся, например, журнал "Exponenta Pro. Математика в приложениях", сайт http://matlab.exponenta.ru для общения русскоязычных пользователей системы MatLab, образовательный математический портал http://exponenta.ru. Сайты содержат большое количество учебных пособий, приме-

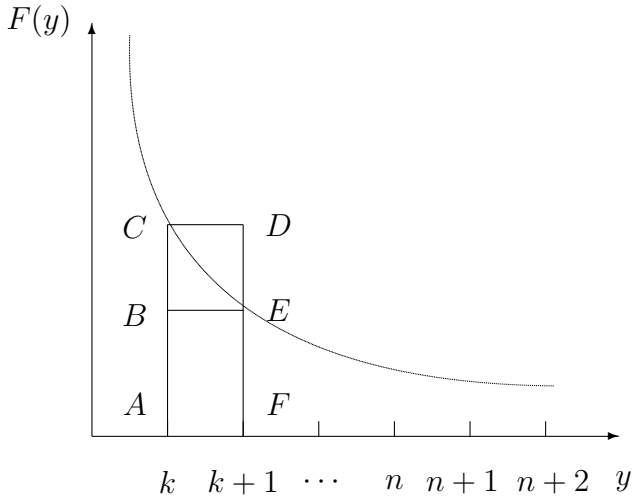


Рис. 1 Оценка членов ряда (1)

ров решения задач с помощью указанных пакетов, научных статей, советов по использованию пакетов. Регулярно организуются конкурсы пользователей систем.

Чтобы решить задачу можно поискать на этих сайтах свою или похожую задачу по математическому анализу, линейной алгебре, аналитической геометрии, обыкновенным дифференциальным уравнениям, теории вероятностей, вычислительной математике, теории функций комплексного переменного среди разобранных примеров. Есть возможность найти в банке решенных студенческих задач свою, задать вопрос для обсуждения на форуме.

Естественно может возникнуть мнение, что при таком обилии различных математических пакетов изучение вычислительной математики излишне. Однако богатством пакетов надо еще уметь воспользоваться. Проблемы из каждого раздела вычислительной математики настолько разнообразны, что не могут быть решены с помощью стандартных процедур. Постоянно будут возникать вопросы о выборе метода, шага, точности, начального приближения. К этому добавляются вопросы, связанные со спецификой решаемой задачи, с особенностью машинной арифметики. Поэтому, для успешного использования пакетов изучение вычислительной математики необходимо.

Даже при решении простейших задач могут возникнуть трудности, которые на первый взгляд не заметны. Проиллюстрируем это на следующем примере, предложенном Р.В. Хэмингом. Пусть требуется составить таблицу значений функции

$$f(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+x)} \tag{1}$$

на отрезке $[0, 1]$ с шагом 0.1 и точностью $\varepsilon = 10^{-12}$.

Казалось бы нет никаких проблем. Вычисляем последовательно слагаемые и суммируем их. Оценим количество слагаемых, которые необходимо учесть для получения заданной точности.

Пусть $F(y) = \dfrac{1}{y(y+x)}$. Из рисунка 1 видно, что при любом $k > 0$ площадь S_{1k} прямоугольника $ACDF$ больше площади S_{2k} криволинейной трапеции $ACEF$, которая в свою очередь больше площади S_{3k} прямоугольника $ABEF$.

$$S_{1k} = \frac{1}{k(k+x)}, \quad S_{2k} = \int_{k}^{k+1} \frac{dy}{y(y+x)}, \quad S_{3k} = \frac{1}{(k+1)(k+1+x)}.$$

Просуммировав площади по k в пределах от n до ∞, имеем для любого целого положительного n

$$\sum_{k=n+1}^{\infty} \frac{1}{k(k+x)} < \int_{n}^{\infty} \frac{dy}{y(y+x)} < \sum_{k=n}^{\infty} \frac{1}{k(k+x)}. \tag{2}$$

Таким образом, из (2) следует, что если при вычислении суммы ряда (1) ограничиться n-ой частичной суммой, то для погрешности

$$r_n = \sum_{k=n+1}^{\infty} \frac{1}{k(k+x)}$$

получается оценка

$$\int_{n+1}^{\infty} \frac{dy}{y(y+x)} < r_n < \int_{n}^{\infty} \frac{dy}{y(y+x)} = \begin{cases} \frac{1}{n}, & x = 0, \\ \frac{1}{x} \ln \frac{n+x}{n}, & x \in (0,1]. \end{cases} \tag{3}$$

При больших значениях n справедливо приближенное равенство $\ln \dfrac{n+x}{n} \approx \dfrac{x}{n}$, поэтому $r_n < \dfrac{1}{n}$ при всех $x \in [0,1]$. Следовательно, при $n = 10^{12}$ будет выполнено неравенство $r_n < \varepsilon$. При меньшем же значении n, по крайней мере для $x = 0$, требуемое неравенство перестает быть справедливым. Столь большое число слагаемых, которое необходимо просуммировать, делает задачу трудноразрешимой.

Покажем как можно преодолеть возникшие трудности. Прежде всего заметим, что

$$g_m = \sum_{k=1}^{\infty} \frac{1}{k(k+1)\dots(k+m)} = \frac{1}{m\,m!} \tag{4}$$

Действительно,

$$\begin{aligned} g_m = \sum_{k=1}^{\infty} \frac{1}{k(k+1)\dots(k+m)} &= \lim_{K\to\infty} \sum_{k=1}^{K} \frac{1}{k(k+1)\dots(k+m)} = \\ &= \lim_{K\to\infty} \sum_{k=1}^{K} \frac{1}{m} \left(\frac{1}{k(k+1)\dots(k+m-1) - (k+1)(k+2)\dots(k+m)} \right) = \\ &= \lim_{K\to\infty} \frac{1}{m} \left(\frac{1}{1\cdot 2\dots m} - \frac{1}{(K+1)(K+2)\dots(K+m)} \right) = \frac{1}{m\,m!}. \end{aligned}$$

Теперь можно перейти к преобразованию искомой функции $f(x)$. Так как

$$f(1) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = g_1 = 1,$$

имеем

$$f(x) = f(1)+(f(x)-f(1)) = 1+\sum_{k=1}^{\infty} \frac{1}{k} \left(\frac{1}{k+x} - \frac{1}{k+1} \right) = 1+(1-x) \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+x)} = 1+(1-x) f_1(x), \tag{5}$$

где

$$f_1(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+x)}. \tag{6}$$

Так как

$$f_1(2) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)} = g_2 = \frac{1}{4},$$

получаем

$$\begin{aligned} f_1(x) = f_1(2) + (f_1(x) - f_1(2)) &= \frac{1}{4} + \sum_{k=1}^{\infty} \frac{1}{k(k+1)} \left(\frac{1}{k+x} - \frac{1}{k+2} \right) = \\ &= \frac{1}{4} + (2-x) \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)(k+x)} = \frac{1}{4} + (2-x) f_2(x). \end{aligned} \tag{7}$$

Здесь

$$f_2(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)(k+x)}. \tag{8}$$

Из (5),(7) следует, что

$$f(x) = 1 + (1-x)\Big(\frac{1}{4} + (2-x)f_2(x)\Big). \tag{9}$$

Можно продолжить эти рассуждения.

$$f_2(3) = g_3 = \frac{1}{18}, \quad f_2(x) = f_2(3) + (f_2(x) - f_2(3)) = \frac{1}{18} + \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)} \left(\frac{1}{k+x} - \frac{1}{k+3} \right) =$$

$$= \frac{1}{18} + (3-x) \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)(k+3)(k+x)} = \frac{1}{18} + (3-x) f_3(x),$$

$$f_3(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)(k+3)(k+x)}, \quad f(x) = 1 + (1-x)\left(\frac{1}{4} + (2-x)\Big(\frac{1}{18} + (3-x) f_3(x) \Big) \right).$$

По аналогии

$$f_4(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)(k+3)(k+4)(k+x)},$$

$$f(x) = 1 + (1-x)\left(\frac{1}{4} + (2-x)\left(\frac{1}{18} + (3-x)\Big(\frac{1}{96} + (4-x) f_4(x) \Big) \right) \right) \tag{10}$$

и так далее.

Чтобы вычислить $f(x)$ по формуле (10) придется суммировать ряд для функции $f_4(x)$. Оценим какое теперь надо взять количество слагаемых, причем оценку получим достаточно грубую, не очень заботясь и том, чтобы определить как можно меньшее число слагаемых, обеспечивающее необходимую точность. Поэтому рассмотрим остаток ряда для случая $x = 0$. Для остальных значений x остаток ряда, очевидно меньше. При вычислении $f(0)$ по формуле (10) имеем

$$f(0) = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + 24 f_4(0).$$

Потребуем, чтобы

$$24 \sum_{k=n+1}^{\infty} \frac{1}{k^2(k+1)(k+2)(k+3)(k+4)} \le 10^{-12}. \tag{11}$$

Так же, как это было сделано выше, легко получить, что

$$24 \sum_{k=n+1}^{\infty} \frac{1}{k^2(k+1)(k+2)(k+3)(k+4)} < \int_n^{\infty} \frac{24dy}{y^2(y+1)(y+2)(y+3)(y+4)} < \int_n^{\infty} \frac{24dy}{y^6} = \frac{24}{5n^5}.$$

Поэтому, для получения (11) достаточно чтобы $24/(5n^5) < 10^{-12}$, то есть достаточно взять $n \approx 100\sqrt[5]{500} \approx 347$.

Приведенный пример показывает, что "грубой силой" удается решить далеко не любую задачу. Решению задачи должен предшествовать ее тщательный анализ. Наличие же мощных пакетов прикладных программ, зачастую существенно облегчают работу вычислителя, однако не освобождает его от всестороннего изучения задачи, выбора рационального метода ее решения, анализа полученных в процессе вычисления результатов и, быть может, корректировки выбранного метода вычислений.

Как и любая другая математическая дисциплина, вычислительная математика требует для своего успешного освоения решения задач. Поэтому, в конце каждой главы в учебное пособие включен небольшой список задач для самостоятельного решения. Каждая глава снабжена также примером тестов, которые используются для текущего контроля знаний студентов. Кроме того, учебное пособие содержит краткое описание лабораторных работ, которые каждый студент должен выполнить в течении семестра.

К сегодняшнему дню издано большое количество учебников и задачников по вычислительной математике различного уровня сложности. В списке литературы приведены тех из них, которые использовались при написании данного учебного пособия.

Во втором издании учебного пособия были добавлены примеры, задачи, рассмотрены дополнительные вопросы, в частности, методы нахождения собственных чисел не симметрических матриц, исследована проблема обусловленности задачи нахождения собственных чисел и собственных векторов.

1 МАШИННАЯ АРИФМЕТИКА И ПОГРЕШНОСТИ ВЫЧИСЛЕНИЙ

1.1 ПОГРЕШНОСТЬ РЕЗУЛЬТАТА ЧИСЛЕННОГО РЕШЕНИЯ ЗАДАЧ

1.1.1 Источники и классификация погрешностей

Процесс исследования исходного объекта математическими методами неизбежно носит приближенный характер, так как на каждом этапе вычислительного эксперимента вносятся те или иные погрешности. Эти погрешности обусловливаются следующими причинами:

- математическое описание задачи является неточным, так как сама математическая модель описывает лишь приближенно физическую модель и, кроме того, приближенно определяются исходные данные, входящие в описание задачи, так как они берутся из физического эксперимента;
- применяемый метод численного решения заменяет непрерывную модель дискретной;
- при вводе чисел в ЭВМ, при выполнении ею арифметических операций возникают ошибки округления.

В соответствии с этими причинами появляющиеся ошибки делят на:

- неустранимые погрешности;
- погрешности метода;
- вычислительные погрешности.

Неустранимые погрешности часто, в свою очередь, делят на:

- погрешности, которые являются следствием неточного задания числовых данных, входящих в математическое описание задачи **(погрешности исходных данных);**
- погрешности, связанные с несоответствием математического описания задачи реальности **(погрешности математической модели).**

Как уже отмечалось во введении, после выбора численного метода возникает необходимость выбора алгоритма численной реализации этого метода. Например, если исходная математическая модель является системой дифференциальных уравнений, то численным методом может служить построенная по определенным правилам система алгебраических уравнений, а алгоритм — метод решения этой системы.

Вычислительный алгоритм называют **устойчивым**, если в процессе работы погрешности округлений возрастают незначительно, и **неустойчивым** в противном случае[1]. Применение неустойчивого алгоритма зачастую приводит к появлению в процессе вычислений чисел, выходящих за пределы чисел, представимых на ЭВМ.

Зачастую типичной является такая ситуация, когда неустранимая погрешность больше погрешности метода, а погрешностью округлений в случае использования устойчивых алгоритмов можно пренебречь по сравнению с погрешностью метода. В общем случае надо стремиться, чтобы все указанные погрешности имели один и тот же порядок. Например, бессмысленно стремиться построить численный метод, который обеспечивал бы точность порядка 10^{-10}, когда сама неустранимая погрешность имеет порядок 10^{-1}.

Далее в этом разделе будут обсуждаться вопросы, связанные с природой вычислительной погрешности и оценкой ее величины.

1.1.2 Абсолютная и относительная погрешности

Определение 1.1.1 *Если a — точное значение некоторой величины, а a^* — известное приближение к нему, то* ***абсолютной погрешностью*** *приближенного значения a^* называют некоторую величину $\Delta(a^*)$, про которую известно, что*

$$|a^* - a| \le \Delta(a^*).$$

Задание величин a^* и $\Delta(a^*)$ означает, что определен отрезок $[a^* - \Delta(a^*), a^* + \Delta(a^*)]$, внутри которого находится число a.

Иногда вместо термина "абсолютная погрешность" используют термин **предельная абсолютная погрешность**. Число $\Delta(a^*)$ определяется неоднозначно: его можно увеличить. Обычно стараются указать возможно меньшее значение $\Delta(a^*)$.

Определение 1.1.2 ***Относительной погрешностью*** *приближенного значения называют некоторую величину $\delta(a^*)$, такую что*

$$\left|\frac{a^* - a}{a^*}\right| \le \delta(a^*).$$

Из приведенных определений следует, что абсолютная и относительная погрешности неотрицательны.

Если a — известное число, например π, то иногда говорят об абсолютной $\Delta(a)$ и относительной $\delta(a)$ погрешностях заданного числа.

Замечание. Некоторые авторы абсолютной и относительной погрешностями называют соответственно величины $a^* - a$ и $\dfrac{a^* - a}{a^*}$.

Определение 1.1.3 ***Значащими цифрами*** *числа называют все цифры в его записи, начиная с первой ненулевой слева.*

Пример. Пусть $a^* = 0.0\underline{2534}$, $b^* = 0.0\underline{253400}$. Здесь значащие цифры подчеркнуты. Тогда у первого числа 4, а у второго 6 значащих цифр.

[1]Примеры устойчивых и неустойчивых алгоритмов будут приведены в последующих параграфах.

Определение 1.1.4 *Значащая цифра называется* ***верной в широком смысле****, если абсолютная погрешность числа не превосходит единицы разряда, соответствующего этой цифре.*

Значащую цифру называют ***верной в узком смысле****, если абсолютная погрешность не превосходит половины единицы разряда, соответствующего этой цифре.*

Если все значащие цифры верны, то говорят, что число записано **со всеми верными цифрами**. При этом, если, например, число записано со всеми верными значащими цифрами и абсолютная погрешность не указана, то ее принимает равной единице разряда последней значащей цифры.

Пример. $a^* = 0.0\underline{2534},\ \Delta(a^*) = 0.000002$ — это число со всеми верными цифрами как в широком, так и в узком смысле.

$a^* = 0.0\underline{2534}00,\ \Delta(a^*) = 0.000007$ — подчеркнутые значащие цифры верны в широком смысле. В узком же смысле цифра 4 не будет верной.

$a^* = 12.34$, все значащие цифры в записи числа верны в узком смысле. Тогда полагают $\Delta(a^*) = 0.005$.

Часто информация о какой-либо величине задается пределами ее изменения: $a_1 \le a \le a_2 \quad 1.2543 \le a \le 1.2552$, причем обычно числа a_1, a_2 записывают с одинаковым числом десятичных знаков.

Как правило, абсолютную и относительную погрешности записывают в виде числа, содержащего одну – две значащие цифры.

Тот факт, что a^* — приближенное значение числа a с абсолютной погрешностью $\Delta(a^*)$ иногда записывают так

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

причем a^* и $\Delta(a^*)$ принято записывать с одинаковым числом десятичных знаков ($a = 2.538 \pm 0.003$).

Тот факт, что a^* — приближенное значение a с относительной погрешностью $\delta(a^*)$ записывают так: $a = a^*(1 \pm \delta(a^*))$.

В обиходе часто употребляют термин: найти решение с погрешностью 10^{-n}. Под этим имеют в виду не вышеприведенное определение, а то, что погрешность имеет *такой порядок*, то есть, если решение будет найдено, например, с погрешностью $1.5 \cdot 10^{-n}$, то такой результат будет признан удовлетворительным.

Зачастую приходится округлять приближенные или точные числа. Обычно для этого используется правило округления по дополнению, суть которого в следующем. Чтобы округлить число до n значащих цифр, отбрасывают все его цифры, стоящие справа от n-ой значащей цифры. При этом:

1) если первая из отброшенных цифр меньше 5, то оставшиеся десятичные цифры оставляют без изменения;

2) если первая из отброшенных цифр больше 5, то к последней оставшейся десятичной цифре прибавляется 1;

3) если первая из отброшенных цифр равна 5 и среди оставшихся отброшенных цифр есть ненулевые, то к последней оставшейся десятичной цифре прибавляется 1;

4) если первая из отброшенных цифр равна 5 и все отброшенные цифры равны нулю, то последняя оставшаяся цифра остается неизменной, если она четная, и увеличивается на единицу, если она нечетная.

Очевидно, что при таком правиле округления, погрешность округления не превосходит половины единицы десятичного разряда, определяемого последней оставленной значащей цифрой.

Пример. Округляя число 12500 до двух значащих цифр, получим $12 \cdot 10^3$.

1.2 МАШИННАЯ АРИФМЕТИКА

В этом параграфе будут введены некоторые основные понятия, относящиеся к вычислениям с плавающей точкой. Вначале обсудим обычный способ аппроксимации системы действительных чисел в вычислительной машине, а затем рассмотрим источники ошибок в машинных вычислениях.

В большинстве компьютеров внутреннее представление чисел двоичное, то есть в виде последовательностей нулей и единиц. Этого требуют соображения технологии, однако для человека такое представление чисел неестественно. Чтобы яснее показать особенности машинной арифметики, отделив их от деталей машинной реализации, в дальнейшем будем в большей части обсуждения опираться на десятичную, а не на двоичную арифметику.

Рассмотрим сначала **машинное представление целых чисел.**

Машинные целые числа представляются конечным количеством разрядов. Для иллюстрации будем считать, что у нас целые числа имеют шесть десятичных знаков. Каждое целое число характеризуется также знаком + или −. Все это означает, что количество машинных целых чисел конечно. В нашем примере наименьшим будет число −999999, наибольшим число 999999. Целые числа вне этой области для данного компьютера не существуют.

Если работать с целыми числами, далекими от границ числовой области компьютера, то машинная арифметика дает правильные результаты, например $5 + 7 = 12$, $8 - 27 = -19$ и $27 * 3 = 81$. Деление целых чисел приводит снова к целому числу: в качестве результата операции принимается частное, а остаток отбрасывается. Это значит, что $1/3 = 0$, $4/2 = 2$, $7/(-3) = -2$.

Если результат операции над целыми числами слишком велик или слишком мал для данного компьютера, то последствия трудно предсказать. На одних компьютерах будет выдано сообщение об ошибке, а выполнение текущего вычисления прекращено, на других результат будет заменен по циклическому правилу, а вычисление продолжено без какого-либо указания об ошибке. Так, для нашего примера было бы $999999 + 1 = -999999$. Таким образом, нельзя полагаться на результаты вычислений, выходящих за пределы числовой области компьютера.

Эти замечания в равной степени применимы к операциям с целыми числами, выполняемым в двоичной арифметике, хотя значения наибольшего и наименьшего числа будут иными. В компьютерах ряда популярных моделей для хранения целого числа отводится 32 двоичных разряда, причем один разряд — для знака числа. В этом случае наибольшим целым числом является $2^{31} - 1 = 2147483647$, а наименьшим — число -2^{31}.

Перейдем теперь к вопросу **машинного представления вещественных чисел**. Много разных методов было предложено для аппроксимации вещественных чисел посредством конечных машинных представлений. Метод, принятый в настоящее время почти на всех машинах, — это **числа с плавающей точкой**. Множество $\mathcal{F}$ чисел с плавающей точкой характеризуется четырьмя параметрами: **основанием системы счисления** β, **точностью** или **разрядностью** t и **интервалом показателей** $[L, U]$. Каждое число x с плавающей точкой, принадлежащее $\mathcal{F}$, имеет значение

$$x = \pm\Big(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t}\Big)\beta^e,$$

где целые числа $d_1, \ldots, d_t$ удовлетворяют неравенствам $0 \le d_i \le \beta - 1, (i = 1, \ldots, t)$

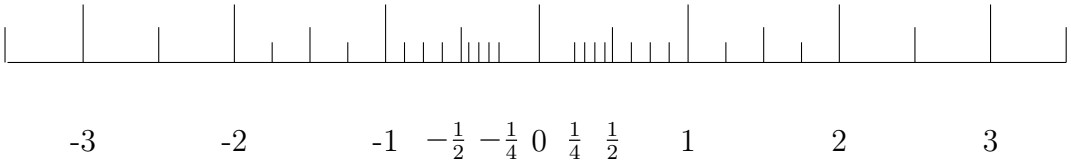


Рис. 1.1 33-точечное множество $\mathcal{F}$

и $L \leq e \leq U$. Если для каждого ненулевого x из $\mathcal{F}$ справедливо $d_1 \neq 0$, то система чисел $\mathcal{F}$ называется **нормализованной**. Целое число e называется **показателем**, а число

$$f = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t}$$

— **дробной частью** или **мантиссой**.

Множество $\mathcal{F}$ не является бесконечным множеством. В нем ровно

$$2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$$

чисел. Они расположены неравномерно, равномерность расположения имеет место лишь при фиксированном показателе.

На рисунке 1.1 показано 33-точечное множество $\mathcal{F}$ для небольшой иллюстративной системы с параметрами $\beta = 2$, $t = 3$, $L = -1$, $U = 2$.

Поскольку $\mathcal{F}$ — конечное множество, невозможно сколько-нибудь детально отобразить континуум вещественных чисел. Более того, вещественные числа с модулем, большим максимального элемента из $\mathcal{F}$, не могут быть отображены вообще. То же самое верно в отношении ненулевых вещественных чисел, меньших по абсолютной величине по сравнению с наименьшим положительным числом из $\mathcal{F}$. Наконец, каждое число из $\mathcal{F}$ должно представлять целый интервал вещественных чисел. Если x — вещественное число, не выходящее за границы множества $\mathcal{F}$, то будем обозначать через $fl(x)$ элемент из $\mathcal{F}$, ближайший к x. Заметим, что если x равноудалено от двух чисел из $\mathcal{F}$, то $fl(x)$ может принимать любое из двух соседних с x значений.

На множестве $\mathcal{F}$, являющемся моделью системы вещественных чисел, определены арифметические операции в соответствии с тем, как они выполняются вычислительной машиной. Предположим, что x и y — числа с плавающей точкой. Тогда обычная сумма $x+y$ зачастую уже не принадлежит $\mathcal{F}$. К примеру, возьмем 33-точечную систему, введенную выше и положим $x = 5/4, y = 3/8$. Следовательно, операция сложения сама должна моделироваться в машине посредством приближения.

Если $x, y \in \mathcal{F}$ и число $x+y$ не выходит за границы множества $\mathcal{F}$, то идеалом было бы выполнение равенства $x \oplus y = fl(x+y)$, где $x \oplus y$ обозначает операцию сложения чисел машинной. В большинстве вычислительных машин этот идеал достигается или почти достигается для всех таких x и y. Поэтому в нашем игрушечном 33-точечном множестве $\mathcal{F}$ следует ожидать, что $5/4 \oplus 3/8$ равно $3/2$ либо $7/4$. Разность между $x \oplus y$ и $x + y$ (для x, y из $\mathcal{F}$) есть ошибка округления, совершенная при сложении на ЭВМ. Аналогичные свойства верны для операций вычитания, умножения и деления.

Причина, по которой число $5/4 + 3/8$ не принадлежит 33-точечному множеству $\mathcal{F}$, связана с расположением элементов $\mathcal{F}$. С другой стороны, сумма типа $7/2+7/2$ не принадлежит $\mathcal{F}$ потому, что 7 больше максимального элемента из $\mathcal{F}$. Попытка образовать такую сумму вызовет на большинстве машин сигнал переполнения операции и данное вычисление на этом заканчивается.

Очень редко обычное произведение $x \cdot y$ принадлежит $\mathcal{F}$, поскольку, как правило, оно записывается посредством $2t$ либо $2t-1$ значащих цифр. Кроме того, переполнение гораздо более вероятно при умножении. Наконец, при умножении на ЭВМ возможно возникновение машинного нуля, когда два ненулевых числа x и y имеют ненулевое произведение, меньшее по абсолютной величине наименьшего ненулевого элемента из $\mathcal{F}$. В этом случае часто говорят, что происходит **исчезновение порядка**. Это событие обычно не имеет таких катастрофических последствий, как переполнение, и многие компьютеры заменяют результат нулем без какого-либо указания на то, что случилось нечто из ряда вон выходящее. Тем не менее есть вычисления, для которых факт появления машинного нуля важен.

Операции сложения и умножения на ЭВМ коммутативны, но не ассоциативны. Дистрибутивный закон для них также не выполняется. Пусть, например, $\beta = 10,\ t = 4,\ U = -L = 10$. Тогда

$$(0.2305 \oplus 0.4595) \oplus 1001 = 0.6900 \oplus 1001 = fl(1001.6900) = 1002,$$

в то время как

$$0.2305 \oplus (0.4595 \oplus 1001) = 0.2305 \oplus fl(1001.4595) = 0.2305 \oplus 1001 = fl(1001.2305) = 1001.$$

Поскольку эти алгебраические законы имеют фундаментальное значение для математического анализа, анализ вычислений произведенных на ЭВМ сложен.

В компьютере с двоичной арифметикой числа с плавающей точкой обычно представляются в памяти 32 двоичными разрядами, или битами. Стандарт IEEE (Institute of Electrical and Electronics Engineers) отводит 24 бита для мантиссы и 8 битов для показателя. Сюда входят также биты для хранения знаков мантиссы и показателя. Предполагается, что самый левый бит мантиссы любого числа равен 1, что позволяет не хранить этот бит. По техническим причинам показатель хранится как целое число в интервале $[0, 255]$. Чтобы получить фактическое значение показателя, нужно вычесть 127 из хранимого числа. Значение 255 резервируется для представления бесконечности, а также указания незаконных результатов, например квадратных корней из отрицательных чисел. Подобные незаконные результаты называются "не-числами". Из сказанного следует, что наибольшее число с плавающей точкой M_∞ равно [2] $+0.1...1_2 \times 2^{127} \approx 10^{38}$ Наименьшее число с плавающей точкой M_0 равно приблизительно 10^{-38}. Мантисса из 24 битов соответствует примерно 7 десятичным разрядам.

Между каждыми двумя степенями двойки равномерно расположены 2^{22} числа с плавающей точкой. Например, 2^{22} числа находится между 2^{-128} и 2^{-127} и столько же чисел между 2^{126} и 2^{127}. Таким образом, числа с плавающей точкой гуще расположены вблизи нуля.

Стандарт IEEE рекомендует также, чтобы компьютеры выполняли арифметику с большей разрядностью, несмотря на то, что результаты операций записываются в память с 32 битами. В машинах, поддерживающих этот стандарт, арифметика с плавающей точкой нередко реализована с внутренней разрядностью 80 битов.

Все компиляторы с языков Паскаль, Си, Фортран допускают числа с плавающей точкой удвоенной точности. Для их представления используется вдвое большее количество битов, чем для обычных чисел с плавающей точкой. Если вычисления удвоенной точности не поддержаны аппаратно, то они гораздо медленнее вычислений с одинарной точностью. Большинство компьютеров, соответствующих стандарту

[2] Здесь индекс 2 внизу обозначает систему счисления

IEEE, обладают аппаратно реализованным режимом удвоенной точности, по крайней мере, как опцией. Фортран допускает также арифметику комплексных чисел с плавающей точкой одинарной и удвоенной точности, при этом комплексное число представляется парой вещественных чисел с плавающей точкой, обычной или удвоенной точности. Комплексная арифметика редко поддерживается аппаратно, как правило, она реализуется программным путем.

Проектирование процессоров, обеспечивающих действительно надежную арифметику с плавающей точкой, оказалось нелегким делом. Имеется масса примеров обычных вычислений, дающих неправильные результаты, последствия которых могут быть весьма серьезными. Стандарт IEEE точно определяет, как должно выполняться округление и что следует делать, если вычисления приводят к переполнению, исчезновению порядка или к необходимости извлечения квадратного корня из отрицательного числа. В распространенных моделях персональных компьютеров и рабочих станций используются несколько микропроцессоров, поддерживающих этот стандарт. Среди них — Intel начиная с серий 8087/80287/80387 и далее, а также процессоры серий Motorola 6888X.

Как можно было заметить выше, при сложении машинных чисел различной величины результат может оказаться точно равен одному из слагаемых. В этом случае вполне можно было бы, не изменяя результата, заменить меньшее число нулем.

Наименьшее число с плавающей точкой, которое при сложении с числом 1.0 дает результат, больший чем 1.0, называется **машинным эпсилоном** и обозначается $\varepsilon_{\text{маш}}$.

Машинный эпсилон определяет относительную погрешность арифметики компьютера. Если x и y два положительных числа с плавающей точкой и $x > y$, то их сумму можно записать в виде

$$x + y = x\Big(1 + \frac{y}{x}\Big).$$

Очевидно, что при $y/x < \varepsilon_{\text{маш}}$ сумма с плавающей точкой чисел x и y совпадает с x. Более тщательное исследование показывает, что относительная погрешность сложения чисел с плавающей точкой ограничена величиной $\varepsilon_{\text{маш}}$.

Для 32-битовой арифметики с плавающей точкой, удовлетворяющей стандарту IEEE, $\varepsilon_{\text{маш}} = 2^{-22} \approx 1.2 \times 10^{-7}$ при использовании округлений. Поэтому бессмысленно рассчитывать более, чем на семь верных десятичных знаков в любом результате вычисления с плавающей точкой или на то, что мы сумеем разрешить относительные различия, меньшие этого уровня.

Многие методы имеют входной параметр ε, задающий желаемую точность. Неразумно присваивать ему значение, меньшее чем $\varepsilon_{\text{маш}}$.

Оценим в заключение границу относительной погрешности при приближении числа x, числом x^* из множества $\mathcal{F}$, при условии, что основание системы счисления $\beta = 2$. Для простоты будем считать, что число x^* получается отбрасыванием "лишних" разрядов из мантиссы. Так как

$$x = \pm\Big(\frac{d_1}{2} + \frac{d_2}{2^2} + \cdots + \frac{d_t}{2^t} + \frac{d_{t+1}}{2^{t+1}} + \ldots\Big)2^e,$$

где $d_i = 0$ или 1, причем $d_1 = 1$, а

$$x^* = \pm\Big(\frac{d_1}{2} + \frac{d_2}{2^2} + \cdots + \frac{d_t}{2^t}\Big)2^e,$$

имеем

$$|x - x^*| = \Big(\frac{d_{t+1}}{2^{t+1}} + \frac{d_{t+2}}{2^{t+2}} \ldots\Big)2^e \le \Big(1 + \frac{1}{2} + \ldots\Big)2^e 2^{-(t+1)} = 2^{e-t}.$$

Заметим, что $|x| \geq \frac{1}{2}2^e$. Тогда

$$\left|\frac{x - x^*}{x}\right| \leq \frac{2^{e-t}}{2^{e-1}} = 2^{1-t}.$$

Можно показать, что при более точном способе округления, то есть при замене числа x числом $fl(x)$, справедлива более точная оценка $|x - fl(x)| \leq |x|2^{-t}$.

1.3 ПОГРЕШНОСТЬ ФУНКЦИИ

Пусть искомая величина $y = y(a_1, \ldots, a_n)$ зависит от параметров $a_1, \ldots, a_n$, которые принадлежат некоторой области G, а y^* — приближенное значение величины y.

Определение 1.3.1 ***Предельной абсолютной погрешностью*** $A(y^*)$ *назовем величину*

$$A(y^*) = \sup_{(a_1,\ldots,a_n)\in G} |y(a_1, \ldots, a_n) - y^*|.$$

Величина $\frac{A(y^*)}{|y^*|}$ *называется* ***предельной относительной погрешностью.***

Будем считать, что $G = \{(a_1, \ldots, a_n) : \ |a_i - a_i^*| \leq \Delta(a_i^*), \ i = 1, \ldots, n\}$ и $y^* = y(a_1^*, \ldots, a_n^*)$. Известно, что для гладкой функции $y(a_1, \ldots, a_n)$ справедливо равенство:

$$y(a_1, \ldots, a_n) - y^* = \sum_{j=1}^{n} \frac{\partial y}{\partial a_j}(a_j - a_j^*),$$

где частные производные вычислены в точке

$$(a_1^* + \theta(a_1 - a_1^*), \ldots, a_n^* + \theta(a_n - a_n^*)), \quad 0 \leq \theta \leq 1.$$

Тогда

$$|y(a_1, \ldots, a_n) - y^*| \leq A_0(y^*) = \sum_{j=1}^{n} \sup_G \left|\frac{\partial y}{\partial a_j}\right| \Delta(a_j^*). \tag{1.1}$$

При малых $\Delta(a_j^*)$ получим

$$\sup_G \left|\frac{\partial y}{\partial a_j}\right| \approx \left|\frac{\partial y(a_1^*, \ldots, a_n^*)}{\partial a_j}\right|,$$

поэтому обычно пользуются более простой по сравнению с (1.1) оценкой, которая однако не является совсем строгой:

$$|y(a_1, \ldots, a_n) - y^*| \leq \sum_{j=1}^{n} \left|\frac{\partial y(a_1^*, \ldots, a_n^*)}{\partial a_j}\right| \Delta(a_j^*). \tag{1.2}$$

Эту оценку называют **линейной оценкой погрешности**.

Рассмотрим некоторые примеры оценки погрешности при вычислении функций.
1) Пусть $y = y(a_1, \ldots, a_n) = \gamma_1 a_1 + \cdots + \gamma_n a_n$, где $\gamma_i = \pm 1$. Тогда

$$\left|\frac{\partial y}{\partial a_j}\right| = |\gamma_j| \equiv 1.$$

Следовательно, оценка (1.1) примет вид

$$|y(a_1,\dots,a_n)-y^*|\le \Delta(a_1^*)+\dots+\Delta(a_n^*)=A(y^*).$$

Отсюда следует правило: **предельная абсолютная погрешность суммы или разности равна сумме абсолютных погрешностей слагаемых.**

2) Рассмотрим теперь функцию $y=y(a_1,\dots,a_n)=a_1^{\lambda_1}\dots a_2^{\lambda_2}\dots a_n^{\lambda_n}$. Имеем тогда

$$\frac{\partial y(a_1^*,\dots,a_n^*)}{\partial a_j}=\lambda_j\frac{y^*}{a_j^*}.$$

Следовательно,

$$A(y^*)\approx\sum_{j=1}^{n}|\lambda_j||y^*|\frac{\Delta(a_j^*)}{|a_j^*|}.$$

Разделив это равенство на $|y^*|$, получим

$$\Big|\frac{y(a_1,\dots,a_n)-y^*}{y^*}\Big|\le\sum_{j=1}^{n}|\lambda_j|\delta(a_j^*).$$

В частности, если $y=a_1\cdot a_2$, или $y=a_1\cdot a_2^{-1}$, получим следующее правило: **предельная относительная погрешность произведения или частного равна сумме предельных относительных погрешностей.**

Иногда возникает так называемая **обратная задача:** с какой точностью надо задать приближенные значения аргументов $a_1^*,\dots,a_n^*$ функции $y=y(a_1,\dots,a_n)$, чтобы погрешность приближенного значения $y(a_1^*,\dots,a_n^*)$ не превосходила заданной величины ε?

Пусть истинное значение аргумента $(a_1,\dots,a_n)$ и приближенное $(a_1^*,\dots,a_n^*)$ лежат в некоторой выпуклой области G. Положим

$$C_j=\sup_G\Big|\frac{\partial y}{\partial a_j}\Big|.$$

Тогда справедлива оценка

$$|y(a_1,\dots,a_n)-y(a_1^*,\dots,a_n^*)|\le\sum_{j=1}^{n}C_j\Delta(a_j^*).$$

Следовательно, $\Delta(a_1^*),\dots,\Delta(a_n^*)$ достаточно выбрать так, чтобы

$$\sum_{j=1}^{n}C_j\Delta(a_j^*)\le\varepsilon.$$

Например, если взять $\Delta(a_1^*)=\dots=\Delta(a_n^*)=\delta$, то $\delta\le\epsilon/(C_1+\dots+C_n)$. Можно поступить по другому, потребовав, чтобы при всех j выполнялось равенство $C_j\Delta(a_j^*)=\varepsilon/n$. Тогда $\Delta(a_j^*)\le\varepsilon/(nC_j)$.

1.4 УМЕНЬШЕНИЕ ПОГРЕШНОСТИ ВЫЧИСЛЕНИЙ

В этом параграфе будут рассмотрены некоторые причины возникновения существенных погрешностей и случаи, когда можно избежать потери точности путем правильной организации вычислений.

Найдем предельную относительную погрешность разности чисел $y = a_1 - a_2$.

$$\delta(y^*) = \frac{\Delta(a_1^*) + \Delta(a_2^*)}{|y^*|} = \frac{\Delta(a_1^*) + \Delta(a_2^*)}{|a_1^* - a_2^*|}.$$

Если a_1^* и a_2^* близки, то $\delta(y^*)$ может оказаться довольно большой величиной. Например, пусть $a_1^* = 2000, \quad a_2^* = 2001$, и все значащие цифры в записи этих чисел верны в широком смысле. Тогда абсолютная погрешность $\Delta(a_j^*),\ j = 1, 2$ не превосходит единицы последнего разряда, то есть $\Delta(a_1^*) = \Delta(a_2^*) = 1$. Следовательно, $\delta(a_1^*) \approx \delta(a_2^*) = 1/2000 = 5 \cdot 10^{-4}$, а $\delta(a_1^* - a_2^*) = 2$, то есть относительная погрешность разности велика, хотя относительная погрешность исходных данных мала. Причина столь существенного возрастания относительной погрешности заключается в том, что пришлось вычитать близкие большие (относительно разности) числа. Так как числа большие, они были округлены с большой абсолютной погрешностью, которая оказалась сравнимой с разностью этих чисел. Такое явление иногда называют **катастрофической потерей верных знаков.** Поэтому при организации вычислительных алгоритмов следует по возможности избегать вычитания близких чисел. Покажем, как это можно сделать, на примере решения квадратного уравнения $ax^2 + bx + c = 0$, предполагая, что $a \neq 0$. Его корни определяются соотношениями

$$x_1 = \frac{-b - \sqrt{D}}{2a}, \quad x_2 = \frac{-b + \sqrt{D}}{2a}, \quad D = b^2 - 4ac. \tag{1.3}$$

Если b много больше чем $4ac$, то возникает опасность вычитания близких чисел в числителе одного из выражений (1.3) из-за того, что $\sqrt{D} \approx |b|$. Действительно, пусть в квадратном уравнении коэффициенты имеют значения: $a = 1,\ b = -320,\ c = 16$. Решая это уравнение "точно" получим $x_1 = 319.950,\ x_2 = 0.0500078$. Рассмотрим теперь множество чисел с плавающей точкой $\mathcal{F}$, характеризующееся следующими параметрами: $\beta = 10,\ t = 4,\ L = -10,\ U = 10$. Производя расчеты на ЭВМ, на которой реализована такая система чисел $\mathcal{F}$, получим

$$x_{1,2} = \frac{0.3200 \cdot 10^3 \pm \sqrt{0.1024 \cdot 10^6 - 0.6400 \cdot 10^2}}{0.2000 \cdot 10^1} = \frac{0.3200 \cdot 10^3 \pm 0.3198 \cdot 10^3}{0.2000 \cdot 10^1}.$$

Следовательно, $x_1 = 319.9,\ x_2 = 0.1$. Таким образом, если для первого корня относительная ошибка мала, то для второго она очень большая. Возможен другой метод решения квадратного уравнения, который позволяет избежать подобной потери точности. Один корень находится путем использования той формулы (1.3), которая не содержит операцию вычитания. Эту формулу можно записать в виде

$$x_1 = \frac{-b - \mathrm{sign}(b)\sqrt{D}}{2a}.$$

Для вычисления второго корня можно применить теорему Виета, согласно которой

$$x_2 = \frac{c}{a\, x_1}.$$

В заключение сделаем одно замечание относительно метода вычисления середины c отрезка $[a, b]$. Для точки c можно предложить две формулы

$$c = \frac{a+b}{2}, \tag{1.4}$$

$$c = a + \frac{b-a}{2}. \tag{1.5}$$

В том случае, когда из-за особенностей машинной арифметики проходится делать округления, это формулы могут дать различные результаты. Рассмотрим примеры, проясняющие ситуацию. Как и выше будем считать, что на ЭВМ реализована система чисел $\mathcal{F}$, и вычисления ведутся с округлениями. Если $a = 0.8882$, а $b = 0.8884$, то $a + b = 1.777$ и результат вычисления по формуле (1.4) равен 0.8885, то есть лежит вне отрезка $[a, b]$. В то же время $a + (b - a)/2 = 0.8882 + 0.0001 = 0.8883$, результат точен. Заметим, что если попытаться найти по формуле (1.4) середину отрезка $[0.8882, 0.8885]$, то получим правый конец отрезка.

С другой стороны, если $a = -0.8882$, а $b = 0.8884$, то $(a + b)/2 = 0.0001$. Таким образом, получается точный результат. Применение же формулы (1.5) дает $a + (b - a)/2 = -0.8882 + 0.8885 = 0.0003$. Примеры показывают, что в том случае, когда числа a и b имеют одинаковые знаки, предпочтительнее формула (1.5), а если знаки разные — (1.4).

В параграфе 5.1.1 будет рассматриваться метод, в котором потребуется находить середины отрезков в том случае, когда их концы a_n и b_n изменяются и сходятся к одному пределу. Тогда, если предел не равен нулю, числа a_n и b_n примут один знак и, поэтому, формула (1.5) предпочтительнее. Общий принцип состоит в том, что лучше строить формулы, в которых нужная величина представима в виде небольшой поправки к хорошему приближению.

Другой источник возникновения ошибок рассмотрим на следующем примере. Пусть требуется найти сумму чисел:

$$S = 0.2764 + 0.3944 + 1.475 + 26.46 + 1364.$$

Если сложить все числа, а затем округлить полученный результат до четырех значащих цифр, получим $S = 1393$. Будем считать теперь, что вычисления происходят на машине, в которой реализована описанная в предыдущем примере система с плавающей точкой $\mathcal{F}$. При вычислении на машине округления происходят после каждого сложения. Складывая на машине числа в порядке записи имеем $0.2764 + 0.3944 = 0.6708$, $0.6708 + 1.475 = 2.146$, $2.146 + 26.46 = 28.61$, $28.61 + 1364 = 1393$, то есть верный результат. Если же складывать числа в обратном порядке, получим $1364+26.46 = 1390$, $1390+1.1475 = 1391$, $1391+0.3944 = 1391$, $1391+0.2764 = 1391$. Этот пример еще раз иллюстрирует тот факт, что в "машинной арифметике" не выполняется закон ассоциативности операции сложения.

Анализ процесса вычислений показывает, что потеря точности возникла из-за того, что прибавления к большому числу малого (в случае, когда оно много меньше) не происходит, то есть сумма оказывается равной большему числу. Даже если таких малых чисел много, то на результат они все равно не повлияют, так как прибавляются по одному. Отсюда следует правило: сложение чисел надо проводить по мере их возрастания.

Еще более наглядный пример дает использование рядов для вычисления значения функции. Известно, что

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots. \tag{1.6}$$

Таблица 1.4

	$\left(\frac{2-\sqrt{3}}{2+\sqrt{3}}\right)^2$	$(2-\sqrt{3})^4$	$(7-4\sqrt{3})^2$	$97-56\sqrt{3}$
$\sqrt{3}\approx 1.7$	0.00657	0.00810	0.04000	1.8000
$\sqrt{3}\approx 1.73$	0.00524	0.00523	0.00640	0.12000
$\sqrt{3}\approx 1.732$	0.00516	0.00516	0.00518	0.00800
$\sqrt{3}\approx 1.7321$	0.00515	0.00515	0.00513	0.00240

Этот ряд знакочередующийся и сходится абсолютно при всех значениях x.

Вычислим, используя эту формулу, значение функции $\sin x$ при $x=\pi/6\approx 0.5236$ с точностью до 10^{-4}.

$$\sin 0.5236 \approx 0.5236 - 0.2393\cdot 10^{-1} + 0.3281\cdot 10^{-3} = 0.5000.$$

Здесь воспользовались тем, что в соответствии с признаком Лейбница для сходящегося знакочередующегося ряда, члены которого убывают по модулю, остаток ряда по абсолютной величине не превосходит модуля первого отброшенного члена. Таким образом, вычисленное значение синуса совпадает с точным значением (в пределах принятой точности).

Возьмем теперь $x=4\pi+\pi/6\approx 25.66$. Даже если учитывать члены ряда до 10^{-8} и приводить вычисления с восемью значащими цифрами, в результате аналогичных вычислений (то есть суммирование слева направо по формуле (1.6)) получим абсурдный результат $\sin 25.66\approx 24$.

В программах, использующих ряды могут быть приняты различные меры для предотвращения подобной потери точности. Для тригонометрических функций — использование формулы приведения позволяет свести аргумент к промежутку [0, 1]. Для экспоненты — выделение целой и дробной части степени $e^x = e^{n+a} = e^n\cdot e^a$, где n — целое, $0<a<1$. Тогда e^a можно вычислить с помощью ряда, а e^n — умножением.

Метод вычисления $\sin 25.66$ показывает, как плохо продуманный алгоритм может привести к неудовлетворительному результату. Трудность в данном случае удалось преодолеть путем изменения алгоритма.

Предыдущие примеры показывают, что, в отличие от "классической" математики, выбор алгоритма влияет на результаты вычислений. Рассмотрим еще один пример, показывающий, как форма записи выражения влияет на результат вычисления его значения.

Пусть требуется найти приближенное значение величины $a=\left(\frac{2-\sqrt{3}}{2+\sqrt{3}}\right)^2$. Избавляясь от иррациональности в знаменателе, и, проведя очевидные преобразования, получим четыре представления для a:

$$a=\left(\frac{2-\sqrt{3}}{2+\sqrt{3}}\right)^2 = (2-\sqrt{3})^4 = (7-4\sqrt{3})^2 = 97-56\sqrt{3}.$$

Значения величины a, вычисленные при использовании различных форм ее представления, и различных приближений числа $\sqrt{3}$ приведены в таблице 1.4. Более точные вычисления дают $a\approx 0.005154776$. Анализ приведенных результатов вычислений показывает, что самой простой, является последняя формула, но она при этом дает самые неточные результаты.

Рассмотрим еще один пример "плохого" алгоритма. Пусть требуется вычислить интеграл

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

где n — целое число. Интегрируя по частям, получаем рекуррентную формулу

$$I_n = 1 - nI_{n-1}, \quad n = 1, 2, \ldots, \quad I_0 = 1 - e^{-1}.$$

Казалось бы, что по указанной формуле возможно провести вычисления интеграла для различных значений n. Однако, с точки зрения вычислительной математики она совершенно непригодна. Например, в результате расчета по рекуррентной формуле получено $I_{20} = -30.1923949$, что, очевидно, не имеет ничего общего с истинным значением интеграла, которое положительно.

Причина в данном случае заключается в том, что погрешность задания начального значения I_0, которую избежать невозможно, при подсчете I_n возрастет в $n!$ раз. Алгоритм, в котором небольшое изменение исходных данных приводит к существенному изменению результата вычислений принято называть **неустойчивым алгоритмом**.

В данном случае, для вычисления интеграла при конкретном значении n можно применить специальные методы, так называемые квадратурные формулы, которые будут рассмотрены ниже.

Можно было поступить иначе. Рекуррентную формулу перепишем в виде

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

Положив при большом n значение интеграла равным 0, например, $I_{30} = 0$, найдем значения I_n по этой рекуррентной формуле. В результате получим, например, $I_{20} = 0.045545$, $I_{10} = 0.083877$, $I_1 = 0.367879$. У приведенных чисел все значащие цифры оказались точными. Таким образом, ошибка в I_{30} оказалась полностью подавленной. Это пример устойчивого алгоритма. На сей раз при вычислении I_0 ошибка, возникшая при задании I_n, убывает в $n!$ раз.

Однако для некоторых задач "правильные" ответы нельзя получить никаким алгоритмом, потому что сами решения этих задач чувствительны к малым ошибкам, допущенным при представлении исходных данных. Такие задачи называют **неустойчивыми** или **плохо обусловленными** задачами.

Рассмотрим опубликованный в 1963 году Уилкинсоном пример такой задачи. Пусть

$$P(x) = (x-1)(x-2)\ldots(x-19)(x-20) = x^{20} - 210x^{19} + \ldots.$$

Корнями этого полинома являются числа $1, 2, \ldots, 19, 20$. Предположим, что коэффициент при x^{19} изменен с -210 на $-210 + 2^{-23}$. Какое воздействие это малое изменение произведет на корни полинома? В результате очень точных вычислений (11 разрядов) были получены следующие результаты, округленные до трех знаков:

$$1.00, \quad 2.00, \quad 3.00, \quad 4.00, \quad 5.00, \quad 6.00, \quad 7.00, \quad 8.01, \quad 8.92,$$
$$10.1 \pm 0.64i, \quad 11.8 \pm 1.65i, \quad 14.0 \pm 2.52i, \quad 16.7 \pm 2.81i, \quad 19.5 \pm 1.94i, \quad 20.8.$$

Таким образом, малое изменение в одном из коэффициентов привело к тому, что десять корней стали комплексными, причем два из них отодвинулись от действительной оси более, чем на 2.81. Отметим еще раз, что причина, по которой эти корни

так сильно изменились, лежит не в ошибках округления и не связана с выбором алгоритма вычислений. Суть в чувствительности самой задачи.

Еще один пример неустойчивой задачи. У системы

$$\begin{cases} x_1 + 10x_2 = 11, \\ 1000x_1 + 10001x_2 = 11001 \end{cases} \tag{1.7}$$

имеется единственное решение $x_1 = x_2 = 1$. В том случае, когда в правой части первого уравнения допущена небольшая погрешность равная $0,001$, система приобретает вид

$$\begin{cases} x_1 + 10x_2 = 11.001, \\ 1000x_1 + 10001x_2 = 11001 \end{cases} \tag{1.8}$$

и ее решение $x_1 = 11.001,\ x_2 = 0$ существенно отличается от решения системы (1.7).

1.5 ЗАДАЧИ К ГЛАВЕ 1

1.5.1 Примеры решения задач

1. Пусть положительное число a^* имеет n верных десятичных знаков в узком смысле и является приближенным значением точного числа a. Показать, что для относительной погрешности $\delta(a^*)$ справедлива оценка

$$\delta(a^*) \le \frac{10^{1-n}}{2\alpha_m},$$

где α_m — первая значащая цифра числа a^*. Кроме того,

$$\frac{|a - a^*|}{a} \le \frac{10^{1-n}}{\alpha_m}.$$

Решение. Ограничимся доказательством второго неравенства. Пусть число $a^* = \alpha_m 10^m + \alpha_{m-1}10^{m-1} + \cdots + \alpha_{m-n+1}10^{m-n+1} + \ldots,\ \ \alpha_m \ge 1$ является приближенным значением точного числа a и имеет n верных знаков. Тогда по определению имеем: $|a - a^*| \le \Delta(a^*) = 0.5 \cdot 10^{m-n+1}$. Отсюда следует, что

$$a \ge a^* - 0.5\cdot 10^{m-n+1} \ge \alpha_m 10^m - 0.5\cdot 10^{m-n+1} = 0.5\cdot 10^m(2\alpha_m - 10^{1-n}) \ge 0.5\cdot 10^m(2\alpha_m - 1). \tag{1.9}$$

Так как $2\alpha_m - 1 = \alpha_m - (\alpha_m - 1) \ge \alpha_m$, из (1.9) следует $a \ge 0.5 \cdot 10^m \alpha_m$. Тогда

$$\frac{|a - a^*|}{a} \le \frac{0.5 \cdot 10^{m-n+1}}{0.5 \cdot 10^m \alpha_m} = \frac{10^{1-n}}{\alpha_m}.$$

Утверждение доказано.

2. Необходимо вычислить объем сферического слоя, если известны радиус внутренней сферы r и толщина слоя h. Предложите формулу для вычисления, если $h \ll r$. Приведите пример, подтверждающий ваши выводы.

Решение. Очевидно, что искомый объем V равен

$$V = \frac{4}{3}\pi\big((r+h)^3 - r^3\big).$$

Так как по условию h мало по сравнению с r, вычитание двух близких чисел может привести к большой потере точности. Поэтому выгоднее вычислять результат по формуле

$$V = \frac{4}{3}\pi\big(3r^2h + 3rh^2 + h^3\big).$$

Пусть, например, вычисления проводятся с четырьмя десятичными знаками, $r = 1.000,\ h = 1.000 \cdot 10^{-2}$. Будем для простоты вычислять $\widetilde{V} = \frac{3}{4\pi}V$. Точное значение величины $\widetilde{V} = 0.030301$. Вычисление по первой формуле даст $1.010 \cdot 1.010 = 1.0201 \approx 1.020,\ \ 1.020 \cdot 1.01 = 1.0302 \approx 1.030,\ \ \widetilde{V^*} = 1.030 - 1.000 = 0.030$. В итоге, абсолютная ошибка, полученная по первой формуле $\Delta(\widetilde{V^*}) = 3.01 \cdot 10^{-4}$.

Применение второй формулы дает $3 \cdot 1.000^2 \cdot (1.000 \cdot 10^{-2}) = 3.000 \cdot 10^{-2},\ \ 3 \cdot 1.000 \cdot (1.000 \cdot 10^{-2})^2 = 3.000 \cdot 10^{-4},\ \ (1.000 \cdot 10^{-2})^3 = 1.000 \cdot 10^{-6},\ \ 3.000 \cdot 10^{-2} + 3.000 \cdot 10^{-4} = 3.030 \cdot 10^{-2},\ \ 3.030 \cdot 10^{-2} + 1.000 \cdot 10^{-6} = 3.0301 \cdot 10^{-2} \approx 3.030 \cdot 10^{-2}$. Таким образом, вторая формула дает $\Delta(\widetilde{V^*}) = 1.00 \cdot 10^{-6}$, то есть абсолютная ошибка, полученная по второй формуле, примерно в 300 раз меньше.

3. Показать, что если $y = \ln x$, то $\Delta(y^*) = \delta(x^*)$.

Решение. Заметим, что если $y = f(x)$ — дифференцируемая функция и абсолютная погрешность аргумента мала, то по формуле Лагранжа

$$\Delta(y^*) = |f(x) - f(x^*)| = |f'(\xi)||x - x^*| \approx |f'(x^*)||x - x^*| = |f'(x^*)|\Delta(x^*),$$

где ξ — некоторая точка, лежащая между x и x^*. Так как $(\ln x)' = 1/x$, имеем

$$\Delta(y^*) = \Delta(x^*)/x^* = \delta(x^*).$$

4. Вычислить значение $z = \ln(10.3 + \sqrt{4.4})$, считая верными в узком смысле все знаки приближенных чисел $x^* = 10.3,\ y^* = 4.4$.

Решение. Число y^* имеет относительную погрешность $\delta(y^*) = \frac{0.05}{4.4} \approx 0.012$, поэтому $\sqrt{y^*}$ имеет относительную погрешность 0.006. Тогда имеем $\sqrt{y^*} = \sqrt{4.4} \approx 2.1,\ \Delta(\sqrt{y^*}) = 2.1 \cdot 0.006 = 0.013$.

Абсолютная погрешность суммы $x^* + \sqrt{y^*} = 10.3 + 2.1 = 12.4$ оценивается величиной $0.05 + 0.013 = 0.063$ и ее относительная погрешность равна $0.063/12.4 \approx 0.005$. В соответствии с результатом предыдущей задачи, такой же будет абсолютная погрешность логарифма, то есть $\Delta(z^*) = 0.005$, а $z^* = \ln(12.4) = 2.517$. Здесь результат имеет только три верных знака, последняя цифра 7 является сомнительной.

1.5.2 Задачи

1. Докажите, что если приближенные числа имеют один знак, то относительная погрешность суммы не превосходит наибольшей из относительных погрешностей слагаемых.

2. Докажите, что если $y = a^x,\ a > 0$, то $\delta(y^*) = \Delta(x^*)\ln a$.

3. Рассмотрим воображаемую систему с плавающей точкой, состоящую из следующих чисел:

$$\mathcal{F} = \{\pm b_1 b_2 b_3 \cdot 2^{\pm y}\},$$

где каждое число b_2, b_3 и y принимает одно из значений 0 или 1, а b_1 всегда равно 1 за исключением случая, когда $b_1 = b_2 = b_3 = y = 0$.

а) Покажите, что множество $\mathcal{F}$ содержит 25 элементов.

б) Изобразите фрагмент вещественной оси с нанесенными на нее элементами множества $\mathcal{F}$.

в) Чему равно $\varepsilon_{\text{маш}}$, каковы наибольшее и наименьшее числа среди положительных чисел множества $\mathcal{F}$?

4. При разработке программ для ЭВМ, включаемых в библиотеки стандартных программ, возникает немало проблем. Проиллюстрируем одну из них на примере вычисления евклидовой нормы вектора $\mathbf{x} = (x_1, \ldots, x_m)$, которая определяется по формуле

$$\|\mathbf{x}\| = \Big(\sum_{i=1}^{m} x_i^2\Big)^{1/2}. \tag{1.10}$$

Покажите, что при непосредственном вычислении на ЭВМ по формуле (1.10) может оказаться, что $\mathbf{x} \neq 0$, в то время как $\|\mathbf{x}\| = 0$. Для иллюстрации можно воспользоваться системой чисел из предыдущей задачи.

Изменится ли результат, если для ненулевого вектора вычисления проводить по формуле

$$\|\mathbf{x}\| = x_{max}\Big(\sum_{i=1}^{m} \bar{x}_i^2\Big)^{1/2},$$

где $x_{max} = \max\limits_{i=1,\ldots,m} |x_i|, \quad \bar{x}_i = x_i/x_{max}$?

5. Всегда ли верно утверждение, что $fl\Big(\frac{a+b}{2}\Big) \in [a, b]$?

Ответ. Нет. В задаче 3 взять $a = 0.51,\ b = 0.52$. Тогда

$$fl\Big(\frac{a+b}{2}\Big) = 0.5$$

.

6. Пусть $y = \sqrt{2} - 1$. Тогда эту величину можно записать в виде $y = (\sqrt{2} + 1)^{-1}$. Какая из этих двух формул более чувствительна к погрешности при приближенном задании $\sqrt{2}$ в виде конечной десятичной дроби?

Указание. Сравнить модули производных функций $x - 1$ и $(x + 1)^{-1}$.

7. Пусть ищется наименьший корень уравнения $x^2 - 140x + 1 = 0$. Вычисления проводятся в десятичной системе счисления, причем в мантиссе числа после округления удерживаются 4 разряда. Какая из двух формул

$$x = 70 - \sqrt{4899} \quad \text{или} \quad x = \frac{1}{70 + \sqrt{4899}}$$

дает более точный результат?

8. С каким числом верных знаков надо взять $\lg 2$, для того чтобы вычислить корни уравнения $x^2 - 2x + \lg 2 = 0$ с четырьмя верными знаками?

9. Найти абсолютную погрешность определителя,

$$\begin{vmatrix} 2.5 \pm 0.1 & -1.1 \pm 0.2 \\ 4.2 \pm 0.2 & 3.0 \pm 0.1 \end{vmatrix}.$$

10. Пусть на ЭВМ вычисляется сумма

$$S_{1000000} = \sum_{i=1}^{1000000} \frac{1}{i^2}.$$

По какому алгоритму

$$S_0 = 0, \quad S_n = S_{n-1} + \frac{1}{n^2}, \quad n = 1, \ldots, 1000000,$$

или

$$\tilde{S}_{1000000} = 0, \quad \tilde{S}_{n-1} = \tilde{S}_n + \frac{1}{n^2}, \quad n = 1000000, \ldots, 1,$$

следует считать, чтобы суммарная вычислительная погрешность была меньше?

11. Пусть $|x| < 1$. В каком порядке лучше всего вычислять сумму $\sum\limits_{i=1}^{n} x^i$ с точки зрения уменьшения вычислительной погрешности.

1.5.3 Примеры тестовых вопросов к главе 1

1. Укажите число значащих цифр числа 0.0543210

2. Пусть $\mathcal{F}$ нормализованная система чисел с плавающей точкой. Она характеризуется четырьмя параметрами: основанием системы счисления β, точностью или разрядностью t и интервалом показателей $[L, U]$. Каждое число x с плавающей точкой, принадлежащее $\mathcal{F}$, имеет значение

$$x = \pm\Big(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t}\Big)\beta^e,$$

где целые числа $d_1, \ldots, d_t$ удовлетворяют неравенствам $0 \le d_i \le \beta - 1, (i = 1, \ldots, t)$, $d_1 \neq 0$ и $L \le e \le U$.

Пусть $\beta = 2$, $t = 3$, $L = -3$, $U = 3$. Чему равно в этой системе наименьшее положительное число? Ответ привести в виде десятичной дроби.

3. Требуется вычислить величину $z = 0.5\sin(x+y)$ с абсолютной погрешностью не превосходящей 0.004. Предполагается, что максимальная допустимая погрешность каждого из аргументов одинакова. Какова при этом максимально допустимая погрешность аргументов? Ответ привести в виде десятичной дроби.

4. В каких из приведенных ниже случаях относительная погрешность значения функции может быть больше относительных погрешностей значений аргументов, если аргументы заданы так, что их относительные погрешности равны?

а) $a + b, \quad a, b > 0;$

б) $ab;$

в) $\dfrac{a}{b}, \quad b \neq 0;$

г) $\dfrac{\sqrt{a}}{b^3}, \quad b \neq 0;$

д) $a + b, \quad a > 0, b < 0;$

е) $a - b, \quad a > 0, b < 0.$

5. Какое из чисел $a = 33.3 \pm 0.1$ или $b = 2.22 \pm 0.01$ задано точнее?

6. Пусть коэффициенты p и q квадратного уравнения $x^2 + 2px + q = 0$ заданы с абсолютной погрешностью $\triangle$. Известно, что уравнение имеет действительные корни. Какое из приведенных ниже выражений позволяет оценить абсолютную погрешность $\triangle(x_{1,2})$ корней $x_{1,2}$ этого уравнения?

а) $\triangle(x_{1,2}) = \left(1 + \dfrac{2|p| + 1}{2\sqrt{p^2 - q}}\right) \cdot \triangle;$

б) $\triangle(x_{1,2}) = \triangle;$

в) $\triangle(x_{1,2}) = \left(|p| + \dfrac{|p| + 1}{\sqrt{p^2 - q}}\right) \cdot \triangle;$

г) $\triangle(x_{1,2}) = \dfrac{|q|}{|p|} \cdot \triangle;$

д) $\triangle(x_{1,2}) = \left(|p| + \sqrt{p^2 - q}\right) \cdot \triangle;$

е) $\triangle(x_{1,2}) = \dfrac{|p|}{|q|} \cdot \triangle.$

2 ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ ЛИНЕЙНОЙ АЛГЕБРЫ

2.1 ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

В этом параграфе будет рассмотрен один класс численных методов решения систем линейных алгебраических уравнений:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{2.1}$$

где $\mathbf{A}$ — квадратная матрица размера $m \times m$, причем определитель $\det(\mathbf{A}) \neq 0$, $\mathbf{x} = (x_1, \ldots, x_m)^T$, $\mathbf{b} = (b_1, \ldots, b_m)^T$. Здесь индекс "T" — означает операцию транспонирования.

Для большинства задач характерна ситуация, когда m велико, поэтому не любой метод решения системы может быть использован для нахождения решения. Действительно, предположим, что решение системы находится с помощью правила Крамера, согласно которому $x_i = \Delta_i/\Delta$. Здесь $\Delta = \det \mathbf{A}$, а Δ_i — определитель матрицы, полученной из матрицы $\mathbf{A}$ заменой i-го столбца столбцом $\mathbf{b}$. Таким образом, для нахождения решения потребуется вычислить $m + 1$ определитель. Если при этом определитель вычислять, исходя из того, как он был введен в курсе линейной алгебры, то есть как сумму произведений элементов, взятых с учетом определенного знака из различных строк и различных столбцов, то придется вычислить $m!$ слагаемых. Значит, для вычисления только одного определителя потребуется порядка $(m-1)m!$ арифметических операций ($m-1$ умножение для вычисления $m!$ слагаемых). Пусть $m = 30$. Тогда, используя формулу Стирлинга[1] для оценки факториала, имеем:

$$(m-1)m! \approx (m-1)\left(\frac{m}{e}\right)^m \cdot \sqrt{2\pi m} \approx 29\left(\frac{30}{2.7}\right)^{30} \cdot \sqrt{60\pi} > 29 \cdot 10^{30} \cdot \sqrt{180} > 365 \cdot 10^{30}.$$

Если ЭВМ выполняет даже 10^{16} арифметических операций в секунду[2], то для вычисления только одного определителя потребуется

$$\frac{365 \cdot 10^{30}}{365 \cdot 24 \cdot 3600 \cdot 10^{16}} = \frac{10^{12}}{864} > 10^9 \text{ лет.}$$

Методы численного решения системы делят на два класса: **прямые и итерационные.**

[1] $m! \approx \left(\frac{m}{e}\right)^m \cdot \sqrt{2\pi m}$ при больших m.

[2] На момент написания этой книги наилучшие суперкомпьютеры имели теоретическую пиковую производительность менее 3-х петафлопс, то есть менее $3 \cdot 10^{15}$ операций с плавающей точкой. Персональный компьютер, например, на базе Intel Core i7-975 XE 3,33 ГГц мог обеспечить только около 50 гигафлопс, то есть $5 \cdot 10^{10}$ операций с плавающей точкой.

В **прямых методах** решение $\mathbf{x}$ системы находится за конечное число арифметических операций. Заметим, что при реализации на ЭВМ, как правило, точное решение не получается из-за наличия ошибок округления.

Сопоставление различных методов производится по числу арифметических операций, необходимых для получения решения. Предпочтение отдается тому методу, который требует меньшего числа операций.

Итерационные методы или **методы последовательных приближений** состоят в том, что решение $\mathbf{x}$ системы находится как предел при $n \to \infty$ последовательности $\mathbf{x}^{(n)}$, где n - номер итерации. Как правило, за конечное число шагов предел не достигается. Обычно задается некоторая точность $\varepsilon > 0$ и вычисления проводятся до тех пор, пока $\|\mathbf{x} - \mathbf{x}^{(n)}\| < \varepsilon$.[3]

В настоящее время прямые методы применяют обычно для решения систем до порядка 10^3, а итерационные — до порядка 10^6.

2.1.1 Метод Гаусса

Прямые методы типа метода Гаусса не требуют от матрицы системы специального вида и применяются для систем умеренного порядка.

Метод Гаусса основан на приведении матрицы системы к треугольному виду.

Вычтем из второго уравнения системы $\mathbf{Ax} = \mathbf{b}$ первое, умноженное на такое число, чтобы уничтожился коэффициент при x_1, то есть на $C_{21} = a_{21}/a_{11}$. Затем, таким же образом вычтем из третьего уравнения первое, умножив его на $C_{31} = a_{31}/a_{11}$ и так далее. Тогда обратятся в ноль все элементы первого столбца матрицы системы, лежащие ниже главной диагонали. Затем, при помощи второго уравнения исключим из третьего, четвертого и так далее уравнений коэффициенты, образующие второй столбец матрицы системы. Последовательно продолжая этот процесс, обратим в ноль все элементы матрицы системы, лежащие ниже главной диагонали.

Запишем общие формулы процесса. Пусть проведено исключение коэффициентов из $k-1$ столбца. Тогда ненулевые элементы ниже главной диагонали могут быть только в уравнениях, начиная с k-го:

$$\sum_{j=k}^{m} a_{ij}^{(k)} x_j = b_i^{(k)}, \quad k \le i \le m. \tag{2.2}$$

Умножая k - строку на число

$$C_{pk} = \frac{a_{pk}^{(k)}}{a_{kk}^{(k)}}, \quad p > k \tag{2.3}$$

и вычитая из p-й строки получим, что k-ый элемент p-ой строки обратится в 0, а остальные изменятся по формулам:

$$a_{pl}^{(k+1)} = a_{pl}^{(k)} - C_{pk} a_{kl}^{(k)}, \tag{2.4}$$

$$b_p^{(k+1)} = b_p^{(k)} - C_{pk} b_k^{(k)}, \quad k < p, l \le m. \tag{2.5}$$

Произведя вычисления по этим формулам, при всех указанных индексах, обратим в ноль все элементы k-го столбца, начиная с $k+1$-го. Назовем такое исключение **циклом** процесса. В результате выполнения всех циклов получим треугольную матрицу. Это так называемый **прямой ход исключения** или **прямой ход метода Гаусса** .

[3]Обычно, в реальных расчетах проверка этого условия затруднительна. Поэтому, как будет показано далее, на практике используются другие критерии прекращения итераций.

Таким образом, после проведения прямого хода исключения, система (2.1) примет вид

$$\sum_{j=k}^{m} a_{ij}^{(k)} x_j = b_i^{(k)}, \quad 1 \le k \le m, \quad a_{1j}^{(1)} = a_{1j}, \quad b_1^{(1)} = b_1. \tag{2.6}$$

Заметим, что в полученной матрице ниже главной диагонали стоят нули. На эти места можно записать числа C_{pk}, которые, как увидим ниже, могут понадобиться. Полученная треугольная система легко решается **обратным ходом метода Гаусса** по формулам:

$$x_k = \frac{\left(b_k^{(k)} - \sum\limits_{j=k+1}^{m} a_{kj}^{(k)} x_j\right)}{a_{kk}^{(k)}}, \quad k = m, m-1, \ldots, 1.$$

Оценим приближенно количество операций, которые надо совершить для нахождения решения системы уравнений по методу Гаусса, считая размерность системы достаточно большой. Первый шаг прямого хода, то есть процесс обращения в нуль всех элементов первого столбца, лежащих ниже главной диагонали, потребует $m-1$ деление, $m(m-1)$ умножений и столько же вычитаний. Таким образом, общее число операций на первом шаге равно $Q_1 = 2m(m-1) + m - 1 = 2(m-1)^2 + 3(m-1)$. Аналогично на k-ом шаге прямого хода потребуется $Q_k = 2(m-k)^2 + 3(m-k)$ операций. Общее число операций прямого хода составляет

$$Q = \sum_{k=1}^{m-1} Q_k = 2\sum_{k=1}^{m-1}(m-k)^2 + 3\sum_{k=1}^{m-1}(m-k) =$$
$$= 2\sum_{k=1}^{m-1} k^2 + 3\sum_{k=1}^{m-1} k = \frac{2m(m-1)(2m-1)}{6} + \frac{3m(m-1)}{2} \approx \frac{2}{3}m^3.$$

Легко получить, что обратный ход метода Гаусса требует всего $O(m^2)$ арифметических операций, что при больших m пренебрежимо мало по сравнению с прямым ходом. Таким образом, для реализации метода Гаусса потребуется приблизительно $2m^3/3$ арифметических операций, причем основное их количество приходится на прямой ход.

Прямой ход метода Гаусса, описанный выше, нельзя будет произвести, если в процессе расчетов на главной диагонали окажется нулевой элемент $a_{kk}^{(k)} = 0$. В том случае, когда в k-ом столбце матрицы промежуточной системы (2.2) все элементы, расположенные на и ниже главной диагонали равны нулю, имеем $\det \mathbf{A} = 0$. Если же матрица $\mathbf{A}$ не вырождена, то в части столбца начиная с главной диагонали и ниже обязательно найдется ненулевой элемент. Поэтому перестановкой строк можно переместить ненулевой элемент на главную диагональ и продолжить расчет.

Таким образом, с математической точки зрения метод Гаусса гарантированно приводит к решению системы в случае невырожденной матрицы. Однако на практике, в связи с наличием в процессе вычислений ошибок округления, возможно возникновение катастрофических ошибок. Для того, чтобы разобраться в сути дела рассмотрим, следуя [27], следующий пример.

Выберем систему

$$\begin{cases} -10^{-5} x_1 + x_2 = 1, \\ 2x_1 + x_2 = 0, \end{cases} \tag{2.7}$$

решение которой равно $x_1 = -0.4999975\ldots, \quad x_2 = 0.999995\ldots$

Решим эту систему методом Гаусса, предполагая, что компьютер выполняет операции над десятичными числами, причем под мантиссу отводятся четыре разряда. Имеем

$$C_{21} = \frac{0.2 \cdot 10^1}{-0.1 \cdot 10^{-4}} = -0.2 \cdot 10^6,$$

$$a_{22}^{(2)} = a_{22} - C_{21}a_{12} = 0.1 \cdot 10^1 - (-0.2 \cdot 10^6)(0.1 \cdot 10^1) = 0.2000|01 \cdot 10^6 = 0.2 \cdot 10^6,$$

$$b_2^{(2)} = 0 - (-0.2 \cdot 10^6)(0.1 \cdot 10^1) = 0.2 \cdot 10^6.$$

Здесь | отделяет знаки в записи числа, которые отбрасываются компьютером в связи с размером мантиссы. Следовательно, ошибка округления возникла только при определении коэффициента $a_{22}^{(2)}$. Обратный ход метода Гаусса дает

$$x_2 = \frac{b_2^{(2)}}{a_{22}^{(2)}} = \frac{0.2 \cdot 10^6}{0.2 \cdot 10^6} = 0.1 \cdot 10^1, \quad x_1 = \frac{b_1 - a_{12}x_2}{a_{11}} = \frac{0.1 \cdot 10^1 - 0.1 \cdot 10^1}{-0.1 \cdot 10^{-4}} = 0.$$

Найденное значение x_2 хорошо согласуется с точным значением, в то время как x_1 не имеет с точным значением ничего общего. Заметим, что в процессе вычислений была допущена только одна ошибка округления в шестом десятичном знаке мантиссы при нахождении $a_{22}^{(2)}$, в то время как все остальные операции выполнены точно.

Для того, чтобы ответить на вопрос, почему сталь малая ошибка округления привела к катастрофическому искажению решения, воспользуемся принципом **обратного анализа ошибок**, который заключается в следующем. Анализируется не какая допущена ошибка, а какая задача решалась на самом деле в предположении, что в процессе решения ошибки округления отсутствуют. В исследуемом примере величина $0.000001 \cdot 10^6$, которая отброшена при нахождении $a_{22}^{(2)}$, совпадает с коэффициентом a_{22}. Таким образом, из формулы для вычисления $a_{22}^{(2)}$ следует, что сосчитанное значение этого коэффициента, а, значит и решения, было бы точно таким же, если бы значение a_{22} равнялось нулю. Следовательно, при вычислении на нашем компьютере было найдено точное решение системы

$$\begin{cases} -10^{-5}x_1 + x_2 = 1, \\ 2x_1 = 0, \end{cases} \tag{2.8}$$

которая имеет совершенно другое решение.

Причиной ошибки явился большой множитель C_{21}, который из-за малой длины мантиссы не позволил коэффициенту a_{22} внести свой вклад в сумму, определяющую $a_{22}^{(2)}$. В свою очередь, этот множитель возник из-за малости коэффициента a_{11} по сравнению с a_{21}. Заметим, что если переставить местами уравнения, то есть рассмотреть систему

$$\begin{cases} 2x_1 + x_2 = 0, \\ -10^{-5}x_1 + x_2 = 1, \end{cases}$$

то метод Гаусса дает

$$C_{21} = \frac{-0.1 \cdot 10^{-4}}{0.2 \cdot 10^1} = -0.5 \cdot 10^{-5},$$

$$a_{22}^{(2)} = a_{22} - C_{21}a_{12} = 0.1 \cdot 10^1 - (-0.5 \cdot 10^{-5})(0.1 \cdot 10^1) = 0.1000|05 \cdot 10^1 = 0.1 \cdot 10^1,$$

$$b_2^{(2)} = 0.1 \cdot 10^1 - (-0.5 \cdot 10^{-5})0 = 0.1 \cdot 10^1,$$

$$x_2 = \frac{0.1 \cdot 10^1}{0.1 \cdot 10^1} = 1, \quad x_1 = \frac{0 - (0.1 \cdot 10^1)1}{0.2 \cdot 10^1} = -0.5.$$

Теперь решение довольно хорошо согласуется с точным решением.

Таким образом, если элемент на главной диагонали $a_{kk}^{(k)}$ мал по сравнению с другими элементами k-го столбца, расположенными ниже главной диагонали, в процессе исключения k-я строка умножается на большие числа C_{pk}, что может привести к большим ошибкам округления. Поэтому, для избежания подобного эффекта, метод Гаусса модифицируют следующим образом. Каждый цикл процесса вычислений начинают с перестановки строк. Среди элементов $a_{sk}^{(k)}$, $s = k, k+1, \ldots, m$ находят наибольший по модулю, который называют **главным** или **ведущим**, и перестановкой строк выводят его на главную диагональ, после чего выполняют цикл исключения. Такая модификация алгоритма называется **методом Гаусса с выбором главного элемента**.

В методе Гаусса с выбором главного элемента погрешности округлений обычно существенно меньше, чем в случае, когда выбор главного элемента отсутствует.

Погрешность можно еще уменьшить, если выбирать на каждом шаге максимальный элемент не только в столбце, но и по всей матрице. Однако, программа при этом существенно усложняется, хотя обычно точность растет не очень существенно. Следует подчеркнуть, что не всегда выбор главного элемента в столбце обеспечивает удовлетворительную точность. Чтобы проиллюстрировать это достаточно рассмотреть следующий пример

$$\begin{cases} 10x_1 - 10^6 x_2 = -10^6, \\ 2x_1 + x_2 = 0. \end{cases} \tag{2.9}$$

Выбор главного элемента в столбце для этой системы не требуется. Легко убедиться, что в результате вычислений на описанном выше компьютере придется столкнуться с той же самой проблемой, с которой встретились при решении системы (2.7). Это не удивительно, если заметить, что система (2.9) получилась из системы (2.7) умножением первого уравнения на -10^6. Выбор главного элемента во всей матрице системы (2.9) приведет к перестановке первого и второго столбцов матрицы [4]. В результате вычислений теперь получится результат близкому к правильному.

Стратегия выбора главного элемента в столбце дает приемлемые результаты, если матрица коэффициентов масштабирована так, что максимальные по модулю элементы в каждой строке и каждом столбце имеют одинаковый порядок величин. В настоящее время не существует алгоритма масштабирования, однако в некоторых случаях известно какие строки и столбцы нуждаются в предварительном масштабировании. Например, для системы (2.9) следует пронормировать первую строку так, чтобы ее максимальный элемент стал примерно равным единице. Тогда элемент a_{11} станет малым и стратегия выбора главного элемента в столбце приведет в перестановке строк, после чего будет получен достаточно хороший результат.

Контроль вычислений можно вести по величине **невязки**:

$$r_k = b_k - \sum_{i=1}^{m} a_{ki} x_i, \quad 1 \le k \le m.$$

Если величины r_k велики, то, как правило, это означает, что решение найдено неверно[5].

[4]Перестановке столбцов матрицы соответствует переименование переменных. x_2 станет теперь первой переменной, а x_1 — второй.

[5]Ниже в параграфе 2.1.4 будет приведен пример, показывающий, что не всегда малая величина невязки означает, что решение найдено достаточно точно, а большая величина невязки свидетельствует о плохом приближении к решению. Таким образом, к выводам, сделанным по значению невязки, следует относиться с осторожностью.

Метод Гаусса позволяет легко найти определитель и обратную матрицу.

Действительно, так как все описанные действия по преобразованию матрицы системы не меняют модуля определителя, а результирующая матрица имеет треугольный вид и, значит, ее определитель равен произведению диагональных элементов, получим $\det \mathbf{A} = \pm \prod_{k=1}^{m} a_{kk}^{(k)}$. Для определения знака надо только подсчитать число перестановок строк. Так как каждая перестановка строк меняет знак определителя, в случае четного числа перестановок выбирается знак плюс, в противном случае — минус.

Заметим, что при таком вычислении определителя число арифметических операций примерно равно $\frac{2}{3}m^3$.

Для нахождения обратной матрицы $\mathbf{A}^{-1}$ воспользуемся равенством $\mathbf{A}\mathbf{A}^{-1} = \mathbf{E}$, где $\mathbf{E}$ — единичная матрица. Из него следует, что если рассматривать каждый столбец обратной матрицы как вектор, то k-ый столбец является решением системы (2.1) со специальной правой частью. В ней все компоненты равны нулю, кроме одной k-ой, которая равна единице. Значит, надо решить m систем с одинаковой матрицей $\mathbf{A}$ и различными правыми частями. При этом приведение матрицы $\mathbf{A}$ к диагональному виду делают один раз. В дальнейшем при помощи чисел C_{sk} по формулам (2.5) преобразуются правые части и затем совершают обратный ход метода Гаусса.

Замечание В том случае, когда из-за округлений невязка $\mathbf{r}$ получается большой, ее можно уменьшить, основываясь на следующих соображениях. Обозначим полученное в результате применения метода Гаусса приближенное значение решения через $\mathbf{x}^{(1)}$ и соответствующую невязку через $\mathbf{r}^{(1)}$. Постараемся определить разность $\Delta\mathbf{x}^{(1)}$ между точным решением $\mathbf{x}$ и приближенным решением $\mathbf{x}^{(1)}$. Имеем

$$\mathbf{A}\Delta\mathbf{x}^{(1)} = \mathbf{A}(\mathbf{x} - \mathbf{x}^{(1)}) = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^{(1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(1)} = \mathbf{r}^{(1)}.$$

Таким образом, вектор $\Delta\mathbf{x}^{(1)}$ является решением системы линейных алгебраических уравнений $\mathbf{A}\Delta\mathbf{x}^{(1)} = \mathbf{r}^{(1)}$. Решая эту систему, находим с точностью до ошибок округления поправку $\Delta\mathbf{x}^{(1)}$ к решению $\mathbf{x}^{(1)}$. Тогда, в качестве нового приближения к решению берем $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \Delta\mathbf{x}^{(1)}$. В случае необходимости, это решение можно уточнить, используя аналогичные рассуждения.

2.1.2 Метод квадратного корня

Метод предназначен для решения системы уравнений (2.1) с симметричной матрицей. Он основан на представлении матрицы $\mathbf{A}$ в виде произведения

$$\mathbf{A} = \mathbf{S}^T\mathbf{D}\mathbf{S}, \tag{2.10}$$

где $\mathbf{S}$ - верхняя треугольная матрица [6] с положительными элементами на главной диагонали, $\mathbf{S}^T$ - транспонированная к ней, $\mathbf{D}$ — диагональная матрица с элементами на главной диагонали равными ± 1.

Если разложение (2.10) матрицы $\mathbf{A}$ получено, то решение системы (2.1) сводится к последовательному решению трех систем уравнений:

$$\mathbf{S}^T\mathbf{z} = \mathbf{b}, \quad \mathbf{D}\mathbf{y} = \mathbf{z}, \quad \mathbf{S}\mathbf{x} = \mathbf{y} \tag{2.11}$$

[6] Матрица называется верхней треугольной, если все ее элементы, расположенные ниже главной диагонали, равны нулю.

Каждая из этих систем легко решается, так как $\mathbf{S}$ и $\mathbf{S}^T$ — треугольные матрицы, а $\mathbf{D}$ — диагональная.

Получим расчетные формулы метода.

Обозначим через d_{ij}, s_{ij} и s^*_{ij} элементы матриц $\mathbf{D}$, $\mathbf{S}$ и $\mathbf{S}^T$ соответственно. Так как матрица $\mathbf{D}$ — диагональная, $d_{ij} = 0$ при $i \neq j$. Из того, что $\mathbf{S}$ — верхняя треугольная матрица следует, что $s_{ij} = 0$ при $i > j$. Кроме того, $s_{ij} = s^*_{ji}$.

Имеем:

$$(\mathbf{DS})_{ij} = \sum_{l=1}^{m} d_{il} s_{lj} = d_{ii} s_{ij},$$

где $(\mathbf{DS})_{ij}$ — элемент, стоящий на пересечении i-ой строки и j-го столбца матрицы $\mathbf{DS}$. Тогда

$$(\mathbf{S}^T\mathbf{DS})_{ij} = \sum_{l=1}^{m} s_{li} d_{ll} s_{lj} = a_{ij} \quad i, j = 1, 2, \ldots, m. \tag{2.12}$$

Так как матрица $\mathbf{A}$ симметрична, достаточно рассмотреть случай $i \leq j$. Перепишем (2.12) в виде:

$$a_{ij} = \sum_{l=1}^{i-1} s_{li} d_{ll} s_{lj} + s_{ii} s_{ij} d_{ii} + \sum_{l=i+1}^{m} s_{li} d_{ll} s_{lj}.$$

Поскольку $s_{li} = 0$ при $l > i$, последнее слагаемое в правой части пропадает. В результате получим:

$$a_{ij} = s_{ii} s_{ij} d_{ii} + \sum_{l=1}^{i-1} s_{li} d_{ll} s_{lj}, \quad i \leq j. \tag{2.13}$$

В частности, при $i = j$ получим:

$$s_{ii}^2 d_{ii} = a_{ii} - \sum_{l=1}^{i-1} s_{lj}^2 d_{ll}.$$

Отсюда следует, что знак элемента d_{ii} должен совпадать со знаком выражения, стоящего в правой части.

$$d_{ii} = \operatorname{sign}\Big(a_{ii} - \sum_{l=1}^{i-1} s_{li}^2 d_{ll}\Big), \quad s_{ii} = \Big|a_{ii} - \sum_{l=1}^{i-1} s_{li}^2 d_{ll}\Big|^{\frac{1}{2}}. \tag{2.14}$$

Из (2.13) при $i < j$ имеем

$$s_{ij} = \frac{a_{ij} - \sum\limits_{l=1}^{i-1} s_{li} s_{lj} d_{ll}}{s_{ii} d_{ii}}.$$

Полученные формулы позволяют найти сначала d_{11}, s_{11}, а затем последовательно все элементы первой строки матрицы $\mathbf{S}$. После этого берем $i = 2$, находим d_{22}, все элементы во второй строки матрицы $\mathbf{S}$ и т.д.

Теперь можно найти решения систем (2.11).

$$
\begin{aligned}
z_1 &= \frac{b_1}{s_{11}}, \\
z_i &= \frac{b_i - \sum\limits_{l=1}^{i-1} z_l s_{li}}{s_{ii}}, \quad i = 2, 3, \ldots, m, \\
y_i &= \frac{z_i}{d_{ii}}, \quad i = 1, 2, \ldots, m, \\
x_m &= \frac{y_m}{s_{mm}}, \\
x_i &= \frac{y_i - \sum\limits_{l=i+1}^{m} s_{il} x_l}{s_{ii}}, \quad i = m-1, m-2, \ldots, 1.
\end{aligned}
$$

Метод квадратного корня требует примерно $m^3/3$ арифметических операций, то есть при больших m он вдвое быстрее метода Гаусса.

В процессе нахождения элементов матрицы $\mathbf{S}$ может оказаться, что при некотором значении индекса i элемент $s_{ii} = 0$. Тогда дальнейший расчет станет невозможен. От этого можно избавиться, переставляя на место a_{ii} другой диагональный элемент a_{jj} , то есть переставляя i-ый и j-ый столбцы и i-ую и j-ую строки матрицы $\mathbf{A}$. Заметим, что перестановка столбцов матрицы равносильна изменению нумерации неизвестных в системе (2.1).

2.1.3 Метод прогонки

Метод прогонки является модификацией метода Гаусса для случая системы уравнений с **трехдиагональной матрицей**. Трехдиагональными называют матрицы, у которых ненулевые элементы могут стоять только на главной диагонали, на одной диагонали непосредственно над главной диагональю и одной непосредственно под главной диагональю. Другими словами, в i-ой строке трехдиагональной матрицы ненулевыми могут быть $(i-1)$-ый, i-ый и $(i+1)$-ый элементы. Системы такого типа часто встречаются при численном решении дифференциальных уравнений, причем размерность системы обычно велика и может достигать нескольких тысяч уравнений.

Запишем систему в виде

$$a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i, \quad 1 \le i \le m, \quad a_1 = c_m = 0. \tag{2.15}$$

Будем искать решение этой системы в виде:

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \tag{2.16}$$

где α_i, β_i — коэффициенты, которые подлежат определению. Их называют **прогоночными коэффициентами**. Уменьшим в (2.16) индекс на 1 и подставим полученное выражение в (2.15):

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

или

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

Сравнивая эту формулу с (2.16), получим:

$$\alpha_{i+1} = \frac{c_i}{b_i - a_i\alpha_i}, \quad \beta_{i+1} = \frac{a_i\beta_i - d_i}{b_i - a_i\alpha_i}. \tag{2.17}$$

Так как $a_1 = 0$, первое уравнение системы (2.15) можно записать в виде:

$$x_1 = \frac{c_1}{b_1}x_2 - \frac{d_1}{b_1},$$

откуда следует, что

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

По формулам (2.17) находим все значения $\alpha_i,\ \beta_i$ для $i = 3, \ldots, m+1$ (**прямой ход прогонки**).

Так как $c_m = 0$, из (2.17) имеем $\alpha_{m+1} = 0$. Из (2.16) теперь следует, что $x_m = \beta_{m+1}$. По формулам (2.16), двигаясь в направлении убывания индекса, находим все значения x_i (**обратный ход прогонки**).

Вычисления по формулам прогонки требуют всего $8m$ арифметических действий, причем $5m$ операций умножения и деления и $3m$ — сложения и вычитания.

Покажем, что в формулах прямого хода отсутствует деление на ноль [7], если выполнены условия: $|a_i| > 0, \quad |c_i| > 0$, при $i = 2, \ldots, m-1$ и

$$|b_i| \geq |a_i| + |c_i|, \quad i = 1, \ldots, m, \tag{2.18}$$

причем, хотя бы при одном значении индекса i выполняется строгое неравенство. Условие (2.18) называется **условием диагонального преобладания**.

Предположим сначала, что $|\alpha_i| \leq 1$ при некотором значении i. Тогда

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

Если же $|\alpha_i| < 1$, либо в (2.18) выполняется строгое неравенство, аналогично рассуждая получим, что $|\alpha_{i+1}| < 1$. Поскольку при $i = 1$ из (2.18) следует, что $|b_1| \geq |c_1|$, получим $|\alpha_2| \leq 1$ и, значит, $|\alpha_i| \leq 1$ для $i = 2, \ldots, m-1$. Однако, поскольку в (2.18) хотя бы один раз выполняется строгое неравенство, то начиная с этого момента $|\alpha_{i+1}| < 1$ и, значит, $|\alpha_m| < 1$. Тогда из этого неравенства и (2.18) следует, что

$$|b_m - a_m\alpha_m| \geq |b_m| - |\alpha_m||a_m| > |b_m| - |a_m| \geq 0,$$

то есть $b_m - a_m\alpha_m \neq 0$. Кроме того для $i = 2, \ldots, m-1$

$$|b_i - a_i\alpha_i| \geq |b_i| - |a_i||\alpha_i| \geq |a_i|(1 - |\alpha_i|) + |c_i| \geq |c_i| > 0.$$

Что и требовалось доказать.

При выполнении условий (2.18) ошибки, возникающие при округлении, не накапливаются. Действительно, пусть вместо x_i в результате ошибки получили величину $\tilde{x}_i = x_i + \delta_i$. Тогда

$$\tilde{x}_{i-1} = \alpha_i(x_i + \delta_i) + \beta_i = x_{i-1} + \alpha_i\delta_i$$

и так как $|\alpha_i| < 1$, то ошибка

$$|\delta_{i-1}| = |\tilde{x}_{i-1} - x_{i-1}| = |\alpha_i\delta_i| \leq |\delta_i|.$$

[7] Это в свою очередь означает, что решение системы существует

Напомним, что метод, в котором возникающие ошибки не возрастают существенно, называется **устойчивым**. Заметим, что если бы модули прогоночных коэффициентов α_i были бы больше $\widetilde{\alpha} > 1$, то из предыдущей оценки следовало бы, что $|\delta_1| > \widetilde{\alpha}^{i-1}\delta_i$. Поэтому $|\delta_1|$ может оказаться очень большим при больших значениях i. В связи с этим, прогонку называют **устойчивой**, если для всех прогоночных коэффициентов выполняется неравенство $|\alpha_i| \leq 1$.

Заметим, что (2.18) является лишь достаточным и не является необходимым условием корректной разрешимости системы уравнений. Поэтому, если условие (2.18) не выполнено, то это еще не означает, что методом прогонки пользоваться нельзя.

2.1.4 Обусловленность матрицы системы линейных алгебраических уравнений

При использовании численных методов для решения математических задач, как уже было отмечено ранее, необходимо различать свойства самой задачи и свойства вычислительного алгоритма, предназначенного для ее решения. Для математической задачи принято рассматривать вопрос о ее корректности. Будем говорить, что некоторая задача **корректна,** если при любых входных данных из определенного класса она имеет единственное решение, которое непрерывно зависит от этих входных данных. Последнее свойство называется **устойчивостью** и означает, что малому изменению (малому возмущению, малым ошибкам) входных данных соответствует малое изменение решения.

Естественно, что есть смысл решать численно только те задачи, у которых есть решение. Свойство единственности необходимо для того, чтобы знать все ли решения найдены. Требование устойчивости связано с тем, что исходные данные задачи обычно берутся из эксперимента, поэтому они известны только приближенно. Устойчивость означает, что полученное по приближенным данным решение мало отличается от точного.

Рассмотрим вопросы корректности исходной задачи и численных алгоритмов ее решения на примере системы линейных алгебраических уравнений.

Как известно, система линейных алгебраических уравнений (2.1) имеет единственное решение для любого вектора $\mathbf{b}$ тогда и только тогда, когда определитель $\det\mathbf{A} \neq 0$. В этом случае существует матрица $\mathbf{A}^{-1}$ и решение записывается в виде $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Следовательно, для установления корректности задачи о разрешимости системы линейных алгебраических уравнений надо установить непрерывную зависимость от входных данных (устойчивость). Различают **устойчивость по правой части**, когда ошибка (возмущение) присутствует только в правой части $\mathbf{b}$ и **коэффициентную устойчивость** (возмущается матрица $\mathbf{A}$).

Рассмотрим устойчивость по правой части. Нас интересует насколько сильно изменится $\mathbf{x}$ в результате изменения $\mathbf{b}$.

Наряду с основной системой запишем возмущенную систему, то есть систему с измененной правой частью

$$\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}. \tag{2.19}$$

Введем обозначения: $\Delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x},\ \Delta\mathbf{b} = \tilde{\mathbf{b}} - \mathbf{b}$.

Предположим, что при всех $\Delta\mathbf{b}$ справедлива оценка

$$\|\Delta\mathbf{x}\| \leq M_1\|\Delta\mathbf{b}\|, \tag{2.20}$$

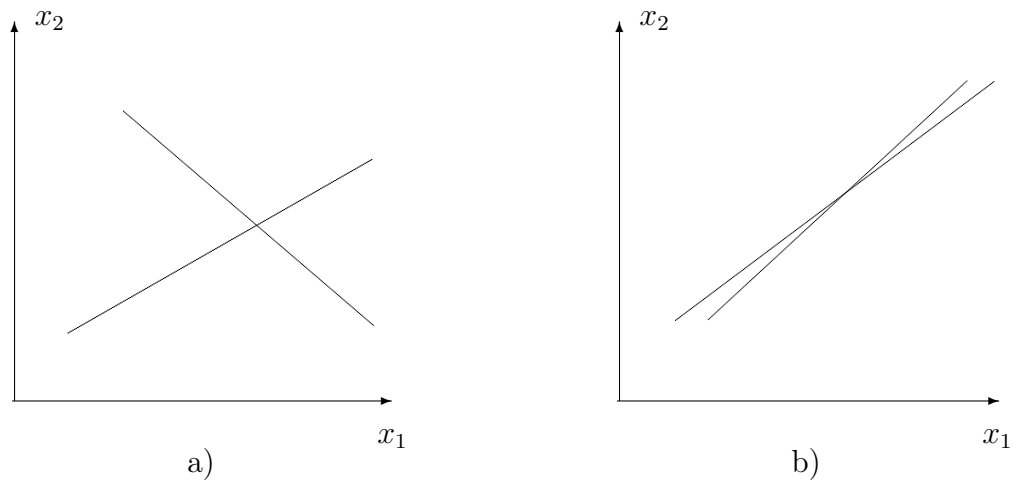


Рис. 2.1 Решение системы двух уравнений графическим методом

где постоянная $M_1 > 0$, не зависящая от $\Delta\mathbf{b}$. Тогда $\Delta\mathbf{x} \to 0$ при $\Delta\mathbf{b} \to 0$, то есть задача устойчива по правой части. В связи с этим, при исследовании системы (2.1), неравенство (2.20) часто называют **условием устойчивости по правой части**.

Если $\det A \neq 0$, то система устойчива по правой части. Действительно, вычитая из системы (2.19) систему (2.1), получим:

$$\mathbf{A}\Delta\mathbf{x} = \Delta\mathbf{b}.$$

Тогда

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}$$

и

$$\|\Delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\| \cdot \|\Delta\mathbf{b}\|. \tag{2.21}$$

Отсюда следует, что для всех $\Delta\mathbf{b}$ справедлива оценка (2.20) с $M_1 = \|\mathbf{A}^{-1}\| > 0$.

Наличие устойчивости очень важно при численном решении, поскольку вектор $\mathbf{b}$ зачастую нельзя задать точно. Заметим, что чем больше M_1, тем сильнее погрешность правой части может исказить искомое решение. Рассмотрим геометрическую иллюстрацию этой ситуации. Пусть решается система двух уравнений с двумя неизвестными. Тогда графически решение системы — это нахождение точки пересечения двух прямых. Каждая прямая задается соответствующим уравнением системы, а изменение правой части приводит к перемещению прямой вверх, вниз (вправо, влево). На рисунке 2.1 изображено два типичных случая. В первом случае (рисунок 2.1 a) угол между прямыми велик, поэтому небольшое перемещение прямых приведет к небольшому смещению точки пересечения. Во втором же случае (рисунок 2.1 b) прямые почти параллельны и небольшое перемещение прямых приведет к существенному смещению точки пересечения. Численное решение таких систем вызывает определенные трудности, поскольку вычисления приходится проводить с высокой точностью, то есть такие системы являются "плохими" для решения на ЭВМ.

Так как во втором случае прямые почти параллельны, определитель матрицы системы близок к нулю. Однако, величина определителя матрицы $\mathbf{A}$ является не очень хорошей мерой "качества" системы уравнений. Для примера рассмотрим диагональную матрицу порядка 100 с числом 0.1 на главной диагонали. Ее определитель равен 10^{-100}, то есть весьма мал. В то же время решение системы запишется в виде $x_i = 10b_i, \quad i = 1, \ldots, 100$ и ошибки в задании b_i мало повлияют на решение системы.

Посмотрим как оценивается относительная погрешность решения. Из (2.1) следует $\|\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$. Умножив это неравенство на неравенство (2.21) получим после

простых преобразований:

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}. \tag{2.22}$$

Число $M_A = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ называют **числом обусловленности матрицы**. Его принято считать мерой "качества" системы уравнений. Матрицы с большим числом M_A называются **плохо обусловленными**. При численном решении систем с плохо обусловленными матрицами возможно сильное накопление относительной погрешности решения. Заметим, что для приведенного выше примера диагональной матрицы с числами 0.1 на диагонали, число обусловленности равно 1, так как $\mathbf{A} = 0.1\mathbf{E}$, $\mathbf{A}^{-1} = 10\mathbf{E}$, где $\mathbf{E}$ — единичная матрица.

В примере, приведенном в конце параграфа 1.4, для матрицы коэффициентов

$$\mathbf{A} = \begin{pmatrix} 1 & 10 \\ 1000 & 10001 \end{pmatrix}$$

существует обратная

$$\mathbf{A}^{-1} = \begin{pmatrix} 10001 & -10 \\ -1000 & 1 \end{pmatrix}.$$

Если в пространстве двумерных векторов $\mathbf{x} = (x_1, x_2)$ ввести норму

$$\|\mathbf{x}\| = \max\{|x_1|, |x_2|\},$$

то норма матрицы $\mathbf{A} = (a_{ij})_{i,j=1}^n$ вычисляется по формуле [20]

$$\|\mathbf{A}\| = \max_{i=1,2} \sum_{j=1}^{2} |a_{ij}|.$$

Поэтому $\|\mathbf{A}\| = 11001$, $\|\mathbf{A}^{-1}\| = 10011$ и $M_A > 10^8$.

Другим примером плохо обусловленной матрицы является матрица Гильберта

$$\mathbf{H}_n = \Big(\frac{1}{i+j-1}\Big)_{i,j=1}^{n}.$$

У этой матрицы число обусловленности растет с ростом n и уже при $n = 8$ превышает 10^{10}.

Отметим некоторые свойства числа обусловленности.

1. $M_A \ge 1$.

Действительно, пусть $\mathbf{E}$ — единичная матрица. Тогда $\mathbf{E} = \mathbf{A}\mathbf{A}^{-1}$ и

$$1 = \|\mathbf{E}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \le \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = M_A.$$

2. $M_A \ge \dfrac{|\lambda_{max}(\mathbf{A})|}{|\lambda_{min}(\mathbf{A})|}$, где $\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})$ — минимальное и максимальное по модулю собственные числа матрицы $\mathbf{A}$.

Для доказательства достаточно заметить, что для любой матрицы модуль ее собственного числа не превосходит нормы матрицы и, кроме того, если λ — собственное число матрицы $\mathbf{A}$, то λ^{-1} — собственное число матрицы $\mathbf{A}^{-1}$. Поэтому $|\lambda_{max}(\mathbf{A})| \le \|\mathbf{A}\|$, $|\lambda_{min}^{-1}(\mathbf{A})| \le \|\mathbf{A}^{-1}\|$, откуда следует требуемое утверждение.

В некоторых случаях $M_A = \dfrac{|\lambda_{max}(\mathbf{A})|}{|\lambda_{min}(\mathbf{A})|}$. Пусть, например, в пространстве задано скалярное произведение по формуле $(\mathbf{y}, \mathbf{z}) = \sum\limits_{i=1}^{m} y_i z_i$, где $y_i,\ z_i$ — координаты векторов $\mathbf{y}$ и $\mathbf{z}$ соответственно и $\mathbf{A}$ — симметричная матрица. Из линейной алгебры известно, что в этом случае в пространстве m-мерных векторов существует ортонормированный базис, состоящий из собственных векторов матрицы $\mathbf{A}$. Покажем, что при сделанных предположениях $\|\mathbf{A}\| = |\lambda_{max}(\mathbf{A})|$. Действительно, если $\mathbf{e}_i$ — ортонормированный базис, состоящий из собственных векторов матрицы $\mathbf{A}$, соответствующих собственным числам λ_i, то есть $\mathbf{A}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, то любой вектор $\mathbf{x}$ представим в виде $\mathbf{x} = \sum\limits_{i=1}^{m} c_i \mathbf{e}_i$, где c_i — коэффициенты разложения вектора по базису. Тогда

$$\mathbf{A}\mathbf{x} = \mathbf{A}\sum_{i=1}^{m} c_i \mathbf{e}_i = \sum_{i=1}^{m} c_i \mathbf{A}\mathbf{e}_i = \sum_{i=1}^{m} c_i \lambda_i \mathbf{e}_i.$$

Следовательно, согласно обобщенной теореме Пифагора,

$$\|\mathbf{A}\mathbf{x}\|^2 = \sum_{i=1}^{m} c_i^2 \lambda_i^2 \le \lambda_{max}^2(\mathbf{A}) \sum_{i=1}^{m} c_i^2 = {\lambda_{max}}^2(\mathbf{A})\|\mathbf{x}\|^2.$$

Поэтому $\|\mathbf{A}\| \le |\lambda_{max}(\mathbf{A})|$ и так как $\|\mathbf{A}\| \ge |\lambda_{max}(\mathbf{A})|$ получаем, что $\|\mathbf{A}\| = |\lambda_{max}(\mathbf{A})|$.

Аналогично доказывается, что $\|\mathbf{A}^{-1}\| = |\lambda_{min}^{-1}(\mathbf{A})\|$.

3. $M_{AB} \le M_A \cdot M_B$.

Это неравенство легко следует из того, что $\|\mathbf{A}\mathbf{B}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$.

Несколько сложнее получить полную оценку относительной погрешности решения, то есть найти оценку для того случая, когда погрешности могут быть как при задании правой части, так и при задании матрицы системы. Справедлива следующая теорема

Теорема 2.1.1 *Пусть* $\mathbf{x}$ *– решение системы (2.1) с невырожденной матрицей* $\mathbf{A}$, $\mathbf{x} + \Delta\mathbf{x}$ *– решением системы линейных алгебраических уравнений*

$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b} \tag{2.23}$$

и $\Delta\mathbf{A}$ *таково, что* $1 - M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} > 0$. *Тогда справедлива оценка*

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{M_A}{1 - M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}} \left(\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \right). \tag{2.24}$$

Доказательство. Раскрывая скобки в (2.23) и учитывая (2.1), получим

$$\mathbf{A}\Delta\mathbf{x} + \Delta\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \Delta\mathbf{b}.$$

Умножим это равенство на $\mathbf{A}^{-1}$, тогда

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b} - \mathbf{A}^{-1}\Delta\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}).$$

Отсюда следует, что

$$\begin{gathered}\|\Delta\mathbf{x}\| = \|\mathbf{A}^{-1}\Delta\mathbf{b} - \mathbf{A}^{-1}\Delta\mathbf{A}(\mathbf{x} + \Delta\mathbf{x})\| \le \|\mathbf{A}^{-1}\Delta\mathbf{b}\| + \|\mathbf{A}^{-1}\Delta\mathbf{A}(\mathbf{x} + \Delta\mathbf{x})\| \le \\ \le \|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\| + \|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|(\|\mathbf{x}\| + \|\Delta\mathbf{x}\|) = \\ = M_A \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{A}\|} + M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\|\mathbf{x}\| + M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\|\Delta\mathbf{x}\|.\end{gathered}$$

Имеем далее, учитывая, что $\|\mathbf{b}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$

$$\left(1 - M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) \|\Delta\mathbf{x}\| \le M_A \|\mathbf{x}\| \left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{A}\|\|\mathbf{x}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) \le M_A \|\mathbf{x}\| \left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right).$$

Отсюда следует (2.24). Теорема доказана.

Замечание. Если в (2.24) положить $\Delta\mathbf{A} = 0$, получим неравенство (2.22).

Следующий пример показывает какой может быть получен результат при нахождении решения системы с плохо обусловленной матрицей. Предположим, что мы хотим решить систему, в которой $b_1 = 0.1$, а все остальные элементы матрицы $\mathbf{A}$ и вектора $\mathbf{b}$ — целые числа. Предположим далее, что у нас двоичный компьютер с 24 битами для мантиссы и что мы каким-то образом умеем вычислять точное решение для системы, уже записанной в память машины. Тогда единственная ошибка будет связана с двоичным представлением числа 0.1, которая (см. параграф 1.2) оценивается величиной $0.1 \cdot 2^{-24}$. Пусть норма вектора вычисляется по формуле $\|\mathbf{x}\| = \max\limits_{i=1,\ldots,m} |x_i|$ и $M_A = 10^5,\ \|\mathbf{b}\| = 1$. В этом случае

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le M_A \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \le 10^5 \cdot 2^{-24} \cdot 0.1 \approx 6 \cdot 10^{-4}.$$

Следовательно, простой акт записи правой части системы в машине может вызвать изменения в четвертой значащей цифре компонент правильного решения.

Попытаемся ответить теперь на вопрос, как влияет погрешность округления при решении систем линейных алгебраических уравнений. Для большинства вычислительных процессов влияние погрешности округления можно учесть, рассматривая возмущенную систему $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$. Это означает, что процесс решения системы (2.1), искаженный погрешностями округления, эквивалентен точному решению некоторой возмущенной системы. Предположим для простоты, что правая часть системы задана точно, тогда возмущенная система имеет вид $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \mathbf{b}$. Матрица $\Delta\mathbf{A} = \tilde{\mathbf{A}} - \mathbf{A}$ называется **матрицей эквивалентных возмущений**. Для каждого метода эта матрица своя. Можно показать, что для метода Гаусса

$$\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} = O(m \cdot 2^{-t}),$$

где t — число разрядов мантиссы в двоичном представлении чисел в ЭВМ с плавающей точкой. Тогда

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{M_A}{1 - M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}} \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} = O(M_A \cdot m \cdot 2^{-t}).$$

Отметим еще одно свойство плохо обусловленных систем. Малость величины невязки $\mathbf{r} = \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}}$ для плохо обусловленной системы не всегда свидетельствует о близости приближенного решения $\widetilde{\mathbf{x}}$ к точному решению $\mathbf{x}$. Действительно, вернемся еще раз к примеру, приведенному в конце параграфа 1.4. Рассмотрим два приближения к решению системы (1.7): $\widetilde{\mathbf{x}}_1 = (11.001, 0)$ и $\widetilde{\mathbf{x}}_2 = (1, 1.001)$. Тогда $\mathbf{r}_1 = -(0.001, 0)$, а $\mathbf{r}_2 = -(0.01, 10.001)$. Таким образом, вектор $\widetilde{\mathbf{x}}_2$ существенно ближе к решению $\mathbf{x} = (1, 1)$, чем $\widetilde{\mathbf{x}}_1$, в то время, как $\|\mathbf{r}_1\| \ll \|\mathbf{r}_2\|$.

Следуя [27] рассмотрим еще один пример, когда информация о невязке может ввести в заблуждение. Если требуется найти обратную матрицу к матрице $\mathbf{A}$ и в результате вычисления $\mathbf{A}^{-1}$ получено приближение $\mathbf{B}$, то попытаться оценить точность

вычисления можно, анализируя матрицу невязки $\mathbf{R}_r = \mathbf{AB} - \mathbf{E}$. Если бы матрица $\mathbf{B}$ совпадала с матрицей $\mathbf{A}^{-1}$, то матрица $\mathbf{R}_r$ была бы нулевой. Обратная матрица удовлетворяет также соотношению $\mathbf{A}^{-1}\mathbf{A} = \mathbf{E}$. Поэтому невязку можно ввести по-другому $\mathbf{R}_l = \mathbf{BA} - \mathbf{E}$.

Пусть

$$\mathbf{A} = \begin{pmatrix} 9999 & 9998 \\ 10000 & 9999 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 9999.9999 & -9997.0001 \\ -10001 & 9998 \end{pmatrix}.$$

Тогда вычисление невязок дает

$$\mathbf{R}_r = \begin{pmatrix} 0.0001 & 0.0001 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{R}_l = \begin{pmatrix} 19998 & 19995 \\ -199999 & -19996 \end{pmatrix}.$$

Матрица $\mathbf{R}_r$ указывает на то, что $\mathbf{B}$ хорошо приближает $\mathbf{A}^{-1}$, в то время как $\mathbf{R}_l$ свидетельствует о плохом приближении. На самом деле

$$\mathbf{A}^{-1} = \begin{pmatrix} 9999 & -9998 \\ -10000 & 9999 \end{pmatrix}.$$

Таким образом, в рассмотренном примере относительная ошибка в определении каждого элемента обратной матрицы приблизительно равна 0.0001.

Для того, чтобы понять, почему анализ невязок $\mathbf{R}_r$ и $\mathbf{R}_l$ дает столь противоречивые результаты, найдем число обусловленности исходной матрицы. Если вычислять норму матрицы $\mathbf{A}$ по формуле $\|\mathbf{A}\| = \max\limits_{i=1,2}(|a_{i1}| + |a_{i2}|)$, то $\|\mathbf{A}\| = \|\mathbf{A}^{-1}\| = 19999$ и $M_A = 19999^2$. Следовательно, матрица $\mathbf{A}$ является плохо обусловленной, что и объясняет сложившуюся ситуацию.

Подчеркнем еще раз, что число обусловленности не связано с какой-либо точностью вычислений или численным алгоритмом, а характеризуется только свойствами матрицы исходной системы.

2.2 ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Существует большая группа задач, приводящая к линейным системам (2.1) с сотнями и тысячами неизвестных. Зачастую у этих систем матрица $\mathbf{A}$ слабо заполнена, причем элементы матрицы задаются простой формулой, и, следовательно, могут вычисляться по мере необходимости. Желательно решать такие системы методами, которые вообще не меняют матрицу $\mathbf{A}$ и требуют хранения лишь нескольких векторов размерности m. Методы, отвечающие этим требованиям, существуют и называются **итерационными**. В них начинают с какого-нибудь приближения $\mathbf{x}^{(0)}$ и выполняют некоторый процесс, использующий $\mathbf{A}$, $\mathbf{b}$, $\mathbf{x}^{(0)}$ и приводящий к новому вектору $\mathbf{x}^{(1)}$. Затем процесс повторяют. На $k+1$-ом шаге итерационного процесса по $\mathbf{A}$, $\mathbf{b}$ и $\mathbf{x}^{(k)}$ получают $\mathbf{x}^{(k+1)}$. При соответствующих предположениях вектора $\mathbf{x}^{(k)}$ сходятся к решению при $k \to \infty$.

Окончание итераций определяется либо заданием максимального числа итераций K, либо одним из условий:

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \varepsilon, \qquad \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \varepsilon\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|,$$
$$\|\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}\| < \varepsilon, \qquad \|\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}\| < \varepsilon\|\mathbf{A}\mathbf{x}^{(0)} - \mathbf{b}\|,$$

где ε — заданная допустимая погрешность.

2.2.1 Метод простой итерации

Систему (2.1) перепишем в виде эквивалентной системы

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}. \tag{2.25}$$

Решение системы (2.25) найдем как предел последовательности

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c}. \tag{2.26}$$

Существует большое число способов, позволяющих перейти от (2.1) к (2.25). Например, если все диагональные элементы матрицы $\mathbf{A}$ не равны нулю, то систему можно получить, разделив i-ое уравнение ($i = 1, \ldots, m$) на a_{ii} и записав его в виде

$$x_i = -\sum_{\substack{j=1 \\ j \neq i}}^{m} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}}.$$

Получившийся при этом итерационный метод называется **методом Якоби**.

В функциональном анализе была доказана теорема Банаха о неподвижной точке [20], согласно которой, если $\|\mathbf{B}\| < 1$, то итерационный метод (2.26) сходится при любом начальном приближении $\mathbf{x}^{(0)}$. При этом

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|\mathbf{B}\|^k}{1 - \|\mathbf{B}\|} \|\mathbf{x}^{(0)} - \mathbf{x}^{(1)}\|. \tag{2.27}$$

Если в качестве нормы вектора выбрать величину

$$\|\mathbf{x}\| = \max_{i=1,\ldots,m} |x_i|,$$

то норма матрицы $\mathbf{B}$ [8] вычисляется по формуле (см. [20]) :

$$\|\mathbf{B}\| = \max_{i=1,\ldots,m} \sum_{j=1}^{m} |b_{ij}|,$$

где b_{ij} — элементы матрицы $\mathbf{B}$. Тогда получается следующее **достаточное условие сходимости метода Якоби**:

$$\|\mathbf{B}\| = \max_{i=1,\ldots,m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Это означает, что достаточно, чтобы было выполнено условие диагонального преобладания для матрицы $\mathbf{A}$, то есть

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{m} |a_{ij}|, \quad i = 1, \ldots, m.$$

Докажем теперь теорему о необходимом и достаточном условии сходимости метода простой итерации

[8] Имеется в виду норма оператора, порождаемого матрицей $\mathbf{B}$.

Теорема 2.2.1 *Пусть система (2.25) имеет единственное решение. Итерационный процесс (2.26) сходится к решению системы (2.25) при любом начальном приближении тогда и только тогда, когда все собственные числа матрицы* $\mathbf{B}$ *по модулю меньше* 1.

Доказательство. Необходимость. Проведем доказательство от противного. Пусть λ — собственное число матрицы $\mathbf{B}$, модуль которого больше или равен 1, а $\mathbf{e}$ — соответствующий этому числу собственный вектор. Выберем начальное приближение $\mathbf{x}^{(0)} = \mathbf{x} + \mathbf{e}$, где $\mathbf{x}$ — решение системы (2.25). Тогда

$$\mathbf{x}^{(1)} = \mathbf{B}\mathbf{x}^{(0)} + \mathbf{c} = \mathbf{B}(\mathbf{x} + \mathbf{e}) + \mathbf{c} = \mathbf{B}\mathbf{x} + \mathbf{c} + \mathbf{B}\mathbf{e} = \mathbf{x} + \lambda\mathbf{e}.$$

Рассуждая аналогично, получим $\mathbf{x}^{(k)} = \mathbf{x} + \lambda^k\mathbf{e}$. Отсюда следует, что $\mathbf{x}^{(k)}$ не сходится к $\mathbf{x}$ при $k \to \infty$.

Достаточность. Для простоты, при доказательстве достаточности предположим, что в пространстве векторов существует базис, состоящий из собственных векторов $\mathbf{e}_i$ $(i = 1, \ldots, m)$ матрицы $\mathbf{B}$ и λ_i — соответствующие им собственные числа. Такое предположение выполняется, например, для симметричных матриц.

Пусть $\mathbf{y}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$. Вычитая (2.25) из (2.26) получим, что для $\mathbf{y}^{(k)}$ выполняется соотношение

$$\mathbf{y}^{(k+1)} = \mathbf{B}\mathbf{y}^{(k)}. \tag{2.28}$$

Разложим вектор $\mathbf{y}^{(k)}$ по собственным векторам матрицы $\mathbf{B}$

$$\mathbf{y}^{(k)} = \sum_{i=1}^{m} \gamma_i^{(k)}\mathbf{e}_i \tag{2.29}$$

и подставим это разложение в (2.28). В результате получим:

$$\sum_{i=1}^{m} \gamma_i^{(k+1)}\mathbf{e}_i = \mathbf{B}\Big(\sum_{i=1}^{m} \gamma_i^{(k)}\mathbf{e}_i\Big) = \sum_{i=1}^{m} \gamma_i^{(k)}\mathbf{B}\mathbf{e}_i = \sum_{i=1}^{m} \gamma_i^{(k)}\lambda_i\mathbf{e}_i.$$

В силу линейной независимости векторов $\mathbf{e}_i$ отсюда следует, что коэффициенты разложения $\gamma_i^{(k)}$ удовлетворяют соотношениям:

$$\gamma_i^{(k+1)} = \lambda_i\gamma_i^{(k)}, \quad i = 1, \ldots, m.$$

Отсюда легко получить, что

$$\gamma_i^{(k)} = \lambda_i^k\gamma_i^{(0)}, \quad i = 1, \ldots, m. \tag{2.30}$$

Поскольку $|\lambda_i| < 1$, из (2.30) следует, что $\gamma_i^{(k)} \to 0$ при $k \to \infty$. Учитывая (2.29), получаем теперь что $\mathbf{y}^{(k)} \to 0$ при $k \to \infty$, что и требовалось доказать.

Замечание. Из (2.29), (2.30) следует, что скорость сходимости определяется максимальным по модулю собственным числом матрицы $\mathbf{B}$, поскольку его степени медленнее всего стремятся к нулю.

Получим, используя теорему 2.2.1, необходимое и достаточное сходимости метода Якоби. Представим матрицу системы $\mathbf{A}$ в виде $\mathbf{A} = \mathbf{A}_0 + \mathbf{D}$. Здесь $\mathbf{A}_0$ — матрица, которая отличается от исходной матрицы $\mathbf{A}$ тем, что элементы главной диагонали заменены нулями, а $\mathbf{D}$ — диагональная матрица с элементами на диагонали, совпадающими с элементами матрицы $\mathbf{A}$. Тогда для метода Якоби матрица $\mathbf{B} = -\mathbf{D}^{-1}\mathbf{A}_0$.

Из равенств

$$\det\left(-\mathbf{D}^{-1}\mathbf{A}_0 - \lambda\mathbf{E}\right) = \det\left(-\mathbf{D}^{-1}\mathbf{A}_0 - \lambda\mathbf{D}^{-1}\mathbf{D}\right) = \det\left(-\mathbf{D}^{-1}\right)\det\left(\mathbf{A}_0 + \lambda\mathbf{D}\right) = 0$$

следует, что собственные числа матрицы $\mathbf{B} = -\mathbf{D}^{-1}\mathbf{A}_0$ совпадают с корнями уравнения $\det\left(\mathbf{A}_0 + \lambda\mathbf{D}\right) = 0$.

Таким образом, необходимое и достаточное условие сходимости метода Якоби можно сформулировать следующим образом: все корни уравнения

$$\begin{vmatrix} \lambda a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & \lambda a_{22} & a_{13} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & \lambda a_{mm} \end{vmatrix} = 0$$

по модулю меньше 1.

Изучим теперь влияние ошибок округления на результат вычисления решения методом простой итерации в предположении, что $\|\mathbf{B}\| < 1$. Для этого будем трактовать суммарный эффект ошибок округления при выполнении одной итерации как возмущение правой части итерационного процесса (2.26). Если $\mathbf{d}^{(k+1)}$ — суммарная погрешность округления на $(k+1)$-ой итерации, то вектор $\widetilde{\mathbf{x}}^{(k+1)}$, который получается в результате вычислений на этой итерации вместо вектора $\mathbf{x}^{(k+1)}$, равен

$$\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{B}\widetilde{\mathbf{x}}^{(k)} + \mathbf{c} + \mathbf{d}^{(k+1)}.$$

Тогда

$$\begin{aligned} \|\widetilde{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\| &= \|(\mathbf{B}\widetilde{\mathbf{x}}^{(k-1)} + \mathbf{c} + \mathbf{d}^{(k)}) - (\mathbf{B}\mathbf{x}^{(k-1)} + \mathbf{c})\| = \\ &= \|\mathbf{B}(\widetilde{\mathbf{x}}^{(k-1)} - \mathbf{x}^{(k-1)}) + \mathbf{d}^{(k)}\| \le \|\mathbf{B}\|\|\widetilde{\mathbf{x}}^{(k-1)} - \mathbf{x}^{(k-1)}\| + \|\mathbf{d}^{(k)}\| \le \\ &\le \|\mathbf{B}\|^2\|\widetilde{\mathbf{x}}^{(k-2)} - \mathbf{x}^{(k-2)}\| + \|\mathbf{B}\|\|\mathbf{d}^{(k-1)}\| + \|\mathbf{d}^{(k)}\| \le \\ &\le \cdots \le \|\mathbf{B}\|^k\|\widetilde{\mathbf{x}}^{(0)} - \mathbf{x}^{(0)}\| + \|\mathbf{B}\|^{k-1}\|\mathbf{d}^{(1)}\| + \cdots + \|\mathbf{B}\|\|\mathbf{d}^{(k-1)}\| + \|\mathbf{d}^{(k)}\|. \end{aligned} \tag{2.31}$$

Заметим, что, вообще говоря, $\widetilde{\mathbf{x}}^{(0)} \neq \mathbf{x}^{(0)}$. Это связано с тем, что не любое число точно представимо в машинном виде. Пусть $\delta = \max(\|\widetilde{\mathbf{x}}^{(0)} - \mathbf{x}^{(0)}\|, \max\limits_{i=1,\cdots,k} \mathbf{d}^{(i)})$. Тогда, учитывая, что $\|\mathbf{B}\| < 1$, из (2.31) имеем

$$\|\widetilde{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\| \le \delta(1 + \|\mathbf{B}\| + \|\mathbf{B}\|^2 + \cdots + \|\mathbf{B}\|^k) = \delta\frac{1 - \|\mathbf{B}\|^{k+1}}{1 - \|\mathbf{B}\|} \le \frac{\delta}{1 - \|\mathbf{B}\|}.$$

Таким образом, рост числа итераций не вызывает неограниченного роста погрешности, связанной с ошибками, вносимыми округлениями. В то же время, в том случае, когда $\|\mathbf{B}\|$ близка к 1, суммарное влияние округлений может оказаться довольно большим.

2.2.2 Выбор оптимального значения параметра итерации

Перепишем систему уравнений (2.1) в виде

$$\mathbf{x} = \mathbf{x} - \tau(\mathbf{A}\mathbf{x} - \mathbf{b}), \tag{2.32}$$

где τ — некоторый числовой параметр.

Таким образом, система (2.1) преобразовалась к виду (2.25), с матрицей $\mathbf{B}$, равной $\mathbf{E} - \tau\mathbf{A}$, где $\mathbf{E}$ — единичная матрица. Постараемся выбрать τ так, чтобы метод сходился, причем как можно скорее, то есть чтобы вектор $\mathbf{y}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$ как можно скорее сходился к 0.

Заметим прежде всего, что если λ_i — собственные числа матрицы $\mathbf{A}$, а $\lambda_i(\mathbf{B})$ — собственные числа матрицы $\mathbf{B}$, то $\lambda_i(\mathbf{B}) = 1 - \tau\lambda_i$. При этом собственные вектора матриц $\mathbf{A}$ и $\mathbf{B}$ совпадают. Тогда, в соответствии с утверждением теоремы 2.2.1, для сходимости метода необходимо, чтобы при всех i выполнялось условие $|1-\tau\lambda_i| < 1$, то есть $-1 < 1 - \tau\lambda_i < 1$. Очевидно, что если среди чисел λ_i есть действительные числа разных знаков, то ни при каком значении τ это условие не выполняется. Поэтому в дальнейшем в этом параграфе будем считать, что все числа λ_i положительны.

Предположим, что известны границы, в которых находятся собственные числа, то есть известны такие положительные числа μ и Λ, что $\mu \le \lambda_i \le \Lambda$ при всех i. Тогда для выполнения условия сходимости должны выполняться условия: $\tau > 0$ и $\tau\lambda_i < 2$ при всех i. Отсюда следует, что итерационный метод сойдется при любом τ из промежутка $(0,\, 2/\Lambda)$.

Так как из замечания к теореме 2.2.1 следует, что скорость сходимости метода определяется наибольшим по модулю собственным числом матрицы $\mathbf{B} = \mathbf{E} - \tau\mathbf{A}$, параметр τ следует выбрать так, чтобы наибольшее по модулю собственное число было как можно меньше. Это означает, что следует решить задачу о нахождении $\min\limits_{\tau} \max\limits_{\mu\le\lambda\le\Lambda} |1 - \tau\lambda|$.

Для решения этой проблемы рассмотрим рисунок 2.2, на котором при различных значениях τ изображен график функции $f_\tau(\lambda) = 1 - \tau\lambda$. Пусть τ_0 таково, что

$$1 - \tau_0\mu = -(1 - \tau_0\Lambda). \tag{2.33}$$

Тогда

$$f_{\tau_0}(\mu) < f_\tau(\mu) \ \text{ при } \tau < \tau_0$$

и

$$|f_{\tau_0}(\Lambda)| < |f_\tau(\Lambda)| \ \text{ при } \tau > \tau_0.$$

Это означает, что $\tau = \tau_0$ является искомым оптимальным значением. Из уравнения (2.33) получим, что $\tau_0 = 2/(\mu + \Lambda)$.

Как отмечалось, скорость сходимости определяется наибольшим по модулю собственным числом матрицы $\mathbf{B} = \mathbf{E} - \tau_0\mathbf{A}$. Это число равно

$$1 - \frac{2}{\mu + \Lambda}\mu = \frac{\Lambda - \mu}{\Lambda + \mu}.$$

2.2.3 Метод Зейделя

Описанный в параграфе 2.2.1 метод Якоби можно записать в виде

$$\begin{aligned}
a_{11}x_1^{(k+1)} + a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \cdots + a_{1m}x_m^{(k)} &= b_1,\\
a_{21}x_1^{(k)} + a_{22}x_2^{(k+1)} + a_{23}x_3^{(k)} + \cdots + a_{2m}x_m^{(k)} &= b_2,\\
&\cdots\\
a_{m1}x_1^{(k)} + a_{m2}x_2^{(k)} + a_{m3}x_3^{(k)} + \cdots + a_{mm}x_m^{(k+1)} &= b_m.
\end{aligned} \tag{2.34}$$

Эти формулы означают, что при известном векторе $\mathbf{x}^{(k)}$, первая компонента вектора $\mathbf{x}^{(k+1)}$ находится из первого уравнения, вторая компонента — из второго уравнения и

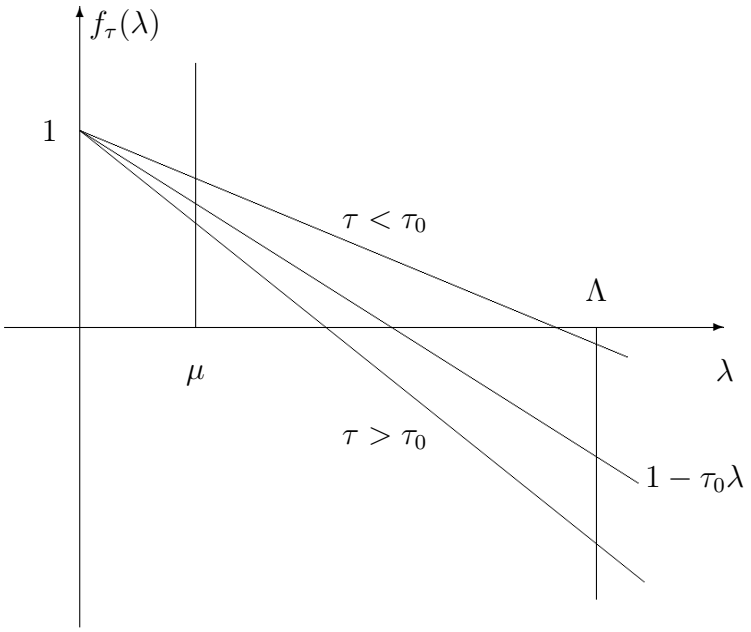


Рис. 2.2 График функции $f_\tau(\lambda) = 1 - \tau\lambda$ при различных значениях τ

так далее. Заметим, что к моменту вычисления $x_i^{(k+1)}$, уже найдены $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$, которые можно было бы использовать в расчетах. Поэтому модернизируем формулы (2.34) следующим образом:

$$\begin{aligned} a_{11}x_1^{(k+1)} + a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \cdots + a_{1m}x_m^{(k)} &= b_1, \\ a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} + a_{23}x_3^{(k)} + \cdots + a_{2m}x_m^{(k)} &= b_2, \\ &\cdots \\ a_{m1}x_1^{(k+1)} + a_{m2}x_2^{(k+1)} + a_{m3}x_3^{(k+1)} + \cdots + a_{mm}x_m^{(k+1)} &= b_m. \end{aligned} \tag{2.35}$$

Полученный итерационный метод нахождения решения системы (2.1) называется **методом Зейделя.**

Для решения вопроса о сходимости метода заметим, что систему (2.35) можно представить в матричном виде

$$\mathbf{A}_1\mathbf{x}^{(k+1)} + \mathbf{A}_2\mathbf{x}^{(k)} = \mathbf{b},$$

где

$$\mathbf{A}_1 = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mm} \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1m} \\ 0 & 0 & a_{23} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Отсюда следует, что

$$\mathbf{x}^{(k+1)} = -\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{x}^{(k)} + \mathbf{A}_1^{-1}\mathbf{b}.$$

Таким образом, метод Зейделя эквивалентен методу простой итерации с матрицей $\mathbf{B} = -\mathbf{A}_1^{-1}\mathbf{A}_2$. Тогда для сходимости метода Зейделя необходимо и достаточно, чтобы все собственные числа этой матрицы по модулю были меньше 1. Из равенств

$$\det\left(-\mathbf{A}_1^{-1}\mathbf{A}_2 - \lambda\mathbf{E}\right) = \det\left(-\mathbf{A}_1^{-1}\mathbf{A}_2 - \lambda\mathbf{A}_1^{-1}\mathbf{A}_1\right) = \det\left(-\mathbf{A}_1^{-1}\right)\det\left(\mathbf{A}_2 + \lambda\mathbf{A}_1\right)$$

следует, что собственные числа матрицы $-\mathbf{A}_1^{-1}\mathbf{A}_2$ совпадают с корнями уравнения $\det(\mathbf{A}_2 + \lambda\mathbf{A}_1) = 0$.

Таким образом, необходимое и достаточное условие сходимости метода Зейделя можно сформулировать следующим образом: все корни уравнения

$$\begin{vmatrix} \lambda a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ \lambda a_{21} & \lambda a_{22} & a_{13} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda a_{m1} & \lambda a_{m2} & \lambda a_{m3} & \cdots & \lambda a_{mm} \end{vmatrix} = 0$$

по модулю меньше 1.

Можно доказать, что метод Зейделя сходится, если $\mathbf{A}$ — симметричная, положительно определенная матрица[9].

Получим критерий определения числа итераций, которые достаточно совершить, для получения методом Зейделя приближенного решения системы линейных алгебраических уравнений (2.1) с заданной точностью ε. Для этого перепишем систему (2.1) как это было сделано в пункте 2.2.1 в виде (2.25), где элементы b_{ij} матрицы $\mathbf{B}$ и компоненты c_i вектора $\mathbf{c}$ определяются следующим образом

$$c_i = \frac{b_i}{a_{ii}}, \quad b_{ii} = 0, \quad b_{ij} = -\frac{a_{ij}}{a_{ii}}, \ i, j = 1 \ldots, m, \ i \neq j.$$

Матрицу $\mathbf{B}$ представим в виде суммы двух матриц $\mathbf{B_1}$ и $\mathbf{B_2}$, где

$$\mathbf{B}_1 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ b_{21} & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & b_{m3} & \cdots & 0 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 0 & b_{12} & b_{13} & \cdots & b_{1m} \\ 0 & 0 & b_{23} & \cdots & b_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Таким образом, система (2.1) перепишется в эквивалентном виде

$$\mathbf{x} = \mathbf{B}_1\mathbf{x} + \mathbf{B}_2\mathbf{x} + \mathbf{c}, \tag{2.36}$$

а метод Зейделя примет вид

$$\mathbf{x}^{(k+1)} = \mathbf{B}_1\mathbf{x}^{(k+1)} + \mathbf{B}_2\mathbf{x}^{(k)} + \mathbf{c}. \tag{2.37}$$

Введем вектор $\mathbf{y}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$, где $\mathbf{x}$ — решение системы (2.1), а, значит, и (2.36). Тогда, вычитая (2.36) из (2.37), получим

$$\mathbf{y}^{(k+1)} = \mathbf{B}_1\mathbf{y}^{(k+1)} + \mathbf{B}_2\mathbf{y}^{(k)}.$$

Отсюда следует, что

$$\mathbf{y}^{(k+1)} = \mathbf{B}_1\mathbf{y}^{(k+1)} + \mathbf{B}_2\mathbf{y}^{(k+1)} + \mathbf{B}_2(\mathbf{y}^{(k)} - \mathbf{y}^{(k+1)}). \tag{2.38}$$

Учитывая, что $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$, $\mathbf{y}^{(k)} - \mathbf{y}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}$, и обозначая через $\mathbf{E}$ единичную матрицу, получаем из (2.38)

$$(\mathbf{E} - \mathbf{B})\mathbf{y}^{(k+1)} = \mathbf{B}_2(\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}). \tag{2.39}$$

[9]Напомним, что симметричная матрица называется положительно определенной, если для любого ненулевого вектора $\mathbf{x}$ выполняется неравенство $\sum_{i,j=1}^{m} a_{ij}x_ix_j > 0$

В [20] доказывалось, что если $\|\mathbf{B}\| < 1$, то существует $(\mathbf{E}-\mathbf{B})^{-1}$, причем $\|(\mathbf{E}-\mathbf{B})^{-1}\| \le 1/(1-\|\mathbf{B}\|)$. Поэтому из (2.39) получим

$$\begin{gathered}\|\mathbf{x}^{(k+1)}-\mathbf{x}\| = \|\mathbf{y}^{(k+1)}\| = \|(\mathbf{E}-\mathbf{B})^{-1}\mathbf{B}_2(\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)})\| \le \\ \le \|(\mathbf{E}-\mathbf{B})^{-1}\|\|\mathbf{B}_2\|\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\| \le \frac{\|\mathbf{B}_2\|}{1-\|\mathbf{B}\|}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\|.\end{gathered} \tag{2.40}$$

Следовательно, если k таково, что

$$\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\| \le \frac{1-\|\mathbf{B}\|}{\|\mathbf{B}_2\|}\varepsilon, \tag{2.41}$$

то в силу (2.40) выполняется неравенство

$$\|\mathbf{x}^{(k+1)}-\mathbf{x}\| \le \varepsilon.$$

Таким образом, доказано следующее утверждение.

Теорема 2.2.2 *Если $\|\mathbf{B}\| < 1$ и число k таково, что выполнено неравенство (2.41), то $\mathbf{x}^{(k+1)}$ приближает решение $\mathbf{x}$ уравнения (2.1) с абсолютной погрешностью ε.*

Согласно теореме для получение приближенного решения с заданной точностью ε вычисления следует прекратить тогда, когда будет выполнено неравенство (2.41).

Напомним еще раз, что если, например, $\|\mathbf{x}\| = \max\limits_{i=1,\dots,m}|x_i|$, то

$$\|\mathbf{B}\| = \max_{i=1,\dots,m}\sum_{\substack{j=1\\ j\ne i}}^{m}|b_{ij}| = \max_{i=1,\dots,m}\sum_{\substack{j=1\\ j\ne i}}^{m}\frac{|a_{ij}|}{|a_{ii}|}, \quad \|\mathbf{B}_2\| = \max_{i=1,\dots,m}\sum_{j=i+1}^{m}|b_{ij}| = \max_{i=1,\dots,m}\sum_{j=i+1}^{m}\frac{|a_{ij}|}{|a_{ii}|}.$$

2.3 АЛГЕБРАИЧЕСКАЯ ПРОБЛЕМА СОБСТВЕННЫХ ЧИСЕЛ

Многие задачи, возникающие в механике, физике, математике, требуют нахождения собственных чисел матрицы. При этом иногда исследователей интересуют только некоторые собственные числа. Например, выбор оптимального параметра при итерационном методе решения системы линейных алгебраических уравнений требует знания только наибольшего и наименьшего собственных чисел. Исследование резонансных явлений предполагает нахождение собственного числа, ближайшего к заданному числу λ_0. В других случаях требуется знать все собственные числа матрицы.

Задачи определения собственных чисел делятся на **частичные проблемы**, когда речь идет о нахождении одного или нескольких чисел, и **полные**, когда ищутся все собственные числа. Конечно, умея решать полную проблему, мы решим и частичную. Однако зачастую такой подход может привести к неоправданно большому объему вычислений.

2.3.1 Обусловленность проблемы нахождения собственных чисел и собственных векторов

Прежде чем переходить к изучению методов нахождения собственных чисел, рассмотрим вопрос о том как погрешность при задании матрицы влияет на погрешность

собственных чисел. Напомним, что под обусловленностью задачи понимается чувствительность ее решения к малым погрешностям исходных данных. Задачу называют **хорошо обусловленной**, если малым погрешностям исходных данных соответствуют малые погрешности решения, и **плохо обусловленной** если возможны большие изменения решения.

Будут рассматриваться матрицы, элементы которых действительные числа. Введем определение

Определение 2.3.1 *Матрица называется* **нормальной**, *если она перестановочна со своей транспонированной, то есть* $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A}$.

Очевидно, что любая симметричная матрица является нормальной. Справедливо следующее утверждение.

Теорема 2.3.1 *Пусть* $\mathbf{A}$ *и* $\widetilde{\mathbf{A}}$ *симметричные матрицы размера* $m \times m$ *с элементами* a_{ij} *и* $\widetilde{a}_{ij}$ *соответственно, а* λ_i *и* $\widetilde{\lambda}_i$ *их собственные числа, занумерованные в порядке неубывания модулей (с учетом кратности). Тогда*

$$\left(\sum_{i=1}^{m} |\lambda_i - \widetilde{\lambda}_i|^2\right)^{1/2} \leq \left(\sum_{i=1}^{m}\sum_{j=1}^{m} |a_{ij} - \widetilde{a}_{ij}|^2\right)^{1/2}. \tag{2.42}$$

Доказательство этого факта следует из двух утверждений, приведенных в [10] под номерами 20.9 и 16.24:

Теорема 2.3.2 *(Виландт-Гофман). Пусть* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *— нормальные матрицы размера* $m \times m$, *причем* $\mathbf{C} = \mathbf{A} + \mathbf{B}$. *Обозначим через* λ_i, β_i, γ_i *их собственные числа, занумерованные в порядке неубывания модулей. Тогда*

$$\sum_{i=1}^{m} |\gamma_i - \lambda_i|^2 \leq \sum_{i=1}^{m} |\beta_i|^2.$$

Теорема 2.3.3 *Для любой матрицы* $\mathbf{B}$ *размера* $m \times m$ *с элементами* b_{ij} *и собственными числами* β_i *справедливо неравенство*

$$\sum_{i=1}^{m} |\beta_i|^2 \leq \sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij}^2,$$

причем равенство достигается тогда и только тогда, когда матрица нормальная.

Достаточно заметить, что в силу условия симметрии матрицы $\mathbf{A}$ и $\widetilde{\mathbf{A}}$, а также их разность являются нормальными матрицами. Положим $\mathbf{B} = \widetilde{\mathbf{A}} - \mathbf{A}$. Тогда $b_{ij} = \widetilde{a}_{ij} - a_{ij}$, $\mathbf{C} = \mathbf{A} + \mathbf{B} = \widetilde{\mathbf{A}}$ и $\gamma_i = \widetilde{\lambda}_i$. Чтобы получить требуемое утверждение достаточно воспользоваться теперь приведенными теоремами.

Неравенство (2.42) означает, что малым погрешностям при задании матрицы $\mathbf{A}$ соответствуют малые погрешности при определении собственных чисел. Таким образом, задача нахождения собственных чисел для симметричных матриц хорошо обусловлена.

Для матриц, не являющихся симметричными, ситуация зачастую является прямо противоположной. Среди них существуют матрицы, у которых собственные числа

чрезвычайно чувствительны к погрешностям элементов матриц. Рассмотрим следующий пример, иллюстрирующий это утверждение. Пусть $\mathbf{A}$ и $\widetilde{\mathbf{A}}$ — матрицы размера 10×10

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad \widetilde{\mathbf{A}} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 10^{-10} & 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Матрица $\widetilde{\mathbf{A}}$ отличается от матрицы $\mathbf{A}$ одним элементом, стоящим в 10-ой строке и 1-ом столбце. Собственные числа матриц будем находить путем отыскания корней характеристического уравнения, то есть уравнения $\det(\mathbf{A} - \lambda\mathbf{E}) = 0$, где $\mathbf{E}$ — единичная матрица. Тогда, вычисляя определители путем разложения по первому столбцу, легко получить, что $\det(\mathbf{A} - \lambda\mathbf{E}) = \lambda^{10}$, а $\det(\widetilde{\mathbf{A}} - \lambda\mathbf{E}) = \lambda^{10} - 10^{-10}$. Это означает, что у матрицы $\mathbf{A}$ все собственные числа равны нулю. У матрицы же $\widetilde{\mathbf{A}}$ собственные числа равны $0.1,\ 0.1e^{i(\pi/5)},\ 0.1e^{i(2\pi/5)}, \cdots, 0.1e^{i(9\pi/5)}$, где $e^{i\varphi} = \cos\varphi + i\sin\varphi,\ i^2 = -1$. Следовательно, вычисляя выражения, входящие в неравенство (2.42) имеем

$$\left(\sum_{i=1}^{10} \sum_{j=1}^{10} (a_{ij} - \widetilde{a}_{ij})^2 \right)^{1/2} = 10^{-10},$$

в то время как

$$\left(\sum_{i=1}^{10} |\lambda_i - \widetilde{\lambda}_i|^2 \right)^{1/2} = \sqrt{0.1} \approx 0.316.$$

Таким образом, малая добавка только к одному из элементов матрицы $\mathbf{A}$ привела к существенному, по сравнению с этой добавкой, изменению собственных чисел.

Обозначим теперь через $\mathbf{x}$ собственный вектор матрицы $\mathbf{A}$, соответствующий числу λ, а через $\widetilde{\mathbf{x}}$ — собственный вектор матрицы $\widetilde{\mathbf{A}}$, соответствующий числу $\widetilde{\lambda}$. Собственных векторов соответствующих одному и тому же собственному числу бесконечно много, они могут различаться, например, ненулевым множителем. Поэтому выбор в качестве меры близости собственных векторов величины $\|\mathbf{x} - \widetilde{\mathbf{x}}\|$ является неудачным решением. Даже если брать нормированные вектора, они могут отличаться знаком. Поэтому для собственных векторов целесообразно сравнивать не близость нормы, а близость направления. В качестве меры близости можно использовать величину $\sin\varphi$, где φ — угол между векторами. Вспоминая, что скалярное произведение векторов равно произведению длин векторов на $\cos\varphi$, получаем в качестве меры близости величину

$$\sin\varphi = \sqrt{1 - \cos^2\varphi} = \left(1 - \left(\frac{(\mathbf{x}, \widetilde{\mathbf{x}})}{\|\mathbf{x}\| \|\widetilde{\mathbf{x}}\|} \right)^2 \right)^{1/2}, \quad \text{где} \ \ \|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2}.$$

Можно показать, что в том случае, когда матрицы $\mathbf{A}$ и $\widetilde{\mathbf{A}}$ симметрические и у матрицы $\mathbf{A}$ нет близких собственных чисел, задача вычисления собственных векторов такой матрицы хорошо обусловлена. Среди не симметричных матриц существуют такие, для которых задача о вычислении собственных векторов плохо обусловлена.

2.3.2 Частичная проблема собственных чисел

Как правило, при решении частичной проблемы ставится задача найти наибольшее (наименьшее) собственное число или собственное число, ближайшее к заданному числу. Покажем, что эти задачи могут быть легко разрешены, если имеется алгоритм для поиска максимального по модулю собственного числа матрицы.

Для того, чтобы найти максимальное (минимальное) собственное число матрицы $\mathbf{A}$, можно рассмотреть матрицу $\mathbf{B} = \mathbf{A} + c\mathbf{E}$. Тогда $\lambda_B = \lambda_A + c$, где λ_A и λ_B — собственные числа матриц $\mathbf{A}$ и $\mathbf{B}$ соответственно. Соответствующие этим числам собственные вектора матриц $\mathbf{A}$ и $\mathbf{B}$ совпадают. Ясно, что при большом положительном числе c максимальное по модулю собственное число матрицы $\mathbf{B}$ положительно и равно сумме c и максимального собственного числа матрицы $\mathbf{A}$. При большом по модулю отрицательном числе c, максимальное по модулю собственное число матрицы $\mathbf{B}$ отрицательно и равно сумме c и минимального собственного числа матрицы $\mathbf{A}$. Следовательно, найдя соответствующее значение λ_B и зная c, можно определить максимальное (минимальное) собственное число матрицы $\mathbf{A}$.

Если надо найти собственное число матрицы $\mathbf{A}$ ближайшее к числу λ_0, то для этого достаточно найти максимальное по модулю собственное число матрицы $\mathbf{B} = \mathbf{E} - c(\mathbf{A} - \lambda_0\mathbf{E})^2$ при некотором малом значении c. Так как $\lambda_B = 1 - c(\lambda_A - \lambda_0)^2$ и при малых c значение $\lambda_B > 0$, максимальному по модулю значению λ_B соответствует λ_A ближайшее к λ_0.

Можно было бы поступить по-другому, взяв матрицу $\mathbf{B} = (\mathbf{A} - \lambda_0\mathbf{E})^{-1}$. Тогда $\lambda_B = (\lambda_A - \lambda_0)^{-1}$ и λ_B максимально по модулю, когда величина $|\lambda_A - \lambda_0|$ минимальна, то есть число λ_A ближе всего к λ_0.

Итак, необходимо научиться находить максимальное по модулю собственное число матрицы.

Будем считать, что в пространстве m-мерных векторов существует полная система $\mathbf{e}_1, \ldots, \mathbf{e}_m$, состоящая из собственных векторов матрицы $\mathbf{A}$, которым соответствуют собственные числа $\lambda_1, \ldots, \lambda_m$. Пусть $(\mathbf{e}_i, \mathbf{e}_i) = 1$. Предположим также, что

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|.$$

Из этого предположения следует, что λ_1 — действительное число. Если бы λ_1 было мнимым, то комплексно сопряженное к нему число тоже было бы собственным и равным ему модулю, то есть $|\lambda_1| = |\lambda_2|$.

Возьмем произвольный вектор $\mathbf{x}^{(0)}$ и последовательно вычислим вектора

$$\mathbf{x}^{(n+1)} = \mathbf{A}\mathbf{x}^{(n)}, \quad n = 0, 1 \ldots.$$

Если

$$\mathbf{x}^{(0)} = \sum_{i=1}^{m} \gamma_i \mathbf{e}_i,$$

то

$$\mathbf{x}^{(n)} = \sum_{i=1}^{m} \gamma_i \mathbf{A}^n \mathbf{e}_i = \sum_{i=1}^{m} \gamma_i \lambda_i^n \mathbf{e}_i = \gamma_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n).$$

Тогда для скалярного произведения векторов имеем:

$$(\mathbf{x}^{(n)}, \mathbf{x}^{(n)}) = \gamma_1^2 \lambda_1^{2n} + O(|\lambda_1|^n |\lambda_2|^n),$$
$$(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)}) = \lambda_1 \left(\gamma_1^2 \lambda_1^{2n} + O(|\lambda_1|^n |\lambda_2|^n)\right).$$

Возьмем

$$\lambda_1^{(n)} = \frac{(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)})}{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})} = \lambda_1 \frac{\gamma_1^2 \lambda_1^{2n} + O(|\lambda_1|^n |\lambda_2|^n)}{\gamma_1^2 \lambda_1^{2n} + O(|\lambda_1|^n |\lambda_2|^n)} = \lambda_1 \frac{1 + O\left(\frac{1}{\gamma_1^2} \left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}{1 + O\left(\frac{1}{\gamma_1^2} \left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}. \tag{2.43}$$

Из условия $|\lambda_1| > |\lambda_2|$ следует, что

$$\lim_{n\to\infty} \left|\frac{\lambda_2}{\lambda_1}\right|^n = 0,$$

поэтому из (2.43) имеем $\lim_{n\to\infty} \lambda_1^{(n)} = \lambda_1$.

Таким образом, вычисляя числа

$$\lambda_1^{(n)} = \frac{(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)})}{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})}$$

получим последовательность, которая сходится к искомому собственному числу. Вычисления следует проводить до тех пор, пока числа $\lambda_1^{(n)}$ не перестанут меняться в пределах заданной точности. Напомним, что скалярное произведение двух векторов может быть вычислено как сумма произведений соответствующих координат.

Если $|\lambda_1| > 1$, то $\|\mathbf{x}^{(n)}\| = \sqrt{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})} \to \infty$ при $n \to \infty$, что при больших n может привести к переполнению. Если же $|\lambda_1| < 1$, то $\|\mathbf{x}^{(n)}\| \to 0$ при $n \to \infty$ и при больших n может оказаться, что в машинном представлении $\mathbf{x}^{(n)} = 0$. Чтобы избежать этих неприятностей время от времени необходимо нормировать вектор $\mathbf{x}^{(n)}$, путем деления его на $\|\mathbf{x}^{(n)}\|$. В результате получится новый вектор, норма которого равна 1. Очевидно, что такая операция нормирования не меняет значения числа $\lambda_1^{(n)}$, так как

$$\lambda_1^{(n)} = \frac{(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)})}{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})} = \frac{(\mathbf{A}\mathbf{x}^{(n)}, \mathbf{x}^{(n)})}{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})} = \frac{(\mathbf{A}(k\mathbf{x}^{(n)}), k\mathbf{x}^{(n)})}{(k\mathbf{x}^{(n)}, k\mathbf{x}^{(n)})},$$

где $k = \|\mathbf{x}^{(n)}\|^{-1}$.

Для нахождения собственного вектора, соответствующего собственному числу λ_1 заметим, что при $\gamma_1 \neq 0$

$$\mathbf{e}_1^{(n)} = \frac{\mathbf{x}^{(n)}}{\|\mathbf{x}^{(n)}\|} = \frac{\sum_{i=1}^{m} \gamma_i \lambda_i^n \mathbf{e}_i}{|\gamma_1 \lambda_1^n| + O(|\lambda_2|^n)} = \frac{\frac{\gamma_1 \lambda_1^n}{|\gamma_1 \lambda_1^n|}\mathbf{e}_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}{1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = \frac{\gamma_1 \lambda_1^n}{|\gamma_1 \lambda_1^n|}\mathbf{e}_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right).$$

Следовательно, $\mathbf{e}_1^{(n)}$ при $n \to \infty$ сходится к вектору, который быть может знаком отличается от $\mathbf{e}_1$.

Замечание 2.10.1 Если условие $|\lambda_1| > |\lambda_2|$ не выполнено, итерационный процесс может не сойтись или сойтись к числу отличному от λ_1.

Рассмотрим примеры. Пусть

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}, \quad \mathbf{x}^{(0)} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Тогда

$$\mathbf{x}^{(1)} = \mathbf{A}\mathbf{x}^{(0)} = \begin{pmatrix} 2x_1 \\ -2x_2 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \mathbf{A}\mathbf{x}^{(1)} = \begin{pmatrix} 4x_1 \\ 4x_2 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \mathbf{A}\mathbf{x}^{(2)} = \begin{pmatrix} 8x_1 \\ -8x_2 \end{pmatrix}, \quad \dots$$

и

$$\lambda_1^{(0)} = \frac{(\mathbf{x}^{(1)}, \mathbf{x}^{(0)})}{(\mathbf{x}^{(0)}, \mathbf{x}^{(0)})} = 2\frac{x_1^2 - x_2^2}{x_1^2 + x_2^2},$$
$$\lambda_1^{(1)} = \frac{(\mathbf{x}^{(2)}, \mathbf{x}^{(1)})}{(\mathbf{x}^{(1)}, \mathbf{x}^{(1)})} = 2\frac{x_1^2 - x_2^2}{x_1^2 + x_2^2},$$

и так далее.

Таким образом, предел существует,

$$\lim_{n\to\infty} \lambda_1^{(n)} = 2\frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}$$

зависит от начального вектора и может получиться любым. Если взять $x_1 = x_2$, то $\lim\limits_{n\to\infty} \lambda_1^{(n)} = 0$, при $x_1 = \sqrt{3}/2,\ x_2 = 1/2$ получим $\lim\limits_{n\to\infty} \lambda_1^{(n)} = 1$. В любом из приведенных случае предел отличен от собственных чисел матрицы $\mathbf{A}$.

Рассмотрим теперь пример, когда сходимости нет. Пусть

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & -2 \end{pmatrix}, \quad \mathbf{x}^{(\mathbf{0})} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Тогда

$$\mathbf{x}^{(1)} = \mathbf{A}\mathbf{x}^{(0)} = \begin{pmatrix} 2x_1 + x_2 \\ -2x_2 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \mathbf{A}\mathbf{x}^{(1)} = \begin{pmatrix} 4x_1 \\ 4x_2 \end{pmatrix} = 4\mathbf{x}^{(\mathbf{0})}.$$

Отсюда следует, что

$$\lambda_1^{(0)} = \frac{(\mathbf{x}^{(1)}, \mathbf{x}^{(0)})}{(\mathbf{x}^{(0)}, \mathbf{x}^{(0)})} = \frac{2x_1^2 + x_1x_2 - 2x_2^2}{x_1^2 + x_2^2},$$
$$\lambda_1^{(1)} = \frac{(\mathbf{x}^{(2)}, \mathbf{x}^{(1)})}{(\mathbf{x}^{(1)}, \mathbf{x}^{(1)})} = 4\frac{2x_1^2 + x_1x_2 - 2x_2^2}{4x_1^2 + 4x_1x_2 + 5x_2^2},$$
$$\lambda_1^{(2k)} = \lambda_1^{(0)}, \quad \lambda_1^{(2k+1)} = \lambda_1^{(1)}, \quad k = 1, 2, \dots$$

Так как, вообще говоря, $\lambda_1^{(0)} \neq \lambda_1^{(1)}$, последовательность расходится.

Поскольку заранее не известно выполнено ли предположение о том, что $|\lambda_1| > |\lambda_2|$, необходимо в случае сходимости итераций после их завершения произвести проверку, вычисляя невязку $\mathbf{r} = \mathbf{A}\mathbf{e}_1 - \lambda_1\mathbf{e}_1$.

Если у матрицы $\lambda_1 = -\lambda_2 > |\lambda_3| \geq \dots$, для отыскания собственных чисел достаточно рассмотреть матрицы $\mathbf{B}_\pm = \mathbf{A} \pm c\mathbf{E}$, где c — некоторое фиксированное положительное число. Применив описанный выше метод к матрице $\mathbf{B}_+$, получим $\lambda_1 + c$. Если же взять матрицу $\mathbf{B}_-$, получим $\lambda_2 - c$.

Замечание 2.10.2 В связи с тем, что вектор $\mathbf{x}^{(0)}$ выбирается произвольным образом, может оказаться, что в его разложении по базису коэффициент при $\mathbf{e}_1$ равен нулю. Это, вообще говоря, не означает, что итерационный процесс даст приближения, которые не сойдутся к λ_1. Из-за присутствия в вычислениях округлений, в процессе итераций может появиться компонента, пропорциональная $\mathbf{e}_1$, вследствие чего требуемый результат будет получен. Однако, поскольку нет уверенности в том, что

влияние вычислительной погрешности оказалось существенным, в то время как в пределах выбранной точности итерационный процесс сошелся, желательно провести один или несколько расчетов с различными значениями $\mathbf{x}^{(\mathbf{0})}$.

Замечание 2.10.3 Описанный метод может сходиться очень медленно, если матрица такова, что число $|\lambda_2/\lambda_1|$ близко к 1. Действительно, из (2.43) следует, что

$$\frac{|\lambda_1^{(n)} - \lambda_1|}{|\lambda_1|} = O\left(\frac{1}{\gamma_1^2}\left|\frac{\lambda_2}{\lambda_1}\right|^n\right).$$

Поэтому, если, например, $|\lambda_2/\lambda_1| = 0.999$, то понадобится более 2300 итераций, чтобы уменьшить относительную ошибку в 10 раз.

Рассмотрим теперь как можно найти $\mathbf{e}_2$ и λ_2 в случае, когда у матрицы $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \cdots$. Если бы начальное приближение было таким, что в его разложении по базису коэффициент при $\mathbf{e}_1$ был бы равен нулю и в процессе итераций не появлялась бы компонента пропорциональная $\mathbf{e}_1$, то итерации сходились бы к λ_2.

Покажем, как можно построить такое начальное приближение. Для этого заметим, что у исходной матрицы $\mathbf{A}$ и транспонированной матрицы $\mathbf{A}^{(T)}$ собственные числа совпадают, в то время как собственные вектора, соответствующие различным собственным числам ортогональны [20]. Применяя описанную выше процедуру к матрицам $\mathbf{A}$ и $\mathbf{A}^{(T)}$, найдем приближенно $\mathbf{e}_1$ и $\mathbf{g}_1$, где $\mathbf{g}_1$ — собственный вектор матрицы $\mathbf{A}^{(T)}$, соответствующий числу λ_1. Вектор $\mathbf{g}_1$ ортогонален $\mathbf{e}_2, \ldots, \mathbf{e}_m$. Поэтому

$$(\mathbf{x}^{(0)}, \mathbf{g}_1) = \sum_{i=1}^{m} \gamma_i (\mathbf{e}_i, \mathbf{g}_1) = \gamma_1 (\mathbf{e}_1, \mathbf{g}_1).$$

Отсюда следует, что

$$\gamma_1 = \frac{(\mathbf{x}^{(0)}, \mathbf{g}_1)}{(\mathbf{e}_1, \mathbf{g}_1)}.$$

Таким образом, из начального приближения $\mathbf{x}^{(0)}$ можно исключить компоненту, пропорциональную $\mathbf{e}_1$, взяв для итераций в качестве начального вектор

$$\mathbf{y}^{(0)} = \mathbf{x}^{(0)} - \frac{(\mathbf{x}^{(0)}, \mathbf{g}_1)}{(\mathbf{e}_1, \mathbf{g}_1)} \mathbf{e}_1.$$

Далее проводим итерации по формуле $\mathbf{y}^{(n+1)} = \mathbf{A}\mathbf{y}^{(n)}$. Необходимо только учесть замечание 2.10.2. Для этого в процессе итераций время от времени следует повторять процедуру исключения из разложения векторов $\mathbf{y}^{(n)}$ слагаемого, пропорционального $\mathbf{e}_1$, то есть время от времени вектор $\mathbf{y}^{(n)}$ следует заменять на вектор

$$\mathbf{y}^{(n)} - \frac{(\mathbf{y}^{(n)}, \mathbf{g}_1)}{(\mathbf{e}_1, \mathbf{g}_1)} \mathbf{e}_1.$$

В заключение вернемся к вопросу о нахождении собственного числа λ_A матрицы $\mathbf{A}$ ближайшего к заданному числу λ_0. Как отмечалось выше, для этого достаточно найти наибольшее по модулю собственное число λ_B матрицы $\mathbf{B} = (\mathbf{A} - \lambda_0 \mathbf{E})^{-1}$. Согласно описанному в этом параграфе алгоритму, надо вычислять вектора

$$\mathbf{x}^{(n+1)} = \mathbf{B}\mathbf{x}^{(n)} \tag{2.44}$$

после чего находить отношение скалярных произведений. Чтобы проводить вычисления по формуле (2.44) надо найти матрицу $\mathbf{B}$. Равенство (2.44) можно переписать иначе

$$(\mathbf{A} - \lambda_0 \mathbf{E})\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} \tag{2.45}$$

Тогда, для нахождения $\mathbf{x}^{(n+1)}$ надо решить систему линейных алгебраических уравнений. После чего положить

$$\lambda_A^{(n+1)} = \lambda_0 + \frac{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})}{(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)})}. \tag{2.46}$$

Такие итерации принято называть **обратными**, а число λ_0 — **сдвигом**. Заметим, что матрица системы уравнений на каждой итерации одна и та же.

Можно увеличить скорость сходимости, если известно достаточно хорошее приближение λ_0 к искомому собственному числу λ_A. Для этого итерации (2.45),(2.46) достаточно переписать в виде

$$(\mathbf{A} - \lambda_A^{(n)}\mathbf{E})\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)}, \quad \lambda_A^{(n+1)} = \lambda_A^{(n)} + \frac{(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})}{(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)})}. \tag{2.47}$$

Таким образом, на каждой итерации величина сдвига $\lambda_A^{(n)}$ меняется. При этом, для того, чтобы избежать переполнение, после каждой итерации вектор $\mathbf{x}^{(n+1)}$ надо заменить на нормированный вектор. Так матрица $(\mathbf{A} - \lambda_A^{(n)}\mathbf{E})$ при больших n становится плохо обусловленной, при малом значении модуля разности $|\lambda_A^{(n+1)} - \lambda_A^{(n)}|$ итерации следует прекратить или перестать изменять величину сдвига.

2.3.3 Полная проблема собственных чисел

Рассмотрим как решается проблема на примере **метода вращений**, который носит еще название **метод Якоби**.

Напомним сначала некоторые факты из линейной алгебры.

Определение 2.3.2 *Матрицы* $\mathbf{A}$ *и* $\mathbf{B}$ *называются* **подобными**, *если существует такая невырожденная матрица* $\mathbf{D}$, *что* $\mathbf{B} = \mathbf{D}^{-1}\mathbf{A}\mathbf{D}$.

Теорема 2.3.4 *Если матрицы подобны, то множества их собственных чисел совпадают.*

Доказательство. Заметим, что

$$\begin{aligned}\det(\mathbf{B} - \lambda\mathbf{E}) &= \det(\mathbf{D}^{-1}\mathbf{A}\mathbf{D} - \lambda\mathbf{D}^{-1}\mathbf{D}) = \\ &= \det(\mathbf{D}^{-1}(\mathbf{A} - \lambda\mathbf{E})\mathbf{D}) = \det\mathbf{D}^{-1}\det(\mathbf{A} - \lambda\mathbf{E})\det\mathbf{D} = \det(\mathbf{A} - \lambda\mathbf{E}),\end{aligned}$$

так как $\det\mathbf{D}^{-1}\det\mathbf{D} = 1$. Таким образом, $\det(\mathbf{B} - \lambda\mathbf{E}) = 0$ тогда и только тогда, когда $\det(\mathbf{A} - \lambda\mathbf{E}) = 0$. Отсюда следует утверждение теоремы.

Выясним теперь как связаны между собой собственные вектора подобных матриц $\mathbf{A}$ и $\mathbf{B}$. Пусть $\mathbf{e}$ — собственный вектор матрицы $\mathbf{A}$, соответствующий числу λ. Тогда $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$. Перепишем это равенство в виде $\mathbf{A}\mathbf{D}\mathbf{D}^{-1}\mathbf{e} = \lambda\mathbf{D}\mathbf{D}^{-1}\mathbf{e}$. Умножим полученное равенство слева на матрицу $\mathbf{D}^{-1}$. В результате получим $(\mathbf{D}^{-1}\mathbf{A}\mathbf{D})\mathbf{D}^{-1}\mathbf{e} = \lambda(\mathbf{D}^{-1}\mathbf{D})\mathbf{D}^{-1}\mathbf{e}$ или $\mathbf{B}(\mathbf{D}^{-1}\mathbf{e}) = \lambda(\mathbf{D}^{-1}\mathbf{e})$. Это означает, что $\mathbf{D}^{-1}\mathbf{e}$ — собственный вектор матрицы $\mathbf{B}$, соответствующее тому же собственному числу λ. Пусть теперь $\mathbf{f}$ — собственный вектор матрицы $\mathbf{B}$, соответствующий числу λ. Тогда $\mathbf{B}\mathbf{f} = \lambda\mathbf{f}$. Перепишем это равенство в виде $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}\mathbf{f} = \lambda\mathbf{f}$. Умножим слева это равенство на матрицу $\mathbf{D}$. Тогда $\mathbf{A}(\mathbf{D}\mathbf{f}) = \lambda(\mathbf{D}\mathbf{f})$. Следовательно, $\mathbf{D}\mathbf{f}$ — собственный вектор матрицы $\mathbf{A}$, соответствующий числу λ.

Если удастся подобрать матрицу $\mathbf{D}$ так, чтобы матрица $\mathbf{B}$ оказалась диагональной, то задача о нахождении собственных чисел матрицы $\mathbf{A}$ будет решена. Действительно, по теореме матрицы $\mathbf{A}$ и $\mathbf{B}$ имеют одни и те же собственные числа, а собственные числа матрицы $\mathbf{B}$ — это ее диагональные элементы. Легко найти и в этом случае и собственные вектора. Для этого достаточно заметить, что для матрицы $\mathbf{B}$ собственный вектор, соответствующий числу, стоящему на диагонали в k-ой строке имеет все компоненты равные нулю, кроме k-ой , которую можно взять равной 1. Тогда, для получения соответствующего собственного вектора матрицы $\mathbf{A}$, необходимо умножить матрицу $\mathbf{D}$ на этот вектор. В результате получится k-ый столбец матрицы $\mathbf{D}$. Таким образом, собственными векторами матрицы $\mathbf{A}$ являются столбцы матрицы $\mathbf{D}$.

Метод Якоби как раз и позволяет построить матрицу $\mathbf{D}$.

Пусть $\mathbf{A}$ — симметричная матрица и пусть $S_A = \sum\limits_{i,j=1}^{m} a_{ij}^2$ Обозначим через $\mathbf{U}_{kl}$ ортогональную матрицу [10]

$$\mathbf{U}_{kl} = \begin{pmatrix} 1 & & & & & & & 0 \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & \alpha & \cdots & \cdots & \cdots & -\beta \\ & & & \vdots & 1 & & & \vdots \\ & & & \vdots & & \ddots & & \vdots \\ & & & \vdots & & & 1 & \vdots \\ & & & \beta & \cdots & \cdots & \cdots & \alpha \\ 0 & & & & & & & 1 \end{pmatrix}, \quad \alpha^2 + \beta^2 = 1.$$

В этой матрице число α стоит на местах (k,k) [11] и (l,l), остальные диагональные элементы равны 1, число β расположено на месте (l,k), число $-\beta$ — на (k,l), а остальные элементы — нули.

Такая матрица задает поворот в двумерной плоскости, проходящей через оси, соответствующие k-ой и l-ой координатам.

Если $\mathbf{C} = \mathbf{A}\mathbf{U}_{kl}$, то

$$c_{ik} = a_{ik}\alpha + a_{il}\beta, \quad c_{il} = -a_{ik}\beta + a_{il}\alpha,$$
$$c_{ij} = a_{ij} \quad \text{при} \quad 1 \le i,j \le m \quad j \ne k,l.$$

Таким образом, умножение матрицы $\mathbf{A}$ справа на матрицу $\mathbf{U}_{kl}$ изменяет у матрицы $\mathbf{A}$ только элементы k-го и l-го столбцов. Однако, легко видеть, что при этом

$$c_{ik}^2 + c_{il}^2 = a_{ik}^2 + a_{il}^2, \quad 1 \le i \le m,$$

то есть попарные суммы квадратов модулей элементов k-го и l-го столбцов не меняются. Следовательно, $S_A = S_C$.

Аналогичное свойство сохраняется при умножении матрицы $\mathbf{C}$ на $\mathbf{U}_{kl}^{(T)}$ слева (как и ранее индекс "T" означает операцию транспонирования). Матрица $\mathbf{B} = \mathbf{U}_{kl}^{(T)}\mathbf{C}$ отличается от матрицы $\mathbf{C}$ только элементами k-ой и l-ой строк:

$$b_{ki} = c_{ki}\alpha + c_{li}\beta, \quad b_{li} = -c_{ki}\beta + c_{li}\alpha,$$
$$b_{ji} = c_{ji} \quad \text{при} \quad 1 \le i,j \le m \quad j \ne k,l.$$

[10] Напомним, что матрица называется ортогональной, если ее обратная матрица совпадает с транспонированной

[11] Первое число обозначает номер строки матрицы, второе — номер столбца.

Следовательно, матрица $\mathbf{B}$ отличается от матрицы $\mathbf{A}$ лишь двумя строками и двумя столбцами. При этом $S_A = S_B$. Переход от матрицы $\mathbf{A}$ к матрице $\mathbf{B}$ будем в дальнейшем называть вращением.

Разобьем S_A на два слагаемых

$$S_A^{(1)} = \sum_{i=1}^{m} a_{ii}^2, \quad S_A^{(2)} = \sum_{\substack{i,j=1 \\ i \neq j}}^{m} a_{ij}^2$$

и посмотрим, как изменяются $S_A^{(1)}$ и $S_A^{(2)}$ после умножения $\mathbf{A}$ на $\mathbf{U}_{kl}$ и $\mathbf{U}_{kl}^{(T)}$.

Недиагональные элементы a_{ik}, a_{il} и a_{ki}, a_{li} при $i \neq k, l$ меняются так, что парные суммы квадратов их модулей сохраняются. Кроме этих элементов вне диагонали есть еще один изменяющийся элемент — a_{kl}. Поэтому $S_A^{(2)}$ меняется настолько, насколько меняется a_{kl}^2. Будем подбирать матрицу $\mathbf{U}_{kl}$ так, чтобы $S_A^{(2)}$ уменьшилось.

Чтобы максимально уменьшить $S_A^{(2)}$, подберем $\mathbf{U}_{kl}$ так, чтобы аннулировать элемент a_{kl}. Из формул для элементов матриц $\mathbf{C}$ и $\mathbf{B}$ следует, что

$$b_{kl} = c_{kl}\alpha + c_{ll}\beta = a_{kl}(\alpha^2 - \beta^2) + (a_{ll} - a_{kk})\alpha\beta.$$

Полагая $b_{kl} = 0$, получим

$$a_{kl}(\alpha^2 - \beta^2) = (a_{kk} - a_{ll})\alpha\beta.$$

Возводя это равенство в квадрат и учитывая условие $\alpha^2 + \beta^2 = 1$, исключим из него β. В результате получим биквадратное уравнение для нахождения α:

$$\alpha^4 - \alpha^2 + a_{kl}^2[4a_{kl}^2 + (a_{kk} - a_{ll})^2]^{-1} = 0.$$

Решая это уравнение и выбирая один из корней, имеем

$$\alpha = \sqrt{\frac{1}{2}\Big(1 + \frac{1}{\sqrt{1+\mu^2}}\Big)}, \quad \text{где} \quad \mu = \frac{2a_{kl}}{a_{kk} - a_{ll}},$$

$$\beta = (\mathrm{sign}\mu)\sqrt{\frac{1}{2}\Big(1 - \frac{1}{\sqrt{1+\mu^2}}\Big)}.$$

Если $a_{kk} = a_{ll}$, то $\alpha = \beta = \sqrt{1/2}$.

Итак, $S_B^{(2)} < S_A^{(2)}$, в то время как $S_B^{(1)} > S_A^{(1)}$, причем $S_B^{(1)} + S_B^{(2)} = S_A^{(1)} + S_A^{(2)}$.

Если подобрать такие вращения, что $S^{(2)} \to 0$, то недиагональные элементы после определенного числа вращений станут пренебрежительно малы, и матрица превратится почти в диагональную. Диагональные элементы и будут искомыми собственными значениями.

Еще не рассматривался вопрос о выборе k и l, то есть о выборе того элемента, который следует аннулировать. Конечно, выгоднее всего аннулировать максимальный недиагональный элемент. Но тогда при каждом вращении потребуется нахождение этого элемента, что повлечет за собой большие затраты машинного времени, при больших m, так как придется перебрать все элементы матрицы, а это потребует $O(m^2)$ операций. Поэтому поступим следующим образом. Составим суммы недиагональных элементов каждой строки:

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} a_{ij}^2,$$

из них выберем наибольшую, а в ней наибольшее слагаемое. Соответствующий элемент матрицы и будем аннулировать. Поиск такого элемента потребует просмотра двух строк: строки, составленной из чисел r_i, и для выбранного i строки $|a_{ij}|$. Вычисление сумм r_i экономично, так как при каждом вращении меняются только две суммы r_k и r_l.

Для доказательства сходимости метода заметим, что аннулируемый элемент составляет не менее $1/(m-1)$ суммы r_k, которая, в свою очередь, не менее $1/m$ части суммы $S_A^{(2)}$. Таким образом, поскольку обращаются в ноль два симметричных элемента, за одно вращение сумма $S_A^{(2)}$ уменьшится не менее, чем на $2S_A^{(2)}/(m(m-1))$. Следовательно,

$$S_B^{(2)} \le S_A^{(2)}\Big(1 - \frac{2}{m(m-1)}\Big).$$

Значит, после n вращений сумма квадратов элементов матрицы, не стоящих на главной диагонали, уменьшится не менее чем в

$$\Big(1 - \frac{2}{m(m-1)}\Big)^n$$

раз и, следовательно, эта сумма стремится к нулю при $n \to \infty$.

Из построения следует, что матрица $\mathbf{D} = \prod \mathbf{U_{kl}}$. Таким образом, вычисляя в процессе вращений $\prod \mathbf{U_{kl}}$, получим при завершении итераций собственные вектора, являющиеся столбцами матрицы $\mathbf{D}$.

Рассмотрим теперь вкратце на примере LR- и QR-алгоритмов как можно решить полную проблему собственных чисел для несимметричных матриц. В этих алгоритмах также как и в методе вращения используется специальным образом подобранное преобразование подобия, которое приводит исходную матрицу к треугольному виду. А как известно, у треугольной матрицы диагональные элементы являются собственными числами.

Начнем с LR-алгоритма. Пусть данная квадратная матрица $\mathbf{A}$ представима в виде произведения $\mathbf{A} = \mathbf{LR}$, где $\mathbf{L}$ левая (нижняя) треугольная матрица, на главной диагонали которой стоят числа 1, а $\mathbf{R}$ — правая (верхняя) треугольная матрица. В этом случае говорят, что матрица $\mathbf{A}$ допускает LR-разложение. При определенных условиях такое разложение существует (см. задачу 3 к этой главе). Тогда $\mathbf{R} = \mathbf{L^{-1}A}$. Пусть $\mathbf{A_1} = \mathbf{RL}$. Подставляя в последнее равенство выражение для $\mathbf{R}$, получаем $\mathbf{A_1} = \mathbf{L^{-1}AL}$. Следовательно, матриц $\mathbf{A}$ и $\mathbf{A_1}$ подобны и поэтому их собственные числа совпадают.

Если матрицу $\mathbf{A_1}$ подобно матрице $\mathbf{A}$ можно разложить на произведение левой и правой треугольных матриц $\mathbf{A_1} = \mathbf{L_1R_1}$, то, положив $\mathbf{A_2} = \mathbf{R_1L_1}$, получим как и ранее, что собственные числа матриц $\mathbf{A_1}$ и $\mathbf{A_2}$ совпадают. Значит, у матриц $\mathbf{A}$ и $\mathbf{A_2}$ одинаковые собственные числа.

Продолжая этот процесс, если только на каждой итерации возможно LR-разложение, получим последовательность матриц $\mathbf{A_n}$, $n = 0, 1, \ldots$, $\mathbf{A_0} = \mathbf{A}$. При определенных ограничениях на исходную матрицу доказано, что матрицы $\mathbf{A_n}$ сходятся к правой треугольной матрице при $n \to \infty$. Примером условия применимости метода является (см. [27]) требование наличия у исходной матрицы m различных по модулю собственных чисел и возможность LR-разложения матриц $\mathbf{P}$ и $\mathbf{P^{-1}}$, где $\mathbf{P}$ и $\mathbf{P^{-1}}$ такие матрицы, что $\mathbf{A} = \mathbf{P^{-1}DP}$, а $\mathbf{D}$ — диагональная матрица.

Заметим, что условие сходимости метода означает, что у матрицы отсутствуют не только кратные собственные числа, но, в случае матрицы из вещественных элементов, комплексные собственные числа. К сожалению, условия сходимости являются

трудно проверяемыми, поэтому они носят скорее теоретический, чем практический характер.

Более распространенным является QR-алгоритм, предложенный в 1961 году. Главное его отличие от LR-алгоритма состоит в том, что матрица $\mathbf{A}$ представляется в виде $\mathbf{A} = \mathbf{QR}$, где $\mathbf{Q}$ ортогональная матрица, а $\mathbf{R}$ — правая треугольная матрица. Доказано, что такое разложение существует для любой квадратной матрицы. Таким образом, алгоритм описывается набором формул

$$\mathbf{A_0} = \mathbf{A}, \quad \mathbf{A_n} = \mathbf{Q_n R_n}, \quad \mathbf{A_{n+1}} = \mathbf{R_n Q_n}, \quad n = 0, 1, \ldots$$

В случае различных по модулю собственных чисел, матрицы $\mathbf{A_n}$ сходятся к правой треугольной матрице. При этом для элементов $a_{ij}^{(n)}$ матрицы $\mathbf{A_n}$, стоящих ниже главной диагонали, имеет место равенство

$$a_{ij}^{(n)} = O\left(\left|\frac{\lambda_i}{\lambda_j}\right|^n\right), \quad i > j,$$

где собственные числа λ_k матрицы занумерованы в порядке убывания модуля. Это означает, что быстрее всего сходится к нулю элементы, стоящие на месте $(m, 1)$. В каждой строке (столбце) медленнее всего сходятся к нулю элементы непосредственно примыкающие к главной диагонали.

В случае кратных или комплексных собственных чисел сходимость матриц в определенном смысле тоже имеется, но при этом предельная матрица имеет более сложный вид. Сходимость теперь, вообще говоря, не поэлементная, а **по форме**. Сходимость по форме означает, что определенные элементы матрицы могут существенно изменяться от итерации к итерации но "предельная"матрица имеет клеточный правый треугольный вид. Характеристический многочлен "предельной"матрицы равен произведению характеристических многочленов ее диагональных клеток. Поэтому, найдя корни характеристических многочленов клеток, получают собственные числа исходной матрицы. Более подробное описание этой ситуации можно найти в [5], [27].

В описанном выше виде, ни LR ни QR-алгоритмы обычно не применяются. Это, прежде всего, связано с тем, что LR- и QR-разложения требуют $O(m^3)$ арифметических операций. Поэтому каждый шаг процесса выполняется слишком медленно. Для сокращения числа арифметических операций на каждой итерации, матрицу $\mathbf{A}$ предварительно приводят к, так называемому, правому почти диагональному виду путем применения преобразования подобия. Говорят, что матрица является **правой почти диагональной** или **матрицей Хессенберга**, если у нее ниже главной диагонали имеется только одна ненулевая диагональ, непосредственно примыкающая к главной. Другими словами, если матрица с элементами $a_{i\,j}$ является матрицей Хессенберга, то $a_{i\,j} = 0$ при $j < i - 1$. Исходная матрица приводится к матрице Хессенберга с помощью, так называемых, **преобразований Хаусхолдера**.

Для описания преобразования Хаусхолдера рассмотрим сначала матрицу $\mathbf{U} = \mathbf{E} - 2\mathbf{w}\mathbf{w}^T$, где $\mathbf{w}$ — некоторый вектор-столбец единичной длины с элементами w_i, $\mathbf{w}^T$ соответствующая вектор-строка с теми же элементами. В соответствии с правилами перемножения матриц $\mathbf{w}\mathbf{w}^T$ — квадратная симметричная матрица с элементами $w_{i\,j} = w_i w_j$, а $\mathbf{w}^T\mathbf{w} = \sum\limits_{i=1}^{m} w_i^2 = 1$. Матрица $\mathbf{U}$ называется **матрицей отражения** или **матрицей Хаусхолдера**. Так как единичная матрица $\mathbf{E}$ и $\mathbf{w}\mathbf{w}^T$ — симметричные матрицы, матрица отражения $\mathbf{U}$ симметрична, то есть $\mathbf{U} = \mathbf{U}^T$. Кроме того,

$$\mathbf{U}^T\mathbf{U} = (\mathbf{E} - \mathbf{2ww^T})(\mathbf{E} - \mathbf{2ww^T}) = \mathbf{E} - \mathbf{4ww^T} + \mathbf{4ww^Tww^T} = \mathbf{E} - \mathbf{4ww^T} + \mathbf{4ww^T} = \mathbf{E}.$$

Это равенство означает, что $\mathbf{U} = \mathbf{U}^T = \mathbf{U}^{-1}$, то есть $\mathbf{U}$ — ортогональная матрица.

Положим $\mathbf{A_1} = \mathbf{A}$, $\mathbf{A_2} = \mathbf{U_1 A_1 U_1}$, $\mathbf{U}_1 = \mathbf{E} - 2\mathbf{w_1 w_1}^T$, где $\mathbf{w_1}$ — некоторый вектор-столбец единичной длины. Тогда матрицы $\mathbf{A_1}$ и $\mathbf{A_2}$ подобны и, следовательно, имеют одни и те же собственные числа.

Вектор $\mathbf{w_1}$ подбирается так, чтобы все элементы первого столбца матрицы $\mathbf{A_2}$ кроме, быть может, первых двух, обратились в ноль. Для этого заметим сначала, что

$$\mathbf{A_2} = (\mathbf{E} - 2\mathbf{w_1 w_1}^T)\mathbf{A_1}(\mathbf{E} - 2\mathbf{w_1 w_1}^T) = \mathbf{A_1} - 2\mathbf{w_1 w_1}^T\mathbf{A_1} - 2\mathbf{A_1 w_1 w_1}^T + 4\mathbf{w_1 w_1}^T\mathbf{A_1 w_1 w_1}^T.$$

Если у вектора $\mathbf{w_1}$ первая координата равна нулю, то у матрицы $\mathbf{w_1}\mathbf{w_1}^T$ нулевыми будут первый столбец и первая строка. Тогда у матриц $\mathbf{A_1}\mathbf{w_1}\mathbf{w_1}^T$ и $\mathbf{w_1}\mathbf{w_1}^T\mathbf{A_1}\mathbf{w_1}\mathbf{w_1}^T$ нулевым будет первый столбец. Следовательно, у матрицы $\mathbf{A_2}$ первый столбец равен первому столбцу матрицы $\mathbf{A_1}$ минус первый столбец матрицы $2\mathbf{w_1}\mathbf{w_1}^T\mathbf{A_1}$. Обозначим через $\mathbf{a_1^{(k)}}$ первый столбец матрицы $\mathbf{A_k}$, $k = 1, 2$, и $a_{i\,1}^{(k)}$ координаты этого столбца. Пусть $w_i^{(1)}$ — координаты вектора $\mathbf{w_1}$, тогда

$$\mathbf{a_1^{(2)}} = \mathbf{a_1^{(1)}} - 2\mathbf{w_1}\mathbf{w_1}^T\mathbf{a_1^{(1)}} = \begin{pmatrix} a_{1\,1}^{(1)} \\ a_{2\,1}^{(1)} - 2w_2^{(1)}\mathbf{w_1}^T\mathbf{a_1^{(1)}} \\ \cdots\cdots\cdots\cdots\cdots \\ a_{m\,1}^{(1)} - 2w_m^{(1)}\mathbf{w_1}^T\mathbf{a_1^{(1)}} \end{pmatrix}.$$

Покажем, что если координаты вектора $\mathbf{w_1}$ определить следующим образом:

$$\mathbf{w_1}^T = \alpha(0, a_{2\,1}^{(1)} - s, a_{3\,1}^{(1)}, \ldots, a_{m\,1}^{(1)}), \quad s = \operatorname{sign}(a_{2\,1}^{(1)}) \left(\sum_{i=2}^{m} (a_{i\,1}^{(1)})^2\right)^{1/2}, \quad \alpha = \left(2s(s - a_{2\,1}^{(1)})\right)^{-1/2}, \quad (2.48)$$

то длина этого вектора равна 1 и в столбце $\mathbf{a_1^{(2)}}$ все элементы, начиная с третьего, равны нулю. Действительно,

$$\sum_{i=1}^{m} (w_i^{(1)})^2 = \alpha^2 \left((a_{2\,1}^{(1)} - s)^2 + \sum_{i=3}^{m} (a_{i\,1}^{(1)})^2\right) = \alpha^2 \left((a_{2\,1}^{(1)})^2 - 2a_{2\,1}^{(1)}s + s^2 + s^2 - (a_{2\,1}^{(1)})^2\right) = \alpha^2 2s(s - a_{2\,1}^{(1)}) = 1.$$

Далее имеем,

$$\mathbf{w_1}^T\mathbf{a_1^{(1)}} = \alpha \left((a_{2\,1}^{(1)} - s)a_{2\,1}^{(1)} + \sum_{i=3}^{m} (a_{i\,1}^{(1)})^2\right) = \alpha \left(s^2 - a_{2\,1}^{(1)}s\right) = \frac{1}{2\alpha}.$$

Поэтому

$$a_{2\,1}^{(1)} - 2w_2^{(1)}\mathbf{w_1}^T\mathbf{a_1^{(1)}} = a_{2\,1}^{(1)} - \frac{2\alpha(a_{2\,1}^{(1)} - s)}{2\alpha} = s, \quad a_{i\,1}^{(1)} - 2w_i^{(1)}\mathbf{w_1}^T\mathbf{a_1^{(1)}} = a_{i\,1}^{(1)} - \frac{2\alpha a_{i\,1}^{(1)}}{2\alpha} = 0, \quad i = 3, \ldots, m.$$

Таким образом, первый столбец матрицы $\mathbf{A_2}$ имеет нужную форму.

Положим теперь $\mathbf{A_3} = \mathbf{U_2}\mathbf{A_2}\mathbf{U_2}$, $\mathbf{U}_2 = \mathbf{E} - 2\mathbf{w_2}\mathbf{w_2}^T$, где $\mathbf{w_2}$ — вектор, две первые координаты которого равны нулю. Тогда у матрицы $\mathbf{w_2}\mathbf{w_2}^T$ две первые строки и два первых столбца нулевые. Поэтому у матрицы $\mathbf{U}_2$ в первой строке и первом столбце все элементы равны нулю, кроме первого, который равен 1, во втором столбце и второй строке отличен от нуля только второй элемент, который также равен 1. Поэтому у матрицы $\mathbf{A_3}$ первый столбец совпадает с первым столбцом матрицы $\mathbf{A_2}$. По аналогии с (2.48) положим

$$\mathbf{w_2}^T = \alpha(0, 0, a_{3\,2}^{(2)} - s, a_{4\,2}^{(2)}, \ldots, a_{m\,2}^{(2)}), \quad s = \operatorname{sign}(a_{3\,2}^{(2)}) \left(\sum_{i=3}^{m} (a_{i\,2}^{(2)})^2\right)^{1/2}, \quad \alpha = \left(2s(s - a_{3\,2}^{(2)})\right)^{-1/2},$$

где $a_{i\,2}^{(2)}$ — координаты второго столбца матрицы $\mathbf{A_2}$. Подобно тому, как это сделано выше, доказывается, что вектор $\mathbf{w_2}$ имеет единичную длину и у матрицы $\mathbf{A_3}$ во втором столбце только первые три координаты могут быть отличны от нуля.

Продолжая аналогичные преобразования с использованием векторов $\mathbf{w_k}$, $k = 3, \ldots, m-2$, у которых первые k координат равны нулю, а остальные подобраны соответствующим образом, получим в результате матрицу Хессенберга.

Таким образом показано, что для любой квадратной матрицы $\mathbf{A}$ существует такая ортогональная матрица $\mathbf{U}$, что матрица $\mathbf{B} = \mathbf{U}^T\mathbf{A}\mathbf{U}$ является правой почти диагональной матрицей. Матрица $\mathbf{U} = \mathbf{U_1} \cdots \mathbf{U_{m-2}}$, где $\mathbf{U_k}$ — соответствующим образом построенные матрицы Хаусхолдера.

Замечание 1. Если матрица $\mathbf{A}$ симметрична, то матрица $\mathbf{B}$ также симметрична и, следовательно, является трехдиагональной.

Пусть $\mathbf{B}$ матрица Хессенберга, которая получается из матрицы $\mathbf{A}$ путем применения преобразований Хаусхолдера. Матрица $\mathbf{B}$ подобна матрице $\mathbf{A}$ и, следовательно, у них одни и те же собственные числа. Применим QR-алгоритм к матрице $\mathbf{B}$.

Получим QR-разложение матрицы $\mathbf{B}$. Теперь его осуществить достаточно просто с помощью ортогональных матриц $\mathbf{U}_{k\,l}$, введенных при рассмотрении метода вращения, описанного в начале этого параграфа.

Пусть $b_{2\,1} \neq 0$. Возьмем матрицу $\mathbf{U}_{1\,2}$ и рассмотрим произведение $\mathbf{B_1} = \mathbf{U}_{1\,2}^T\mathbf{B}$. Как отмечалось при изучении метода вращения, у полученной матрицы $\mathbf{B_1}$ только первые две строки отличаются от соответствующих строк матрицы $\mathbf{B}$ и элементы $b_{i\,j}^{(1)}$ матрицы $\mathbf{B_1}$ равны

$$b_{1i}^{(1)} = b_{1i}\alpha + b_{2\,i}\beta, \quad b_{2\,i}^{(1)} = -b_{1i}\beta + b_{2\,i}\alpha,$$
$$b_{j\,i}^{(1)} = b_{j\,i} \quad \text{при} \quad 1 \leq i, j \leq m \quad j \neq 1, 2.$$

Осталось подобрать α и β так, чтобы $\alpha^2 + \beta^2 = 1$ и $b_{2\,1}^{(1)} = 0$. Для этого достаточно положить $\alpha = \sin\varphi$, $\beta = \cos\varphi$, $\varphi = \operatorname{arctg}\left(\dfrac{b_{1\,1}}{b_{2\,1}}\right)$. Таким образом, учитывая, что $\mathbf{B}$ — матрица Хессенберга, получаем, что $\mathbf{B_1}$ также является матрицей Хессенберга, то есть $b_{ij}^{(1)} = 0$ при $j < i - 1$ и, кроме того, $b_{2\,1}^{(1)} = 0$.

Пусть теперь $\mathbf{B_2} = \mathbf{U}_{2\,3}^T\mathbf{B_1}$. Тогда

$$b_{2i}^{(2)} = b_{2i}^{(1)}\alpha + b_{3\,i}^{(1)}\beta, \quad b_{3\,i}^{(2)} = -b_{2i}^{(1)}\beta + b_{3\,i}^{(1)}\alpha,$$
$$b_{j\,i}^{(2)} = b_{j\,i}^{(1)} \quad \text{при} \quad 1 \leq i, j \leq m \quad j \neq 2, 3.$$

При этом

$$b_{2\,1}^{(2)} = b_{2\,1}^{(1)}\alpha + b_{3\,1}^{(1)}\beta = 0, \quad b_{3\,1}^{(2)} = -b_{2\,1}^{(1)}\beta + b_{3\,1}^{(1)}\alpha = 0.$$

Подбирая теперь α и β так, чтобы $\alpha^2 + \beta^2 = 1$ и $b_{3\,2}^{(2)} = 0$, получим, что матрицы $\mathbf{B_2}$ является матрицей Хессенберга, причем в первом и втором столбцах у нее ниже главной диагонали элементы равны нулю. Продолжая аналогичным образом, получим, при соответствующем выборе элементов матриц $\mathbf{U}_{i\,i+1}$, треугольную матрицу $\mathbf{R} = \mathbf{U}_{m-1\,m}^T \ldots \mathbf{U}_{1\,2}^T\mathbf{B}$. Отсюда, учитывая ортогональность матриц $\mathbf{U}_{i\,i+1}$, следует, что $\mathbf{B} = \mathbf{QR}$, где $\mathbf{Q} = \mathbf{U}_{1\,2} \ldots \mathbf{U}_{m-1\,m}$ — ортогональная матрица. Итак, QR-разложение матрицы $\mathbf{B}$ получено. Заметим, что оно требует всего $O(m^2)$ операций.

Полученная матрица $\mathbf{Q}$ является матрицей Хессенберга. Это проверяется путем перемножением матриц $\mathbf{U}_{i\,i+1}$. Учитывая, что $\mathbf{R}$ — верхняя треугольная матрица, заключаем теперь, что $\mathbf{RQ}$ также оказывается матрицей Хессенберга. Этот факт является чрезвычайно важным для QR-алгоритма. Он позволяет заключить, что если алгоритм применяется к матрице Хессенберга, то на каждом шаге алгоритма будут возникать матрицы, которые автоматически имеют форму Хессенберга. Таким образом, процедуру приведения исходной матрицы к форме Хессенберга необходимо провести всего лишь один раз перед началом выполнения QR-алгоритма.

Осталось теперь найти собственные вектора. Как отмечалось в начале этого параграфа, если $\mathbf{B} = \mathbf{U}^{-1}\mathbf{AU}$, и $\mathbf{f}$ — собственный вектор матрицы $\mathbf{B}$ соответствующий собственному числу λ, то $\mathbf{Uf}$ — собственный вектор матрицы $\mathbf{A}$, соответствующий тому же собственному числу λ. Поэтому, учитывая ортогональность матрицы преобразования, достаточно найти собственные вектора матрицы $\mathbf{B}$, которая получается в результате приведения к форме Хессенберга матрицы $\mathbf{A}$, после чего нахождение собственных векторов исходной матрицы не составит труда.

Пусть $\widetilde{\lambda}_k$ — приближение к собственному числу λ_k матрицы $\mathbf{B}$. Если бы $\widetilde{\lambda}_k$ в точности совпадало с λ_k, матрица $\mathbf{B} - \widetilde{\lambda}_k\mathbf{E}$ была бы вырожденной и собственный вектор можно было бы найти как некоторое нетривиально решение системы

$$(\mathbf{B} - \widetilde{\lambda}_k\mathbf{E})\mathbf{x} = \mathbf{0}. \tag{2.49}$$

Если же $\widetilde{\lambda}_k \neq \lambda_k$, то система (2.49) вообще говоря имеет только тривиальное решение. В этом случае возьмем в качестве приближенного собственного вектора решение системы

$$(\mathbf{B} - \widetilde{\lambda}_k\mathbf{E})\mathbf{x} = \mathbf{b}, \tag{2.50}$$

где $\mathbf{b}$ — пока что произвольный вектор. Для того чтобы прояснить ситуацию предположим, что у матрицы $\mathbf{B}$ имеется m линейно независимых собственных векторов $\mathbf{f}_1, \ldots, \mathbf{f}_m$, соответствующих собственным числам $\lambda_1, \ldots, \lambda_m$. Тогда, если $\mathbf{b} = \gamma_1\mathbf{f}_1 + \cdots + \gamma_m\mathbf{f}_m$, то

$$\mathbf{x} = (\mathbf{B} - \widetilde{\lambda}_k\mathbf{E})^{-1}\mathbf{b} = \sum_{i=1}^{m} \gamma_i(\mathbf{B} - \widetilde{\lambda}_k\mathbf{E})^{-1}\mathbf{f}_i = \sum_{i=1}^{m} \frac{\gamma_i}{\lambda_i - \widetilde{\lambda}_k}\mathbf{f}_i = \frac{\gamma_k}{\lambda_k - \widetilde{\lambda}_k}\mathbf{f}_k + \sum_{\substack{i=1 \\ i\neq k}}^{m} \frac{\gamma_i}{\lambda_i - \widetilde{\lambda}_k}\mathbf{f}_i. \tag{2.51}$$

Здесь воспользовались там, что матрица $(\mathbf{B} - \widetilde{\lambda}_k\mathbf{E})^{-1}$ имеет те же самые собственные вектора, что и матрица $\mathbf{B}$ и собственные числа $(\lambda_i - \widetilde{\lambda}_k)^{-1}$. Если коэффициент $|\gamma_k|$ не очень мал, а $\widetilde{\lambda}_k$ достаточно

близко к λ_k, то есть $|\lambda_k - \widetilde{\lambda}_k| \ll |\lambda_i - \widetilde{\lambda}_k|$ при $i \neq k$, то второе слагаемое в правой части равенства (2.50) много меньше первого слагаемого. Поэтому

$$\mathbf{x} \approx \frac{\gamma_k}{\lambda_k - \widetilde{\lambda}_k}\mathbf{f}_k.$$

Таким образом, $\mathbf{x}$ можно приближенно считать собственным вектором. При необходимости его можно пронормировать.

Замечание 2. Может оказаться, что у матрицы два близких собственных числа, например λ_k и λ_{k+1}. Тогда $\widetilde{\lambda}_k$ будет близко не только к λ_k, но и к λ_{k+1}. В результате в правой части (2.51) первое слагаемое уже не будет преобладать над суммой. Из суммы надо будет выделить еще одно слагаемое, соответствующее λ_{k+1}. Вектор $\mathbf{x}$ окажется приблизительно равным некоторой линейной комбинации векторов $\mathbf{f}_k$ и $\mathbf{f}_{k+1}$ и, вообще говоря не будет собственным. Для матриц, у которых есть близкие собственные числа, задача нахождения собственных векторов в общем случае является плохо обусловленной.

Замечание 3. Если собственное число λ_k кратное, то в правой части (2.51) из суммы надо выделить все слагаемые, соответствующие числу λ_k. Далее, проведя аналогичные рассуждения, получим, что $\mathbf{x}$ приблизительно является линейной комбинацией собственных векторов, соответствующих числу λ_k и, следовательно, является собственным вектором (разумеется при том условии, что он не является нулевым). Выбирая различные вектора $\mathbf{b}$ можно будет получить различные линейные комбинации и, значит, различные собственные вектора.

Замечание 4 . Может оказаться, что в разложении вектора $\mathbf{b}$ коэффициент γ_k близок к нулю или даже равен нулю. Тогда первое слагаемое в правой части (2.51) не будет доминирующим. Поэтому полезно еще раз решить систему, взяв в качестве вектора $\mathbf{b}$ только что найденный вектор $\mathbf{x}$. У этого вектора будет уже большее значение γ_k и еще одно решение системы даст лучшее приближение. При необходимости эту итерационную процедуру можно повторить, хоть, как привило, одного дополнительного решения оказывается вполне достаточно.

Выбирая различные значения $\mathbf{b}$ можно избежать ситуации, когда $\gamma_k = 0$, так как вероятность того, что у случайно выбранного вектора $\mathbf{b}$ окажется коэффициент $\gamma_k = 0$ очень мала. Впрочем, из-за ошибок округления, коэффициент при $\mathbf{f}_k$ у вектора $\mathbf{x}$ скорее всего, окажется ненулевым, даже, если у вектора $\mathbf{b}$ он был нулевым, поэтому еще несколько итераций приведут к искомому приближению.

Существует целый ряд других модификаций QR-алгоритма, позволяющих увеличить скорость сходимости см. [1], [5], [27].

2.4 ЗАДАЧИ К ГЛАВЕ 2

Будем говорить, что квадратная **матрица имеет простую структуру**, если ее собственные вектора $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, где m размерность матрицы, образуют базис в m-мерном пространстве.

При оценке границ собственных чисел матрицы зачастую полезными бывают **теоремы Гершгорина**.

Теорема 2.4.1 (Первая теорема Гершгорина) *Все собственные числа комплексной матрицы* $\mathbf{A}$ *принадлежат объединению кругов*

$$|z - a_{i\,i}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{m} |a_{i\,j}|, \qquad i = 1, \ldots, m.$$

Теорема 2.4.2 (Вторая теорема Гершгорина) *Если объединение кругов*

$$|z - a_{i\,i}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{m} |a_{i\,j}|, \qquad i = 1, \ldots, m$$

распадается на несколько связных частей, то каждая такая часть содержит столько собственных чисел, сколько кругов ее составляет.

2.4.1 Примеры решения задач

1. Пусть матрица

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 0 & 1 & -1 & \dots & -1 \\ 0 & 0 & 1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

имеет размерность m. Показать, что $M_A = m2^{m-1}$, если $\|\mathbf{x}\| = \max\limits_{i=1,\dots,m} |x_i|$. Таким образом, обусловленность матрицы не зависит от величины определителя, который в данном случае равен 1.

Решение. Возьмем произвольный вектор $\mathbf{b}$ и решим систему $\mathbf{Ax} = \mathbf{b}$. Так как матрица $\mathbf{A}$ — треугольная, решение системы легко выписать:

$$\begin{aligned} x_m &= b_m, \\ x_{m-1} &= b_{m-1} + b_m, \\ x_{m-2} &= b_{m-2} + b_{m-1} + 2b_m, \\ x_{m-3} &= b_{m-3} + b_{m-2} + 2b_{m-1} + 2^2 b_m, \\ &\dots \qquad\qquad\qquad\qquad \dots\dots\dots \\ x_1 &= b_1 + b_2 + 2b_3 + \dots + 2^{m-3} b_{m-1} + 2^{m-2} b_m. \end{aligned}$$

Отсюда получаем обратную матрицу

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 & \dots & 2^{m-3} & 2^{m-2} \\ 0 & 1 & 1 & 2 & \dots & 2^{m-4} & 2^{m-3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

При выбранном в условии задачи определении нормы, норма матрицы определяется как максимальная из сумм, где i-ая сумма составленных их модулей элементов i-ой строки. Следовательно, $\|A^{-1}\| = 1 + 1 + 2 + 2^2 + \dots + 2^{m-2} = 2^{m-1}$, $\|A\| = m$, откуда и следует требуемое утверждение.

2. При каком условии сходится метод простой итерации $\mathbf{x}^{n+1} = \mathbf{B}\mathbf{x}^n + \mathbf{c}$, если

$$\mathbf{B} = \begin{pmatrix} b & a & 0 \\ a & b & a \\ 0 & a & b \end{pmatrix}?$$

Решение. Необходимое и достаточное условие сходимости метода простой итерации при любом начальном приближении — все собственные числа матрицы $\mathbf{B}$ по модулю меньше 1. Поэтому найдем собственные числа матрицы $\mathbf{B}$.

$$\begin{vmatrix} b-\lambda & a & 0 \\ a & b-\lambda & a \\ 0 & a & b-\lambda \end{vmatrix} = (b-\lambda)\big((b-\lambda)^2 - 2a^2\big) = 0.$$

Отсюда, $\lambda = b$, $\lambda = b \pm \sqrt{2}a$. Значит, метод сходится, если $|b| < 1$ и $|b \pm \sqrt{2}a| < 1$.

3. Пусть $\mathbf{A}$ — матрица второго порядка, причем $a_{ii} \neq 0,\ i = 1, 2$. Докажите, что методы Якоби и Зейделя для системы $\mathbf{Ax} = \mathbf{b}$ сходятся или расходятся одновременно.

Решение. Метод Якоби сходится тогда и только тогда, когда все числа λ, удовлетворяющие уравнению

$$\begin{vmatrix} a_{11}\lambda & a_{12} \\ a_{21} & a_{22}\lambda \end{vmatrix} = 0,$$

по модулю меньше 1 (см. задачу **5.**).

Метод Зейделя сходится тогда и только тогда, когда все числа λ, удовлетворяющие уравнению

$$\begin{vmatrix} a_{11}\lambda & a_{12} \\ a_{21}\lambda & a_{22}\lambda \end{vmatrix} = 0,$$

по модулю меньше 1.

Для метода Якоби из уравнения для λ получаем

$$\lambda = \pm\sqrt{\frac{a_{12}a_{21}}{a_{11}a_{22}}}.$$

Для метода Зейделя —

$$\lambda = 0, \quad \lambda = \frac{a_{12}a_{21}}{a_{11}a_{22}}.$$

Следовательно, условия сходимости для методов Якоби и Зейделя одновременно либо выполняются, либо не выполняются.

Следует отметить, что уже для системы трех уравнений возможна ситуация, когда метод Зейделя сходится, а метом Якоби расходится. Убедитесь в этом для системы с матрицей

$$\begin{vmatrix} 1 & 2 & 2 \\ 2 & 5 & 6 \\ 2 & 6 & 9 \end{vmatrix} = 0.$$

4. Доказать, что если λ_{min} — минимальное, а λ_{max} — максимальное собственное число симметричной матрицы $\mathbf{A}$, то справедливы неравенства

$$\lambda_{min} \leq \min_{1\leq i\leq m} a_{ii}, \quad \lambda_{max} \geq \max_{1\leq i\leq m} a_{ii}.$$

Решение. Из линейной алгебры известно, что для симметрической матрицы существует ортонормированный базис, состоящий из собственных векторов этой матрицы. Пусть $\mathbf{e}^{(i)}$ — вектора базиса, то есть

$$\mathbf{A}\mathbf{e}^{(i)} = \lambda_i\mathbf{e}^{(i)}, \quad (\mathbf{e}^{(i)}, \mathbf{e}^{(j)}) = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases}$$

и произвольный вектор $\mathbf{x}$ представим в виде $\mathbf{x} = \sum\limits_{i=1}^{m} \chi_i\mathbf{e}^{(i)}$. При этом имеем

$$\|\mathbf{x}\|^2 = (\mathbf{x}, \mathbf{x}) = \Big(\sum_{i=1}^{m} \chi_i\mathbf{e}^{(i)}, \sum_{j=1}^{m} \chi_j\mathbf{e}^{(j)}\Big) = \sum_{i=1}^{m}\sum_{j=1}^{m} \chi_i\chi_j(\mathbf{e}^{(i)}, \mathbf{e}^{(j)}) = \sum_{i=1}^{m} \chi_i^2.$$

Тогда

$$(\mathbf{Ax},\mathbf{x}) = \Big(\mathbf{A}\sum_{i=1}^{m}\chi_i\mathbf{e}^{(i)},\sum_{j=1}^{m}\chi_j\mathbf{e}^{(j)}\Big) = \Big(\sum_{i=1}^{m}\chi_i\mathbf{Ae}^{(i)},\sum_{j=1}^{m}\chi_j\mathbf{e}^{(j)}\Big) =$$
$$= \Big(\sum_{i=1}^{m}\chi_i\lambda_i\mathbf{e}^{(i)},\sum_{j=1}^{m}\chi_j\mathbf{e}^{(j)}\Big) = \sum_{i=1}^{m}\sum_{j=1}^{m}\chi_i\lambda_i\chi_j(\mathbf{e}^{(i)},\mathbf{e}^{(j)}) = \sum_{i=1}^{m}\lambda_i\chi_i^2.$$

Отсюда получаем, что для любого вектора $\mathbf{x}$:

$$\lambda_{min}\|\mathbf{x}\|^2 = \lambda_{min}\sum_{i=1}^{m}\chi_i^2 \le \sum_{i=1}^{m}\lambda_i\chi_i^2 = (\mathbf{Ax},\mathbf{x}) \le \lambda_{max}\sum_{i=1}^{m}\chi_i^2 = \lambda_{max}\|\mathbf{x}\|^2.$$

Если в этом неравенстве взять вектор $\mathbf{x}$ таким, что все его координаты, кроме i-ой равны нулю, а i-ая координата равна 1, и учесть, что для выбранного подобным образом вектора $\|\mathbf{x}\| = 1$, $(\mathbf{Ax},\mathbf{x}) = a_{i\,i}$, получим

$$\lambda_{min} \le a_{i\,i} \le \lambda_{max}.$$

Так как индекс i может принимать любые значения между 1 и m, получаем теперь, что требуемое утверждение очевидно.

5. Оцените сверху и снизу число обусловленности M_A матрицы

$$\mathbf{A} = \begin{vmatrix} 1000 & 20 & 200 & 1 \\ 20 & 50 & 6 & -2 \\ 200 & 6 & 570 & -1 \\ 1 & -2 & -1 & 10 \end{vmatrix},$$

если $\|\mathbf{x}\| = \sqrt{\sum\limits_{i=1}^{m} x_i^2}$.

Решение. Было показано, что для симметричной матрицы при оговоренном в условии задачи выборе нормы, число обусловленности равно отношению наибольшего по модулю собственного числа к наименьшему по модулю собственному числу. Так как матрица $\mathbf{A}$ симметрична, все ее собственные числа действительны и, согласно первой теореме Гершгорина, лежат в объединении отрезков

$$[779, 1221],\ [22, 78],\ [363, 777],\ [6, 14].$$

Отсюда следует, что все собственные числа положительны, поэтому наибольшее по модулю собственное число совпадает с наибольшим, а наименьшее по модулю — с наименьшим. Учитывая утверждение предыдущей задачи, получаем

$$6 \le \lambda_{min} \le 10, \quad 1000 \le \lambda_{max} \le 1221.$$

Следовательно,

$$100 = \frac{1000}{10} \le M_A = \frac{\lambda_{max}}{\lambda_{min}} \le \frac{1221}{6} = 203.5.$$

6. Доказать, неравенство

$$\frac{\|(\mathbf{A}+\Delta\mathbf{A})^{-1} - \mathbf{A}^{-1}\|}{\|(\mathbf{A}+\Delta\mathbf{A})^{-1}\|} \le M_A\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}.$$

Решение. Обозначим для краткости $\mathbf{B} = \mathbf{A} + \Delta\mathbf{A}$. Имеем

$$\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\mathbf{B}^{-1} = (\mathbf{E} - \mathbf{A}^{-1}\mathbf{B})\mathbf{B}^{-1} = \mathbf{B}^{-1} - \mathbf{A}^{-1},$$

где $\mathbf{E}$ — единичная матрица. Тогда

$$\|\mathbf{B}^{-1} - \mathbf{A}^{-1}\| = \|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\mathbf{B}^{-1}\| \le \|\mathbf{A}^{-1}\|\|\mathbf{A} - \mathbf{B}\|\|\mathbf{B}^{-1}\|.$$

Отсюда

$$\frac{\|\mathbf{B}^{-1} - \mathbf{A}^{-1}\|}{\|\mathbf{B}^{-1}\|} \le \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{A} - \mathbf{B}\|}{\|\mathbf{A}\|},$$

что и требовалось получить.

2.4.2 Задачи

1. Докажите, что число обусловленности матрицы $\mathbf{A}$ равно числу обусловленности матрицы $\mathbf{B}$, если $\mathbf{B} = c\mathbf{A}$, где $c = const$.

2. Пусть матрица $\mathbf{G}$ имеет вид клетки Жордана

$$\mathbf{G} = \begin{pmatrix} 1 & g & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & g & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & g & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & g \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

и в пространстве векторов введена норма $\|\mathbf{x}\| = \max\limits_{i=1,\ldots,m} |x_i|$. Исследовать зависимость числа обусловленности матрицы $\mathbf{G}$ от числа g.

3. Доказать следующие утверждения:

a) Пусть $\mathbf{L_k^-}$ — нижняя треугольная матрица размера $m \times m$, причем $k < m$, у которой все диагональные элементы равны 1, элемент стоящий на месте $(k+1,\ k)$ равен $-C_{k+1\,k}$, на месте $(k+2,\ k)$ равен $-C_{k+2\,k}$,..., на месте $(m,\ k)$ равен $-C_{m\,k}$, а все остальные элементы равны нулю. Показать, что умножение квадратной матрицы $\mathbf{A}$ слева на матрицу $\mathbf{L_k^-}$ равносильно тому, что у матрицы $\mathbf{A}$ из $(k+1)$-ой строки вычитается k-ая, умноженная на $C_{k+1\,k}$, из $(k+2)$-ой строки вычитается k-ая, умноженная на $C_{k+2\,k}$ и так далее.

b) Обратная матрица к матрице $\mathbf{L_k^-}$, которую обозначим $\mathbf{L_k}$, имеет следующий вид: все диагональные элементы равны 1, элемент стоящий на месте $(k+1,\ k)$ равен $C_{k+1\,k}$, на месте $(k+2,\ k)$ равен $C_{k+2\,k}$,..., на месте $(m,\ k)$ равен $C_{m\,k}$, а все остальные элементы равны нулю.

c) У матрицы $\mathbf{L} = \mathbf{L_1} \cdot \mathbf{L_2} \ldots \mathbf{L_m}$ первый столбец совпадает с первым столбцом матрицы $\mathbf{L_1}$, второй — со вторым столбцом матрицы $\mathbf{L_2}$ и так далее.

d) Если матрица $\mathbf{A}$ такова, что при выполнении прямого хода метода Гаусса на главной диагонали не возникает нулевых элементов, то существуют нижняя (или еще говорят левая) треугольная матрица $\mathbf{L}$, у которой на главной диагонали стоят 1, и верхняя (ее иногда называют правой) треугольная матрица $\mathbf{R}$ такие, что $\mathbf{A}$=$\mathbf{LR}$. Матрица $\mathbf{R}$ совпадает с матрицей, которая получается в результате прямого хода метода Гаусс. Матрица $\mathbf{L}$ образуется, если ниже главной диагонали ее элементы положить равными множителям $C_{p\,k}$ определенным по формуле (2.3).

Замечание 1. Стандартные обозначения **L** и **R** связаны с английскими словами left и right. Другой, часто используемый стандарт обозначений: **L** и **U** от слов lower и upper.

Замечание 2. Представление матрица **A** в виде произведения матриц **L** и **R** называют **LR-разложением матрицы A** или **факторизацией** матрицы **A**. Если для обозначения матриц используются обозначения **L** и **U**, то говорят о **LU**-разложении.

е) Элементы $l_{ij},\ r_{ij}$ матриц **L** и **R** вычисляются по формулам

$$r_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} r_{kj},\ i \le j, \qquad l_{ij} = \frac{1}{r_{jj}}\Big(a_{ij} - \sum_{k=1}^{i-1} l_{ik} r_{kj}\Big),\ i > j.$$

Определите последовательность, в которой следует вычислять эти элементы.

4. Пусть про матрицу **A** известно, что она имеет простую структуру и что ее собственные числа принадлежат отрезку $[\mu, M],\ \mu > 0$. Доказать, что при любом положительном значении итерационного параметра τ сходится следующий итерационный метод

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \tau\Big(\mathbf{b} - \frac{1}{2}\mathbf{A}(\mathbf{x}_{n+1} + \mathbf{x}_n)\Big).$$

К чему сходится этот итерационный процесс? Определить оптимальное значение итерационного параметра.

5. Докажите, что если $\mathbf{Ax} = \mathbf{b}$, а $(\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$, то

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x} + \Delta\mathbf{x}\|} \le M_A \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}.$$

6. Для системы линейных алгебраических уравнений с матрицей

$$\mathbf{A} = \begin{pmatrix} b & a & 0 \\ a & b & a \\ 0 & a & b \end{pmatrix}$$

определить необходимые и достаточные условия сходимости методов Якоби и Зейделя.

Ответ. Для обоих методов условие сходимости $|a/b| < 1/\sqrt{2}$.

7. Система уравнений $\mathbf{Ax} = \mathbf{b}$ с матрицей

$$\mathbf{A} = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$$

решается методом Зейделя. Доказать, что при любом начальном приближении итерации сходятся, если $|a| < 1$. Если же $|a| > 1$, то существует начальное приближение, при котором итерации расходятся.

8. Методом Якоби решается система уравнений с матрицей

$$\mathbf{A} = \begin{pmatrix} 1 & 0.2 & 0.3 & -0.4 \\ -1.2 & 3.4 & 1.2 & -0.5 \\ 2.8 & 0 & 6.2 & 3 \\ 1.1 & 2.1 & 3.3 & 7 \end{pmatrix}.$$

Исследовать метод на сходимость.

9. Вывести формулы левой прогонки, то есть прогонки, при которой решение системы ищется в виде $x_{i+1} = \gamma_i x_i + \delta_i$, где γ_i, δ_i — прогоночные коэффициенты.

10. Задана система уравнений

$$\begin{cases} c_1x_1 + d_1x_2 + e_1x_3 & = f_1, \\ b_2x_1 + c_2x_2 + d_2x_3 + e_2x_4 & = f_2, \\ a_kx_{k-2} + b_kx_{k-1} + c_kx_k + d_kx_{k+1} + e_kx_{k+2} & = f_k, \quad k = 3, \ldots, n-2, \\ a_{n-1}x_{n-3} + b_{n-1}x_{n-2} + c_{n-1}x_{n-1} + d_{n-1}x_n & = f_{n-1}, \\ a_nx_{n-2} + b_nx_{n-1} + c_nx_n & = f_n. \end{cases}$$

Матрица такой системы называется пятидиагональной. Вывести формулы прогонки для решения данной системы.

Указание. Решение системы искать в виде $x_k = \alpha_k x_{k-1} + \beta_k x_{k-2} + \gamma_k$.

11. Следом матрицы называется сумма ее диагональных элементов. Доказать, что след матрицы равен сумме всех ее собственных чисел.

12. Имеется программа, позволяющая найти собственное число матрицы, модуль которого наибольший. Как с помощью этой программы найти наибольшее и наименьшее собственные числа матрицы, если известно, что собственные числа матрицы различны.

13. Пусть матрица $\mathbf{A}$ имеет простую структуру и ее собственные числа таковы, что

$$\lambda_1 = -\lambda_2 > |\lambda_3| \geq \cdots \geq \lambda_m.$$

Предложите модификацию итерационного метода, позволяющую найти λ_1 и $\mathbf{e}_1, \mathbf{e}_2$, где $\mathbf{e}_i$ — собственные вектора, соответствующие собственным числам λ_i.

14. Пусть матрицы $\mathbf{A}$ симметрична и m ее размерность. Будем считать, что собственные вектора матрицы $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ и соответствующие им собственные числа пронумерованы так, что

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|.$$

Множество всевозможных линейных комбинаций векторов $\mathbf{e}_2, \ldots, \mathbf{e}_m$ обозначим через L. Доказать, что при $\mathbf{x}_0 \notin L$ справедлива оценка

$$\frac{(\mathbf{x}_{k+1}, \mathbf{x}_k)}{(\mathbf{x}_k, \mathbf{x}_k)} = \lambda_1 + O(|\lambda_2/\lambda_1|^{2k}),$$

где $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k,\ k = 0, 1, \ldots$

15. Пусть $\mathbf{A}$ — симметричная матрица. Про ее собственные значения известно, что $2 \leq \lambda_i \leq 3,\ i = 1, \ldots, m-1$, а $\lambda_m \approx 1$. Построить итерационный процесс вида $\mathbf{x}_{k+1} = (\mathbf{A} + c\mathbf{E})\mathbf{x}_k,\ \lambda_m^{(k)} = \dfrac{(\mathbf{x}_{k+1}, \mathbf{x}_k)}{(\mathbf{x}_k, \mathbf{x}_k)}$, $\mathbf{E}$ — единичная матрица, c — константа для получения λ_m с наилучшей при данной информации скоростью сходимости.

Ответ: $c = -2.5$

16. Покажите, что матрица

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

имеет LR-разложение, а матрица $\mathbf{A_1} = \mathbf{RL}$ такого разложения не имеет, и, следовательно, в этом случае нельзя применить LR-алгоритм.

2.4.3 Примеры тестовых вопросов к главе 2

1. Какие из перечисленных ниже систем можно решать методом квадратного корня? В ответе в порядка возрастания, через пробел перечислить последовательность номеров. Знаки препинания не ставить.

а) $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$, б) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \\ 3 & 4 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$,

в) $\begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, г) $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

2. Пусть M_A — число обусловленности матрицы

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Какое из утверждений верно?

а) $M_A = 1$;

б) $M_A \leq 2$;

в) $M_A \geq 3$;

г) $M_A = 2.5$.

3. Ищется наибольшее по модулю собственное число матрицы $\mathbf{A}$ методом итераций. Пусть $\mathbf{x^{(k)}} = \mathbf{A}\mathbf{x}^{\mathbf{(k-1)}}$, $\mathbf{x}^{(0)}$ произвольный вектор. Какое из указанных ниже условий может быть выбрано в качестве критерия прекращения итераций?

а) $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\| < \varepsilon$;

б) $\left| \dfrac{(\mathbf{x}^{(k+1)}, \mathbf{x}^{(k)})}{(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})} - \dfrac{(\mathbf{x}^{(k+2)}, \mathbf{x}^{(k+1)})}{(\mathbf{x}^{(k+1)}, \mathbf{x}^{(k+1)})} \right| < \varepsilon$;

в) $\left| \dfrac{(\mathbf{x}^{(k+1)}, \mathbf{x}^{(k)})}{(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})} \right| < \varepsilon$.

4. Пусть $\mathbf{U}$ — матрица вращения. Тогда она обладает свойствами

а) $\mathbf{U} = \mathbf{U}^T$, где индекс "T" означает операцию транспонирования;

б) $\mathbf{U}^{-1} = \mathbf{U}^T$;

в) $\mathbf{U}$ — верхняя треугольная матрица;

г) $\mathbf{U}$ — нижняя треугольная матрица;

д) $\mathbf{U}$ — диагональная матрица.

5. Система уравнений

$$\begin{cases} 3x_1 + x_2 = 1, \\ x_1 + 3x_2 + x_3 = 1, \\ x_2 + 3x_3 + x_4 = 1, \\ \cdots\cdots, \\ x_8 + 3x_9 + x_{10} = 1, \\ x_9 + 3x_{10} = 1 \end{cases}$$

решается методом прогонки. При совершении прямого хода прогонки все действия выполняются точно. В начале же обратного хода прогонки из-за погрешности округления была совершена ошибка при определении x_{10}. Абсолютная величина ошибки равна Δ_{10}. Все дальнейшие действия выполнялись точно. Какое из утверждений относительно полученной абсолютной величины ошибки Δ_1 первой компоненты решения x_1 справедливо?

а) $\Delta_1 = \Delta_{10}$;

б) $\Delta_1 = 3^{10} \cdot \Delta_{10}$;

в) $\Delta_1 < \Delta_{10}$;

г) $\Delta_1 > \Delta_{10}$.

6. Пусть собственные числа матрицы занумерованы в порядке убывания их модуля, $\mathbf{e}_1$ — собственный вектор, соответствующий первому собственному числу матрицы, а $\mathbf{g}_1$ — собственный вектор транспонированной матрицы, соответствующий первому собственному числу. Обозначим через $\mathbf{x}$ произвольный вектор. Для нахождения второго собственного числа в качестве начального вектора следует выбрать

а) $\mathbf{x} - \mathbf{e}_1$;

б) $\mathbf{x} - \mathbf{g}_1$;

в) $\mathbf{x} - \dfrac{(\mathbf{x}, \mathbf{e}_1)}{(\mathbf{e}_1, \mathbf{g}_1)}\mathbf{e}_1$;

г) $\mathbf{x} - \dfrac{(\mathbf{x}, \mathbf{e}_1)}{(\mathbf{e}_1, \mathbf{g}_1)}\mathbf{g}_1$;

д) $\mathbf{x} - (\mathbf{x}, \mathbf{e}_1)\mathbf{e}_1$;

е) $\mathbf{x} - (\mathbf{x}, \mathbf{e}_1)\mathbf{g}_1$;

ж) среди перечисленных выше нет вектора, который следует выбрать в качестве начального.

3 ПРИБЛИЖЕНИЕ ФУНКЦИЙ

Данная глава будет посвящена вопросам приближения функций. Обычно при этом рассматриваются две основные задачи. Первая — **задача интерполирования**, которая состоит в том, что по значениям функции в некоторых точках из области ее определения, восстановить значения функции в других точках области. То есть речь идет о приближении исходной функции, заданной, вообще говоря, таблично, другой функцией из определенного класса так, что в точках таблицы функции совпадают.

Вторая задача — **аппроксимация**. В этом случае функция $f(x)$ заменяется функцией $\varphi(x)$ из некоторого класса так, чтобы отклонение $\|f - \varphi\|$ в некоторой норме было минимальным.

3.1 ИНТЕРПОЛИРОВАНИЕ

Простейшая задача, приводящая к задаче приближения функций заключается в том, что в некоторых точках $x_1, \ldots, x_n$ известны значения функции $f(x)$ и требуется каким-то образом восстановить ее значения в других точках. Иногда при этом известно, что приближенную функцию целесообразно искать в виде $\varphi(x, a_1, \ldots, a_n)$. Если параметры $a_1, \ldots, a_n$ выбирать из соображений $f(x_i) = \varphi(x_i, a_1, \ldots, a_n), \;\; i = 1, \ldots, n$, то точки $x_1, \ldots, x_n$ называют **узлами интерполяции**, а такой способ приближения функции $f(x)$ функцией $\varphi(x, a_1, \ldots, a_n)$ — **интерполяцией** или **интерполированием**. Таким образом, графики интерполирующей и интерполируемой функций в узлах интерполяции совпадают.

Если $x_m = \min\limits_{i=1,\ldots,n} x_i$, $x_M = \max\limits_{i=1,\ldots,n} x_i$, где x_i - узлы интерполяции, а значение функции вычисляется в точке $x \notin [x_m, x_M]$, то вместо термина интерполяция используется термин **экстраполяция**. Например, если ежечасно в течение суток регистрируется температура воздуха, то на основании этих данных можно попытаться рассчитать прогноз значений температуры на последующие сутки. Это прогнозирование и есть решение задачи экстраполяции. Очевидно, что чем дальше узлы экстраполяции от точек, в которых известно значение функции (в данном случае чем больше пройдет времени от тех суток, когда проводились измерения), тем менее точны результаты прогнозирования.

Если функция $\varphi(x, a_1, \ldots, a_n)$ зависит от параметров $a_1, \ldots, a_n$ нелинейно, то интерполяцию называют **нелинейной**. **Линейной** будем считать такую интерполяцию, когда

$$\varphi(x, a_1, \ldots, a_n) = \sum_{i=1}^{n} a_i \varphi_i(x), \tag{3.1}$$

где $\varphi_i(x)$ заданные линейно независимые функции. Тогда задача построения интерполирующей функции, по крайней мере теоретически, легко решается. По определению

интерполяции

$$\sum_{i=1}^{n} a_i \varphi_i(x_k) = f(x_k), \quad k = 1, \ldots, n.$$

Следовательно, для нахождения коэффициентов $a_1, \ldots, a_n$ получена система линейных алгебраических уравнений. Решив эту систему, находим a_i и по формуле (3.1) интерполирующую функцию.

3.1.1 Интерполяционные многочлены Лагранжа и Ньютона

Пусть заданы $n+1$ узел интерполирования $x_i, \quad i = 0, \ldots, n$. Часто в качестве интерполирующей функции выбирается многочлен степени n:

$$L_n(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

который называют **интерполяционным многочленом**. Этот многочлен определяется $(n+1)$-им коэффициентом. Наиболее употребительные формы записи такого многочлена: Лагранжа и Ньютона.

Постараемся представить многочлен $L_n(x)$ в виде линейной комбинации значений функции $f(x)$ в узлах интерполирования:

$$L_n(x) = \sum_{k=0}^{n} C_k(x) f(x_k), \tag{3.2}$$

где $C_k(x)$ — многочлены степени не выше n. Так как по определению интерполяционного многочлена выполняются равенства $L_n(x_i) = f(x_i)$ при $i = 0, 1, \ldots, n$, имеем:

$$\sum_{k=0}^{n} C_k(x_i) f(x_k) = f(x_i), \quad i = 0, \ldots, n.$$

Эти соотношения, очевидно, будут выполнены, если

$$C_k(x_i) = \begin{cases} 1, & \text{при } i = k, \\ 0, & \text{при } i \neq k, \end{cases} \quad i = 0, \ldots, n. \tag{3.3}$$

Равенства (3.3) означают, что $x_0, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$ являются корнями многочлена $C_k(x)$. Поэтому его можно представить в виде

$$C_k(x) = \lambda_k (x - x_0)(x - x_1) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n).$$

Для того, чтобы выполнялось равенство $C_k(x_k) = 1$, необходимо выбрать λ_k в виде

$$\lambda_k = \Big(\prod_{\substack{j=0 \\ j \neq k}}^{n} (x_k - x_j) \Big)^{-1}.$$

Таким образом, подставляя найденное выражение для $C_k(x)$ в (3.2), получим **интерполяционный многочлен Лагранжа** :

$$L_n(x) = \sum_{k=0}^{n} \frac{\prod\limits_{\substack{j=0 \\ j \neq k}}^{n} (x - x_j)}{\prod\limits_{\substack{j=0 \\ j \neq k}}^{n} (x_k - x_j)} f(x_k). \tag{3.4}$$

Формула (3.4) может быть записана в несколько ином виде. Введем для этого функцию

$$\omega(x) = (x - x_0)(x - x_1)\dots(x - x_{n-1})(x - x_n). \tag{3.5}$$

Тогда

$$\prod_{\substack{j=0\\ j\neq k}}^{n}(x - x_j) = \frac{\omega(x)}{(x - x_k)}, \quad \prod_{\substack{j=0\\ j\neq k}}^{n}(x_k - x_j) = \omega'(x_k).$$

Поэтому

$$L_n(x) = \sum_{k=0}^{n} \frac{\omega(x)}{(x - x_k)\omega'(x_k)} f(x_k). \tag{3.6}$$

Получим теперь другое представление для $L_n(x)$, которое напоминает формулу Тейлора. Введем для этого некоторые определения.

Пусть в узлах $x_k,\ k = 0, \dots n$ известны значения функции $f(x)$. **Разделенными разностями первого порядка** называются отношения

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i},\ i, j = 0, 1, \dots, n,\ i \neq j.$$

Разделенные разности второго порядка это

$$f(x_i, x_j, x_k) = \frac{f(x_j, x_k) - f(x_i, x_j)}{x_k - x_i}.$$

Аналогично, если известны разделенные разности k-го порядка

$$f(x_j, x_{j+1}, \dots, x_{j+k}), \quad f(x_{j+1}, x_{j+2}, \dots, x_{j+k+1}),$$

то **разделенная разность $(k+1)$-го порядка** имеет вид:

$$f(x_j, \dots, x_{j+k+1}) = \frac{f(x_{j+1}, \dots, x_{j+k+1}) - f(x_j, \dots, x_{j+k})}{x_{j+k+1} - x_j}.$$

Лемма 3.1.1 *Для разделенной разности k-го порядка справедливо представление:*

$$f(x_j, \dots, x_{j+k}) = \sum_{i=j}^{j+k} \frac{f(x_i)}{\prod\limits_{\substack{l=j\\ l\neq i}}^{j+k}(x_i - x_l)}. \tag{3.7}$$

Доказательство. Воспользуемся методом математической индукции. При $k = 1$ имеем:

$$\sum_{i=j}^{j+1} \frac{f(x_i)}{\prod\limits_{\substack{l=j\\ l\neq i}}^{j+1}(x_i - x_l)} = \frac{f(x_j)}{x_j - x_{j+1}} + \frac{f(x_{j+1})}{x_{j+1} - x_j} = \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} = f(x_j, x_{j+1}).$$

Предположим теперь, что для разделенной разности k-го порядка формула (3.7) справедлива. Покажем, что она верна и для разделенной разности $k+1$-го порядка.

$$f(x_j,\dots,x_{j+k+1}) = \frac{f(x_{j+1},\dots,x_{j+k+1}) - f(x_j,\dots,x_{j+k})}{x_{j+k+1}-x_j} =$$

$$= \frac{1}{x_{j+k+1}-x_j}\left(\sum_{i=j+1}^{j+k+1}\frac{f(x_i)}{\prod\limits_{\substack{l=j+1\\ l\neq i}}^{j+k+1}(x_i-x_l)} - \sum_{i=j}^{j+k}\frac{f(x_i)}{\prod\limits_{\substack{l=j\\ l\neq i}}^{j+k}(x_i-x_l)}\right) =$$

$$= \frac{f(x_j)}{(x_j-x_{j+k+1})\prod\limits_{l=j+1}^{j+k}(x_j-x_l)} +$$

$$+\frac{f(x_{j+1})}{x_{j+k+1}-x_j}\left(\frac{1}{\prod\limits_{l=j+2}^{j+k+1}(x_{j+1}-x_l)} - \frac{1}{\prod\limits_{\substack{l=j\\ l\neq j+1}}^{j+k}(x_{j+1}-x_l)}\right)+\dots+$$

$$+\frac{f(x_{j+k})}{x_{j+k+1}-x_j}\left(\frac{1}{\prod\limits_{\substack{l=j+1\\ l\neq j+k}}^{j+k+1}(x_{j+k}-x_l)} - \frac{1}{\prod\limits_{l=j}^{j+k-1}(x_{j+k}-x_l)}\right)+$$

$$+\frac{f(x_{j+k+1})}{(x_{j+k+1}-x_j)\prod\limits_{l=j+1}^{j+k}(x_{j+k+1}-x_l)} =$$

$$= \frac{f(x_j)}{\prod\limits_{l=j+1}^{j+k+1}(x_j-x_l)} + \frac{f(x_{j+1})}{(x_{j+k+1}-x_j)\prod\limits_{l=j+2}^{j+k}(x_{j+1}-x_l)}\left(\frac{1}{x_{j+1}-x_{j+k+1}} - \frac{1}{x_{j+1}-x_j}\right)+\dots+$$

$$+\frac{f(x_{j+k})}{(x_{j+k+1}-x_j)\prod\limits_{l=j+1}^{j+k-1}(x_{j+k}-x_l)}\left(\frac{1}{x_{j+k}-x_{j+k+1}} - \frac{1}{x_{j+k}-x_j}\right) + \frac{f(x_{j+k+1})}{\prod\limits_{l=j}^{j+k}(x_{j+k+1}-x_l)} =$$

$$= \frac{f(x_j)}{\prod\limits_{l=j+1}^{j+k+1}(x_j-x_l)} + \frac{f(x_{j+1})}{\prod\limits_{\substack{l=j\\ l\neq j+1}}^{j+k+1}(x_{j+1}-x_l)} + \dots + \frac{f(x_{j+k})}{\prod\limits_{\substack{l=j\\ l\neq j+k}}^{j+k+1}(x_{j+k}-x_l)} + \frac{f(x_{j+k+1})}{\prod\limits_{l=j}^{j+k}(x_{j+k+1}-x_l)} =$$

$$= \sum_{i=j}^{j+k+1}\frac{f(x_i)}{\prod\limits_{\substack{l=j\\ l\neq i}}^{j+k+1}(x_i-x_l)}.$$

Из формулы (3.7) в частности следует, что

$$f(x_0,\dots,x_n) = \sum_{i=0}^{n}\frac{f(x_i)}{\prod\limits_{\substack{l=0\\ l\neq i}}^{n}(x_i-x_l)}. \tag{3.8}$$

Интерполяционным многочленом Ньютона называется многочлен

$$P_n(x) = f(x_0) + (x - x_0)f(x_0, x_1) + (x - x_0)(x - x_1)f(x_0, x_1, x_2) + \cdots + \\ + (x - x_0)(x - x_1)\dots(x - x_{n-1})f(x_0, \dots, x_n). \quad (3.9)$$

Покажем, что $P_n(x) = L_n(x)$, то есть интерполяционный многочлен Ньютона является другой формой записи интерполяционного многочлена Лагранжа. Для этого положим $L_0(x) = f(x_0)$ и представим $L_n(x)$ в виде

$$L_n(x) = L_0(x) + \sum_{j=1}^{n}(L_j(x) - L_{j-1}(x)). \quad (3.10)$$

Так как из условий интерполяции $L_j(x_k) = L_{j-1}(x_k) = f(x_k)$ для $k = 0, \dots, j-1$, то $L_j(x) - L_{j-1}(x)$ — многочлен степени j, который обращается в 0 в j точках: $x_0, \dots x_{j-1}$. Поэтому

$$L_j(x) - L_{j-1}(x) = A_j(x - x_0)\dots(x - x_{j-1}), \quad (3.11)$$

где A_j некоторый коэффициент. Из равенства $L_j(x_j) = f(x_j)$ следует, что

$$A_j = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0)(x_j - x_1)\dots(x_j - x_{j-1})}.$$

Используя формулу (3.4) для $L_{j-1}(x_j)$ получим

$$A_j = \frac{f(x_j)}{(x_j - x_0)\dots(x_j - x_{j-1})} - \\ - \sum_{k=0}^{j-1} \frac{f(x_k)}{x_j - x_k} \frac{1}{(x_k - x_0)(x_k - x_1)\dots(x_k - x_{k-1})(x_k - x_{k+1})\dots(x_k - x_{j-1})} = \\ = \sum_{k=0}^{j} \frac{f(x_k)}{(x_k - x_0)\dots(x_k - x_{k-1})(x_k - x_{k+1})\dots(x_k - x_{j-1})(x_k - x_j)}$$

Таким образом, из (3.8) имеем

$$A_j = f(x_0, \dots, x_j).$$

Отсюда и из формул (3.10), (3.11) получаем требуемое представление.

Замечание 1. Формулу Ньютона удобно применять в том случае, когда интерполируется одна и та же функция $f(x)$, но число узлов интерполяции постоянно увеличивается. Тогда к уже полученному результату просто добавляются новые слагаемые. При использовании же формулы Лагранжа добавление каждой точки приводит к полному пересчету всех слагаемых.

Замечание 2. В том случае, когда число узлов интерполяции меняться не будет, формулу Ньютона можно записать с использованием обобщенной **схемы Горнера**

$$P_n(x) = (\cdots((f(x_0, \dots, x_n)(x - x_{n-1}) + f(x_0, \dots, x_{n-1}))(x - x_{n-2}) + \\ + f(x_0, \dots, x_{n-2}))(x - x_{n-3}) + \cdots + f(x_0, x_1))(x - x_0) + f(x_0). \quad (3.12)$$

Вычисление по формуле (3.9), при уже подсчитанных разделенных разностях, требует примерно $3n$ арифметических операций, а по формуле (3.12) – $2n$.

Замечание 3. При выводе формулы Ньютона не предполагалось, что узлы интерполяции $x_0, \ldots, x_n$ расположены в определенном порядке. Поэтому можно, перенумеровав узлы, получить другое представление для интерполяционного многочлена:

$$\widetilde{P}_n(x) = f(x_n) + (x - x_n)f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1})f(x_n, x_{n-1}, x_{n-2}) + \cdots + \\ + (x - x_n)\ldots(x - x_1))f(x_n, x_{n-1}, \ldots, x_0). \quad (3.13)$$

Если узлы интерполяции пронумерованы так, что $x_0 < x_1 < \cdots < x_n$, то формулу (3.9) называют формулой интерполирования вперед, а (3.13) — назад.

3.1.2 Погрешность интерполирования

В этом параграфе будет получена оценка для ошибки при интерполировании многочленом Лагранжа гладкой функции $f(x)$, заданной на отрезке $[a, b]$. При замене $f(x)$ на $L_n(x)$ возникает ошибка $r_n(x) = f(x) - L_n(x)$, которую называют **погрешность интерполирования** или **остаточным членом интрполяционной формулы**.

Очевидно, что в узлах интерполирования $r_n(x) = 0$. Для оценки $r_n(x)$ в остальных точках отрезка $[a, b]$ рассмотрим функцию

$$g(s) = f(s) - L_n(s) - k\omega_n(s),$$

где k — константа, а функция $\omega_n(s) = (s - x_0)\cdots(s - x_n)$. Найдем величину r_n в некоторой фиксированной точке x. Выберем k так, чтобы в выбранной точке x выполнялось равенство $g(x) = 0$. Тогда

$$k = \frac{f(x) - L_n(x)}{\omega_n(x)}. \quad (3.14)$$

Будем считать, что интерполируемая функция $f(x)$ имеет $n + 1$ непрерывную производную. Тогда $g(s)$ имеет не менее $n + 2$ нулей (в точках $x, x_0, \ldots, x_n$). Значит, по теореме Ролля, $g'(x)$ имеет не менее $(n+1)$ нулей, g'' — n нулей и так далее, $g^{(n+1)}$ - по крайней мере один ноль. Следовательно, существует точка ξ такая, что $g^{(n+1)}(\xi) = 0$. Учитывая вид $g(s)$, и, принимая во внимание, что ω_n — многочлен степени $n + 1$, имеем

$$g^{(n+1)}(s) = f^{(n+1)}(s) - k \cdot (n + 1)!.$$

Тогда $f^{(n+1)}(\xi) - k \cdot (n + 1)! = 0$. Подставляя сюда выражение для k (3.14), получим:

$$f^{(n+1)}(\xi) = k \cdot (n + 1)! = \frac{f(x) - L_n(x)}{\omega_n(x)} \cdot (n + 1)!,$$

то есть

$$r_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!}\omega_n(x). \quad (3.15)$$

Тем самым получено представление для погрешности интерполяции. Из этого представления следует оценка погрешности:

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{(n + 1)!}|\omega_n(x)|, \quad (3.16)$$

где

$$M_{n+1} = \max_{x \in [a,b]} |f^{(n+1)}(x)|.$$

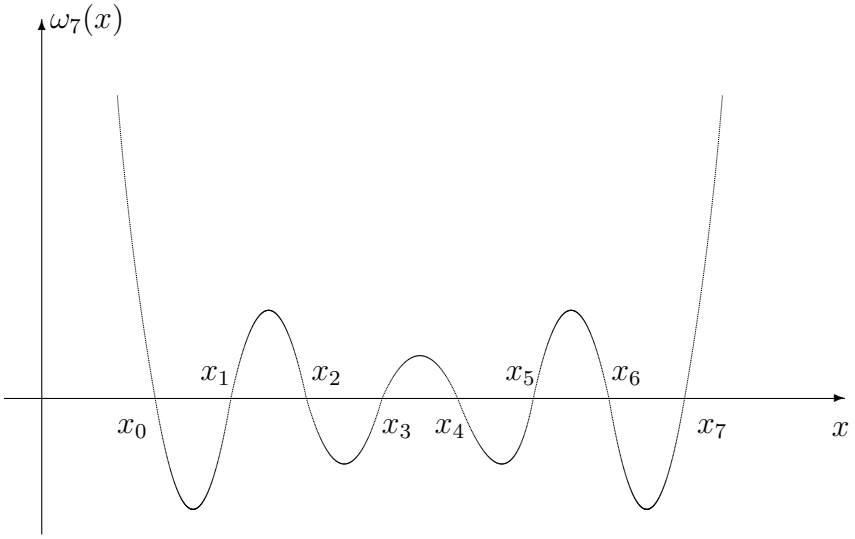


Рис. 3.1 График функции $\omega_7(x)$

Для анализа оценки погрешности предположим, что узлы интерполяции занумерованы в порядке возрастания и расположены равномерно, то есть шаг таблицы задания функции $h = x_{i+1} - x_i = const$. Тогда схематически график функции $\omega_n(x)$ имеет вид, изображенный на рисунке 3.1. Из него видно, что максимальные по модулю значения функции $\omega_n(x)$ на каждом промежутке $[x_i, x_{i+1}]$ растут по мере удаления от середины отрезка $[x_0, x_n]$. За пределами этого отрезка модуль функции $\omega_n(x)$ растет очень быстро по мере удаления от отрезка. Этим объясняется почему при экстраполяции ошибки всегда больше, чем при интерполяции. Поэтому, на практике, при нахождении значения функции в точке x с помощью интерполирования, узлы интерполяции стараются выбрать так, чтобы эта точка была как можно ближе к середине множества узлов.

Если число n нечетное, то есть $n = 2k + 1$, то из симметрии следует, что в центральном интервале экстремум функции $|\omega_n(x)|$ достигается точно в середине. В точке $\tilde{x}$, расположенной в середине центрального интервала и, значит, в середине отрезка $[x_0, x_n]$

$$|\omega_n(\tilde{x})| = \left[\frac{h}{2} \cdot \frac{3h}{2} \cdot \frac{5h}{2} \cdot \ldots \cdot \frac{(2k+1)h}{2}\right]^2 = \left[\frac{(2k+1)!!h^{k+1}}{2^{k+1}}\right]^2,$$

где $(2k+1)!!$ — произведение всех нечетных чисел от 1 до $2k+1$. Тогда для значений аргумента из центрального интервала, то есть для $x \in [x_k, x_{k+1}]$ справедлива оценка

$$|f(x) - L_n(x)| \le M_{n+1}\Big(\frac{h}{2}\Big)^{n+1} \frac{(n!!)^2}{(n+1)!}.$$

Отсюда видно, что погрешность убывает как h^{n+1}. Значит, для повышения точности интерполирования можно пойти по пути уменьшения расстояния между узлами, то есть воспользоваться таблицей с более мелким шагом.

Другой путь повышения точности при фиксированном шаге — увеличение числа узлов. Однако на практике редко пользуются многочленами степени выше 4 - 5. Это связано с тем, что, как правило, не известно как ведет себя величина M_{n+1} в

зависимости от n. У функции вообще может не существовать производных высокого порядка.

Найти $\max\limits_{x\in[x_0,x_n]}|\omega(x)|$ при произвольном значении n в общем случае сложно. Однако, легко видеть, что справедлива грубая оценка

$$\max_{x\in[x_0,x_n]}|\omega_n(x)| < n!h^{n+1}.$$

Поэтому

$$|f(x)-L_n(x)| \le \frac{M_{n+1}}{(n+1)!}|\omega_n(x)| \le \frac{M_{n+1}}{(n+1)}h^{n+1}. \tag{3.17}$$

Это неравенство может быть использовано теперь для определения необходимого числа узлов интерполяции, либо для выбора шаг при фиксированной степени интерполяционного многочлена, если стоит задача приближенного вычисления значения функции f в точке x с заданной точностью ε. При этом предполагается, что точка x лежит между крайними узлами интерполяции и известны оценки для величин M_{n+1}.

Следует отметить, что поскольку многочлены Ньютона и Лагранжа отличаются только формой записи, представление (3.15) и оценка (3.16) справедливы также для формулы Ньютона. Однако, для погрешности интерполирования можно получить другое выражение. Для этого воспользуемся формулой (3.8), записав ее в виде

$$f(x,x_0,\dots,x_n) = \frac{f(x)}{\prod\limits_{l=0}^{n}(x-x_l)} + \sum_{i=0}^{n}\frac{f(x_i)}{(x_i-x)\prod\limits_{\substack{l=0\\l\ne i}}^{n}(x_i-x_l)}.$$

Выразим отсюда $f(x)$. В результате, учитывая формулу (3.4), получим

$$f(x) = \sum_{i=0}^{n}\frac{\prod\limits_{\substack{l=0\\l\ne i}}^{n}(x-x_l)}{\prod\limits_{\substack{l=0\\l\ne i}}^{n}(x_i-x_l)}f(x_i) + f(x,x_0,\dots,x_n)\prod_{l=0}^{n}(x-x_l) =$$

$$= L_n(x) + f(x,x_0,\dots,x_n)\omega_n(x).$$

Таким образом, погрешность интерполирования можно представить в виде

$$f(x) - L_n(x) = f(x,x_0,\dots,x_n)\omega_n(x). \tag{3.18}$$

Сопоставляя (3.18) и (3.15) видим, что на отрезке $[a,b]$ существует такая точка ξ, что

$$f(x,x_0,x_1,\dots,x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3.19}$$

Формула (3.19) устанавливает связь между разделенной разностью $(n+1)$-го порядка и $(n+1)$-ой производной функции $f(x)$.

Из формулы (3.18) можно получить некоторое обоснование следующего правила, применяемого при вычислении приближенного значения функции по формуле Ньютона: вычисляют и суммируют последовательно слагаемые формулы Ньютона до тех пор пока их величина больше допустимой погрешности. Тем самым определяют сколько узлов интерполирования следует подключить в расчет. Такая оценка по первому отброшенному члену формулы Ньютона делается в процессе выполнения вычислений, поэтому ее называют **апостериорной**. Поскольку обычно величины производных искомой функции заранее неизвестны, то на практике редко

удается найти необходимое число узлов на основании **априорной** оценки (то есть сделанной до начала вычислений), полученной из формулы (3.17). Апостериорная оценка не является строгой. Ее применение основано на том, что по формуле (3.19) $(n+1)!f(x_0, x_1, \ldots, x_{n+1}) \approx f^{(n+1)}(\xi)$, поэтому величина первого отброшенного слагаемого формулы Ньютона приблизительно совпадает с погрешностью интерполирования.

Обсудим теперь кратко вопрос о сходимости интерполяционного процесса. Говоря о сходимости интерполяционного процесса, следует сначала сформулировать более точно что под этом понимается.

Множество точек $x_i,\ i = 0, 1 \ldots, n$, таких, что

$$a \le x_0 < x_1 < \cdots < x_n \le b$$

назовем **сеткой** на отрезке $[a, b]$ и обозначим Ω_n. Пусть задана последовательность сеток на отрезке $[a, b]$ с возрастающим числом узлов

$$\Omega_0 = \{x_0^{(0)}\}, \ldots, \Omega_n = \{x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}\}, \ldots$$

и функция $f(x)$ определена на отрезке $[a, b]$. Тогда можно определить последовательность интерполяционных многочленов $L_n(x)$, построенных для этой функции по ее значениям в узлах сетки Ω_n.

Определение 3.1.1 *Говорят, что интерполяционный процесс для функции $f(x)$* **сходится в точке** $x^* \in [a, b]$**,** *если существует предел*

$$\lim_{n\to\infty} L_n(x^*) = f(x^*).$$

Кроме поточечной можно рассматривать сходимость в различных нормах. Например, выбирая в качестве нормы $\|f\| = \max\limits_{x\in[a,b]} |f(x)|$ получим равномерную сходимость.

Определение 3.1.2 *Говорят, что интерполяционный процесс для функции $f(x)$* **сходится равномерно на отрезке** $[a, b]$**,** *если*

$$\lim_{n\to\infty} \max_{x\in[a,b]} |L_n(x) - f(x)| = 0.$$

Ниже приведем без доказательства ряд утверждений, касающихся сходимости интерполяционного процесса. Заметим, что сходимость зависит как от выбора последовательности сеток, так и от интерполируемой функции.

Теорема 3.1.1 (Бернштейн) *Последовательность интерполяционных многочленов, построенных для функции $f(x) = |x|$ по равноотстоящим узлам на отрезке $[-1, 1]$, не сходится к этой функции ни в одной точке отрезка, кроме точек $-1,\ 0, 1$.*

Теорема 3.1.2 (Фабер) *Какова бы ни была последовательность сеток Ω_n, найдется непрерывная на $[a, b]$ функция $f(x)$ такая, что последовательность интерполяционных многочленов $L_n(x)$ не сходится к $f(x)$ равномерно на отрезке $[a, b]$.*

Теорема 3.1.3 (Марцинкевич) *Если $f(x)$ непрерывна на $[a, b]$, то найдется такая последовательность сеток, для которой соответствующий итерполяционный процесс сходится равномерно на $[a, b]$.*

К сожалению построить такие сетки чрезвычайно сложно.

3.1.3 Составление таблиц и обратная интерполяция

В этом параграфе будут рассмотрены некоторые применения полученных ранее результатов.

Рассмотрим следующую задачу. Требуется определить шаг h, с которым необходимо разбить отрезок $[a, b]$, чтобы построить таблицу значений некоторой функции. При этом ставится условие: при нахождении значений функции в промежуточных точках отрезка с помощью интерполяционного многочлена заданной степени m, погрешность не должна превосходить некоторой величины ε. Как правило, строятся таблицы так, чтобы они допускали вычисления с применением линейной интерполяции, то есть функция приближается многочленом степени $m = 1$.

Для вычисления функции $f(x)$ по такой таблице берут узлы x_i и x_{i+1}, которые являются ближайшими к точке x, и полагают по формуле Ньютона

$$f(x) \approx f(x_i) + (x - x_i)f(x_i, x_{i+1}).$$

Как было установлено в предыдущем параграфе погрешность этой формулы имеет вид

$$\frac{f''(\xi)(x - x_i)(x - x_{i+1})}{2}$$

и не превосходит ε, если

$$\frac{1}{2}\max_{\xi\in[a,b]}|f''(\xi)|\cdot\max_{x\in[x_i,x_{i+1}]}\big(|x-x_i||x-x_{i+1}|\big) = \frac{1}{2}\max_{\xi\in[a,b]}|f''(\xi)|\cdot\frac{1}{4}h^2 \le \varepsilon,\ \ h = x_{i+1}-x_i. \quad (3.20)$$

Например, составляется таблица синусов на промежутке $[0, \pi/2]$ так, чтобы $\varepsilon = 0.0001$. Из оценки (3.20) следует (максимальное значение модуля второй производной синуса равно 1), что для подбора шага h достаточно выполнения неравенства

$$\frac{h^2}{8} \le 0.0001.$$

Отсюда имеем: $h \le 0.01 \cdot 2\sqrt{2}$. Таким образом, с учетом удобства вычислений, шаг может быть выбран 0.025.

Рассмотрим теперь так называемую **обратную интерполяцию**.

Пусть известна таблица $y_i = f(x_i)$ значений некоторой функции. В случае монотонности функции $y = f(x)$, существует обратная функция $x = f^{-1}(y)$. При обратной интерполяции приближается функция $x = f^{-1}(y)$. С точки зрения вычислений между прямым и обратным интерполированием нет разницы, надо просто читать таблицу наоборот, то есть считать y_i — аргументом функции, а x_i — ее значением.

Примером применения обратного интерполирования может служить задача о нахождении решения уравнения $f(x) = 0$ для монотонной функции $y = f(x)$. Искомый корень — такое число x, что $x = f^{-1}(0)$. Поэтому достаточно по таблице $y_i = f(x_i)$ приближенно с использованием обратной интерполяции посчитать это значение.

3.1.4 Интерполяционный многочлен Эрмита

В предыдущих параграфах предполагалось, что в каждом узле известно только значение функции. Однако, иногда, помимо значения функции в узле может быть задано еще значение производных. Более общая постановка задачи интерполирования многочленом состоит в следующем.

В узлах $x_k \in [a,b]$, $k = 0, \cdots, m$ заданы значения функции $f(x_k)$ и значения производных $f^{(i)}(x_k)$, $i = 1, 2, \cdots, N_k - 1$. Таким образом, в каждой точке x_k задано N_k величин.

Требуется построить алгебраический многочлен $H_n(x)$ степени не выше n такой, что

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \quad i = 0, \cdots, N_k - 1, \quad k = 0, \cdots, m, \quad n = N_0 + \cdots + N_m - 1.$$

Многочлен $H_n(x)$ называется **интерполяционным многочленом Эрмита** для функции $f(x)$, а число N_k — **кратностью узла** x_k.

Покажем, что многочлен Эрмита всегда существует и единственный. Действительно, $H_n(x)$ содержит $n+1$ коэффициент, для определения которых служит $n+1$ уравнение

$$H_n^{(i)}(x_k) = f^{(i)}(x_k). \tag{3.21}$$

Значит, для того, чтобы эта система линейных алгебраических уравнений имела единственное решение необходимо и достаточно, чтобы однородная система имела только нулевое решение, то есть из равенств $H_n^{(i)}(x_k) = 0$ должно следовать, что все коэффициенты многочлена равны 0. Для доказательства этого предположим, что многочлен не равен тождественно нулю, то есть среди его коэффициентов есть отличные от нуля числа. Равенства $H_n^{(i)}(x_k) = 0$, $i = 0, \cdots, N_k - 1$ означают, что x_k - корни кратности N_k. Тогда, у многочлена $H_n(x)$ имеется, с учетом кратности, $N_0 + \cdots + N_k = n + 1$ корней, а степень многочлена не превосходит n и, значит, он может иметь не более n корней. Полученное противоречие означает, что все коэффициенты многочлена равны 0, что и требовалось доказать.

Так как величины $f^{(i)}(x_k)$ входит только в правую часть системы уравнений (3.21), решение систем зависит от них линейно и поэтому многочлен можно представить в виде линейной комбинации

$$H_n(x) = \sum_{k=0}^{m} \sum_{i=0}^{N_k - 1} C_{ki}(x) f^{(i)}(x_k),$$

где $C_{ki}(x)$ многочлены степени не выше n.

В общем виде формулы для $C_{ki}(x)$ весьма громоздки. Их проще выводить для каждого конкретного случая.

Найдем, например, многочлен $H_3(x)$ такой, что

$$H_3(x_0) = f(x_0), \quad H_3'(x_0) = f'(x_0), \quad H_3(x_1) = f(x_1), \quad H_3'(x_1) = f'(x_1). \tag{3.22}$$

Многочлен $H_3(x)$, будем искать в виде

$$H_3(x) = C_{00}(x) f(x_0) + C_{01}(x) f'(x_0) + C_{10}(x) f(x_1) + C_{11}(x) f'(x_1), \tag{3.23}$$

где C_{ki} — многочлены степени не выше 3. Из (3.23) следует, что для выполнения (3.22) достаточно, чтобы были справедливы равенства:

$$C_{00}(x_0) = 1, \quad C_{00}'(x_0) = 0, \quad C_{00}(x_1) = 0, \quad C_{00}'(x_1) = 0, \tag{3.24}$$
$$C_{01}(x_0) = 0, \quad C_{01}'(x_0) = 1, \quad C_{01}(x_1) = 0, \quad C_{01}'(x_1) = 0, \tag{3.25}$$
$$C_{10}(x_0) = 0, \quad C_{10}'(x_0) = 0, \quad C_{10}(x_1) = 1, \quad C_{10}'(x_1) = 0, \tag{3.26}$$
$$C_{11}(x_0) = 0, \quad C_{11}'(x_0) = 0, \quad C_{11}(x_1) = 0, \quad C_{11}'(x_1) = 1. \tag{3.27}$$

Из третьего и четвертого соотношений равенств (3.24) следует, что x_1 — корень кратности 2 многочлена $C_{00}(x)$. Поэтому этот многочлен имеет вид

$$C_{00}(x) = (Ax + B)(x - x_1)^2,$$

где A и B — некоторые константы. Первое и второе соотношения равенств (3.24) дают:

$$(Ax_0 + B)(x_0 - x_1)^2 = 1, \quad A(x_0 - x_1)^2 + 2(Ax_0 + B)(x_0 - x_1) = 0.$$

Отсюда следует:

$$A = -\frac{2}{(x_0 - x_1)^3}, \qquad B = \frac{1}{(x_0 - x_1)^2} + \frac{2x_0}{(x_0 - x_1)^3} = \frac{3x_0 - x_1}{(x_0 - x_1)^3}.$$

Значит,

$$C_{00}(x) = \frac{(-2x + 3x_0 - x_1)(x - x_1)^2}{(x_0 - x_1)^3}.$$

В силу равенств (3.25) заключаем, что x_1 — корень кратности 2, а x_0 — корень кратности 1 многочлена $C_{01}(x)$. Поэтому $C_{01}(x) = D(x - x_0)(x - x_1)^2$, где D — некоторая константа. Второе из соотношений (3.25) дает $D(x_0 - x_1)^2 = 1$. Значит, $D = 1/(x_0 - x_1)^2$ и

$$C_{01} = \frac{(x - x_0)(x - x_1)^2}{(x_0 - x_1)^2}.$$

Аналогично находим:

$$C_{10}(x) = \frac{(-2x + 3x_1 - x_0)(x - x_0)^2}{(x_1 - x_0)^3}, \quad C_{11} = \frac{(x - x_1)(x - x_0)^2}{(x_1 - x_0)^2}.$$

Подставляя теперь полученные выражения в (3.23) найдем искомый многочлен:

$$H_3(x) = \frac{(-2x + 3x_0 - x_1)(x - x_1)^2}{(x_0 - x_1)^3} f(x_0) + \frac{(x - x_0)(x - x_1)^2}{(x_0 - x_1)^2} f'(x_0) + \\ + \frac{(-2x + 3x_1 - x_0)(x - x_0)^2}{(x_1 - x_0)^3} f(x_1) + \frac{(x - x_1)(x - x_0)^2}{(x_1 - x_0)^2} f'(x_1). \quad (3.28)$$

Замечание. Многочлен Эрмита может быть получен путем применения операции предельного перехода из многочлена Лагранжа (Ньютона). Поясним это на том же примере.

Введем два дополнительных узла x_2 и x_3 и построим по узлам $x_0, \ldots, x_3$ интерполяционный многочлен Лагранжа:

$$L_3(x) = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} f(x_3) + \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} f(x_2) + \\ + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} f(x_1) + \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} f(x_0). \quad (3.29)$$

Зафиксируем теперь точки x, x_0, x_1 и устремим x_2 к x_0, а x_3 к x_1. Сгруппируем в правой части формулы (3.29) первое с третьим и второе с четвертым слагаемые.

Рассмотрим как преобразуются, например, сумма второго и четвертого слагаемых. Обозначим их S_2 и S_4 соответственно.

$$S_2 + S_4 = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} f(x_2) + \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} f(x_0) =$$
$$= \frac{(x-x_1)(x-x_3)}{x_0-x_2}\left(\frac{f(x_0)(x-x_2)}{(x_0-x_1)(x_0-x_3)} - \frac{f(x_2)(x-x_0)}{(x_2-x_1)(x_2-x_3)}\right). \quad (3.30)$$

При x_2 стремящемся к x_0, получаем неопределенность типа $\dfrac{0}{0}$, которую можно раскрыть, применив правило Лопиталя. В результате получим

$$\lim_{x_2\to x_0}(S_2+S_4) =$$
$$= \frac{(x-x_1)(x-x_3)}{-1}\left(\frac{-f(x_0)}{(x_0-x_1)(x_0-x_3)} - \frac{f'(x_0)(x-x_0)}{(x_0-x_1)(x_0-x_3)} + \right.$$
$$\left. + \frac{f(x_0)(x-x_0)(2x_0-x_1-x_3)}{(x_0-x_1)^2(x_0-x_3)^2}\right). \quad (3.31)$$

Если теперь перейти к пределу в (3.31) при x_3 стремящемся к x_1, получим

$$\lim_{x_3\to x_1}\lim_{x_2\to x_0}(S_2+S_4) =$$
$$= (x-x_1)^2\left(\frac{f(x_0)}{(x_0-x_1)^2} + \frac{f'(x_0)(x-x_0)}{(x_0-x_1)^2} - \frac{f(x_0)(x-x_0)(2x_0-2x_1)}{(x_0-x_1)^4}\right) =$$
$$= \frac{(-2x+3x_0-x_1)(x-x_1)^2}{(x_0-x_1)^3} f(x_0) + \frac{(x-x_0)(x-x_1)^2}{(x_0-x_1)^2} f'(x_0). \quad (3.32)$$

Заметим, что выражение, стоящее в правой части формулы (3.32), совпадает с первыми двумя слагаемыми в выражении для $H_3(x)$ из формулы (3.28). Аналогично вычисляется предел суммы первого и третьего слагаемых формулы (3.29).

Можно предложить другой способ построения многочлена Эрмита. Для этого наряду с $H_n(x)$ рассмотрим интерполяционный многочлен Лагранжа $L_m(x)$, принимающий в точках $x_0, x_1, \ldots, x_m$ значения $f(x_0), f(x_1), \ldots, f(x_m)$. Разность $H_n(x) - L_m(x)$ должна быть многочленом степени не выше n, обращающимся в ноль в точках $x_0, x_1, \ldots, x_m$. Следовательно,

$$H_n(x) - L_m(x) = \omega_m(x) H_{n-m}(x),$$

где

$$\omega_m(x) = (x-x_0)(x-x_1)\ldots(x-x_m).$$

При любом многочлене $H_{n-m}(x)$ функция

$$H_n(x) = L_m(x) + \omega_m(x) H_{n-m}(x) \quad (3.33)$$

принимает в узлах интерполирования значения $f(x_i)$. Подберем теперь $H_{n-m}(x)$ так, чтобы были выполнены и остальные условия. Дифференцируя обе части равенства (3.33), получим

$$H_n'(x) = L_m'(x) + \omega_m'(x) H_{n-m}(x) + \omega_m(x) H_{n-m}'(x). \quad (3.34)$$

Полагая здесь $x = x_k$, будем иметь:

$$H'_n(x_k) = L'_m(x_k) + \omega'_m(x_k)H_{n-m}(x_k). \tag{3.35}$$

Так как $\omega'_m(x_k) \neq 0$, в каждой точке, где задано значение $H'_n(x_k)$, мы найдем $H_{n-m}(x_k)$. Дифференцируя равенство (3.34), получим:

$$H''_n(x) = L''_m(x) + \omega''_m(x)H_{n-m}(x) + 2\omega'_m(x)H'_{n-m}(x) + \omega_m(x)H''_{n-m}(x). \tag{3.36}$$

Полагая снова $x = x_k$, найдем:

$$H''_n(x_k) = L''_m(x_k) + \omega''_m(x_k)H_{n-m}(x_k) + 2\omega'_m(x_k)H'_{n-m}(x_k).$$

Из этого равенства мы сумеем найти $H'_{n-m}(x_k)$ в тех точках, в которых заданы значения $H'_n(x_k)$. Продолжим этот процесс далее. Каждый раз коэффициентом при старшей производной от $H_n(x)$ в точках x_k будет $\omega'_n(x_k) \neq 0$. Таким образом, мы сведем нашу задачу об отыскании $H_n(x)$ к задаче об отыскании $H_{n-m}(x)$, удовлетворяющего условиям

$$H^{(i)}_{n-m}(x_k) = \widetilde{f}^{(i)}(x_k), \quad i = 0, \cdots, N_k - 2, \quad k = 0, \cdots, m,$$

где $\widetilde{f}^{(i)}(x_k)$ — заданные числа. К $H_{n-m}(x)$ применим точно такой же прием и так далее. В конце концов, нам потребуется построить интерполяционный многочлен Лагранжа по данным в некоторых точках x_k.

Для оценки погрешности интерполяционной формулы Эрмита можно воспользоваться оценкой (3.16) для многочлена Лагранжа и методом получения многочлена Эрмита из многочлена Лагранжа путем слияния узлов интерполировния. В результате получим

$$|f(x) - H_n(x)| \leq \frac{\max\limits_{x\in[a,b]} |f^{(n+1)}(x)|}{(n+1)!}|\Omega(x)|, \quad \Omega(x) = (x - x_0)^{N_0}(x - x_1)^{N_1} \dots (x - x_m)^{N_m}.$$

Заметим, что если по одной и той же таблице значений функции построить многочлены Лагранжа и Эрмита одинаковой степени, то многочлен Эрмита точнее приближает в точке x, которая расположена как можно ближе к середине множества узлов интерполяции. Это связано с тем, что многочлен $\omega_n(x)$ (см. формулу (3.16)) содержит больше узлов, чем $\Omega(x)$, и поэтому в него входят большие сомножители.

3.1.5 Интерполирование сплайнами

Как уже отмечалось, интерполирование многочленом Лагранжа (Ньютона) на всем отрезке $[a, b]$ с использованием большого числа узлов интерполяции часто приводит к плохим результатам, поэтому для избежания больших погрешностей отрезок разбивают на части и на каждом функцию заменяют многочленом невысокой степени. Такой метод приближения функции называют **кусочно-полиномиальной интерполяцией**.

В настоящее время широкое распространение получила сплайн интерполяция. Сплайном называют кусочно-полиномиальную функцию, определенную на отрезке $[a, b]$ и имеющую на нем некоторое число непрерывных производных.

Одним из наиболее распространенных является интерполяционный кубический сплайн дефекта 1, который в дальнейшем будем называть просто кубическим сплайном.

Пусть на $[a,b]$ задана непрерывная функция $f(x)$. Введем сетку

$$a = x_0 < x_1 < \cdots < x_n = b$$

и пусть $f_i = f(x_i), \quad i = 0, \cdots, n.$

Определение 3.1.3 ***Кубическим сплайном***
[1] *соответствующим функции* $f(x)$ *и узлам* $x_i, \quad i = 0, \ldots, n$ *называется функция* $S(x)$, *определенная на* $[a,b]$ *и удовлетворяющая условиям:*

a) на каждом отрезке $[x_{i-1}, x_i], \quad i = 1, \ldots, n$ *функция* $S(x)$ — *многочлен третьей степени;*

b) функция $S(x)$, *а также ее первая и вторая производные непрерывны на* $[a,b]$;

c) $S(x_i) = f_i, \quad i = 0, 1, \ldots, n.$

Построим сплайн, одновременно показав его существование. Как будет показано ниже, наложенные требования не определяют однозначно сплайн. Необходимы дополнительные условия для его однозначного определения.

На каждом отрезке $[x_{i-1}, x_i], i = 1, 2, ..., n$ будем искать функцию $S(x) = S_i(x)$ в виде многочлена

$$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3, \quad x_{i-1} \le x \le x_i, \quad i = 1, \ldots, n, \tag{3.37}$$

где $a_i,\ b_i,\ c_i,\ d_i$ — коэффициенты, которые требуется найти. Представление многочлена в виде (3.37) удобно тем, что по нему легко найти значения многочлена и его производных в точке x_i. Действительно, имеем:

$$S_i'(x) = b_i + c_i(x - x_i) + \frac{d_i}{2}(x - x_i)^2,$$
$$S_i''(x) = c_i + d_i(x - x_i), \quad S_i'''(x) = d_i.$$

Следовательно,

$$S_i(x_i) = a_i, \quad S_i'(x_i) = b_i, \quad S_i''(x_i) = c_i, \quad S_i'''(x_i) = d_i.$$

Из условия c) следует, что $a_i = f_i, \quad i = 1, \ldots, n$, кроме того положим $a_0 = f_0$.

Требование непрерывности сплайна приводит к равенствам:

$$S_{i-1}(x_{i-1}) = S_i(x_{i-1}), \quad i = 2, \ldots, n.$$

Отсюда, учитывая выражение для S_i, получим:

$$a_{i-1} = a_i + b_i(x_{i-1} - x_i) + \frac{c_i}{2}(x_{i-1} - x_i)^2 + \frac{d_i}{6}(x_{i-1} - x_i)^3, \quad i = 1, \ldots, n. \tag{3.38}$$

Если обозначить $h_i = x_i - x_{i-1}$ и учесть, что a_i уже найдены, то равенства (3.38) перепишутся в виде

$$h_i b_i - \frac{h_i^2}{2} c_i + \frac{h_i^3}{6} d_i = f_i - f_{i-1}, \ i = 1, \ldots, n. \tag{3.39}$$

[1] Точнее интерполяционным кубическим сплайном дефекта 1.

В соответствии с требованием непрерывности первой производной сплайна, должны выполняться равенства

$$S'_{i-1}(x_{i-1}) = S'_i(x_{i-1}), \quad i = 2, \ldots, n,$$

которые приводят к уравнениям:

$$c_i h_i - \frac{d_i}{2} h_i^2 = b_i - b_{i-1}, \quad i = 2, \ldots, n. \tag{3.40}$$

Из условия непрерывности второй производной получим:

$$d_i h_i = c_i - c_{i-1}, \quad i = 2, \ldots, n. \tag{3.41}$$

Таким образом, для определения $3n$ оставшихся неизвестными коэффициентов имеем $3n-2$ уравнения (3.39)—(3.41). Существует несколько способов выделить единственный сплайн. Чаще всего применяется подход, основанный на ограничениях, накладываемых на поведение сплайна в концевых точках $x_0,\ x_n$. Оставшиеся два условия получим, задавая граничные условия для $S(x)$, положив $S''(x_0) = S''(x_n) = 0$. Геометрически это означает, что вне отрезка $[a, b]$ график сплайна представляет собой прямую линию. Другие варианты задания дополнительных условий будут обсуждаться ниже.

Из условия $S''(x_n) = 0$ следует $c_n = 0$, условие $S''(x_0) = 0$ дает $c_1 - d_1 h_1 = 0$. Это равенство можно переписать в виде совпадающим с уравнением (3.41) при $i = 1$:

$$d_1 h_1 = c_1 - c_0,$$

если положить $c_0 = 0$. Таким образом получаем:

$$h_i d_i = c_i - c_{i-1}, \quad i = 1, \ldots, n, \quad c_0 = c_n = 0, \tag{3.42}$$

$$h_i c_i - \frac{h_i^2}{2} d_i = b_i - b_{i-1}, \quad i = 2, \ldots, n, \tag{3.43}$$

$$h_i b_i - \frac{h_i^2}{2} c_i + \frac{h_i^3}{6} d_i = f_i - f_{i-1}, \quad i = 1, \ldots, n. \tag{3.44}$$

Для исследования этой системы исключим из нее величины $b_i,\ d_i$. Для этого уравнения (3.44) перепишем в виде

$$b_i = \frac{h_i}{2} c_i - \frac{h_i^2}{6} d_i + \frac{f_i - f_{i-1}}{h_i},$$
$$b_{i-1} = \frac{h_{i-1}}{2} c_{i-1} - \frac{h_{i-1}^2}{6} d_{i-1} + \frac{f_{i-1} - f_{i-2}}{h_{i-1}},$$

Вычитая второе из этих уравнений из первого и подставляя в (3.43), получим:

$$h_i c_i + h_{i-1} c_{i-1} - \frac{h_{i-1}^2}{3} d_{i-1} - \frac{2h_i^2}{3} d_i = 2\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}} \right) \tag{3.45}$$

Из (3.42) следует

$$h_i^2 d_i = h_i (c_i - c_{i-1}), \quad h_{i-1}^2 d_{i-1} = h_{i-1} (c_{i-1} - c_{i-2}).$$

Подставив эти соотношения в (3.45) и увеличивая все индексы на 1, получим окончательно:

$$\begin{cases} c_0 = 0, \\ h_i c_{i-1} + 2(h_i + h_{i+1})c_i + h_{i+1}c_{i+1} = 6\left(\dfrac{f_{i+1} - f_i}{h_{i+1}} - \dfrac{f_i - f_{i-1}}{h_i}\right), \quad i = 1, \ldots, n-1, \\ c_n = 0. \end{cases} \tag{3.46}$$

Система (3.46) имеет трехдиагональную матрицу с диагональным преобладанием. Поэтому существует единственное решение системы, которое находится методом прогонки. По найденным из системы (3.46) коэффициентам c_i коэффициенты $b_i,\ d_i$ вычисляются теперь по явным формулам:

$$d_i = \frac{c_i - c_{i-1}}{h_i}, \quad b_i = \frac{h_i}{2}c_i - \frac{h_i^2}{6}d_i + \frac{f_i - f_{i-1}}{h_i}, \quad i = 1, \ldots, n.$$

Приведем без доказательства теорему о сходимости процесса интерполирования кубическими сплайнами

Теорема 3.1.4 *Пусть функция $y = f(x)$ определена и четыре раза непрерывно дифференцируема на отрезке $[a, b]$, причем $f''(a) = f''(b) = 0$. Тогда, если $M = \max\limits_{x\in[a,b]} |f^{(4)}(x)|$ и $h = \max\limits_{i=1,\ldots,n} h_i$, то*

$$\max_{x\in[a,b]} |f^{(j)}(x) - S^{(j)}(x)| \le Mh^{4-j}, \quad j = 0, 1, 2.$$

Из теоремы следует, что при $h \to 0$ сплайны $S(x)$ и их производные до второго порядка включительно сходятся к функции $f(x)$ и ее соответствующим производным.

Замечание 1. Выше, при построении сплайна накладывались дополнительные условия равенства в граничных точках отрезка нуля его вторых производных. Такой сплайн называют **естественным**. Существуют другие подходы для задания дополнительных условий.

- Задать производную функции $S(x)$ в точках a и b. Во многих задачах производные в конечных точках известны заранее из физических соображений, поэтому такой подход часто оказывается полезным.

- Оценить производные по исходным данным и использовать эти оценки в качестве значений первых производных. Для этого можно построить по первым (последним) четырем точкам кубический многочлен, затем найти его производную в граничной точке, которую и использовать в качестве недостающего условия.

- Поступить аналогично предыдущему пункту, находя только не первые, а вторые производные.

- Потребовать, чтобы в точке x_1 была непрерывной еще и третья производная сплайна. Это эквивалентно тому, что на первом и втором отрезках кубические функции совпадают. Это также эквивалентно удалению узла x_1 с сохранением требования, чтобы единая на отрезке $[x_0, x_2]$ кубическая функция обеспечивала интерполяцию не только на концах этого отрезка, но и в точке x_1. Аналогично для x_{n-1}.

В любом из перечисленных выше случаев нахождение сплайна сведется к решению системы с трехдиагональной матрицей.

Существуют и другие подходы. Например, задать функцию $S(x)$ периодической, то есть положить

$$S(a) = S(b), \quad S'(a) = S'(b), \quad S''(a) = S''(b).$$

Для большинства задач нет наилучшего способа задания дополнительных условий. С другой стороны, многие наборы данных приводят к сплайнам, которые независимо от сделанного выбора дополнительных условий выглядят почти одинаково.

Замечание 2. Хотя сплайны вошли в математику относительно недавно, они моделируют старое механическое устройство. Чертежники давно пользовались гибкими линейками для того, чтобы провести гладкую кривую, проходящую через заданные точки на плоскости. Линейку ставят на ребро и изгибают так, чтобы ребро проходило сразу через все точки. В механике доказывается, что линейка принимает форму, минимизирующую ее потенциальную энергию и что эта энергия пропорциональна интегралу от квадрата кривизны линейки. Если кривую, по которой изогнута линейка представить функцией $y = s(x)$, то при малых изгибах потенциальная энергия линейки пропорциональна $\int\limits_a^b (s''(x))^2\,dx$. Таким образом, нахождение формы линейки сводится к нахождению такой функции $y = s(x)$, что ее график проходит через заданные точки и интеграл от квадрата второй производной этой функции минимален. Докажем, что сплайн, построение которого описано в этом параграфе, является той функцией, которая определяет форму линейки. Точнее, справедлива следующая теорема.

Теорема 3.1.5 *Пусть $a = x_0 < x_1 < \cdots < x_n = b$ и $y_0, y_1, \cdots, y_n$ — произвольные числа. Обозначим через $S(x)$ естественный кубический сплайн дефекта 1 такой, что его узлами являются точки x_i и $S(x_i) = y_i$, $i = 0, 1, \cdots, n$. Тогда для произвольной дважды непрерывно дифференцируемой функции $f(x)$ такой, что $f(x_i) = y_i$, $i = 0, 1, \cdots, n$ справедливо неравенство*

$$\int\limits_a^b (S''(x))^2\,dx \le \int\limits_a^b (f''(x))^2\,dx.$$

Доказательство. Достаточно показать, что

$$\int\limits_a^b (f''(x))^2\,dx - \int\limits_a^b (S''(x))^2\,dx \ge 0.$$

Имеем

$$\int\limits_a^b (f''(x))^2\,dx - \int\limits_a^b (S''(x))^2\,dx = \int\limits_a^b (f''(x) - S''(x))^2\,dx + 2\int\limits_a^b (f''(x)S''(x) - (S''(x))^2)\,dx.$$

Так как первый из интегралов, стоящих в правой части этого равенства неотрицателен, теорема будет доказана, если показать, что второй интеграл равен нулю. Имеем, использую интегрирование по частям, тот факт, что на промежутке (x_{i-1}, x_i) третья

производная от функции $S(x)$ равна константе, которую обозначим S_i''', и нулевые граничные условия для второй производной от сплайна:

$$\int_a^b (f''(x)S''(x) - (S''(x))^2)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} (f''(x) - S''(x))S''(x)\,dx =$$

$$= \sum_{i=1}^{n} (f'(x) - S'(x))S''(x)\Big|_{x_{i-1}}^{x_i} - \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} (f'(x) - S'(x))S'''(x)\,dx =$$

$$= (f'(b) - S'(b))S''(b) - (f'(a) - S'(a))S''(a) - \sum_{i=1}^{n} S_i''' \int_{x_{i-1}}^{x_i} (f'(x) - S'(x))\,dx =$$

$$= -\sum_{i=1}^{n} S_i'''(f(x) - S(x))\Big|_{x_{i-1}}^{x_i} = 0.$$

Последнее равенство написано на основании того, что в точках x_i значения функций $f(x)$ и $S(x)$ совпадают. Таким образом, теорема доказана.

3.1.6 Многомерная интерполяция

Задача приближения функций существенно усложняется, если перейти к интерполяции функций нескольких переменных. Проиллюстрируем возникающие трудности и подходы для случая функций двух переменных.

Во-первых, не всякое расположение узлов допустимо. Например, при использовании многочленов первой степени $P_1(x, y) = ax + by + c$ узлы интерполяции не должны лежать на одной прямой. В данном случае это легко пояснить геометрически. При интерполяции многочленом первой степени происходит замена поверхности, описываемой уравнением $z = f(x, y)$, плоскостью проходящей через три заданные точки $(x_i, y_i, z_i),\ z_i = f(x_i, y_i),\ i = 1, 2, 3$. Если точки $(x_i, y_i),\ i = 1, 2, 3$ лежат на одной прямой в плоскости OXY, то либо проводимая плоскость определяется однозначно и перпендикулярна плоскости OXY, и, следовательно, не определяет z как функцию от x, y, либо, когда точки $(x_i, y_i, z_i),\ i = 1, 2, 3$ лежат на одной прямой в пространстве, проводимая плоскость не определяется однозначно.

При интерполяции многочленом второй степени необходимо, чтобы узлы не лежали на одной кривой второго порядка и так далее.

Во-вторых, если для многочлена одной переменной его степень взаимно однозначно связана с числом узлов, то уже в двумерном случае не любое число узлов позволяет определить степень многочлена. Многочлен $P_n(x, y) = \sum_{k+l=0}^{n} a_{kl} x^k y^l$ имеет $\frac{(n+1)(n+2)}{2}$ коэффициентов. Если число узлов не соответствует этой формуле, то часть коэффициентов при некоторых степенях должна быть задана принудительно (часто нулями).

В-третьих, меняется понятие экстраполяции. Если соединить все узлы отрезками, крайние отрезки ограничат выпуклое множество. Если искомая точка попадает в это множество, то имеет место интерполяция, если не попадает — экстраполяция.

Рассмотрим некоторые способы приближения значений функции многих переменных.

Пусть на прямоугольной сетке определены значения функции $f_{ij} = f(x_i, y_j)$. Для вычисления приближенного значения функции в точке (x^*, y^*) выберем прямоугольник $[x_{i_1}, x_{i_1+k}] \times [y_{j_1}, y_{j_1+l}]$ из $(k + 1) \cdot (l + 1)$ узлов, в который попадет точка (x^*, y^*).

Сначала проведем одномерную интерполяцию по строкам, то есть при каждом фиксированном значении $j = j_1, \ldots, j_1 + l$ найдем $\varphi(x^*, y_j)$ по значениям $f_{i,j}$. Затем проведем одномерную интерполяцию по столбцу, то есть по значениям $\varphi(x^*, y_j)$ найдем искомое значение функции в точке (x^*, y^*). Таким образом, нахождение значения функции сводится к последовательности одномерных интерполяций.

Очевидно, что для приближенного вычисления значения функции можно было бы поступить иначе. Сначала интерполировать по столбцам, а затем по строкам.

В случае применения лагранжевой одномерной интерполяции легко написать общую формулу, аналогичную одномерной формуле Лагранжа:

$$L_{kl}(x, y) = \sum_{i=i_1}^{i_1+k} \sum_{j=j_1}^{j_1+l} f(x_i, y_j) \prod_{\substack{p=i_1 \\ p \neq i}}^{i_1+k} \prod_{\substack{q=j_1 \\ q \neq j}}^{j_1+l} \frac{(x - x_p)(y - y_q)}{(x_i - x_p)(y_j - y_q)}.$$

Последовательная интерполяция удобна еще тем, что алгоритм легко модифицировать для того случая, когда, например, при каждом фиксированном y_j число узлов вдоль оси OX свое.

Из-за громоздкости формул при многомерной интерполяции многочлены выше второй степени практически не применяются.

Иногда приходится работать с функцией, заданной на нерегулярной сетке, например, с функцией, измеренной экспериментально. Тогда обычно ограничиваются интерполяционным многочленом первой степени $P_{11} = ax+by+c$. Его коэффициенты находят по трем выбранным узлам. Можно не вычислять специально коэффициенты этого многочлена, так как из аналитической геометрии известно, что уравнение плоскости, проходящей через точки $(x_i, y_i, z_i),\ i = 1, 2, 3$ имеет вид

$$\begin{vmatrix} z & 1 & x & y \\ z_1 & 1 & x_1 & y_1 \\ z_2 & 1 & x_2 & y_2 \\ z_3 & 1 & x_3 & y_3 \end{vmatrix} = 0.$$

3.1.7 Тригонометрическая интерполяция. Дискретное и быстрое преобразование Фурье

Дискретное преобразование Фурье применяется при решении многих прикладных задач. Оно стало особенно эффективным методом для различных приложений, после создания быстрого преобразования Фурье.

В математическом анализе доказывалось, что если $f(x)$ — непрерывная периодическая функция с периодом 1, то она может быть разложена в ряд Фурье

$$f(x) = \sum_{q=-\infty}^{\infty} a_q e^{2\pi i q x}, \quad i^2 = -1, \tag{3.47}$$

причем

$$\sum_{q=-\infty}^{\infty} |a_q| < \infty. \tag{3.48}$$

Введем узлы $x_j = j/n$, где n — фиксированное целое положительное число, $j = 0, \ldots, n$ и обозначим $f_j = f(x_j)$. Заметим, что если $q_2 - q_1 = kn$, где k — целое, то число $q_2 x_j - q_1 x_j = knx_j = kj$ также целое. Поэтому, если x — узел, то

$$e^{2\pi i q_2 x} = e^{2\pi i (q_2 - q_1) x} e^{2\pi i q_1 x} = 1 \cdot e^{2\pi i q_1 x} = e^{2\pi i q_1 x}. \tag{3.49}$$

Отсюда следует, что если функцию $f(x)$ рассматривать только в узлах x_j, то в равенстве (3.47) можно привести подобные члены, в результате чего получим

$$f_j = \sum_{q=0}^{n-1} A_q e^{2\pi i q x_j}, \quad j = 0, \ldots, n, \tag{3.50}$$

где

$$A_q = \sum_{l=-\infty}^{\infty} a_{q+ln}. \tag{3.51}$$

Если функция $f(x)$ была известна только в узлах, то, доопределив ее между узлами путем линейной интерполяции, получим кусочно-дифференцируемую функцию. Для нее справедливо (3.47), (3.48), а, значит, и (3.50). При этом коэффициенты A_q могут быть определены без нахождения коэффициентов a_q. Для нахождения коэффициентов A_q определим скалярное произведение

$$(\mathbf{f}, \mathbf{g}) = \frac{1}{n} \sum_{j=0}^{n-1} f_j \bar{g}_j, \ \mathbf{f} = (f_0, \ldots, f_{n-1}), \ \mathbf{g} = (g_0, \ldots, g_{n-1}),$$

где черта означает переход к комплексно сопряженному. Вектора

$$\omega_p = (e^{2\pi i p x_0}, \ldots, e^{2\pi i p x_{n-1}})$$

при $0 \le p \le n-1$ образуют тогда ортонормированную систему. Действительно,

$$(\omega_p, \omega_q) = \frac{1}{n} \sum_{j=0}^{n-1} e^{2\pi i j (p-q)/n}. \tag{3.52}$$

Очевидно тогда, что при $p = q$ выполняется равенство $(\omega_p, \omega_q) = 1$. Если же $p \ne q$, то правая часть равенства (3.52) представляет собой геометрическую прогрессию со знаменателем $e^{2\pi i (p-q)/n}$. Используя формулу суммы геометрической прогрессии и учитывая, что $e^{2\pi i s} = 1$ при любом целом s, имеем

$$(\omega_p, \omega_q) = \frac{1}{n} \frac{e^{2\pi i (p-q)} - 1}{e^{2\pi i (p-q)/n} - 1} = 0.$$

Равенства (3.50) можно переписать в виде

$$\mathbf{f} = \sum_{p=0}^{n-1} A_p \omega_p.$$

Умножая теперь это равенство скалярно на ω_q и учитывая ортонормированность системы векторов ω_q, получим

$$\frac{1}{n} \sum_{s=0}^{n-1} f_s e^{-2\pi i q x_s} = (\mathbf{f}, \omega_q) = (\sum_{j=0}^{n-1} A_p \omega_p, \omega_q) = \sum_{j=0}^{n-1} A_p (\omega_p, \omega_q) = A_q. \tag{3.53}$$

Суммируя полученные в этом параграфе результаты, имеем следующее. Функция

$$\varphi(x) = \sum_{q=0}^{n-1} A_q e^{-2\pi q x},$$

определенная при всех $x \in [0,1]$, где коэффициенты A_q вычисляются по формуле (3.53), обладает тем свойством, что $\varphi(x_l) = f_l,\ l = 0, \ldots, n-1$, то есть интерполирует функцию $f(x)$.

Способ интерполяции функции $f(x)$ функцией $\varphi(x)$ называется **тригонометрической интерполяцией** . Соотношение (3.50) называют **конечным** или **дискретным рядом Фурье** , а коэффициенты A_q — **дискретными коэффициентами Фурье**.

Прямым преобразованием Фурье назовем процедуру вычисления дискретных коэффициентов Фурье по известным в узлах значениям функции $f(x)$. **Обратное преобразованием Фурье** — определение значений в узлах функции $f(x)$ по известным дискретным коэффициентам Фурье.

Осуществление прямого или обратного преобразования Фурье по формулам (3.53) или (3.50) соответственно требуют $O(n^2)$ арифметических операций (вычисляется n величин, причем для вычисления каждой величины необходимо $O(n)$ арифметических операций).

Быстрое преобразование Фурье это такой метод организации вычислений, который позволяет сократить число арифметических операций за счет выделения одинаковых групп слагаемых и вычисления значений этих групп один раз. Рассмотрим идею метода на примере прямого преобразования Фурье. Пусть $n = p_1 p_2$, причем целые числа p_1, p_2 не равны 1. Любые целые числа q и s такие, что $0 \le q, s \le n-1$ можно представить в виде

$$q = q_1 + p_1 q_2, \quad s = s_2 + p_2 s_1, \quad 0 \le q_1, s_1 < p_1, \quad 0 \le q_2, s_2 < p_2.$$

Тогда для коэффициента A_q, который в дальнейшем обозначим $A(q_1, q_2)$, имеем

$$A(q_1, q_2) = \frac{1}{n} \sum_{s=0}^{n-1} f_s e^{-2\pi i q s/n} = \frac{1}{n} \sum_{s_2=0}^{p_2-1} \sum_{s_1=0}^{p_1-1} f_{s_2+p_2 s_1} e^{-2\pi i (q_1+p_1 q_2)(s_2+p_2 s_1)/(p_1 p_2)}.$$

Учитывая, что

$$\frac{(q_1+p_1 q_2)(s_2+p_2 s_1)}{p_1 p_2} = q_2 s_1 + \frac{q_1 s_1}{p_1} + \frac{q s_2}{n},$$

имеем

$$A(q_1, q_2) = \frac{1}{p_1 p_2} \sum_{s_2=0}^{p_2-1} \sum_{s_1=0}^{p_1-1} f_{s_2+p_2 s_1} e^{-2\pi i (q_1 s_1)/p_1} e^{-2\pi i (q s_2)/n} e^{-2\pi i q_2 s_1}.$$

Так как последний множитель в правой части этой формулы равен 1, получим

$$A(q_1, q_2) = \frac{1}{p_2} \sum_{s_2=0}^{p_2-1} A^{(1)}(q_1, s_2) e^{-2\pi i (q s_2)/n}, \quad A^{(1)}(q_1, s_2) = \frac{1}{p_1} \sum_{s_1=0}^{p_1-1} f_{s_2+p_2 s_1} e^{-2\pi i (q_1 s_1)/p_1}.$$

Оценим теперь количество арифметических операций, которые необходимо произвести для нахождения всех коэффициентов $A_q = A(q_1, q_2)$. Вычисление всех величин $A^{(1)}(q_1, s_2)$ потребует $O(p_1^2 p_2)$ арифметических операций, для нахождения теперь $A(q_1, q_2)$ необходимо совершить еще $O(p_1 p_2^2)$ арифметических операций. Поэтому, если $p_1, p_2 = O(\sqrt{n})$, то общее количество операций $O(n^{3/2})$.

Аналогично могут быть получены расчетные формулы при $n = p_1 p_2 \ldots p_r$, использование которых потребуется $O(n(p_1 + \cdots + p_r))$ операций.

Приведем без вывода рекуррентные соотношения для нахождения коэффициентов Фурье в наиболее употребительном случае $n = 2^r$, то есть $p_1 = \cdots = p_r = 2$ Тогда

$$q = \sum_{k=0}^{r} q_k 2^{k-1}, \quad s = \sum_{j=0}^{r} s_{r+1-j} 2^{j-1}, \quad \text{где} \quad q_k,\ s_j = 0,\ 1.$$

Рекуррентные соотношения имеют вид:

$$A^{(0)}(s_1, \ldots, s_r) = f_{s_r + s_{r-1}2 + \cdots + s_1 2^{r-1}},$$

$$A^{(m)}(q_1, \ldots, q_m; s_{m+1}, \ldots, s_r) =$$

$$= \frac{1}{2} \sum_{s_m=0}^{1} \exp\Big(-2\pi i s_m 2^{-m} \sum_{k=1}^{m} q_k 2^{k-1}\Big) A^{(m-1)}(q_1, \ldots, q_{m-1}; s_m, \ldots, s_r), \quad m = 1, \ldots, r,$$

$$A_q = A(q_1, \ldots, q_r) = A^{(r)}(q_1, \ldots, q_r).$$

Переход от каждой совокупности $A^{(m-1)}$ к совокупности $A^{(m)}$ требует $O(n)$ операций, всего таких шагов $r = \log_2 n$, поэтому общее число операций $O(n \log_2 n)$.

3.2 НАИЛУЧШЕЕ ПРИБЛИЖЕНИЕ ФУНКЦИЙ, ЗАДАННЫХ ТАБЛИЧНО

Пусть на отрезке $[a, b]$ заданы функции $\varphi_j(x), \quad j = 0, \ldots, n$ и известны значения функции $f(x)$ в точках $x_k \in [a, b], \quad k = 0, \ldots, m$, причем $m > n$. Тогда задача интерполирования функции $f(x)$ функцией $\varphi(x)$ вида $\varphi(x) = \sum\limits_{j=0}^{n} C_j \varphi_j(x)$, которая была рассмотрена в начале параграфа 3.1, становится переопределенной, так как число условий, которым необходимо удовлетворить, больше числа коэффициентов. Таким образом, в общем случае при $m > n$ задача интерполяции решения не имеет и задача приближения функции $f(x)$ функцией $\varphi(x)$ должна ставиться иначе. Можно рассматривать **задачу о наилучшем приближении**. Она формулируется следующим образом. Обощенный многочлен $\varphi(x) = \sum\limits_{j=0}^{n} C_j \varphi_j(x)$ рассматривается только в узлах x_k. Образуют разности

$$r_k = \varphi(x_k) - f(x_k), \quad k = 0, \ldots, m,$$

которые характеризуют отклонение в узлах x_k точного значения функции от приближенного. Ставится задача так подобрать коэффициенты C_j, чтобы $\|\mathbf{r}\|$ была минимальной, где $\mathbf{r} = (r_0, \ldots, r_m)$. Полученная в результате функция $\varphi(x)$ называется **аппроксимирующей**.

В зависимости от выбора нормы получаются разные задачи:

- при $\|\mathbf{r}\| = \max\limits_{i=0,\ldots,m} |r_i|$ — **задача о наилучшем равномерном приближении**;
- при $\|\mathbf{r}\|^2 = \sum\limits_{k=0}^{m} \rho_k r_k^2$, где $\rho_k > 0$ — весовые коэффициенты, — **задача о наилучшем среднеквадратичном приближении**.

При $m = n$ независимо от выбора нормы задача о наилучшем приближении совпадает с решением задачи об интерполяции, так как наименьшее значение, которое может принять норма — ноль, а требование $\|\mathbf{r}\| = 0$ приводит к равенствам $\varphi(x_k) = f(x_k), \quad k = 0, \ldots, m$.

3.2.1 Метод наименьших квадратов

Рассмотрим как решается задача о наилучшем среднеквадратичном приближении. Условие для нахождения коэффициентов имеет вид:

$$\sum_{k=0}^{m} \rho_k \left(\sum_{j=0}^{n} C_j \varphi_j(x_k) - f(x_k) \right)^2 \to \min .$$

Здесь ρ_k — заданные положительные числа, называемые весовыми коэффициентами. Выражение, которое необходимо минимизировать, является функцией от $C_0, \ldots, C_n$. Приравнивая нулю частные производные по C_i, получим

$$2 \sum_{k=0}^{m} \rho_k \left(\sum_{j=0}^{n} C_j \varphi_j(x_k) - f(x_k) \right) \varphi_i(x_k) = 0. \tag{3.54}$$

Для сокращения записей введем скалярные произведения

$$(f, \varphi) = \sum_{k=0}^{m} \rho_k f(x_k) \varphi(x_k).$$

Тогда равенства (3.54) перепишутся в виде:

$$\sum_{j=0}^{n} C_j (\varphi_j, \varphi_i) = (f, \varphi_i), \quad i = 0, \ldots, n. \tag{3.55}$$

Из этой системы уравнений находятся коэффициенты C_j.

Описанный способ нахождения аппроксимирующей функции называется **методом наименьших квадратов**.

Рассмотрим часто встречающийся случай когда $\varphi_k(x) = x^k,\ k = 0, 1, \ldots, n$, то есть $\varphi(x)$ — многочлен степени не выше n. Тогда система для нахождения коэффициентов C_j многочлена принимает вид

$$\sum_{j=0}^{n} C_j (x^i, x^j) = (f, x^i), \quad (x^i, x^j) = \sum_{k=0}^{m} \rho_k x_k^{i+j}, \quad (f, x^i) = \sum_{k=0}^{m} \rho_k f(x_k) x_k^i.$$

Получающаяся система линейных алгебраических уравнений имеет при больших n плохо обусловленную матрицу, поэтому обычно ограничиваются небольшими степенями многочленов, не более 5.

3.2.2 Сглаживание сеточных функций. Выбор эмпирических зависимостей

Метод наименьших квадратов широко применим при обработке экспериментальных кривых, когда точки измерены с погрешностью τ. При этом весу ρ_i приписывается смысл точности измерений: в тех точках, где точность выше, вес больше. Тогда аппроксимирующая кривая проходит ближе к тем точкам, которые измерены точнее.

Если число n близко к количеству точек, в которых задана экспериментальная кривая, то аппроксимирующая функция близка к интерполирующей, что, очевидно,

неразумно при наличии значительных ошибок эксперимента, так как аппроксимирующая функция должна сгладить ошибки эксперимента. В то же время, если n очень мало, то для описания сложной кривой коэффициентов может не хватить.

Рациональное число коэффициентов можно определить следующим образом. Выбирают некоторое число n, используя (3.54) находят аппроксимирующую функцию φ, после чего определяют среднеквадратичное отклонение $\delta = \|f - \varphi\|/\|1\|$. Если $\delta \gg \tau$, то погрешность аппроксимации много больше погрешности задания исходных данных. Следовательно n мало и его необходимо увеличить. В случае, когда $\delta \ll \tau$, число n велико и часть коэффициентов физически недостоверна. Если же $\delta \approx \tau$, число n выбрано правильно. Однако, если при этом n близко к m и m велико, то следует поискать более подходящий вид аппроксимирующей функции.

К выбору вида аппроксимирующей функции (ее иногда называют **эмпирической формулой**) можно подойти производя анализ исходных данных $(x_i, y_i),\ i = 0, 1, \ldots, m$. Заметим сначала, что зачастую эмпирическую формулу стараются выбрать как можно проще. Такой является линейная зависимость $\varphi(x) = C_1 x + C_0$. Близость экспериментальных данных к линейной зависимости можно оценить вычисляя величины

$$z_i = \Delta y_i/\Delta x_i, \quad \Delta y_i = y_{i+1} - y_i, \quad \Delta x_i = x_{i+1} - x_i, \quad i = 0, \ldots, m-1.$$

Если числа $z_i \approx const$, то точки (x_i, y_i) расположены приблизительно на одной прямой и можно ставить вопрос о нахождении коэффициентов $C_0,\ C_1$.

В ряде случаев к линейной зависимости могут быть сведены и другие экспериментальные данные, когда их график в декартовой системе координат далек от прямой линии. Это может быть сделано путем введения новых переменных $\xi,\ \eta$ вместо $x,\ y$:

$$\xi = \xi(x, y), \quad \eta = \eta(x, y). \tag{3.56}$$

Функции $\xi(x, y),\ \eta(x, y)$ выбираются так, чтобы точки ξ_i, η_i лежали на некоторой прямой линии в плоскости ξ, η. Такое преобразование называется **выравниванием данных**.

Для получения линейной зависимости $\eta = C_1\xi + C_0$ с помощью преобразования (3.56) исходная формула должна быть записана в виде $\eta(x, y) = C_1\xi(x, y) + C_0$. К такому виду легко сводится, например, степенная зависимость $y = ax^b$. Логарифмируя эту формулу, получим $\log y = b \log x + \log a$. Полагая $\xi = \log x,\ \eta = \log y,\ C_1 = b,\ C_0 = \log a$, получаем линейную зависимость.

Как уже отмечалось, опытные данные содержат случайные ошибки, что является причиной разброса этих данных. Во многих случаях целесообразно провести их **сглаживание** для получения более плавного характера исследуемой зависимости.

Пусть в результате эксперимента получена таблица $(x_i, y_i),\ i = 0, \ldots, m,\ x_0 < \cdots < x_m$. Предположим, что узлы равноотстоящие и искомая функция на произвольной части отрезка $[x_0, x_m]$ может быть достаточно хорошо приближена многочленом степени n.

Способ сглаживания состоит в следующем. Для нахождения сглаженного значения $\bar{y}_i$ в точке x_i выбираем по обе стороны от нее k $(n \le 2k)$ значений аргумента из имеющихся в таблице $x_{i-k}, \ldots, x_{i+k}$. По опытным значениям рассматриваемой функции в этих точках строим многочлен степени n с помощью метода наименьших квадратов. Значение полученного многочлена в точке x_i и будет искомым (сглаженным) значением. Процесс повторяется для всех внутренних точек. Сглаживание значений вблизи точек x_0 и x_m производится с помощью крайних точек. Иногда сглаживание повторяют.

Приведем несколько формул для вычисления сглаженных значений опытных данных при различных значениях n и k:

$$n = 1:$$

$$\bar{y}_i = \frac{1}{3}(y_{i-1} + y_i + y_{i+1}), \quad k = 1,$$

$$\bar{y}_i = \frac{1}{5}(y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2}), \quad k = 2;$$

$$n = 3:$$

$$\bar{y}_i = \frac{1}{35}(-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 3y_{i+2}), \quad k = 2,$$

$$\bar{y}_i = \frac{1}{21}(-2y_{i-3} + 3y_{i-2} + 6y_{i-1} + 7y_i + 6y_{i+1} + 3y_{i+2} - 2y_{i+3}), \quad k = 3.$$

3.3 ПРИБЛИЖЕНИЕ ФУНКЦИЙ В ЛИНЕЙНЫХ НОРМИРОВАННЫХ ПРОСТРАНСТВАХ

В предыдущем параграфе решалась задача о наилучшем приближении функции, заданной таблично. В этом параграфе будет рассмотрена задача аппроксимации в более общем виде.

Пусть в некотором нормированном пространстве $\mathcal{B}$ задана конечная система линейно независимых элементов $\varphi_i,\ i = 1, \ldots, n$. Назовем **обобщенным многочленом** линейную комбинацию

$$\varphi = c_1\varphi_1 + \cdots + c_n\varphi_n. \tag{3.57}$$

Задачей о наилучшем приближении будем считать проблему нахождения для заданного элемента $f \in \mathcal{B}$ такого обобщенного многочлена φ, для которого отклонение $\|f - \varphi\|$ минимально. Этот многочлен называется **элементом наилучшего приближения.**

3.3.1 Наилучшее приближение в произвольном линейном нормированном пространстве

Основным результатом этого пункта будет теорема существования.

Теорема 3.3.1 *Элемент наилучшего приближения существует.*

Доказательство. Доказательство теоремы основано на известном из математического анализа факте, что непрерывная функция, зависящая от n переменных, заданная на замкнутом, ограниченном множестве, принимает на этом множестве минимальное значение.

Рассмотрим функцию

$$F(c_1, \ldots, c_n) = \left\| f - \sum_{i=1}^{n} c_i\varphi_i \right\|.$$

Покажем прежде всего, что при любом $f \in \mathcal{B}$ она является непрерывной функцией своих аргументов c_i. Для этого отметим, что в силу свойств нормы, модуль разности

норм не превосходит нормы разности. Поэтому

$$|F((\widetilde{c}_1,\dots,\widetilde{c}_n)-F(c_1,\dots,c_n)=\left|\left\|f-\sum_{i=1}^{n}\widetilde{c}_i\varphi_i\right\|-\left\|f-\sum_{i=1}^{n}c_i\varphi_i\right\|\right|\le$$
$$\le\left\|\sum_{i=1}^{n}(\widetilde{c}_i-c_i)\varphi_i\right\|\le\sum_{i=1}^{n}|\widetilde{c}_i-c_i|\|\varphi_i\|.$$

Данное неравенство означает непрерывность функции F. Так как $f\in\mathcal{B}$ произвольно, непрерывной является и функция

$$F_0(c_1,\dots,c_n)=\left\|\sum_{i=1}^{n}c_i\varphi_i\right\|,$$

полученная из функции F заменой f на нулевой элемент пространства $\mathcal{B}$.

В n-мерном пространстве векторов $\mathbf{c}=(c_1,\dots,c_n)$ введем норму

$$|\mathbf{c}|=\sqrt{\sum_{i=1}^{n}c_i^2}.$$

В силу того, что $F_0(c_1,\dots,c_n)$ непрерывная функция, на единичной сфере

$$\{\mathbf{c}\ :\ |\mathbf{c}|=1\}$$

она достигает своей нижней грани, значение которой обозначим μ. Заметим, что $\mu>0$. Действительно, если нижняя грань достигается в точке $(\widehat{c}_1,\dots,\widehat{c}_n)$ и равна нулю, имеем

$$\mu=\left\|\sum_{i=1}^{n}\widehat{c}_i\varphi_i\right\|=0.$$

Это означает, что φ_i линейно зависимы, что противоречит предположению. Заметим также, что при любом $\mathbf{c}$ отличном от нулевого вектора выполняется неравенство

$$F_0(c_1,\dots,c_n)=\left\|\sum_{i=1}^{n}c_i\varphi_i\right\|=|\mathbf{c}|\left\|\sum_{i=1}^{n}\frac{c_i}{|\mathbf{c}|}\varphi_i\right\|=|\mathbf{c}|F_0(\frac{c_1}{|\mathbf{c}|},\dots,\frac{c_n}{|\mathbf{c}|})\ge|\mathbf{c}|\mu.$$

Возьмем теперь ε таким, что $2\|f\|/\mu<\varepsilon$ и рассмотрим функцию F в шаре $S=\{\mathbf{c}\ :\ |\mathbf{c}|\le\varepsilon\}$. В силу непрерывности она достигает в некоторой точке $\mathbf{c}^0$ этого шара наименьшее значение, которое обозначим m. Заметим, что согласно определению m выполняется неравенство $\|f\|=F(0,\dots,0)\ge m$.

Вне шара S выполняются соотношения $|\mathbf{c}|>\varepsilon$,

$$F(c_1,\dots,c_n)=\left\|f-\sum_{i=1}^{n}c_i\varphi_i\right\|\ge\left\|\sum_{i=1}^{n}c_i\varphi_i\right\|-\|f\|=$$
$$=F_0(c_1,\dots,c_n)-\|f\|\ge|\mathbf{c}|\mu-\|f\|>\varepsilon\mu-\|f\|\ge(2\|f\|/\mu)\mu-\|f\|=\|f\|>m.$$

Таким образом, наименьшее значение принимается внутри шара в точке $\mathbf{c}^0$. Теорема доказана.

Заметим, что в теореме говорится только о существовании элемента наилучшего приближения. Элементов наилучшего приближения может быть, вообще говоря, несколько.

3.3.2 Наилучшее приближение в гильбертовом пространстве

Задача о наилучшем приближении в вещественном гильбертовом пространстве $\mathcal{H}$ подробно рассматривалась в [20], поэтому здесь напомним только основные результаты. Они заключаются в следующем.

Теорема 3.3.2 *Для любого элемента $f \in \mathcal{H}$ существует единственный элемент наилучшего приближения φ. При этом элемент $f - \varphi$ ортогонален всевозможным линейным комбинациям элементов φ_i.*

Из теоремы следует, что для нахождения коэффициентов c_i элемента наилучшего приближения достаточно воспользоваться свойством ортогональности. Обозначим (f, g) — скалярное произведение элементов f и g из $\mathcal{H}$. Тогда

$$0 = (f - \varphi, \varphi_i) = (f - \sum_{k=1}^{n} c_k \varphi_k, \varphi_i).$$

Отсюда следует, что

$$\sum_{k=1}^{n} c_k (\varphi_k, \varphi_i) = (f, \varphi_i), \quad i = 1, \ldots, n. \tag{3.58}$$

Таким образом, для нахождения элемента наилучшего приближения необходимо, решив систему (3.58), определить коэффициенты c_k после чего воспользоваться формулой (3.57).

Заметим, что система (3.58) имеет единственное решение[2]. Действительно, существование решения следует из того, что элемент наилучшего приближения существует. Единственность же вытекает из следующих соображений. Если бы существовало два решения системы, то это означало бы, что нашлось два обобщенных многочлена $\varphi^{(1)}$ и $\varphi^{(2)}$ таких, что для всех $i = 1, \ldots, n$ выполняется равенство

$$(f - \varphi^{(1)}, \varphi_i) = (f - \varphi^{(2)}, \varphi_i).$$

Отсюда следует, что

$$(\varphi^{(1)} - \varphi^{(2)}, \varphi_i) = 0. \tag{3.59}$$

Так как $\varphi^{(1)}$ и $\varphi^{(2)}$ — обобщенные многочлены, их разность $\widetilde{\varphi} = \varphi^{(1)} - \varphi^{(2)}$ также обобщенный многочлен. Пусть его коэффициенты равны d_i. Тогда умножая (3.59) на d_i и суммируя по i, получим

$$(\widetilde{\varphi}, \widetilde{\varphi}) = \|\widetilde{\varphi}\|^2 = 0.$$

Следовательно, $\widetilde{\varphi} = 0$, что невозможно, так как φ_i линейно независимы.

Для примера возьмем в качестве пространства $\mathcal{H}$ пространство $L_2(a, b)$. Напомним, что в нем скалярное произведение определяется по формуле

$$(f, g) = \int_a^b f(x) g(x)\, dx.$$

В этом случае система (3.58) называется **нормальной системой метода наименьших квадратов** и приближение называется **среднеквадратичным**. Возьмем

[2]Напомним, что система элементов $\varphi_i,\ i = 1, \ldots, n$ предполагалась линейно независимой.

функции[3] $\varphi_i(x) = x^i,\ i = 0, 1, \ldots, n$, то есть будем приближать в $L_2(a, b)$ функцию $f(x)$ полиномами. Выберем для определенности $[a, b] = [0, 1]$. Тогда

$$(\varphi_k, \varphi_i) = \int\limits_0^1 x^k \cdot x^i \, dx = \int\limits_0^1 x^{k+i} \, dx = \frac{1}{k+i+1}$$

и система (3.58) принимает вид

$$\begin{cases} c_0 + \frac{1}{2}c_1 + \cdots + \frac{1}{n+1}c_n & = \int\limits_0^1 f(x)\, dx, \\ \frac{1}{2}c_0 + \frac{1}{3}c_1 + \cdots + \frac{1}{n+2}c_n & = \int\limits_0^1 x f(x)\, dx, \\ \cdots & \\ \frac{1}{n+1}c_0 + \frac{1}{n+2}c_1 + \cdots + \frac{1}{2n+1}c_n & = \int\limits_0^1 x^n f(x)\, dx. \end{cases} \tag{3.60}$$

Матрица этой системы

$$\mathbf{H}_{n+1} = \Big(\frac{1}{i+j-1}\Big)_{i,j=1}^{n+1}$$

уже встречалась в параграфе 2.1.4. Это матрица Гильберта. Как отмечалось, она является плохо обусловленной. Для $\mathbf{H}_6$ число обусловленности порядка 10^7, для $\mathbf{H}_9$ — порядка 10^{13}. Поэтому среднеквадратичное приближение многочленами высоких степеней не используется.

Пример. Построим многочлен $P_1^{\text{ск}}(x) = c_0 + c_1 x$ приближающий в пространстве $L_2(0, 1)$ наилучшим образом функцию $f(x) = \sqrt{x}$.

Для решения воспользуемся системой (3.60)

$$\begin{cases} c_0 + \frac{1}{2}c_1 & = \int\limits_0^1 \sqrt{x}\, dx = \frac{2}{3}, \\ \frac{1}{2}c_0 + \frac{1}{3}c_1 & = \int\limits_0^1 x\sqrt{x}\, dx = \frac{2}{5}. \end{cases}$$

Решая эту систему, имеем $c_0 = 4/15,\ c_1 = 4/5$. Таким образом, искомый многочлен $P_1^{\text{ск}}(x) = 4/15 + (4/5)x$. При этом

$$\|\sqrt{x} - P_1^{\text{ск}}\|_{L_2(0,1)} = \left(\int\limits_0^1 \Big(\sqrt{x} - \frac{4}{15} - \frac{4}{5}x\Big)^2 dx\right)^{1/2} = \frac{\sqrt{2}}{30}.$$

Система (3.58) приобретает простой вид, если элементы φ_i попарно ортогональны и их норма равна 1. Тогда получаем

$$c_i = (f, \varphi_i), \quad i = 1, \ldots, n \tag{3.61}$$

и элемент наилучшего приближения φ равен

$$\varphi = \sum_{i=1}^{n} (f, \varphi_i)\varphi_i. \tag{3.62}$$

Числа c_i, определенные по формуле (3.61), называются **коэффициентами Фурье** элемента f, а обобщенный многочлен (3.62) — **многочленом Фурье.**

[3]Здесь для удобства нумерация начинается с 0, а не 1.

3.3.3 Равномерное приближение функций

В том случае, когда при рассмотрении задачи о наилучшем приближении выбирается норма

$$\|f\| = \max_{x\in[a,b]} |f(x)|,$$

говорят о **равномерном приближении** или **приближении Чебышева**.

Чебышевские приближения играют большую роль в приложениях, например, при задании функций в вычислительных машинах. В самом деле, вводить функцию в машину в виде таблицы невыгодно, так как с одной стороны,таблицы требуют много места в памяти и, с другой стороны, поиск нужного значения также отнимает много времени. Обычно подлежащую вычислению функцию $f(x)$ заменяют некоторой другой функцией $g(x)$, значения которой находятся проще. При этом функцию $g(x)$ часто определяют так, чтобы она на рассматриваемом отрезке отклонялась от функции $f(x)$ не более чем на заданную величину ε, то есть $g(x)$ определяют как наилучшее равномерное приближение функции $f(x)$. Правда, найти наилучшее приближение существенно сложнее, чем среднеквадратичное.

Будем изучать равномерное приближение непрерывной функции $f(x)$ на отрезке $[a,b]$ многочленами $P(x) = \sum\limits_{i=0}^{n} a_i x^i$ степени не выше n. Согласно теореме 3.3.1 наилучшее приближение существует. Изучим его свойства.

Теорема 3.3.3 (Валле-Пуссен) *Если для некоторого многочлена $P_0(x)$ степени не выше n функция $f(x) - P_0(x)$ в $n+2$ точках $a \le x_1 < x_2 < \cdots < x_{n+2} \le b$ принимает значения положительного и отрицательного знака поочередно, то для величины*

$$\rho(f) = \min_{(a_0,\ldots,a_n)} \max_{x\in[a,b]} |f(x) - P(x)| = \min_{(a_0,\ldots,a_n)} \|f - P\|.$$

справедлива оценка

$$\min_i |f(x_i) - P_0(x_i)| \le \rho(f) \le \|f - P_0\|. \tag{3.63}$$

Доказательство. Так как $P_0(x)$ один из многочленов степени не выше n, из определения величины $\rho(f)$ следует справедливость правой части неравенства (3.63).

Перейдем к доказательству левой части неравенства. Предположим, что оно не справедливо, значит

$$\min_i |f(x_i) - P_0(x_i)| > \rho(f).$$

Пусть $Q(x)$ — многочлен степени не выше n, наилучшим образом приближающий $f(x)$, то есть $\|f - Q\| = \rho(f)$. Тогда многочлен $Q(x) - P_0(x) = (f(x) - P_0(x)) - (f(x) - Q(x))$ в точках x_i, $i = 1,\ldots,(n+2)$ принимает значения того же знака, что и $f(x) - P_0(x)$, то есть положительного и отрицательного знака поочередно. Следовательно, этот многочлен имеет по крайней мере $n+1$ ноль, что противоречит основной теореме алгебры.

Критерий для определения многочлена наилучшего равномерного приближения формулируется в следующей теореме, которая называется **теоремой Чебышева** или **теоремой об альтернансе**.

Теорема 3.3.4 *Для того, чтобы многочлен $P_n(x)$ степени не выше n был многочленом наилучшего равномерного приближения непрерывной функции $f(x)$, заданной на отрезке $[a,b]$, необходимо и достаточно существование на $[a,b]$ по крайней мере $n+2$ точек $x_0 < x_1 < \cdots < x_{n+1}$ таких, что*

$$f(x_i) - P_n(x_i) = \sigma(-1)^i \|f - P_n\|, \quad i = 0, 1, \ldots, n+1.$$

Здесь σ — число, не зависящее от i, равное -1 или 1.

Точки $x_0, x_1, \ldots, x_{n+1}$, удовлетворяющие условиям теоремы называются **точками чебышевского альтернанса**.

Доказательство. Достаточность. В соответствии с теоремой Валле-Пуссена

$$\min_i |f(x_i) - P_n(x_i)| \le \rho(f) \le \|f - P_n\|.$$

Но по условию теоремы при всех значениях i выполняются равенства

$$|f(x_i) - P_n(x_i)| = \|f - P_n\|,$$

поэтому $\rho(f) = \|f - P_n\|$. Это означает, что $P_n(x)$ — многочлен наилучшего равномерного приближения.

Необходимость. Пусть $P_n(x)$ — многочлен наилучшего равномерного приближения для функции $f(x)$, а это означает, что $\|f - P_n\| = \rho(f)$. Пусть

$$U([a,b]) = \{x \ : \ x \in [a,b], \ |f(x) - P_n(x)| = \rho(f)\},$$

то есть $U([a,b])$ — множество точек отрезка $[a,b]$, где модуль разности $f(x) - P_n(x)$ принимает свое наибольшее значение. Пусть $x_1 = \inf\limits_{x \in U([a,b])} x$. В силу непрерывности функции $f(x) - P_n(x)$ выполняется равенство $|f(x_1) - P_n(x_1)| = \rho(f)$. Для определенности будем считать, что $f(x_1) - P_n(x_1) = \rho(f)$.

Введем теперь множество

$$U((x_1,b]) = \{x \ : \ x \in (x_1,b], \ f(x) - P_n(x) = -\rho(f)\}$$

и пусть $x_2 = \inf U((x_1,b])$. Из непрерывности следует, что $f(x_2) - P_n(x_2) = -\rho(f)$. Затем определим множество

$$U((x_2,b]) = \{x \ : \ x \in (x_2,b], \ f(x) - P_n(x) = \rho(f)\},$$

число $x_3 = \inf U((x_2,b])$ и так далее. Будем продолжать этот процесс до тех пор, пока не получим $x_m = b$ или $|f(x) - P_n(x)| < \rho(f)$ при всех $x \in (x_m, b]$. В результате имеем точки $x_1, \ldots, x_m$ такие, что $f(x_i) - P_n(x_i) = (-1)^{i-1}\rho(f)$. Если окажется, что $m \ge n+2$, то утверждение теоремы справедливо.

Покажем, что неравенство $m < n+2$ невозможно. Предположим, что это неравенство выполнено. Положим $y_0 = a, \ y_m = b$. Для каждого целого числа $i = 2, \ldots, m$ определим число y_{i-1} такое, что при всех $x \in [y_{i-1}, x_i)$ выполняется неравенство $|f(x) - P_n(x)| < \rho(f)$. Очевидно, что в силу непрерывности функции $f(x) - P_n(x)$ числа y_{i-1} существуют. Тогда, по построению, на каждом промежутке $[y_{i-1}, y_i)$, $i = 1, \ldots, m$ есть точки, например, x_i, где $f(x) - P_n(x) = (-1)^{i-1}\rho(f)$ и нет точек, где $f(x) - P_n(x) = (-1)^i\rho(f)$.[4]

[4]Для того, чтобы проиллюстрировать введенные здесь точки и множества рассмотрим рисунок 3.2. На нем множество $U([a,b])$ состоит из точек x_1, x_1^1, x_3 и отрезка $[x_2, x_2^1]$, $U((x_1,b]) = [x_2, x_2^1]$, а в $U((x_2,b])$ входит одна точка x_3.

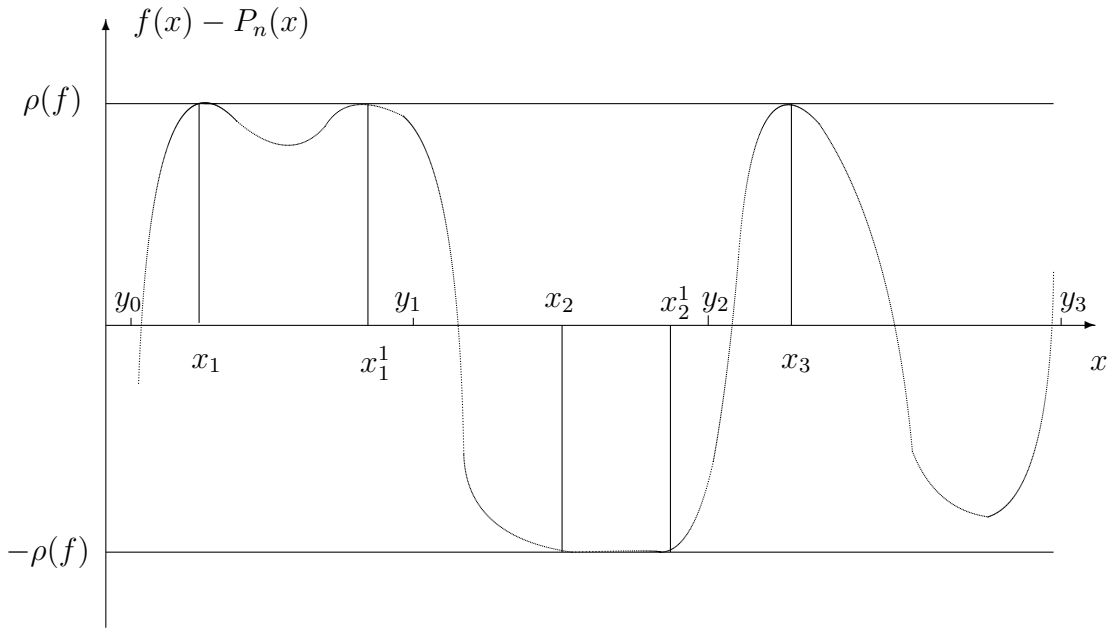


Рис. 3.2 Иллюстрация к доказательству теоремы Чебышева

Введем многочлен

$$\omega(x) = \prod_{j=1}^{m-1} (y_j - x).$$

Этот многочлен обладает той особенностью, что

$$\operatorname{sign} \omega(x) = (-1)^j \quad \text{при} \quad x \in (y_j, y_{j+1}), \quad j = 0, 1, \ldots, m-1.$$

Кроме того, степень многочлена равна $m-1$ и, следовательно, по предположению не превосходит n.

Покажем теперь, что найдется такое число $\varepsilon > 0$, что расстояние от многочлена $P_\varepsilon(x) = P_n(x) + \varepsilon\omega(x)$ до функции $f(x)$ меньше $\rho(f)$. Это будет противоречить тому, что $P_n(x)$ — многочлен наилучшего равномерного приближения.

Возьмем произвольный отрезок из тех, на которые точки y_j делят отрезок $[a, b]$, например, $[y_0, y_1]$. Так как $\omega(x) > 0$ при $x \in [y_0, y_1)$, имеем для этих значений x

$$f(x) - P_\varepsilon(x) = f(x) - P_n(x) - \varepsilon\omega(x) \le \rho(f) - \varepsilon\omega(x) < \rho(f).$$

С другой стороны

$$\begin{aligned} f(x) - P_\varepsilon(x) = f(x) - P_n(x) + \rho(f) - \varepsilon\omega(x) - \rho(f) \ge \\ \ge \min_{x\in[y_0,y_1]} |f(x) - P_n(x) + \rho(f)| - \varepsilon \max_{x\in[y_0,y_1]} \omega(x) - \rho(f) > -\rho(f), \end{aligned}$$

если взять ε таким, что

$$\min_{x\in[y_0,y_1]} |f(x) - P_n(x) + \rho(f)| > \varepsilon \max_{x\in[y_0,y_1]} \omega(x).$$

Необходимое значение ε подобрать можно, так как по построению, на промежутке $[y_0, y_1]$ нет точек x, где $f(x) - P_n(x) = -\rho(f)$. Значит, число

$$\mu = \min_{x\in[y_0,y_1]} |f(x) - P_n(x) + \rho(f)| > 0$$

и достаточно взять

$$0 < \varepsilon < \frac{\mu}{\max\limits_{x\in[y_0,y_1]} \omega(x)}.$$

В точке $x = y_1$ имеем $|f(y_1) - P_\varepsilon(y_1)| = |f(y_1) - P_n(y_1)| < \rho(f)$. Следовательно, на всем отрезке $[y_0, y_1]$ выполняется неравенство $|f(x) - P_\varepsilon(x)| < \rho(f)$.

Проведя подобные рассуждения для произвольного отрезка $[y_j, y_{j+1}]$, получим что $\|f - P_\varepsilon\| < \rho(f)$ на всем отрезке $[a, b]$, а это противоречит определению $rho(f)$. Противоречие доказывает, что предположение о том, что $m < n+2$ не верно. Теорема доказана.

Теорема 3.3.5 *Для непрерывной функции многочлен наилучшего равномерного приближения единствен.*

Доказательство. Предположим противное, то есть что существуют по крайней мере два различных многочлена $P_n^1(x)$ и $P_n^2(x)$ степени не выше n такие, что

$$\|f - P_n^1\| = \|f - P_n^2\| = \rho(f).$$

Тогда

$$\left\|f - \frac{P_n^1 + P_n^2}{2}\right\| \le \frac{1}{2}\|f - P_n^1\| + \frac{1}{2}\|f - P_n^2\| = \rho(f).$$

Из этого неравенства следует, что многочлен $1/2(P_n^1 + P_n^2)$, степень которого не выше n также является многочленом наилучшего равномерного приближения. Тогда, в соответствии с теоремой Чебышева, у него есть по крайней мере $n + 2$ точки альтернанса. Обозначим их $x_0, \ldots, x_{n+1}$. В этих точках выполняются равенства

$$\left|f(x_i) - \frac{1}{2}(P_n^1(x_i) + P_n^2(x_i))\right| = \frac{1}{2}|(f(x_i) - P_n^1(x_i)) + (f(x_i) - P_n^2(x_i))| = \rho(f). \quad (3.64)$$

Но при все значениях $x \in [a, b]$ справедливы неравенства

$$|f(x_i) - P_n^1(x_i)| \le \rho(f), \quad |f(x_i) - P_n^2(x_i)| \le \rho(f),$$

поэтому равенства (3.64) возможны только тогда, когда

$$f(x_i) - P_n^1(x_i) = f(x_i) - P_n^2(x_i).$$

Отсюда следует, что в $n + 2$ точках отличный от тождественного нуля многочлен $P_n^1 - P_n^2$ степени не выше n равен нулю, что противоречит основной теореме алгебры. Полученное противоречие доказывает теорему.

Рассмотрим теперь некоторые примеры нахождения многочлена наилучшего равномерного приближения для непрерывной функции $f(x)$ на отрезке $[a, b]$. Следует отметит, что не существует общих методов построения такого многочлена, однако его можно найти с помощью итераций.

Для построения многочлена нулевой степени достаточно положить

$$P_0 = \frac{1}{2}(\max_{x\in[a,b]} f(x) + \min_{x\in[a,b]} f(x)).$$

Если обозначить через x_1, x_2 точки, в которых функция принимает соответственно максимальное и минимальное значения, то очевидно, что

$$f(x_1) - P_0 = -(f(x_2) - P_0) = \|f - P_0\| = \frac{1}{2}(\max_{x\in[a,b]} f(x) - \min_{x\in[a,b]} f(x)).$$

Таким образом, эти точки являются точками чебышевского альтернанса и то теореме Чебышева P_0 — искомый многочлен (см. рисунок 3.3).

При нахождении многочлена первой степени будем дополнительно предполагать, что функция $f(x)$ дважды непрерывно дифференцируема на интервале (a, b) и вторая производная не меняет знак. По теореме Чебышева многочлен $P_1(x) = a_0 + a_1 x$ тогда и только тогда является многочленом наилучшего равномерного приближения, когда имеются три точки альтернанса x_1, x_2, x_3. Если считать, что точки упорядочены в порядке возрастания, то точка x_2 заведомо должна лежать внутри отрезка. Так как в ней функция $f(x) - P_1(x)$ принимает экстремальное значение,

$$0 = f'(x_2) - P_1'(x_2) = f'(x_2) - a_1.$$

Так как вторая производная функции $f(x)$ не меняет знак, первая производная либо возрастает, либо убывает. Следовательно, первая производная значение a_1 может принимать только один раз. Поэтому функция $f(x) - P_1(x)$ не может принять в других точках интервала (a, b) экстремальное значение. Значит экстремум достигается на концах отрезка, то есть $x_1 = a,\ x_3 = b$. В соответствии со свойствами точек альтернанси имеем

$$f(a) - P_1(a) = -(f(x_2) - P_1(x_2)) = f(b) - P_1(b). \tag{3.65}$$

Отсюда имеем $f(a) - a_0 - a_1 a = f(b) - a_0 - a_1 b$ или

$$a_1 = \frac{f(a) - f(b)}{a - b}. \tag{3.66}$$

Из первых двух уравнений (3.65) получаем

$$a_0 = \frac{1}{2}\Big((f(a) + f(x_2)) - (a + x_2)a_1\Big) = \frac{1}{2}\Big((f(a) + f(x_2)) - (a + x_2)\frac{f(a) - f(b)}{a - b}\Big). \tag{3.67}$$

Для определения x_2 имеем уравнение

$$f'(x_2) = a_1 = \frac{f(a) - f(b)}{a - b}.$$

Это уравнение означает, что касательная в точке x_2 параллельна хорде, соединяющей точки $A(a, f(a)),\ B(b, f(b))$. Геометрически процедура построения графика многочлена $P_1(x)$ сводится к тому, что сначала проводится хорда AB, затем с этим же наклоном проводится касательная к кривой $y = f(x)$ и, наконец, проводится прямая посредине между хордой и касательной (см. рисунок 3.4).

Пример 1. Найдем многочлен наилучшего равномерного приближения $P_1^{\text{рав}}(x) = a_0 + a_1 x$ для функции $f(x) = \sqrt{x}$ на отрезке $[0, 1]$.

Из (3.66) имеем $a_1 = \sqrt{1} - \sqrt{0} = 1$. x_2 находится из уравнения

$$1 = a_1 = f'(x_2) = \frac{1}{2\sqrt{x_2}},$$

откуда следует, что $x_2 = 1/4$. Тогда, согласно (3.67)

$$a_0 = \frac{1}{2}\Big(\sqrt{\frac{1}{4}} - \frac{1}{4}a_1\Big) = \frac{1}{8}.$$

Таким образом, $P_1^{\text{рав}}(x) = 1/8 + x$. При этом для того, чтобы найти расстояние между $\sqrt{x}$ и $P_1^{\text{рав}}(x)$ достаточно вычислить модуль их разности в одной из точек альтернанса, например, в нуле. Получим $\rho(f) = 1/8$.

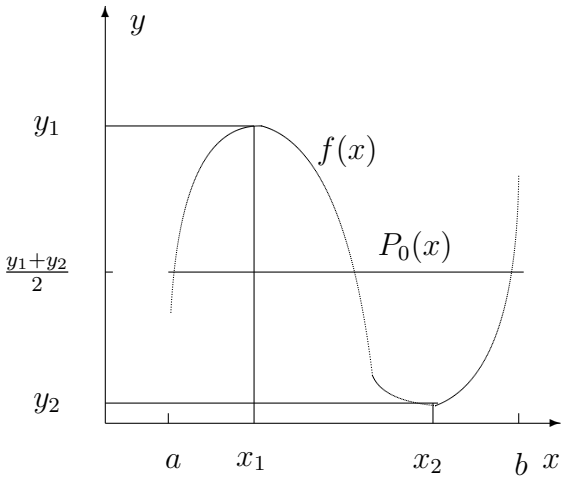


Рис. 3.3 Построение многочлена наилучшего равномерного приближения нулевой степени

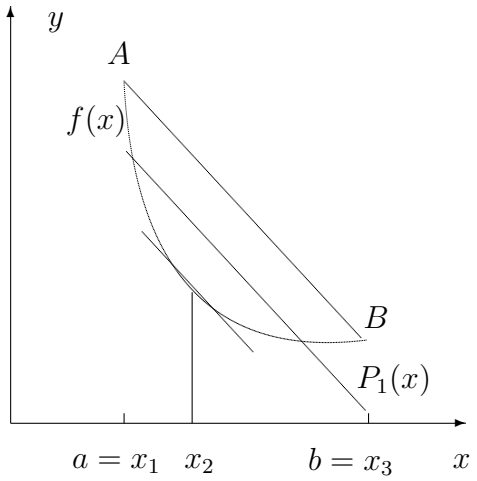


Рис. 3.4 Построение многочлена наилучшего равномерного приближения первой степени

Интересно сравнить полученный многочлен со среднеквадратичным приближением $P_1^{\text{ск}} = 4/15 + (4/5)x$, построенном в конце параграфа 3.3.2 для этой же функции. Имеем

$$\max_{x\in[0,1]}\left|\sqrt{x} - \frac{4}{15} - \frac{4}{5}x\right| = \frac{4}{15} > \frac{1}{8}.$$

С другой стороны,

$$\|\sqrt{x} - P_1^{\text{рав}}\|_{L_2(0,1)} = \left(\int\limits_0^1 \left(\sqrt{x} - \frac{1}{8} - x\right)^2 dx\right)^{1/2} = \sqrt{\frac{7}{960}} > \|\sqrt{x} - P_1^{\text{ск}}\|_{L_2(0,1)} = \frac{\sqrt{2}}{30}.$$

Пример 2. Получим простую формулу для приближенного вычисления значений функции $y = \sin x$ [5].

Заметим прежде всего, что из формулы $\sin 3x = 3\sin x - 4\sin^3 x$ следует, что любые значения синуса можно получить, зная их на отрезке $[0, \pi/6]$. На указанном отрезке знак второй производной не меняется, поэтому применим алгоритм построения многочлена наилучшего равномерного приближения первой степени. Секущая, соединяющая точки $(0,0)$ и $(\pi/6, \sin(\pi/6)) = (\pi/6, 1/2)$, имеет уравнение $y = (3x)/\pi$. Для определения второй точки альтернанса x_2 имеем уравнение $\cos(x) = 3/\pi$, откуда $x_2 = \arccos(3/\pi)$. Уравнение касательной, проходящей через точку

$$(\arccos(3/\pi), \sin(\arccos(3/\pi)))$$

будет

$$y = \frac{3}{\pi}x + \frac{1}{\pi}\left(\sqrt{\pi^2 - 9} - 3\arccos\frac{3}{\pi}\right) \approx 0.955x + 0.010.$$

Уравнение прямой, параллельной этим секущей и касательной и равноудаленной от них, таково[6]:

$$P_1(x) = \frac{3}{\pi}x + \frac{1}{2\pi}\left(\sqrt{\pi^2 - 9} - 3\arccos\frac{3}{\pi}\right) \approx 0.955x + 0.005.$$

[5]Если воспользоваться формулами приведения, то зная значения синуса, легко найти значения косинуса, а, значит, тангенса и котангенса.

[6]Можно было просто воспользоваться формулами (3.66),(3.67)

Таким образом,

$$\sin x \approx 0.955x + 0.005$$

при этом справедлива оценка

$$|\sin x - 0.955x - 0.005| \le 0.005.$$

Для нахождении многочлена наилучшего равномерного приближения в некоторых случаях удобно использовать следующую теорему.

Теорема 3.3.6 *При приближении непрерывной четной (соответственно нечетной) функции $f(x)$ на симметричном относительно нуля отрезке многочленами степени не выше n, наилучшее приближение также будет четной (соответственно нечетной) функцией.*

Доказательство. Пусть $f(x)$ — четная функция для всех $x \in [-a, a]$ и $P_n(x)$ — наилучшее равномерное приближение. Тогда

$$\rho(f) = \max_{x\in[-a,a]} |f(x) - P_n(x)| = \max_{x\in[-a,a]} |f(-x) - P_n(-x)| = \max_{x\in[-a,a]} |f(x) - P_n(-x)|.$$

Это равенство означает, что $P_n(-x)$ также является многочленом наилучшего равномерного приближения. Тогда, согласно теореме единственности, $P_n(x) = P_n(-x)$.

Для нечетной функции доказательство проводится аналогично.

Из приведенной теоремы следует, что если, например, необходимо построить многочлен не выше второй степени, наилучшим образом равномерно приближающий функцию x^3 на отрезке $[-1, 1]$, то многочлен имеет вид $y = ax$. В силу симметрии относительно начала координат очевидно, что число точек альтернанса будет четным и, значит, не менее четырех. Две точки должны быть внутри отрезка и две на концах. Для внутренних точек должно выполняться равенство $(x^3 - ax)' = 0$. Отсюда следует, что среди точек альтернанса есть точки $x = \pm\sqrt{a/3}$. Из условия равенства модуля разности функции и многочлена в точках альтернанса получаем

$$1 - a = -\Big(\sqrt{\frac{a}{3}}\Big)^3 + a\sqrt{\frac{a}{3}}.$$

Этому уравнению удовлетворяет $a = 3/4$. Значит $P_2(x) = (3/4)x$.

Другой подход к решению этой задачи будет рассмотрен в следующем параграфе.

3.3.4 Многочлены Чебышева

Можно по разному ставить задачи, приводящие к многочленам Чебышева. Сформулируем некоторые из них.

При изучении интерполяции $n+1$ раз непрерывно дифференцируемой на отрезке $[a, b]$ функции $f(x)$ многочленом Лагранжа, была получена следующая оценка для погрешности

$$|f(x) - L_n(x)| \le \frac{\max\limits_{x\in[a,b]} |f^{(n+1)}(x)|}{(n+1)!} \Big|\prod_{i=0}^{n}(x - x_i)\Big|, \tag{3.68}$$

где x_i — узлы интерполяции. В связи с этим возникает задача, как выбрать узлы интерполяции, чтобы величина

$$\max_{x\in[a,b]} |\omega_n(x)| = \max_{x\in[a,b]} \Big|\prod_{i=0}^{n}(x - x_i)\Big| \tag{3.69}$$

была минимальной, то есть как минимизировать правую часть оценки погрешности. $\omega_n(x)$ — многочлен степени $n+1$ с коэффициентом 1 при старшей степени. Таким образом, пришли к задаче нахождения многочлена степени $n+1$ с коэффициентом 1 при старшей степени, который наименее отклоняется от нуля, то есть который является многочленом наилучшего равномерного приближения функции тождественно равной нулю. Искомые узлы интерполяции будут его корнями.

Можно по другому сформулировать проблему поиска $\omega_n(x)$: найти многочлен наилучшего равномерного приближения для функции x^{n+1}. Подобная задача решалась в конце предыдущего параграфа для для отрезка $[0,1]$ и $n=2$.

Рассмотрим сначала отрезок $[-1,1]$. **Многочлены Чебышева** $T_n(x),\ n \geq 0$ на отрезке $[-1,1]$ задаются формулой

$$T_n(x) = \cos(n \cdot \arccos x). \tag{3.70}$$

В частности, при $n = 0, 1$ имеем

$$T_0(x) = \cos(0 \cdot \arccos x) = 1, \tag{3.71}$$
$$T_1(x) = \cos(\arccos x) = x. \tag{3.72}$$

Если воспользоваться формулой

$$\cos(n+1)\alpha = 2\cos\alpha\cos n\alpha - \cos(n-1)\alpha,$$

полагая в ней $\alpha = \arccos x$, и, учитывая (3.70), получим

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2\ldots. \tag{3.73}$$

Из формул (3.71)–(3.73) следует теперь, что $T_n(x)$ действительно является многочленом степени n.

Полагая теперь $T_0(x) = 1,\ T_1(x) = x$ на всей числовой оси и распространяя рекуррентную формулу (3.73) на всю числовую ось, последовательно находим по этой формуле

$$T_2(x) = 2x^2 - 1, \qquad T_3 = 4x^3 - 3x,$$
$$T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \ \ldots$$

Рассмотрим **свойства многочленов Чебышева**.

1.*Многочлен Чебешева при четном n является четной функцией, а при нечетном n — нечетной.*

Доказательство этого утверждения очевидным образом следует из (3.71)–(3.73).

2. *Старший коэффициент многочлена $T_n(x)$ при $n > 0$ равен 2^{n-1}.*

Это свойство также непосредственно следует их (3.71)–(3.73).

3. *Многочлен $T_n(x)$ имеет n действительных корней на интервале $(-1,1)$, выражаемых формулой*

$$x_i = \cos\frac{(2i+1)\pi}{2n}, \quad i = 0, 1, \ldots, n-1.$$

Доказательство следует из равенств,

$$T_n(x) = \cos(n\arccos x_i) = \cos\frac{(2i+1)\pi}{2} = 0, \quad i = 0, 1, \ldots, n-1.$$

4. $\max\limits_{x\in[-1,1]} |T_n(x)| = 1$, *причем для* $x_j = \cos\dfrac{j\pi}{n},\ j = 0, 1, \ldots, n$

$$T_n(x_j) = (-1)^j. \tag{3.74}$$

Действительно, учитывая (3.70), имеем $|T_n(x)| \le 1$ при $x \in [-1, 1]$, и $T_n(x_j) = \cos j\pi = (-1)^j$.

5. *Многочлен*

$$\widetilde{T}_n(x) = 2^{1-n}T_n(x), \quad n > 0, \tag{3.75}$$

среди всех многочленов n*-ой степени со старшим коэффициентом, равным единице, имеет на отрезке* $[-1, 1]$ *наименьшее значение максимума модуля, то есть не существует такого многочлена* $\widetilde{P}_n(x)$ n*-ой степени со старшим коэффициентом, равным единице, что*

$$\max_{x\in[-1,1]} |\widetilde{P}_n(x)| < \max_{x\in[-1,1]} |\widetilde{T}_n(x)| = 2^{1-n}, \quad n > 0. \tag{3.76}$$

Для доказательства заметим, что если бы существовал многочлен $\widetilde{P}_n(x)$ со свойствами, указанными в условии утверждения, то в силу (3.76) он не был бы равен тождественно многочлену $\widetilde{T}_n(x)$. Кроме того, разность $\widetilde{P}_n(x) - \widetilde{T}_n(x)$ являлась бы алгебраическим многочленом степени не выше $n - 1$, причем на основании (3.74)-(3.76) в $n + 1$ точках $x_j = \cos\dfrac{j\pi}{n},\ j = 0, 1, \ldots, n$ разность $\widetilde{P}_n(x) - \widetilde{T}_n(x)$ принимала бы отличные от нуля значения чередующихся знаков. Это означало бы, что между этими точками многочлен $\widetilde{P}_n(x) - \widetilde{T}_n(x)$ обращается в ноль, то есть у него есть n действительных корней, что противоречит основной теореме алгебры.

Благодаря свойству 5 многочлены Чебышева называют **многочленами, наименее отклоняющимися от нуля.**

Пусть требуется найти наилучшее равномерное приближение функции x^n на отрезке $[-1, 1]$ многочленом $P_{n-1}(x)$ степени не выше $n - 1$. Из предыдущего параграфа следует, что такое приближение существует и определяется однозначно. Тогда $\widetilde{P}_n(x) = x^n - P_{n-1}(x)$ — многочлен степени n со старшим коэффициентом равным 1. Согласно свойству 5 многочленов Чебышева

$$\max_{x\in[-1,1]} |\widetilde{P}_n(x)| \ge \max_{x\in[-1,1]} |\widetilde{T}_n(x)|. \quad n > 0,$$

а по определению наилучшего равномерного приближения

$$\max_{x\in[-1,1]} |\widetilde{P}_n(x)| \le \max_{x\in[-1,1]} |\widetilde{T}_n(x)|, \quad n > 0.$$

Значит, должно выполняться равенство

$$\max_{x\in[-1,1]} |\widetilde{P}_n(x)| = \max_{x\in[-1,1]} |\widetilde{T}_n(x)| = 2^{1-n}, \quad n > 0, \tag{3.77}$$

Из этого равенства следует, что многочлен $x^n - \widetilde{T}_n(x)$, степень которого не выше $n-1$, также наилучшим образом приближает x^n. В силу единственности наилучшего приближения, отсюда делаем вывод, что

$$P_{n-1}(x) = x^n - \widetilde{T}_n(x). \tag{3.78}$$

Из приведенного рассуждения легко следует свойство 6.

6. *Если* $\overline{P}_n(x)$ *— многочлен степени* $n > 0$ *со старшим коэффициентом равным 1 и* $\max\limits_{x\in[-1,1]} |\overline{P}_n(x)| = 2^{1-n}$*, то* $\overline{P}_n(x) = 2^{1-n}T_n(x) = \widetilde{T}_n(x)$*, то есть существует единственный многочлен степени n со старшим коэффициентом 1, наименее уклоняющийся от нуля.*

Действительно, из условия и равенства (3.77) следует, что

$$\max_{x\in[-1,1]} |\widetilde{P}_n(x)| = \max_{x\in[-1,1]} |\overline{P}_n(x)| = 2^{1-n}. \;\; n > 0.$$

Тогда, проводя аналогичные рассуждения, получим, что $P_{n-1}(x) = x^n - \overline{P}_n(x)$. На основании (3.78) убеждаемся теперь в справедливости утверждения.

Рассмотрим теперь случай произвольного отрезка $[a, b]$. Линейной заменой независимой переменной

$$x = \frac{b+a}{2} + \frac{b-a}{2}t, \quad t = \frac{2x-b-a}{b-a} \tag{3.79}$$

отрезок $[a, b]$ переводится в $[-1, 1]$. Многочлен $T_n(t)$ преобразуется в $T_n\Big(\frac{2x-b-a}{b-a}\Big)$ со старшим коэффициентом $2^{2n-1}(b-a)^{-n}$. Тогда, в соответствии со свойством 5 можно утверждать, что многочлен

$$\widetilde{T}_n^{[a,b]}(x) = (b-a)^n 2^{1-2n} T_n\Big(\frac{2x-b-a}{b-a}\Big), \tag{3.80}$$

старший коэффициент которого равен 1, является многочленом, наименее уклоняющимся от нуля на отрезке $[a, b]$, а многочлен $x^n - \widetilde{T}_n^{[a,b]}(x)$ является многочленом степени не выше $n-1$, наилучшим образом равномерно приближающий функцию x^n.

Перейдем теперь к проблеме, сформулированной в начале параграфа — минимизации правой части оценки (3.68). Как отмечалось, необходимо минимизировать максимальное значение функции $|\omega_n(x)| = |(x-x_0)(x-x_1)\dots(x-x_n)|$. В соответствии со свойствами 5,6 должно выполняться равенство $\omega_n(x) = \widetilde{T}_{n+1}^{[a,b]}(x)$. Корнями функции $\widetilde{T}_{n+1}^{[a,b]}(x)$ являются числа

$$x_i = \frac{1}{2}\Big((b-a)\cos\frac{(2i+1)\pi}{2n+2} + b + a\Big), \;\; i = 0, 1, \dots, n, \tag{3.81}$$

следовательно, их необходимо выбрать узлами интерполяции. При этом

$$\max_{x\in[a,b]} |\omega_n(x)| = \max_{x\in[a,b]} |\widetilde{T}_{n+1}^{[a,b]}(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Тогда при выборе узлов интерполяции в соответствии с формулами (3.81) получаем

$$|f(x) - L_n(x)| \le \frac{\max\limits_{x\in[a,b]} |f^{(n+1)}(x)|}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{3.82}$$

Интересно сравнить оценки приближения гладкой функции с помощью отрезка ряда Тейлора $Q_n(x)$ и с помощью интерполяционного многочлена $L_n(x)$ с узлами в точках (3.81). При построении формулы Тейлора целесообразно взять в качестве точки, в окрестности которой ведется разложение, середину отрезка, то есть $\overline{x} = (a+b)/2$. Тогда, записывая остаточный член формулы Тейлора в виде Лагранжа, получим

$$f(x) = Q_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-\overline{x})^{n+1}, \quad \xi \in (a, b).$$

Отсюда

$$|f(x) - Q_n(x)| \le \frac{\max\limits_{x\in[a,b]} |f^{(n+1)}(x)|}{(n+1)!} \frac{(b-a)^{n+1}}{2^{n+1}}.$$

Таким образом, оценка погрешности многочлена Тейлора в 2^n раза больше оценки погрешности (3.82) интерполяционного многочлена Лагранжа $L_n(x)$ с оптимальным выбором узлов (3.81).

Конечно, было произведено сравнение не самих погрешностей, а только их оценок. Сами же погрешности могут различаться не столь существенно, более того, из сравнения оценок еще не следует, что те же неравенства выполняются для самих погрешностей. Однако, обычно фактическая максимальная погрешность интерполяционного многочлена меньше, чем у многочлена Тейлора. При этом построение интерполяционного многочлена обладает еще тем преимуществом, что не требует вычисление производных функции f.

В каждом конкретном случае можно провести более точное сравнение. Например, на отрезке $[-1, 1]$ для функции e^x сравним отрезок ряда Тейлора $Q_5(x) = \sum\limits_{i=1}^{5} \frac{x^i}{i!}$ и многочлен $L_5(x)$ с узлами, вычисленными по формуле (3.81). Имеем

$$\max_{x\in[-1,1]} |e^x - Q_5(x)| = \max_{x\in[-1,1]} \Big|\sum_{i=6}^{\infty} \frac{x^i}{i!}\Big| > \frac{1}{6!} + \frac{1}{7!} > 0.15 \cdot 10^{-2}.$$

Для многочлена Лагранжа справедлива оценка (3.82), в соответствии с которой

$$\max_{x\in[-1,1]} |e^x - L_5(x)| \le \frac{e}{6!2^5} < 0.12 \cdot 10^{-3}.$$

Таким образом, в рассмотренном примере оценка сверху погрешности интерполяционного многочлена $L_5(x)$ значительно меньше, чем оценка снизу максимальной погрешности многочлена Тейлора $Q_5(x)$.

А теперь проведем сравнение погрешностей при приближении функции $f(x)$ многочленом Лагранжа с оптимальным расположением узлов и многочленом наилучшего равномерного приближения, который будем обозначить $P_n(x)$. Сделаем одно дополнительное предположение: будем считать, что производная $f^{(n+1)}(x)$ не меняет знак на отрезке $[a, b]$. В силу теоремы Чебышева разность $f(x) - P_n(x)$ меняет знак при переходе от одной точки альтернанса к другой, поэтому она обращается в ноль в $n + 1$ точке $y_0, y_1, \ldots, y_n$. Значит, многочлен $P_n(x)$ можно рассматривать, как интерполяционный многочлен с узлами интерполяции $y_0, y_1, \ldots, y_n$. Для погрешности интерполяции согласно (3.15) имеем

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega_n(x), \tag{3.83}$$

где $\xi(x) \in [a, b]$, $\omega_n(x) = (x - y_0)(x - y_1)\ldots(x - y_n)$. Пусть точка x^* такова, что

$$\max_{x\in[a,b]} |\omega_n(x)| = |\omega_n(x^*)|.$$

Тогда

$$\rho(f) = \max_{x\in[a,b]} |f(x) - P_n(x)| \ge |f(x^*) - P_n(x^*)| =$$

$$= \Big|\frac{f^{(n+1)}(\xi(x^*))}{(n+1)!}\omega_n(x^*)\Big| \ge \frac{1}{(n+1)!} \max_{x\in[a,b]} |\omega_n(x)| \min_{x\in[a,b]} |f^{(n+1)}(x)|$$

Согласно свойству 5 имеем

$$\max_{x\in[a,b]}|\omega_n(x)| \geq \max_{x\in[a,b]}|\widetilde{T}_{n+1}^{[a,b]}(x)| = (b-a)^{n+1}2^{-1-2n}.$$

Тогда

$$\rho(f) \geq \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}\min_{x\in[a,b]}|f^{(n+1)}(x)|. \tag{3.84}$$

Для интерполяционного многочлена Лагранжа с оптимальным расположением узлов в силу (3.82)

$$\rho(f) \leq |f(x)-L_n(x)| \leq \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}\max_{x\in[a,b]}|f^{(n+1)}(x)|. \tag{3.85}$$

Сравнивая (3.84) и (3.85) видим, что если $f^{(n+1)}(x)$ сохраняет знак и меняется не очень сильно, то разность между погрешностью многочлена наилучшего равномерного приближения и интерполяционного многочлена с узлами интерполяции, совпадающими с нулями многочлена Чебышева, несущественна.

3.4 ЗАДАЧИ К ГЛАВЕ 3

3.4.1 Примеры решения задач

1. Найти приближенное значение функции $y = \log_2 x$ в точке $x = 1.9$ с точностью $\varepsilon = 0.1$.

Решение. Воспользуемся интерполяционным многочленом Лагранжа первой степени $L_1(x)$, учитывая, что $y(1) = 0,\ y(2) = 1$. Имеем $L_1(x) = x - 1,\ L_1(1.9) = 0.9$. Если положить $y(1.9) \approx L_1(1.9) = 0.9$, то по формуле оценки погрешности (3.15) имеем

$$|y(1.9)-L_1(1.9)| \leq \frac{|y''(\xi)(1.9-1)(1.9-2)|}{2} \leq 0.045\log_2 e\max_{x\in[1,2]}\frac{1}{x^2} < 0045\cdot 1.5\cdot 1 < 0.07$$

Таким образом, если взять $y(1.9) \approx 0.9$, то ошибка будет при этом меньше 0.07.

2. Пусть

$$\psi_k(x) = \frac{\prod\limits_{\substack{j=0\\ j\neq k}}^{n}(x-x_j)}{\prod\limits_{\substack{j=0\\ j\neq k}}^{n}(x_k-x_j)}, \quad k = 0,1,\ldots,n.$$

Доказать, что

$$\psi_0(x)+\psi_1(x)+\cdots+\psi_n(x) = 1, \tag{3.86}$$

$$(x_0-x)^j\psi_0(x)+(x_1-x)^j\psi_1(x)+\cdots+(x_n-x)^j\psi_n(x) = 0, \quad j = 1,\ldots,n. \tag{3.87}$$

Решение. Заметим прежде всего, что

$$L_n(x) = x_0^j\psi_0(x)+x_1^j\psi_1(x)+\cdots+x_n^j\psi_n(x)$$

— интерполяционный многочлен Лагранжа для функции $f(x) = x^j$ и, следовательно, совпадает с ней, если $j = 0, 1, \ldots, n$. Поэтому

$$x_0^j\psi_0(x) + x_1^j\psi_1(x) + \cdots + x_n^j\psi_n(x) = x^j. \tag{3.88}$$

При $j = 0$ получаем равенство (3.86).

Перепишем левую часть формулы (3.87) при $j = 1$ в виде

$$\begin{aligned}(x_0 - x)\psi_0(x) + (x_1 - x)\psi_1(x) + \cdots + (x_n - x)\psi_n(x) = \\ = \big(x_0\psi_0(x) + x_1\psi_1(x) + \cdots + x_n\psi_n(x)\big) - x\big(\psi_0(x) + \psi_1(x) + \cdots + \psi_n(x)\big).\end{aligned}$$

Тогда из (3.86), (3.88) следует, что при $j = 1$ равенство (3.87) справедливо.

Доказательство равенства (3.87) для других значений j аналогично. Используя бином Ньютона, (3.86), (3.88), получаем:

$$\begin{aligned}\sum_{k=0}^{n}(x_k - x)^j\psi_k(x) = \sum_{k=0}^{n}\left(\sum_{l=0}^{j}C_j^l(-x)^l x_k^{j-l}\right)\psi_k(x) = \sum_{l=0}^{j}C_j^l(-x)^l\left(\sum_{k=0}^{n}x_k^{j-l}\psi_k(x)\right) = \\ = \sum_{l=0}^{j}C_j^l(-x)^l x^{j-l} = x^j\sum_{l=0}^{j}C_j^l(-1)^l 1^{j-l} = x^j(1-1)^j = 0,\end{aligned}$$

что и требовалось доказать.

3. Пусть $x_0, x_1, \ldots, x_n,\ c_0, c_1, \ldots, c_n$ — произвольные наборы чисел. Доказать, что существует единственный многочлен $P(x)$ степени не выше n такой, что

$$P(x_0) = c_0, \quad P'(x_1) = c_1, \quad \ldots, \quad P^{(n)}(x_n) = c_n.$$

Решение. Покажем, что коэффициенты многочлена могут быть найдены, причем единственным образом. Пусть

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Тогда

$$c_n = P^{(n)}(x_n) = n!a_n.$$

Следовательно, $a_n = \dfrac{c_n}{n!}$. Далее,

$$c_{n-1} = P^{(n-1)}(x_{n-1}) = \frac{n!}{1!}\,a_n x_{n-1} + (n-1)!a_{n-1}.$$

Отсюда

$$a_{n-1} = \frac{1}{(n-1)!}\Big(c_{n-1} - \frac{n!}{1!}\,a_n x_{n-1}\Big).$$

Рассуждая аналогично, получим

$$a_k = \frac{1}{k!}\Big(c_k - \frac{n!}{(n-k)!}\,a_n x_k - \frac{(n-1)!}{(n-1-k)!}\,a_{n-1}x_k - \cdots - \frac{(k+1)!}{1!}\,a_{k+1}x_k\Big).$$

4. Имеется набор точек $(x_i, y_i),\ i = 0, \ldots, m$, представляющих экспериментальные данные. Известно, что при $x \to \infty$ величина y стремится к некоторому значению y_∞. Такому условию удовлетворяет, например, эмпирическая формула вида $y = \dfrac{x}{ax + b}$

Сформулировать метод, позволяющий подобрать для этих данных коэффициенты эмпирической формулы.

Решение. Коэффициенты a, b входят в формулу, задающую функцию $y = \dfrac{x}{ax+b}$ нелинейно, поэтому преобразуем сначала формулу и исходные данные. Перепишем формулу в виде

$$\frac{1}{y} = a + b\frac{1}{x}$$

и введем новые переменные $\eta = \dfrac{1}{y},\ \xi = \dfrac{1}{x}$. Точки (ξ_i, η_i) связаны уже линейной зависимостью $\eta = a + b\xi$ (совершена процедура выравнивания исходных данных). Применим теперь к ним метод наименьших квадратов, в соответствии с которым коэффициенты a, b находятся из системы (3.55). Выберем веса одинаковыми, поскольку нет никакой информации о точности эксперимента при различных x_i. Тогда для данного конкретного случая система запишется в виде

$$\begin{cases} b\sum\limits_{i=0}^{m}\xi_i + a(m+1) & = \sum\limits_{i=0}^{m}\eta_i, \\ b\sum\limits_{i=0}^{m}\xi_i^2 + a\sum\limits_{i=0}^{m}\xi_i & = \sum\limits_{i=0}^{m}\xi_i\eta_i. \end{cases}$$

Заметим, что выравнивание исходных данных можно было бы произвести другими способами: переписать функцию в виде $\dfrac{x}{y} = ax + b$ и ввести новые переменные $\xi = x,\ \eta = \dfrac{x}{y}$; представить функцию в виде $\dfrac{y}{x} = \dfrac{1}{b} - \dfrac{a}{b}y$ и ввести переменные $\xi = y,\ \eta = \dfrac{y}{x}$.

3.4.2 Задачи

1. Построить многочлен Лагранжа $L_2(x)$ для следующих случаев:

$$x_0 = -1,\ y(x_0) = 3,\quad x_1 = 0,\ y(x_1) = 2,\quad x_2 = 1,\ y(x_2) = 5;$$
$$x_0 = -1,\ y(x_0) = 3,\quad x_1 = 0,\ y(x_1) = 4,\quad x_2 = 2,\ y(x_2) = 6.$$

2. Пусть $\omega_n(x) = (x - x_0)(x - x_1)\dots(x - x_n)$. Вычислить значения выражения

$$\sum_{i=0}^{n} x_i^k \frac{\omega_n(x)}{(x - x_i)\omega'(x_i)} \quad \text{при} \quad k = 0, 1, \dots, n+1.$$

Указание. Для нахождения значения выражения при $k = n+1$ воспользоваться представлением для погрешности интерполирования.

3. Пусть $x_i = a + i(b-a)/n,\ i = 0, \dots, n.$. Для $n = 1, 2, 3$ найти $\max\limits_{x\in[a,b]} |\omega_n(x)|$.

4. Найти оценку приближения функции $f(x)$ на отрезке $[a, b]$ по узлам $x_i = a + i(b-a)/n,\ i = 0, \dots, n.$:

$$a)\ [0, 0.1],\ f(x) = \sin(x),\ n = 1; \qquad b)\ [-1, 0],\ f(x) = e^x,\ n = 2.$$

5. Оценить погрешность приближения функции e^x в точке $x = 0.05$ интерполяционным многочленом Лагранжа, построенным по узлам $x_0 = 0,\ x_1 = 0.1,\ x_2 = 0.2$.

6. Построить многочлен $P_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, удовлетворяющий условиям: $P_3(-1) = 0,\ P_3(1) = 1,\ P_3(2) = 2,\ a_3 = 1$.

7. Построить многочлен $P_4(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, удовлетворяющий условиям: $P_4(-1) = P_4(1) = P_4'(0) = P_4"(0) = 0,\ P_4(0) = 1$.

8. Функция $\ln(x)$ приближается на отрезке $[1, 2]$ интерполяционным многочленом третьей степени по узлам $1, 4/3, 5/3, 2$. Доказать, что погрешность интерполяции в каждой точке отрезка не превосходит $1/300$.

9. Какой должен быть шаг h таблицы значений функции $f(x) = \cos x$, чтобы погрешность при квадратичной интерполяции не превосходила 10^{-6}?

10. Пусть функция $f(x)$ задана на $[a, b]$ и на этом отрезке модуль ее второй производной не превосходит числа M. Оценить погрешность приближения этой функции ломаной, построенной на равномерной сетке с шагом h.

11. Функция двух переменных $F(x_1, x_2)$ приближается интерполяционным многочленом $P(x_1, x_2) = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2$. Известно, что $F(0, 0) = 1,\ F(1, 0) = 2,\ F(0, 1) = 4,\ F(1, 1) = 3$. Найти $P(1/2, 1/2)$.

12. Пусть на отрезке $[a, b]$ фиксированы узлы $x_0, \ldots, x_n$. Обозначим через $\mathcal{P}_n$ оператор, определенный на множестве $(n+1)$-мерных векторов и принимающий значения в пространстве $C(a, b)$ [7]. Оператор $\mathcal{P}_n$ вектору $\mathbf{f} = (f_0, \ldots, f_n)$ ставит в соответствие интерполяционный многочлен Лагранжа $L_n(x)$ такой, что $L_n(x_i) = f_i,\ i = 0, \ldots, n$. Введем в пространстве векторов норму по формуле $\|\mathbf{f}\| = \max\limits_{i=0,\ldots,n} |f_i|$. Тогда норма оператора $\mathcal{P}_n$ называется **константой Лебега**. Доказать, что

$$\|\mathcal{P}_n\| = \max_{x\in[a.b]} \sum_{i=0}^{n} \Big| \frac{\omega_n(x)}{(x - x_i)\omega_n'(x_i)} \Big|,$$

где $\omega_n(x) = (x - x_0) \ldots (x - x_n)$ [8].

13. Показать, что в обозначениях предыдущей задачи при $n = 1$ и $x_0 = a,\ x_1 = b$, выполняется равенство $\|\mathcal{P}_1\| = 1$. Если же $x_0 \neq a,\ x_1 \neq b$, то

$$\|\mathcal{P}_1\| \geq \frac{b - a}{2|x_1 - x_0|}$$

и, значит, может быть сколь угодно большой, если точки $x_0,\ x_1$ близки.

14. Пусть значения интерполируемой функции $f(x)$ заданы в узлах интерполяции с ошибкой δf_i. Показать, что если $L_n(x)$ и $\bar{L}_n(x)$ — многочлены Лагранжа, построенные по значениям f_i и $f_i + \delta f_i$ соответственно, то справедлива оценка

$$\|L_n(x) - \bar{L}_n(x)\| \leq \|\mathcal{P}_n\| \|\delta f\|.$$

[7] Напомним, что $C(a, b)$ — пространство, элементами которого являются непрерывные функции $\psi(x)$, определенные на $[a, b]$, и $\|\psi\| = \max\limits_{x\in[a.b]} |\psi(x)|$.

[8] В случае равноотстоящих узлов $x_i = a + ih,\ i = 0, \ldots, n,\ h = (b-a)/n$ и $n > 1$, можно показать, что

$$2^{n-3} \frac{1}{\sqrt{n-1}} \frac{1}{n - 1.5} < \|\mathcal{P}_n\| < 2^{n-1},$$

то есть $\|\mathcal{P}_n\|$ растет с ростом n.

15. Пусть заданы целые числа $k,\ l$. Функция

$$\varphi_{kl}(x) = \frac{a_k x^k + a_{k-1}x^{k-1} + \cdots + a_0}{x^l + b_{l-1}x^{l-1} + \cdots + b_0}$$

называется **рациональной функцией**. Если заданы значения функции $f(x)$ в узлах $x_i,\ i = 0, \ldots, k+l$, то нахождение функции $\varphi_{kl}(x)$, совпадающей с функцией $f(x)$ в узлах, называется **интерполированием рациональными функциями**. В частности, если $k = l = 1$, то получается так называемая **дробно-линейная интерполяция**. Построить дробно-линейную интерполирующую функцию, полагая, что $x_2 - x_1 = x_1 - x_0 = h$.

16. Каковы первые 5 многочленов Чебышева для отрезка $[0, 1]$?

17. Найти многочлен первой степени, который дает наилучшее равномерное приближение для функции $f(x) = x^3$ на отрезке $[0, 1]$. Найти максимальное значение погрешности.

Ответ $P_1(x) = x - \dfrac{1}{3\sqrt{3}}$.

18. Отыскать многочлен второй степени, который наилучшим образом приближает на отрезке $[0, 1]$ функцию $f(x) = x^3$. Определить максимальное значение погрешности.

Ответ $P_2(x) = \dfrac{48x^2 - 18x + 1}{32}$.

19. Показать, что два многочлена Чебышева $T_n(x)$ и $T_m(x),\ n \neq m$ ортогональны на отрезке $[-1, 1]$ с весом $\rho(x) = 1/\sqrt{1-x^2}$, то есть, что

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} 0, & m \neq n, \\ \pi, & m = n = 0, \\ \frac{\pi}{2}, & m = n > 0. \end{cases}$$

Указание. Сделать замену $x = \cos t$.

20. Пусть функция $f(x)$ раскладывается в ряд по многочленам Чебышева:

$$f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} a_j T_j(x). \tag{3.89}$$

Показать, что в этом случае коэффициенты вычисляются по формуле

$$a_j = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_j(x)}{\sqrt{1-x^2}}\,dx.$$

Указание. Воспользоваться результатом предыдущей задачи, умножив ряд (3.89) на $\dfrac{T_n(x)}{\sqrt{1-x^2}}$ и проинтегрировав по отрезку $[-1, 1]$.

21. Разложить функцию $f(x) = |x|$ на отрезке $[-1, 1]$ в ряд по многочленам Чебышева, оборвав этот ряд на втором члене, сравнить полученное приближение $S_2(x)$ с наилучшим равномерным приближением многочленом $P_2(x)$ порядка не выше 2.

Ответ $P_2(x) = x^2 + \frac{1}{8}$, $S_2(x) = \dfrac{2}{\pi} + \dfrac{4}{3\pi}(2x^2 - 1) = \dfrac{8}{3\pi}x^2 + \dfrac{2}{3\pi}$.

22. Методом обратной интерполяции найти приближенное значение корня уравнения $e^{-2x} - 3x = 0$, если известно, что $e \approx 2.72,\ e^{-1} \approx 0.37$.

В задачах 23-25 предполагается, что узлы $x_0, \ldots, x_n$ занумерованы в порядке возрастания и что $x_{i+1} - x_i = h,\ i = 0, \ldots, n-1$.

Определим конечную разность первого порядка $\Delta y_i = y_{i+1} - y_i$, второго порядка $\Delta^2 y_i = \Delta(\Delta y_i) = \Delta y_{i+1} - \Delta y_i = y_{i+2} - 2y_{i+1} + y_i$, k-го порядка $\Delta^k y_i = \Delta(\Delta^{k-1} y_i)$.

23. Доказать, что

$$\Delta^k y_i = \sum_{j=0}^{k} (-1)^j C_k^j y_{k+i-j}.$$

24. Доказать, что если $y_i = f(x_i)$, то

$$f(x_0, x_1, \ldots, x_k) = \frac{\Delta^k y_i}{k! h^k}.$$

25. Показать, что для равноотстоящих узлов многочлен Ньютона принимает вид

$$P_n(x) = P_n(x_0 + qh) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \cdots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0.$$

26. Имеется набор точек $(x_i, y_i),\ i = 0, \ldots, m$, представляющих экспериментальные данные:

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.66	1.58	1.50	1.44	1.37	1.30	1.22	1.17

Кроме того известно, что при неограниченном увеличении x величина y стремится к нулю. Этому условию удовлетворяют, например, такие простые формулы $y = \dfrac{1}{ax + b}$, $y = Ae^{-kx}$. Подобрать значения параметров, входящих в эти формулы. Пользуясь формулами, получить значения y при $x = 1.25,\ x = 3.75,\ x = -1,\ x = -2$.

27. Пусть задана таблица значений функции $y = f(x)$:

x	-4	-3	-1	0	3	4	5
y	9	7	11	26	56	29	-3

Является ли

$$S(x) = \begin{cases} 1 - 2x, & -4 \le x < -3, \\ 28 + 25x + 9x^2 + x^3, & -3 \le x < -1, \\ 26 + 19x + 3x^2 - x^3, & -1 \le x < 0, \\ 26 + 19x + 3x^2 - 2x^3, & 0 \le x < 3, \\ -163 + 208x - 60x^2 + 5x^3, & 3 \le x < 4, \\ 157 - 32x, & 4 \le x \le 5. \end{cases}$$

естественным сплайном для заданной функции.

3.4.3 Примеры тестовых вопросов к главе 3

1. Пусть интерполяционный многочлен Ньютона строится по точкам $x_i = 0.2i,\ i = 0, 1, \ldots, 5$, в которых известно значение функции $y = f(x)$. Для какой из приведенных ниже функций оценка погрешности интерполяции меньше?

а) $f(x) = \sin \pi x$;

б) $f(x) = \cos \pi x$;

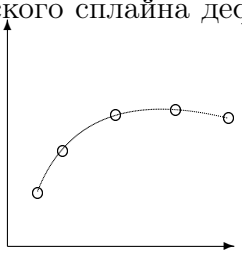
в) $f(x) = e^x$;

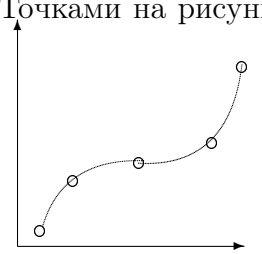
г) $f(x) = x^5$;

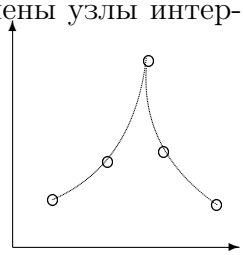
д) $f(x) = x^6$;

е) $f(x) = \ln(1 + x)$.

2. Какой из приведенных ниже графиков не может быть графиком интерполяционного кубического сплайна дефекта 1. Точками на рисунке отмечены узлы интерполяции.

1. 

2. 

3. 

3. Пусть $L_7(x)$ — многочлен Лагранжа построенный для функции заданной таблицей

x_i	1	2	3	4	5	6	7	8
$f(x_i)$	2	5	-1	0	2	2	4	3

, $P_7(x)$ — многочлен Ньютона для этой же функции. Чему равно значение дроби $\dfrac{L_7(3.34)}{P_7(3.34)}$?

4. Известна таблица значений функции $y = f(x)$

x_i	0.1	0.2	0.3
$f(x_i)$	0.4	-0.1	-0.4

.

В какой точке функция обращается в ноль? В ответе указать число с двумя десятичными знаками.

5. Чему равно значение интерполяционного многочлена

$$L_9(x) = \sum_{i=0}^{9} x_i^2 \frac{\prod\limits_{j=0, j\neq i}^{9} (x - x_j)}{\prod\limits_{j=0, j\neq i}^{9} (x_i - x_j)}$$

в точке $x^* = 1.5$?

6. Найти многочлен, интерполирующий функцию $y = f(x)$ такую, что $f(1) = 1,\ f'(1) = 0,\ f(2) = 2$. В ответе записать через пробел коэффициенты многочлена, начиная с коэффициента при старшей степени.

4 ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ И ИНТЕГРИРОВАНИЕ

4.1 ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

Задача численного дифференцирования состоит в приближенном вычислении производных функции $u(x)$ по заданным в конечном числе точек значениям этой функции. Численное дифференцирование применяется тогда, когда функцию $u(x)$ трудно или невозможно продифференцировать аналитически.

Одним из методов получения формул численного дифференцирования является использование формул интерполяции. Для этого достаточно заменить функцию $u(x)$ ее интерполяционным многочленом $L_n(x)$ и вычислить производные многочлена $L_n(x)$, учитывая его явное представление.

Например,

$$L_1(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} u(x_i) + \frac{x - x_i}{x_{i-1} - x_i} u(x_{i-1})$$

— интерполяционный многочлен, проходящий через точки $(x_{i-1}, u(x_{i-1}))$, $(x_i, u(x_i))$. Тогда

$$u'(x_i) \approx L_1'(x) = \frac{u(x_i) - u(x_{i-1})}{x_i - x_{i-1}}.$$

Порядок погрешности формулы численного дифференцирования зависит как от порядка интерполяционного многочлена, так и от расположения узлов. Погрешность определяется следующим образом: пусть $u(x)$ гладкая функция, определенная на отрезке $[a, b]$. Тогда по формуле Тейлора с остаточным членом в форме Лагранжа

$$u(x_{i+k}) = u(x_i) + (x_{i+k} - x_i)u'(x_i) + (x_{i+k} - x_i)^2 \frac{u''(\xi_i)}{2},\ k = 0, \pm 1, \ldots. \tag{4.1}$$

Здесь ξ_i лежит между x_i и x_{i+k}. Подставляя (4.1) в формулу для приближенного вычисления производной, имеем:

$$\frac{u(x_i) - u(x_{i-1})}{x_i - x_{i-1}} =$$

$$= \frac{u(x_i) - \left(u(x_i) - (x_i - x_{i-1})u'(x_i) + (x_i - x_{i-1})^2 \frac{u''(\xi_i)}{2}\right)}{x_i - x_{i-1}} =$$

$$= u'(x_i) - \frac{(x_i - x_{i-1})}{2} u''(\xi_i).$$

Таким образом, погрешность полученной формулы, то есть разность

$$\frac{u(x_i) - u(x_{i-1})}{x_i - x_{i-1}} - u'(x_i)$$

которая имеет порядок $O((x_i - x_{i-1}))$ или, более точно, погрешность не превосходит $M_2 h/2$, где $M_2 = \max\limits_{x\in[a,b]} |u''(x)|, \quad h = x_i - x_{i-1}$.

Для вычисления первой производной можно было построить другой многочлен:

$$L_1(x) = \frac{x - x_i}{x_{i+1} - x_i} u(x_{i+1}) + \frac{x - x_{i+1}}{x_i - x_{i+1}} u(x_i)$$

Тогда для первой производной получим другую формулу:

$$u'(x_i) \approx L_1'(x) = \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i}.$$

Легко проверить, что точность этой формулы такая же как и у предыдущей.

Для получения другой, более точной формулы для вычисления первой производной можно воспользоваться многочленом Лагранжа $L_2(x)$ второй степени. При этом, если $x_{i+1} - x_i = x_i - x_{i-1} = h$, то

$$L_2(x) = \frac{(x - x_i)(x - x_{i+1})}{2h^2} u(x_{i-1}) - \\ - \frac{(x - x_{i-1})(x - x_{i+1})}{h^2} u(x_i) + \frac{(x - x_{i-1})(x - x_i)}{2h^2} u(x_{i+1}). \quad (4.2)$$

Тогда

$$u'(x_i) \approx L_2'(x_i) = \frac{u(x_{i+1}) - u(x_{i-1})}{2h}. \quad (4.3)$$

Из формулы (4.1), взятой при $k = \pm 1$ имеем:

$$\frac{u(x_{i+1}) - u(x_{i-1})}{2h} - u'(x_i) = \\ = \frac{\big(u(x_i) + hu'(x_i) + h^2 u''(x_i)/2 + O(h^3)\big) - \big(u(x_i) - hu'(x_i) + h^2 u''(x_i)/2 + O(h^3)\big)}{2h} - \\ - u'(x_i) = O(h^2).$$

Таким образом, последняя формула, полученная для первой производной, точнее предыдущих.

Приближенное значение производной в точке x_0, то есть в начале таблицы, можно найти следующим образом. Возьмем в формуле (4.2) $i = 1$ и положим

$$u'(x_0) \approx L_2'(x_0) = \frac{-3u(x_0) + 4u(x_1) - u(x_2)}{2h}.$$

Для вывода формулы приближенного вычисления второй производной, очевидно, многочлен $L_1(x)$ не годится. Дважды дифференцируя $L_2(x)$ получим для равноотстоящих узлов

$$u''(x_i) \approx \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2}. \quad (4.4)$$

Используя формулу Тейлора, легко получить теперь, что для гладкой функции $u(x)$ выполняется равенство

$$\frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} = u''(x_i) + \frac{h^2}{12} u^{IV}(x_i) + O(h^4) = u''(x_i) + \frac{h^2}{12} u^{IV}(\xi_i), \quad (4.5)$$

где $\xi_i \in [x_{i-1}, x_{i+1}]$.

Имея формулы для вычисления производных первого порядка, легко получить формулы для нахождения производных более высокого порядка. Поясним это на примере получения формулы (4.4) с помощью формулы (4.3), записанной с шагом $h/2$. Имеем

$$u''(x_i) = (u')'\Big|_{x=x_i} \approx \frac{u'(x_i+h/2) - u'(x_i-h/2)}{h} \approx$$
$$\approx \frac{1}{h}\Big(\frac{u(x_{i+1})-u(x_i)}{h} - \frac{u(x_i)-u(x_{i-1})}{h}\Big) = \frac{u(x_{i-1})-2u(x_i)+u(x_{i+1})}{h^2}.$$

В том случае, когда значения функции заданы с некоторой ошибкой, погрешности, возникающие при вычислении производных, намного превосходят погрешности в задании значений функции и могут даже неограниченно расти при стремлении шага $h_i = x_i - x_{i-1}$ к нулю. Поэтому операцию приближенного вычисления производных называют **некорректной**. Поясним причину некорректности на примере вычисления первой производной. Как было показано, при достаточно малом h выражение $(u_i - u_{i-1})/h$ хорошо приближает $u'(x_i)$. Здесь $u_i = u(x_i),\ u_{i-1} = u(x_{i-1}),\ h = x_i - x_{i-1}$. Тот факт, что h находится в знаменателе, как раз и является причиной некорректности. Действительно, пусть вместо точных величин $u_i,\ u_{i-1}$ имеем $\tilde{u}_i = u_i + \delta_i, \tilde{u}_{i-1} = u_{i-1} + \delta_{i-1}$. Тогда вместо $(u_i - u_{i-1})/h$ будет вычисляться величина

$$\frac{\tilde{u}_i - \tilde{u}_{i-1}}{h} = \frac{u_i - u_{i-1}}{h} + \frac{\delta_i - \delta_{i-1}}{h}.$$

Значит, погрешность при нахождении производной, связанная с неточным заданием u_i (назовем ее погрешность округления и обозначим Δ_i) равна $\Delta_i = (\delta_i - \delta_{i-1})/h$. Если известна граница погрешности δ, то есть $|\delta_i| \le \delta$ при всех i, то $|\Delta_i| \le 2\delta/h$. При этом, если $\delta_i = -\delta_{i-1} = \delta$, то $|\Delta_i| = 2\delta/h$. Отсюда следует, что если δ не зависит от h, то при $h \to 0$ модуль погрешности округления $|\Delta|$ может стремиться к ∞.

Таким образом, при численном дифференцировании погрешность состоит из двух частей: погрешности формулы численного дифференцирования, которая в рассматриваемом примере не превосходит $M_2 h/2$, и погрешности округления.

Естественно выбрать шаг h таким образом, чтобы полная погрешность была минимальной. Для рассматриваемого примера полная погрешность равна $M_2 h/2 + 2\delta/h$. Для нахождения ее минимального значения приравняем нулю ее производную. В результате получим $M_2/2 - 2\delta/h^2 = 0$, откуда $h_{opt} = 2\sqrt{\delta/M_2}$. Заметим, что при этом значении шага $M_2 h/2 = \sqrt{M_2\delta},\ 2\delta/h = \sqrt{M_2\delta}$. Таким образом, в данном случае, при оптимальном значении шага погрешность формулы численного интегрирования равна погрешности, связанной с неточностью задания исходного значения функции.

При вычислении производных более высокого порядка в знаменателе стоит h^k, где $k > 1$. Поэтому влияние неточности задания табличных значений функции сказывается еще сильнее.

4.2 ПРОСТЕЙШИЕ КВАДРАТУРНЫЕ ФОРМУЛЫ

С задачей численного интегрирования приходится довольно часто встречаться на практике. Это связано, прежде всего, с тем, что во многих случаях формула Ньютона-Лейбница, с помощью которой вычисляли определенные интегралы в курсе математического анализа, неприменима. Дело в том, что не для всех элементарных

функций существуют первообразные, выражающиеся через элементарные функции. Например, невозможно выразить с помощью элементарных функций

$$\int e^{-x^2}\,dx, \quad \int \frac{\sin x}{x}\,dx, \quad \int \frac{\cos x}{x}\,dx.$$

Другой часто встречающийся случай — вычисление определенного интеграла от функции, заданной таблично.

Будем изучать методы вычисления интегралов

$$I = \int_a^b f(x)\,dx,$$

основанные на замене интегралов конечными суммами вида

$$I_n = \sum_{i=0}^{n} C_i f(x_i).$$

Здесь C_i - числовые коэффициенты, x_i -точки отрезка $[a, b]$. Приближенное равенство

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} C_i f(x_i) \tag{4.6}$$

называется **квадратурной формулой**, точка x_i — **узлами квадратурной формулы**, коэффициенты C_i — **коэффициентами квадратурной формулы**. Величина $\psi_n = I - I_n$ называется **погрешностью квадратурной формулы**. Она зависит как от выбора узлов x_i, так и от коэффициентов C_i. Если ввести на $[a, b]$ равномерную сетку[1] с шагом h:

$$\{x_i, \quad x_i = a + ih, i = 0, \ldots, N, \quad Nh = b - a\},$$

то

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x)\,dx = \sum_{i=1}^{N} I_i,$$

следовательно, достаточно построить формулу численного интегрирования на отрезке $[x_{i-1}, x_i]$ и затем просуммировать формулы, полученные для каждого отрезка.

При численном вычислении интегралов от функции $y = f(x)$ полагают, что самой трудоемкой операцией является нахождение значений функции в узлах квадратурной формулы. Поэтому, при сравнении формул численного интегрирования, считается лучшей та квадратурная формула, которая позволяет вычислить интеграл с заданной точностью при меньшем числе обращений к процедуре нахождения значений подынтегральной функции.

Рассмотрим сначала простейшие квадратурные формулы, основанные на простейших геометрических соображениях.

Формулы прямоугольников. Геометрически

$$I_i = \int_{x_{i-1}}^{x_i} f(x)\,dx$$

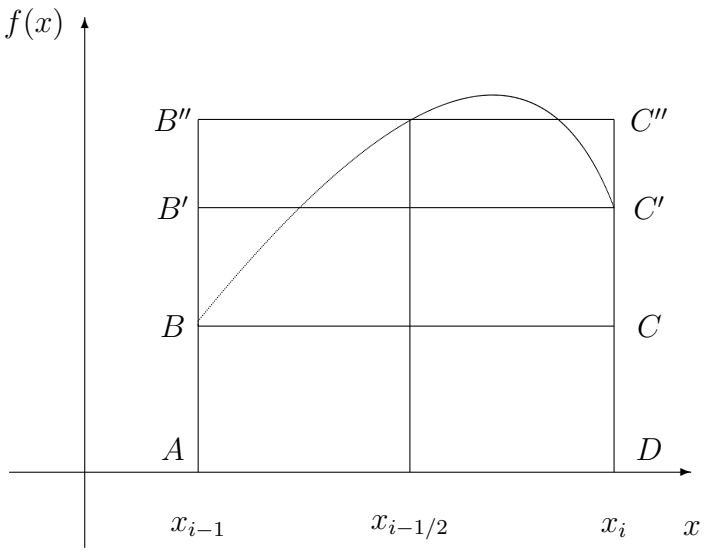


Рис. 4.1 Геометрический смысл формул прямоугольников

— площадь криволинейной трапеции $ABCD$ (см. рис.4.1). Поэтому ее можно приближенно заменить площадью прямоугольника. Обычно выбирается один из прямоугольников $ABCD$, $AB''C''D$, $AB'C'D$. При этом получаются формулы прямоугольников:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \approx \begin{cases} f(x_{i-1})h, \\ f(x_{i-1/2})h, \\ f(x_i)h, \end{cases} \tag{4.7}$$

где $x_{i-1/2} = a + (i - 1/2)h$.

Первая из этих формул носит название формулы **левых** , вторая — **средних**, а третья **правых прямоугольников**.

Погрешность полученных методов численного интегрирования легко оценить, используя формулу Тейлора. Например, для формулы средних прямоугольников имеем:

$$f(x) = f(x_{i-1/2}) + (x - x_{i-1/2})f'(x_{i-1/2}) + \frac{(x - x_{i-1/2})^2}{2} f''(\xi), \quad \xi \in (x_{i-1}, x_i).$$

Следовательно,

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = f(x_{i-1/2})h + \int_{x_{i-1}}^{x_i} \frac{(x_i - x_{i-1/2})^2}{2} f''(\xi)\,dx.$$

Пусть $M_{2,i} = \max\limits_{\xi\in[x_{i-1},x_i]} |f''(\xi)|$. Тогда для погрешности ψ_i имеем:

$$|\psi_i| = \left| \int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-1/2})h \right| \le M_{2,i} \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2}\,dx = \frac{h^3}{24} M_{2,i}.$$

[1]Результаты легко обобщаются на случай неравномерной сетки.

Аналогичные оценки позволяют заключить, что для случаев левых и правых прямоугольников абсолютная величина погрешности не превосходит $\frac{h^2}{2}M_{1,i}$, где

$$M_{1,i} = \max_{\xi\in[x_{i-1},x_i]} |f'(\xi)|.$$

Суммируя формулы (4.7), теперь можно легко получить **составные формулы левых, средних и правых прямоугольников**:

$$\int_a^b f(x)\,dx \approx \begin{cases} \sum\limits_{i=0}^{N-1} f(x_{i-1})h, \\ \sum\limits_{i=1}^{N} f(x_{i-1/2})h, \\ \sum\limits_{i=1}^{N} f(x_i)h. \end{cases} \tag{4.8}$$

Их погрешности равны сумме погрешностей по всем отрезкам. В результате, например, для формулы средних

$$|\psi| = \left|\sum_{i=1}^{N}\psi_i\right| \le \frac{M_2 N h^3}{24} = \frac{h^2(b-a)}{24}M_2 = O(h^2).$$

Здесь $M_2 = \max\limits_{\xi\in[a,b]} |f''(\xi)|$. В этом случае говорят, что формула имеет второй порядок точности. Легко заметить, что формулы левых и правых прямоугольников имеют первых порядок точности, то есть что их погрешность порядка $O(h)$.

Формула трапеций получается, если площадь криволинейной трапеции заменить площадью обычной трапеции:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \approx \frac{f(x_i)+f(x_{i-1})}{2}h.$$

Эту же формулу можно вывести путем замены на отрезке $[x_{i-1}, x_i]$ подынтегральной функции $f(x)$ многочленом Лагранжа первой степени

$$L_{1,i} = \frac{1}{h}[(x-x_{i-1})f(x_i) - (x-x_i)f(x_{i-1})].$$

Так как на основании оценки погрешности формулы Лагранжа

$$f(x) - L_{1,i}(x) = \frac{(x-x_{i-1})(x-x_i)}{2}f''(\xi_i)$$

имеем

$$|\psi_i| = \left|\int_{x_{i-1}}^{x_i} (f(x) - L_{1,i}(x))\,dx\right| = \left|\int_{x_{i-1}}^{x_i} \frac{(x-x_{i-1})(x-x_i)}{2}f''(\xi_i)\,dx\right| \le M_{2,i}\frac{h^3}{12}.$$

Составная формула трапеций имеет вид

$$\int_a^b f(x)\,dx \approx h\sum_{i=1}^{N}\frac{f(x_i)+f(x_{i-1})}{2} = h\Big(\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{N-1}) + \frac{f(x_N)}{2}\Big). \tag{4.9}$$

Ее погрешность

$$|\psi| \le \frac{h^2(b-a)}{12} M_2.$$

Заметим, что оценка погрешности формулы трапеций в два раза больше, чем у формулы средних. Поэтому, если значения функции одинаково легко определяются в любых точках, то лучше вести расчет по более точной формуле средних. Формулу трапеций употребляют в тех случаях, когда функция задана только в узлах сетки, а в серединах интервалов неизвестна.

Формула Симпсона. Для получения большей точности квадратурной формулы, функцию на отрезке $[x_{i-1}, x_i]$ надо заменить многочленом более высокой степени, например, второй:

$$L_{2,i}(x) = \frac{2}{h^2}\Bigg((x - x_{i-1/2})(x - x_i)f(x_{i-1}) - 2(x - x_{i-1})(x - x_i)f(x_{i-1/2}) + \\ + (x - x_{i-1})(x - x_{i-1/2})f(x_i)\Bigg).$$

В результате получается формула, которую называют формулой Симпсона или формулой парабол:

$$\int\limits_{x_{i-1}}^{x_i} f(x)\,dx \approx \int\limits_{x_{i-1}}^{x_i} L_{2,i}(x)\,dx = \frac{h}{6}\big(f(x_{i-1}) + 4f(x_{i-1/2}) + f(x_i)\big).$$

Чтобы не использовать дробные индексы ее можно переписать следующим образом:

$$\int\limits_{x_{i-1}}^{x_{i+1}} f(x)\,dx \approx \frac{h}{3}\big(f(x_{i-1}) + 4f(x_i) + f(x_{i+1})\big). \tag{4.10}$$

Пусть N — четное число. Тогда **составная формула Симпсона** имеет вид

$$\int\limits_a^b f(x)\,dx \approx \frac{h}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{N-2}) + 4f(x_{N-1}) + f(x_N)). \tag{4.11}$$

Оценка погрешности формулы Симпсона легко получается, если использовать разложение по формуле Тейлора

$$f(x) = f(x_i) + (x - x_i)f'(x_i) + \frac{(x - x_i)^2}{2}f''(x_i) + \frac{(x - x_i)^3}{3!}f'''(x_i) + \frac{(x - x_i)^4}{4!}f^{IV}(\xi_i). \tag{4.12}$$

Тогда, учитывая, что интеграл от нечетной функции в симметричных пределах равен

нулю, получим:

$$\int\limits_{x_{i-1}}^{x_{i+1}} f(x)\,dx =$$

$$= \int\limits_{x_{i-1}}^{x_{i+1}} \left(f(x_i) + (x-x_i)f'(x_i) + \frac{(x-x_i)^2}{2} f''(x_i) + \frac{(x-x_i)^3}{3!} f'''(x_i) + \frac{(x-x_i)^4}{4!} f^{IV}(\xi_i) \right) dx =$$

$$= 2hf(x_i) + \frac{h^3}{3} f''(x_i) + \int\limits_{x_{i-1}}^{x_{i+1}} \frac{(x-x_i)^4}{24} f^{IV}(\xi_i)\,dx$$

Используя формулу (4.12) при $x = x_{i-1}$ и при $x = x_{i+1}$, получим

$$\frac{h}{3}\big(f(x_{i+1}) + 4f(x_i) + f(x_{i-1})\big) = 2hf(x_i) + \frac{h^3}{3} f''(x_i) + \frac{h^5}{36} f^{IV}(\tilde{\xi}_i).$$

Тогда, для погрешности ψ_i имеем

$$\psi_i = \int\limits_{x_{i-1}}^{x_{i+1}} f(x)\,dx - \frac{h}{3}\big(f(x_{i+1}) + 4f(x_i) + f(x_{i-1})\big) =$$

$$= \int\limits_{x_{i-1}}^{x_{i+1}} \frac{(x-x_i)^4}{24} f^{IV}(\xi_i)\,dx - \frac{h^5}{36} f^{IV}(\tilde{\xi}_i) \approx \left(\frac{h^5}{60} - \frac{h^5}{36} \right) f^{IV}(x_i) = -\frac{h^5}{90} f^{IV}(x_i).$$

Здесь предполагалось, что $f^{IV}(\xi_i) \approx f^{IV}(\tilde{\xi}_i) \approx const$. При малом h и гладкой функции $f(x)$ такое предположение возможно.

После суммирования по всем парам соседних отрезков, получим для составной формулы Симпсона:

$$|\psi| \le M_4 \frac{h^5 N/2}{90} = M_4 \frac{h^4 (b-a)}{180},$$

где $M_4 = \max\limits_{x\in[a,b]} |f^{IV}(x)|$. Таким образом, составная формула Симпсона имеет 4-й порядок точности.

Заметим, что если $f(x)$ — многочлен третьей степени, то $f^{IV}(x) \equiv 0$. Поэтому из оценки погрешности следует, что формула Симпсона точна для любого многочлена третьей степени.

Квадратурные формулы средних прямоугольников и трапеций при определенных предположениях относительно подынтегральной функции позволяют получить двусторонние оценки для значений интеграла. Справедливо следующее утверждение.

Пусть $y = f(x)$ дважды непрерывно дифференцируемая функция, определенная на отрезке $[a, b]$ и всюду на этом отрезке $f''(x) \ge 0$ (соответственно $f''(x) \le 0$). Приближенные значения интеграла $I = \int\limits_a^b f(x)\,dx$ вычисленные по составной формуле трапеций обозначим через $I_{\text{тр}}$, а по составной формуле средних прямоугольников через $I_{\text{ср}}$. Тогда для интеграла I справедлива оценка $I_{\text{ср}} \le I \le I_{\text{тр}}$ (соответственно $I_{\text{тр}} \le I \le I_{\text{ср}}$).

Для доказательства этого утверждения предположим для определенности, что $f''(x) \ge 0$. Это означает, что функция вогнута. Кроме того, будем считать, что значения функции не отрицательны. Тогда требуемое утверждение получается из следующих соображений. Площадь криволинейной трапеции (см. рисунок 4.2) равная

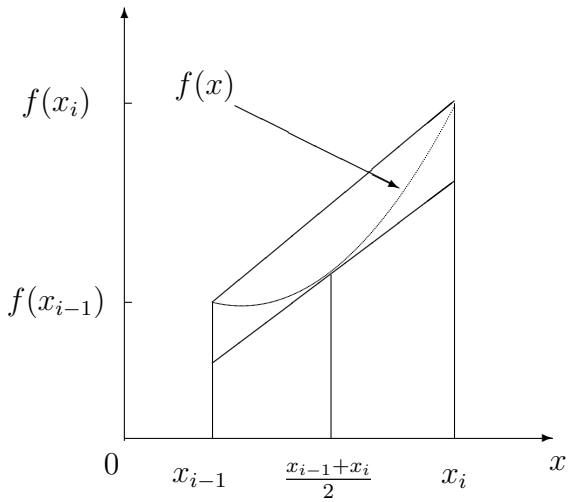


Рис. 4.2 Двусторонняя оценка интеграла

$\int\limits_{x_{i-1}}^{x_i} f(x)\,dx$ не превосходит площади трапеции, ограниченной осью Ox, вертикальными прямыми, проходящими через точки $(0, x_{i-1}), (0, x_i)$ и хордой, соединяющей точки $(x_{i-1}, f(x_{i-1})), (x_i, f(x_i))$. Её площадь равна $0,5(x_i - x_{i-1})(f(x_{i-1}) + f(x_i))$. В то же время площадь криволинейной трапеции не меньше площади трапеции, ограниченной осью Ox, вертикальными прямыми, проходящими через точки $(0, x_{i-1}), (0, x_i)$, и касательной к кривой $y = f(x)$, проведенной через точку $\left(\frac{x_{i-1}+x_i}{2}, f(\frac{x_{i-1}+x_i}{2})\right)$. Площадь этой трапеции равна $(x_i - x_{i-1}) f\left(\frac{x_{i-1}+x_i}{2}\right)$.

Случай, когда функция принимает отрицательные значения рассматривается аналогично.

Из двусторонней оценки следует, что если на отрезке $[a, b]$ вторая производная функции не меняет знак, то можно положить

$$\int\limits_a^b f(x)\,dx \approx \frac{I_{\text{ср}} + I_{\text{тр}}}{2}.$$

При этом очевидно, что

$$\Big|\int\limits_a^b f(x)\,dx - \frac{I_{\text{ср}} + I_{\text{тр}}}{2}\Big| \le \frac{|I_{\text{ср}} - I_{\text{тр}}|}{2}.$$

Полученное неравенство дает оценку погрешности при вычислении интеграла.

4.3 ОЦЕНКА ПОГРЕШНОСТИ КВАДРАТУРНОЙ ФОРМУЛЫ. АВТОМАТИЧЕСКИЙ ВЫБОР ШАГА ИНТЕГРИРОВАНИЯ

Величина погрешности, совершаемой при замене интеграла квадратурной формулой зависит от шага интегрирования и от значения производной функции. Величина

этой производной, как правило, заранее не известна. Рассмотрим как можно апостериорно (то есть после проведения расчета) оценить погрешность. Описываемый ниже метод получения оценки для погрешности называется **методом Рунге**.

Предположим, что необходимо вычислить интеграл $I = \int\limits_a^b f(x)\,dx$. Пусть отрезок $[a,b]$ разбит на частичные отрезки точками $a = x_0 < x_1 < \cdots < x_N = b$ и на каждом частичном отрезке применяется одна и та же квадратурная формула порядка точности $m-1$. Это означает, что

$$I_i - I_{h,\,i} \approx C_i h_i^m, \tag{4.13}$$

где I_i - точное значение интеграла на промежутке $[x_{i-1}, x_i]$, $I_{h,\,i}$ — приближенное, значение, вычисленное по квадратурной формуле, $h_i = x_i - x_{i-1}$. Величина C_i зависит от гладкости интегрируемой функции $f(x)$ и заранее не известна.[2]

Если отрезок $[x_{i-1}, x_i]$ разбить пополам и на каждом из полученных отрезков выполнить расчет тем же методом, то будет найдено, вообще говоря, другое приближенное значение интеграла, которое обозначим $I_{h/2,\,i}$. Согласно (4.13)

$$I_i - I_{h/2,\,i} \approx 2C_i\Big(\frac{h_i}{2}\Big)^m. \tag{4.14}$$

Вычитая (4.14) из (4.13), получим,

$$I_{h/2,\,i} - I_{h,\,i} \approx C_i h_i^m\Big(1 - \frac{1}{2^{m-1}}\Big) = \frac{C_i h_i^m}{2^{m-1}}(2^{m-1}-1) \approx (I_i - I_{h/2,\,i})(2^{m-1}-1)$$

или

$$I_i - I_{h/2,\,i} \approx \frac{I_{h/2,\,i} - I_{h,\,i}}{2^{m-1}-1}. \tag{4.15}$$

Таким образом, мы получаем возможность вычислять погрешность, что в свою очередь позволяет найти интеграл с заданной точностью $\varepsilon > 0$ путем автоматического подбора шага интегрирования h_i. Действительно, пусть $I \approx I_h = \sum\limits_{i=1}^{N} I_{h,\,i}$, причем на каждом частичном отрезке используется одна и та же квадратурная формула. Тогда, если провести на каждом отрезке вычисления дважды: один раз с шагом h_i, второй с $h_i/2$, оценить погрешность по правилу Рунге, и потребовать, чтобы

$$|I_i - I_{h/2,\,i}| \approx \frac{|I_{h/2,\,i} - I_{h,\,i}|}{2^{m-1}-1} \le \frac{\varepsilon h_i}{b-a}, \quad i = 1, \ldots, N. \tag{4.16}$$

то для составной формулы получим:

$$|I - I_h| \le \frac{\varepsilon}{b-a}\sum_{i=1}^{N} h_i = \varepsilon.$$

Это означает, что заданная точность будет достигнута. Если на каком-то отрезке оценка (4.16) не выполнена, то на этом отрезке шаг надо снова измельчить в два раза и снова оценить погрешность. Деление сетки следует проводить до тех пор, пока оценка не будет выполнена, либо пока не будет достигнуто оговоренное заранее число измельчений.

[2]Если разложить погрешность $I_i - I_{h,\,i}$ по степеням h_i, то кроме $C_i h_i^m$ она содержит в своем представлении слагаемые с более высокой степенью h_i. Однако, при малых значениях h_i этими слагаемыми можно пренебречь. Величина $C_i h_i^m$ называется **главным членом погрешности.**

Таким образом, получили возможность вычислять интеграл с крупным шагом там, где функция меняется плавно и с мелким в области быстрого изменения функции. Заметим также, что в соответствии с формулой (4.15) в качестве приближенного значения величины I_i можно взять

$$I_i \approx I_{h/2,\,i} + \frac{I_{h/2,\,i} - I_{h,\,i}}{2^{m-1} - 1}.$$

Идея Рунге допускает некоторое обобщение, называемое **экстраполяцией Ричардсона**. Пусть требуется вычислить некоторую величину Z. Для нахождения этой величины используется приближенная формула $\zeta(h)$, в которую входит некоторый малый параметр h. Например, Z — это значение интеграла, а $\zeta(h)$ — составная формула Симпсона (4.11) или же Z — это значение первой производной в фиксированной точке, $\zeta(h)$ — правая часть формулы (4.3). Предположим далее, что для погрешности приближения известно разложение

$$\zeta(h) - Z = c_1 h^{a_1} + c_2 h^{a_2} + \cdots + c_n h^{a_n} + O(h^{a_{n+1}}). \tag{4.17}$$

Здесь числа $c_i,\ a_i$ не зависят от h, коэффициенты c_i вообще говоря не известны, а значения показателей степени известны и $0 < a_1 < a_2 < \cdots < a_{n+1}$. Обычно разность $a_i - a_{i-1}$ не зависит от i и равна 1 или 2.

Вычислим приближения $\zeta(h)$ при различных значениях параметра h. Будем для определенности считать, что эти значения параметра меняются по закону геометрической прогрессии, то есть вычислим $\zeta(h_i)$, где $h_i = q^i h_0,\ i = 0, 1, \ldots, n$, а q некоторое фиксированное число из промежутка $(0, 1)$. Например, в методе Рунге, рассмотренном выше $q = 1/2$, что соответствовало делению шага пополам. В результате получим

$$\zeta(h_i) - Z = c_1 h_i^{a_1} + c_2 h_i^{a_2} + \cdots + c_n h_i^{a_n} + O(h_i^{a_{n+1}}). \tag{4.18}$$

Имеем в частности,

$$\begin{aligned}
\zeta(h_0) - Z &= c_1 h_0^{a_1} + c_2 h_0^{a_2} + \cdots + c_n h_0^{a_n} + O(h_0^{a_{n+1}}),\\
\zeta(h_1) - Z &= c_1 h_0^{a_1} q^{a_1} + c_2 h_0^{a_2} q^{a_2} + \cdots + c_n h_0^{a_n} q^{a_n} + O(h_0^{a_{n+1}}).
\end{aligned}$$

Исключим отсюда коэффициент c_1. Для этого умножим первое равенство на q^{a_1} и вычтем его из второго. В результате, после деления разности на $1 - q^{a_1}$ получим:

$$\frac{\zeta(h_1) - q^{a_1}\zeta(h_0)}{1 - q^{a_1}} - Z = \frac{(q^{a_2} - q^{a_1})}{1 - q^{a_1}} c_2 h_0^{a_2} + \cdots + \frac{(q^{a_n} - q^{a_1})}{1 - q^{a_1}} c_n h_0^{a_n} + O(h_0^{a_{n+1}}). \tag{4.19}$$

Если ввести обозначения $c_i^{(2)} = \dfrac{(q^{a_i} - q^{a_1})}{1 - q^{a_1}} c_i,\ i = 2, \ldots, n,$

$$\zeta^{(2)}(h_0) = \zeta(h_0) + \frac{\zeta(h_1) - \zeta(h_0)}{1 - q^{a_1}} = \frac{\zeta(h_1) - q^{a_1}\zeta(h_0)}{1 - q^{a_1}},$$

то равенство (4.19) перепишется в виде аналогичном (4.18) при $i = 0$:

$$\zeta^{(2)}(h_0) - Z = c_2^{(2)} h_0^{a_2} + \cdots + c_n^{(2)} h_0^{a_n} + O(h_0^{a_{n+1}}).$$

Подобные выкладки можно провести, записав (4.18) при двух соседних значениях i. В результате получим:

$$\zeta^{(2)}(h_i) - Z = c_2^{(2)} h_i^{a_2} + \cdots + c_n^{(2)} h_i^{a_n} + O(h_i^{a_{n+1}}),\ i = 0, \ldots, n-1, \tag{4.20}$$

где

$$\zeta^{(2)}(h_i) = \zeta^{(1)}(h_i) + \frac{\zeta^{(1)}(h_{i+1}) - \zeta^{(1)}(h_i)}{1 - q^{a_1}}. \tag{4.21}$$

Здесь для последующего единообразия введено обозначение $\zeta^{(1)}(h_i) = \zeta(h_i)$.

Соотношения (4.20) подобны равенствам (4.18). Главное отличие состоит в том, что разложение погрешности по степеням h начинается теперь не с h^{a_1}, а с h^{a_2}. Это означает, что величина $\zeta^{(2)}(h_i)$ точнее приближает Z. Процесс повышения точности можно продолжить, вычисляя $\zeta^{(k)}(h_i)$ с помощью рекуррентных соотношений:

$$\zeta^{(k+1)}(h_i) = \zeta^{(k)}(h_i) + \frac{\zeta^{(k)}(h_{i+1}) - \zeta^{(k)}(h_i)}{1 - q^{a_k}} \quad k = 1, \ldots, n, \quad i = 0, \ldots, n-k. \tag{4.22}$$

В результате $\zeta^{(k)}(h_i)$ будет совпадать с Z с точностью до величин $O(h_i^{a_k})$.

Метод Рунге применим в том случае, когда известен порядок точности квадратурной формулы. Если интегрируемая функция недостаточно гладкая, то реальный порядок точности может отличаться от теоретического и, вообще говоря, заранее не известен. В этом случае можно воспользоваться **алгоритмом Эйткина**. Предположим, что имеется три сетки с постоянными шагами $h_1,\ h_2$ и h_3, причем $h_1 = h,\ h_2 = qh,\ h_3 = q^2h$. Если I_{h_k} — приближенное значение интеграла на k-ой сетке, то ограничиваясь главным членом погрешности, получим

$$I \approx I_{h_k} + Ch_k^p,\ k = 1, 2, 3. \tag{4.23}$$

Здесь $I,\ C,\ p$ — неизвестные величины, которые можно найти из уравнений (4.23). Число p называют **эффективным порядком точности**. Введем обозначения $Ch^p = \phi,\ q^p = \alpha$. Тогда уравнения (4.23) перепишутся в виде

$$I - I_{h_1} \approx \phi,\quad I - I_{h_2} \approx \phi\alpha,\quad I - I_{h_3} \approx \phi\alpha^2. \tag{4.24}$$

Умножая второе из этих равенств на 2, вычитая из полученного первое и третье равенства, получим:

$$2I_{h_2} - I_{h_1} - I_{h_3} \approx \phi(\alpha^2 - 2\alpha + 1) = \phi(\alpha - 1)^2.$$

Вычитая из первого равенства второе и возведя полученный результат в квадрат, имеем:

$$(I_{h_2} - I_{h_1})^2 \approx \phi^2(\alpha - 1)^2.$$

Отсюда следует, что

$$\phi \approx \frac{(I_{h_2} - I_{h_1})^2}{2I_{h_2} - I_{h_1} - I_{h_3}}$$

и, значит, можно уточнить значение интеграла, если воспользоваться формулой:

$$I \approx I_{h_1} + \frac{(I_{h_2} - I_{h_1})^2}{2I_{h_2} - I_{h_1} - I_{h_3}}.$$

Для того, чтобы определить значение эффективного порядка точности заметим, что из (4.24) следует:

$$I_{h_2} - I_{h_1} \approx \phi(1 - \alpha),\quad I_{h_3} - I_{h_2} \approx \phi\alpha(1 - \alpha).$$

Тогда,

$$q^p = \alpha \approx \frac{I_{h_3} - I_{h_2}}{I_{h_2} - I_{h_1}}$$

и

$$p = \frac{\ln(I_{h_3} - I_{h_2}) - \ln(I_{h_2} - I_{h_1})}{\ln q}.$$

Как уже отмечалось выше, эффективный порядок точности, вообще говоря, не совпадает с теоретическим, если подынтегральная функция имеет особенности. Если же особенностей нет, то эффективный порядок точности может только слегка отличаться от теоретического благодаря наличию в погрешности не только главного, но и членов более высокого порядка малости. В этом случае при $h \to 0$ эффективный порядок стремится к теоретическому. Кроме того следует отметить, что эффективный порядок точности не обязательно целое число.

4.4 КВАДРАТУРНЫЕ ФОРМУЛЫ ИНТЕРПОЛЯЦИОННОГО ТИПА

Будем рассматривать формулы приближенного вычисления интегралов

$$I = \int_a^b \rho(x) f(x)\, dx,$$

где $\rho(x) > 0$ — весовая функция, $f(x)$ — достаточно гладкая функция. Формулы приближенного вычисления этих интегралов имеют вид

$$\int_a^b \rho(x) f(x)\,dx \approx \sum_{k=0}^{n} C_k f(x_k) = I_n. \tag{4.25}$$

Здесь $x_k \in [a,b]$, C_k — числа. Если формула (4.25) получается путем замены $f(x)$ интерполяционным многочленом сразу на всем отрезке $[a,b]$, то их называют **квадратурными формулами интерполяционного типа**.

Для определения коэффициентов C_k запишем интерполяционный многочлен Лагранжа в виде

$$L_n(x) = \sum_{k=0}^{n} \frac{\omega(x)}{(x-x_k)\omega'(x_k)} f(x_k),$$

где

$$\omega(x) = \prod_{j=0}^{n}(x-x_j), \quad \omega'(x_k) = \prod_{\substack{j=0\\ j\neq k}}^{n}(x_k - x_j),$$

и подставим его в интеграл вместо функции $f(x)$. Тогда,

$$\int_a^b \rho(x) f(x)\,dx \approx \sum_{k=0}^{n} \int_a^b \frac{\rho(x)\omega(x)}{(x-x_k)\omega'(x_k)} f(x_k)\,dx.$$

Из этого соотношения следует, что формула (4.25) является квадратурной формулой интерполяционного типа тогда и только тогда, когда ее коэффициенты вычисляются по формуле

$$C_k = \int_a^b \frac{\rho(x)\omega(x)}{(x-x_k)\omega'(x_k)}\,dx, \ k = 0, 1, \ldots, n. \tag{4.26}$$

Для оценки погрешности φ_n квадратурной формулы интерполяционного типа запишем $f(x)$ в виде $f(x) = L_n(x) + r_n(x)$, где $r_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega(x)$ — погрешность интерполяции. Тогда

$$\varphi_n = \int_a^b \rho(x) f(x)\,dx - \int_a^b \rho(x) L_n(x)\,dx =$$

$$= \int_a^b \rho(x) r_n(x)\,dx = \frac{1}{(n+1)!}\int_a^b \rho(x) f^{(n+1)}(\xi(x))\omega(x)\,dx \tag{4.27}$$

и

$$|\varphi_n| \le \frac{M_{n+1}}{(n+1)!}\int_a^b \rho(x)|\,\omega(x)|\,dx, \quad M_{n+1} = \max_{x\in[a,b]} |f^{(n+1)}(x)|. \tag{4.28}$$

Так как производная $(n+1)$-го порядка от любого многочлена n степени равна нулю, из оценки погрешности (4.28) следует, что квадратурная формула интерполяционного типа, построенная по узлам $x_0, \ldots, x_n$ из $[a,b]$ является точной для любого многочлена степени n. Это означает, что если $f(x)$ многочлен степени n и C_k вычислены по формуле (4.26), то в (4.25) имеет место точное равенство. Говорят тогда,

что число n является **алгебраической степенью** или **алгебраическим порядком точности** квадратурной формулы.

Можно доказать и обратное утверждение:

Теорема 4.4.1 *Если квадратурная формула*

$$\int_a^b \rho(x)f(x)\,dx \approx \sum_{i=0}^{n} d_i f(x_i)$$

точна для любого многочлена степени n, то она является квадратурной формулой интерполяционного типа.

Доказательство. Многочлены $\psi_k(x) = \dfrac{\omega(x)}{(x-x_k)\omega'(x_k)}$ имеют степень n. Тогда по условию теоремы

$$\int_a^b \rho(x)\psi_k(x)\,dx = \sum_{i=0}^{n} d_i\psi_k(x_i).$$

По формуле (4.26)

$$C_k = \int_a^b \rho(x)\psi_k(x)\,dx.$$

Поэтому

$$C_k = \sum_{i=0}^{n} d_i\psi_k(x_i) = d_k,$$

так как

$$\psi_k(x_i) = \begin{cases} 0, & k \neq i, \\ 1, & k = i, \end{cases}$$

что и требовалось доказать.

Оценка (4.28) в ряде случаев является грубой, так как не учитывает специфики формул. Так, например, формула Симпсона

$$\int_{-1}^{1} f(x)\,dx \approx \frac{1}{3}(f(-1) + 4f(0) + f(1))$$

точна не только для многочлена второй степени, но, как легко убедиться на основании полученной ранее оценки погрешности этой формулы, и для многочлена третьей степени. Можно доказать следующее утверждение: пусть n-четное число, узлы расположены симметрично относительно середины отрезка $[a, b]$ и $\rho(x)$ — четная функция относительно точки $0.5(a + b)$ [3]. Тогда, если квадратурная формула интерполяционного типа точна для любого многочлена степени n, то она точна и для любого многочлена степени $n + 1$.

Если квадратурные формулы интерполяционного типа построены на равномерной сетке, то их называют формулами **Ньютона-Котеса**.

Формулы Ньютона-Котеса с $n \geq 10$, как правило, не используются из-за их численной неустойчивости, приводящей к резкому росту вычислительной погрешности.

[3] $\rho(x)$ является четной относительно точки $0.5(a + b)$, если $\rho(0.5(a + b) + x) = \rho(0.5(a + b) - x)$.

Дело в том, что если коэффициенты C_k положительны, то при вычислении значений функции $f(x_k)$ с ошибкой δ_k имеем

$$\tilde{I}_n = \sum_{k=0}^{n} C_k(f(x_k) + \delta_k) = I_n + \delta I_n,$$

где

$$\delta I_n = \sum_{k=0}^{n} C_k \delta_k.$$

Так как квадратурная формула точна для $f(x) = 1$, получаем

$$\sum_{k=0}^{n} C_k = \int_a^b \rho(x)\,dx = M,$$

причем M не зависит от n. Таким образом, учитывая, что $C_k \geq 0$, имеем

$$|\delta I_n| \leq \sum_{k=0}^{n} |C_k||\delta_k| = \sum_{k=0}^{n} C_k|\delta_k| \leq \max_{k=0,\ldots,n} |\delta_k| \sum_{k=0}^{n} C_k = M \max_{k=0,\ldots,n} |\delta_k|.$$

Значит, при любых n погрешность интеграла имеет тот же порядок, что и погрешность, полученная при вычислении функции. Если же C_k имеют разные знаки, то может оказаться, что величина $\sum\limits_{k=0}^{n} |C_k|$ растет с ростом n, что является причиной неустойчивости. При $\rho(x) \equiv 1$ и $n \geq 10$ среди чисел C_k есть как положительные, так и отрицательные.

Если надо вычислить более точно интеграл, пользуются **составными формулами**, которые получаются при разбиении отрезка $[a, b]$ на части с последующим применением на каждой части формул невысокого порядка. Другой способ — специальным образом подобрать расположение узлов, чтобы получить более точные формулы.

4.5 КВАДРАТУРНЫЕ ФОРМУЛЫ ГАУССА

Во всех предыдущих параграфах узлы квадратурной формулы задавались заранее и формулы определялись только набором своих коэффициентов. В этом параграфе будут подбираться как коэффициенты, так и узлы квадратурной формулы.

Рассмотрим задачу: подобрать узлы и коэффициенты квадратурной формулы

$$\int_a^b \rho(x) f(x)\,dx \approx \sum_{k=1}^{n} C_k f(x_k) \tag{4.29}$$

так, чтобы при заданном n ее алгебраическая степень точности m была максимальной. Это означает, что для произвольного многочлена степени не выше m формула точна и существует многочлен степени $m + 1$, для которого формула не точна. Полученные при таком подходе формулы называют **формулами Гаусса**.

Сформулированное при постановке задачи требование эквивалентно тому, что формула (4.29) точна для любой функции $f(x) = x^i,\ i = 0, 1, \ldots, m$, то есть

$$\sum_{k=1}^{n} C_k x_k^i = \int_a^b \rho(x) x^i\,dx, \quad i = 0, 1, \ldots, m. \tag{4.30}$$

Равенства (4.30) можно рассматривать как систему нелинейных уравнений для определения $2n$ неизвестных $C_k,\ x_k,\ k = 1, \ldots, n$. Таким образом, чтобы число уравнений совпадало с числом неизвестных, надо, чтобы $m = 2n - 1$.

Например, если $n = 1, \rho \equiv 1$, то система (4.30) принимает вид

$$C_1 = \int_a^b 1\,dx = b - a, \quad C_1 x_1 = \int_a^b x\,dx = \frac{b^2 - a^2}{2}.$$

Значит, $x_1 = (a + b)/2$, то есть формула (4.29) имеет вид:

$$\int_a^b f(x)\,dx \approx (b - a) f((a + b)/2).$$

Полученная формула является формулой средних прямоугольников.

Сформулируем критерий, позволяющий судить когда квадратурная формула (4.29) будет точной для любого многочлена степени не выше $2n - 1$. Пусть

$$\omega(x) = (x - x_1)(x - x_2) \ldots (x - x_n).$$

Теорема 4.5.1 *Квадратурная формула (4.29) точна для любого многочлена степени $m = 2n - 1$ тогда и только тогда, когда*

- *многочлен $\omega(x)$ ортогонален с весом $\rho(x)$ любому многочлену $q(x)$ степени меньше n, то есть*
$$\int_a^b \rho(x)\omega(x)q(x)\,dx = 0; \tag{4.31}$$
- *формула (4.29) является формулой интерполяционного типа. Это означает, что*
$$C_k = \int_a^b \rho(x)\frac{\omega(x)}{(x - x_k)\omega'(x_k)}\,dx, \quad k = 1, 2 \ldots, n. \tag{4.32}$$

Необходимость. Пусть формула (4.29) точна для многочлена степени не выше $m = 2n - 1$. Следовательно, она точна для многочлена $\omega(x)q(x)$, степень которого не превышает $2n - 1$. Поэтому выполняется равенство

$$\int_a^b \rho(x)\omega(x)q(x)\,dx = \sum_{k=1}^{n} C_k \omega(x_k)q(x_k) = 0,$$

так как $\omega(x_k) = 0$. Формула (4.32) следует из теоремы 4.4.1, согласно которой если квадратурная формула точна для любого многочлена степени $n - 1$, то она является квадратурной формулой интерполяционного типа.

Достаточность. Пусть выполнены условия теоремы и $f(x)$ - многочлен степени не выше $2n - 1$. Тогда его можно представить в виде

$$f(x) = \omega(x)q(x) + r(x),$$

где $q(x), r(x)$ многочлены степени не выше $n-1$ являющиеся соответственно частным и остатком от деления $f(x)$ на $\omega(x)$. Отсюда следует, что

$$\int_a^b \rho(x)f(x)\,dx = \int_a^b \rho(x)\omega(x)q(x)\,dx + \int_a^b \rho(x)r(x)\,dx = \int_a^b \rho(x)r(x)\,dx.$$

Так как по условию квадратурная формула имеет интерполяционный тип, она точна для любого многочлена степени не выше $n-1$, поэтому

$$\int_a^b \rho(x)r(x)\,dx = \sum_{k=1}^n C_k r(x_k) = \sum_{k=1}^n C_k(f(x_k) - \omega(x_k)q(x_k)) = \sum_{k=1}^n C_k f(x_k),$$

что и требовалось доказать.

Так как условие (4.31) эквивалентно выполнению равенств

$$\int_a^b \rho(x)\omega(x)x^j\,dx = 0, \quad j = 0, \ldots, n-1, \tag{4.33}$$

получаем систему n уравнений относительно n неизвестных x_k, которые входят в определение $\omega(x)$. Отсюда находят x_k и затем, по формулам (4.32) определяют C_k.

Можно показать (однако мы этого делать не будем), что система (4.33) имеет единственное решение, причем все x_k лежат на отрезке $[a, b]$.

Существует многочлен степени $2n$, для которого формула Гаусса не является точной. Этим многочленом будет, например, $\omega^2(x)$. Действительно, если бы формула Гаусса была точной для этого многочлена, имели бы

$$0 < \int_a^b \rho(x)\omega(x)^2\,dx = \sum_{k=1}^n C_k\omega^2(x_k) = 0,$$

что невозможно.

Следует отметить, что коэффициенты C_k всегда положительны. В соответствии с доказанным в предыдущем параграфе это означает, что формулы Гаусса численно устойчивы при любом n. Следовательно, ими можно пользоваться при большом числе узлов. Для доказательства утверждения возьмем многочлен степени $2n-2$

$$\psi_k(x) = \left(\frac{\omega(x)}{(x-x_k)\omega'(x_k)}\right)^2.$$

Он обладает тем свойством, что

$$\psi_k(x_i) = \begin{cases} 0, & k \neq i, \\ 1, & k = i. \end{cases}$$

Так как для этого многочлена формула Гаусса точна, имеем

$$0 < \int_a^b \rho(x)\psi_k(x)\,dx = \sum_{i=1}^n C_i\psi_k(x_i) = C_k \quad k = 1, \ldots, n.$$

Приведем без доказательства представление для погрешности φ_n формулы Гаусса

$$\varphi_n = \frac{1}{(2n)!} \int_a^b \rho(x)\omega^2(x) f^{(2n)}(\zeta)\, dx, \quad \zeta \in [a,b].$$

Для примера рассмотрим случай $\rho(x) \equiv 1$. Линейным преобразованием

$$\xi = \frac{x-a}{b-a} + \frac{x-b}{b-a} \tag{4.34}$$

отрезок $[a,b]$ отображается в отрезок $[-1,1]$. Поэтому, заменив переменную x на переменную ξ по формуле (4.34), получим интеграл на отрезке $[-1,1]$. Следовательно, для того, чтобы посчитать интеграл на промежутке $[a,b]$, достаточно научиться вычислять интеграл на промежутке $[-1,1]$. Известно (см. [20]) что на этом отрезке многочлен Лежандра n-ой степени

$$\mathcal{L}_n(\xi) = \frac{1}{2^n n!} \frac{d^n(\xi^2-1)^n}{d\xi^n}$$

ортогонален любому многочлену степени ниже n. Это означает, что если многочлен Лежандра имеет ровно n действительных, различных корней и все они лежат на промежутке $(-1,1)$, то он лишь множителем отличаются от многочлена $\omega(\xi)$ и корни многочлена Лежандра являются узлами квадратурной формулы Гаусса.

Теорема 4.5.2 *Многочлен Лежандра степени n имеет ровно n действительных различных корней, причем все они расположены на промежутке $(-1,1)$.*

Доказательство. Предположим противное. Пусть на промежутке $(-1,1)$ у многочлена Лежандра находится только m различных действительных корней нечетной кратности $\chi_1, \chi_2, \ldots, \chi_m$, причем $m < n$. Возьмем многочлен $P_m(\xi) = \prod_{i=1}^{m}(\xi - \chi_i)$ при $m > 0$ и $P_0 = 1$ при $m = 0$. Тогда, $P_m(\xi) \cdot \mathcal{L}_n(\xi)$ — многочлен, все корни которого, расположенные на промежутке $(-1,1)$, имеют четную кратность. Значит, этот многочлен тождественно не равен нулю и не меняет на $(-1,1)$ знак, откуда следует, что $\int_{-1}^{1} P_m(\xi)\mathcal{L}_n(\xi)\, d\xi \neq 0$. А это противоречит тому, что многочлен Лежандра $\mathcal{L}_n(\xi)$ ортогонален на $[-1,1]$ любому многочлену меньшей степени.

Если обозначить через ξ_k корни многочлена Лежандра $\mathcal{L}_n(\xi)$, а соответствующие коэффициенты квадратурной формулы через γ_k, [4] то обратное к (4.34) преобразование позволяет вычислить узлы и коэффициенты квадратурной формулы для приближенного нахождения значения исходного интеграла на отрезке $[a,b]$:

$$x_k = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\xi_k, \quad C_k = \frac{1}{2}(b-a)\gamma_k.$$

Существуют таблицы значений узлов (корней многочлена Лежандра) для формулы Гаусса при различных значениях n.

Формулы Гаусса позволяют вычислять и несобственные интегралы. Например, при нахождении интеграла на полупрямой $0 < x < \infty$ и весовой функции $\rho(x) = e^{-x}$, ортогональными будут многочлены Лагерра

$$Z_n(x) = e^x \frac{d^n(x^n e^{-x})}{dx^n},$$

[4] Следует отметить, что значения ξ_k и γ_k зависят от числа узлов n.

корни которых являются узлами квадратурной формулы Гаусса для этого случая. При интегрировании на всей числовой прямой и весе $\rho(x) = e^{-x^2}$, ортогональными будут многочлены Эрмита

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n(e^{-x^2})}{dx^n},$$

следовательно, его корни выбираются узлами квадратурной формулы Гаусса.

4.6 НЕСТАНДАРТНЫЕ ФОРМУЛЫ ИНТЕГРИРОВАНИЯ

Оценки погрешности квадратурных формул предполагали определенную гладкость подынтегральных функций. В этом параграфе будут рассмотрены методы численного интегрирования функций, которые имеют те или иные особенности, а также методы интегрирования несобственных интегралов.

4.6.1 Разрывные функции

Пусть функция и ее производные кусочно-непрерывны, в точках разрыва существуют односторонние производные всех требуемых порядков. Разобьем отрезок интегрирования на части так, чтобы на каждой части функция и ее производные были непрерывны. Представим интеграл в виде суммы интегралов по отрезкам непрерывности. К каждой части затем применяется квадратурная формула. Если порядок точности формулы p, то при одновременном и одинаковом сгущении сетки на каждом отрезке непрерывности, погрешность будет иметь порядок $O(h^p)$.

Если же применять квадратурную формулу к разрывной или не гладкой функции, не выделяя особенности, то при сгущении сетки сходимость хотя и будет, но с невысокой скоростью и без четко выраженного порядка точности.

4.6.2 Интегрирование быстро осциллирующих функций

Часто встречаются интегралы от функций, описывающих высокочастотные колебания, Это быстро изменяющиеся функции, производные которых велики. Примером таких функции является функция вида $f(x) = g(x)e^{i\omega x}$, определенная на отрезке $[a, b]$, где ω - известная частота, причем $(b - a) = O(1)$, $\omega \gg 1$, $i^2 = -1$, а амплитуда колебания $g(x)$ — мало меняющаяся на периоде основного колебания гладкая функция. k-я производная функции $f(x)$ пропорциональна ω^k и, следовательно, велика. Поэтому при интегрировании приходится выбирать столь малый шаг h, чтобы выполнялось условие $h\omega \ll 1$. Это означает, что одна осцилляция содержит много узлов интегрирования, что приводит к большому объему вычислений.

Приближая $g(x)$ несложными полиномиальными интерполяциями, получим квадратурные **формулы Филона**. Выведем одну из них.

Пусть $a = x_0 < x_1 < \cdots < x_n = b$ и $x_{k+1} - x_k = h_k$. Используя, например, формулу

средних, заменим $g(x)$ на интервале (x_k, x_{k+1}) на $g_{k+1/2} = g(x_{k+1/2})$. Тогда

$$\int\limits_{x_k}^{x_{k+1}} f(x)\,dx \approx g_{k+1/2} \int\limits_{x_k}^{x_{k+1}} e^{i\omega x}\,dx = f_{k+1/2} \int\limits_{x_k}^{x_{k+1}} e^{i\omega(x-x_{k+1/2})}\,dx =$$

$$= \frac{1}{i\omega} f_{k+1/2}(e^{i\omega h_k/2} - e^{-i\omega h_k/2}) = \frac{2}{\omega} f_{k+1/2} \sin\Big(\frac{\omega}{2} h_k\Big).$$

Тогда для составной формулы имеем:

$$\int\limits_a^b f(x)\,dx \approx \frac{2}{\omega} \sum_{k=0}^{n-1} f_{k+1/2} \sin\Big(\frac{\omega}{2} h_k\Big). \tag{4.35}$$

Если шаг интегрирования настолько мал, что $h_k\omega \ll 1$, то $\sin\Big(\frac{\omega}{2}h_k\Big) \approx \frac{\omega}{2}h_k$ и формула (4.35) превращается в обычную составную формулу средних прямоугольников.

Для оценки погрешности φ формулы (4.35) заметим, что

$$g(x) - g_{k+1/2} \approx (x - x_{k+1/2}) g'(x_{k+1/2}).$$

Поэтому

$$\varphi = \int\limits_a^b f(x)\,dx - \frac{2}{\omega} \sum_{k=0}^{n-1} f_{k+1/2} \sin\Big(\frac{\omega}{2} h_k\Big) =$$

$$= \sum_{k=0}^{n-1} \int\limits_{x_k}^{x_{k+1}} (g(x) - g_{k+1/2}) e^{i\omega x}\,dx \approx \sum_{k=0}^{n-1} \int\limits_{x_k}^{x_{k+1}} (x - x_{k+1/2}) g'_{k+1/2} e^{i\omega x}\,dx =$$

$$= \frac{2i}{\omega^2} \sum_{k=0}^{n-1} g'_{k+1/2} e^{i\omega x_{k+1/2}} \Big(\sin\frac{\omega h_k}{2} - \frac{\omega h_k}{2}\cos\frac{\omega h_k}{2}\Big).$$

Следовательно, если $\omega h_k \ll \omega$, то $|\varphi| = O(n\omega^{-2} M_1)$, где M_1 — максимальное по модулю значение функции $g'(x)$ на отрезке $[a, b]$. Значит, погрешность будет мала, если M_1 не велико.

4.6.3 Несобственные интегралы

Несобственные интегралы делятся на интегралы по бесконечным или полубесконечным промежуткам и на интегралы в конечных промежутках от бесконечных функций. Ни в том, ни в другом случае не существует методов, обладающих той надежностью и универсальностью, которая присуща большинству методов интегрирования гладких функций на ограниченном промежутке. Выбор наилучшего метода интегрирования существенно зависит от индивидуальных особенностей задачи.

Первый приём вычисления несобственного интеграла с бесконечными пределами заключается в том, чтобы заменой переменной свести его к интегрированию в конечных пределах. Например, для интеграла

$$\int\limits_a^\infty f(x)\,dx, \quad a > 0,$$

замена $x = a/(1-t)$ приводит к интегралу

$$\int\limits_0^1 f\Big(\frac{a}{1-t}\Big)\frac{a}{(1-t)^2}\,dt.$$

Если полученная подынтегральная функция вместе с некоторым числом производных ограничена, то для вычисления этого интеграла применяют стандартные численные методы.

Замену переменной следует производить весьма тщательно, иначе интеграл будет сведён к интегралу по конечному промежутку, но не станет проще. Например, замена $x = -\ln t$ примененная к интегралу

$$I = \int\limits_0^\infty e^{-x}\cos^2(x^2)\,dx \tag{4.36}$$

дает

$$I = \int\limits_0^1 \cos^2(\ln^2 t)\,dt.$$

Здесь подынтегральная функция совершает бесконечно много колебаний на любом отрезке, содержащем точку ноль. Если же использовать замену $x = -2\ln t$, то придем к подынтегральной функции $2t\cos^2(4\ln^2 t)$. Она тоже осциллирует, но ограничена функцией $y = 2t$ и ее предел при $t \to 0$ равен нулю. Поэтому можно пренебречь интегралом по отрезку $[0, \varepsilon]$, а оставшийся интеграл вычислить не сложно.

Второй прием носит название **метода усечения**. В нем бесконечный конец области интегрирования заменяют конечным, после чего интеграл по конечному отрезку вычисляют с помощью какого-нибудь стандартного метода. Часто удается доказать, что вклад, который вносит в значение интеграла отброшенная часть пренебрежимо мал, но еще чаще интегрирование производят по конечным отрезкам разной длины и, сравнивая результаты, убеждаются в том, что приближенное значение интеграла не меняется. Такой способ в ряд случаев может дать неверный результат, но часто, будучи подкреплен должным анализом результатов, он себя оправдывает.

Поясним сказанное на примерах. Рассмотрим снова интеграл (4.36). Обозначая подынтегральную функцию через $f(x)$, запишем

$$I - \int\limits_0^A f(x)\,dx = \int\limits_A^\infty f(x)\,dx < \int\limits_A^\infty e^{-x}\,dx = e^{-A}.$$

Следовательно, если отбросить $\int\limits_A^\infty f(x)\,dx$, то допущенная при этом ошибка не превзойдет e^{-A}.

В том случае, когда сделать подобную оценку затруднительно, можно было бы выбрать достаточно большое число b и вычислять

$$\int\limits_a^\infty f(x)\,dx \approx \int\limits_a^b f(x)\,dx + \int\limits_b^{2b} f(x)\,dx + \int\limits_{2b}^{4b} f(x)\,dx + \ldots,$$

причем слагаемые в правой части добавлять до тех пор, пока очередное слагаемое не станет по модулю меньше заданной точности.

Третий прием заключается в том, что подынтегральная функция представляется в виде $f(x) = \rho(x)g(x)$, где $\rho(x) > 0$ — весовая функция, после чего применяется формула Гаусса. Этот прием годится как в случае интеграла с одним или двумя бесконечными пределами, так и в случае несобственного интеграла в конечных пределах.

Например, для вычисления интеграла

$$I = \int_0^\infty e^{-x} f(x)\,dx$$

воспользуемся формулой Гаусса. Если выбрать два узла $x_1,\ x_2$ и два веса $C_1,\ C_2$, то узлы должны быть корнями многочлена Лагерра второй степени, то есть удовлетворять уравнению

$$e^x(x^2e^{-x})'' = x^2 - 4x + 2 = 0.$$

Отсюда имеем $x_1 = 2 - \sqrt{2},\ x_2 = 2 + \sqrt{2}$. Согласно формуле (4.32)

$$C_1 = \int_0^\infty e^{-x}\frac{x - x_2}{x_1 - x_2}\,dx = \frac{2+\sqrt{2}}{4},\ C_2 = \int_0^\infty e^{-x}\frac{x - x_1}{x_2 - x_1}\,dx = \frac{2-\sqrt{2}}{4}.$$

Поэтому

$$\int_0^\infty e^{-x} f(x)\,dx \approx \frac{2+\sqrt{2}}{4}f(2-\sqrt{2}) + \frac{2-\sqrt{2}}{4}f(2+\sqrt{2}).$$

При интегрировании неограниченной функция в конечных пределах, иногда возможно представить ее в виде $f(x) = g(x) + \varphi(x)$, где $g(x)$ — гладкая ограниченная функция, а $\varphi(x)$ функция, которая имеет особенность, но при этом она интегрируется аналитически. Например, пусть $f(x) = 1/\sqrt{x(1+x^2)}$. В окрестности нуля эта функция ведет себя как $1/\sqrt{x}$. Поэтому положим $\varphi(x) = 1/\sqrt{x}$, тогда

$$g(x) = \frac{1}{\sqrt{x(1+x^2)}} - \frac{1}{\sqrt{x}} = \frac{1}{\sqrt{x}}\Big(\frac{1}{\sqrt{1+x^2}} - 1\Big) = \frac{1}{\sqrt{x}}\,\frac{1-\sqrt{1+x^2}}{\sqrt{1+x^2}} =$$
$$= \frac{1}{\sqrt{x}}\,\frac{-x^2}{\sqrt{1+x^2}(1+\sqrt{1+x^2})} = -x^{3/2}\frac{1}{\sqrt{1+x^2}(1+\sqrt{1+x^2})}.$$

Полученную функцию $g(x)$ легко проинтегрировать каким-нибудь стандартным методом.

В некоторых случаях удается построить нестандартные формулы, которые учитывают характер особенности. Например, для вычисления интеграла

$$\int_{-1}^1 \frac{e^x}{\sqrt{1-x^2}}\,dx$$

на интервале сетки (x_{i-1}, x_i), можно приблизить подынтегральную функцию выражением $e^{x_{i-1/2}}/\sqrt{1-x^2}$, так как числитель меняется слабо. Тогда оставшаяся функция легко интегрируется. В результате получаем квадратурную формулу

$$\int_{-1}^1 \frac{e^x}{\sqrt{1-x^2}}\,dx \approx \sum_{i=1}^n (\arcsin x_i - \arcsin x_{i-1})e^{x_{i-1/2}},\quad x_0 = -1,\quad x_n = 1.$$

4.7 КРАТНЫЕ ИНТЕГРАЛЫ

Формулы приближенного вычисления кратных интегралов по известной таблице значений подынтегральной функции принято называть **кубатурными.** В этом параграфе на примере вычисления двойного интеграла будут рассмотрены проблемы, возникающие при построении вычислительных алгоритмов, и пути их преодоления.

4.7.1 Метод ячеек

Рассмотрим сначала двойной интеграл по прямоугольнику $G = [a, b] \times [\alpha, \beta]$

$$I = \int_{\alpha}^{\beta} \int_{a}^{b} f(x, y)\, dxdy. \tag{4.37}$$

Тогда, по аналогии с формулой средних, этот интеграл приближенно равен

$$I \approx (\beta - \alpha)(b - a) f(\bar{x}, \bar{y}), \tag{4.38}$$

где $\bar{x} = (a + b)/2,\ \bar{y} = (\alpha + \beta)/2$. Оценим погрешность кубатурной формулы (4.38), для чего воспользуемся формулой Тейлора (для сокращения записи аргументы у производных будем опускать, считая, что производные вычисляются в точке $\bar{x}, \bar{y}$):

$$f(x, y) = f(\bar{x}, \bar{y}) + (x - \bar{x}) f_x + (y - \bar{y}) f_y + \\ + \frac{1}{2}(x - \bar{x})^2 f_{xx} + (x - \bar{x})(y - \bar{y}) f_{xy} + \frac{1}{2}(y - \bar{y})^2 f_{yy} + \dots . \tag{4.39}$$

Здесь $f_x = \dfrac{\partial f}{\partial x},\ f_{xx} = \dfrac{\partial^2 f}{\partial x^2}$ и т.д. Подставляя (4.39) в (4.38), имеем после несложных вычислений

$$\int_{a}^{b} \int_{\alpha}^{\beta} f(x, y) dxdy - (b - a)(\beta - \alpha) f(\bar{x}, \bar{y}) \approx \frac{1}{24} S[(b - a)^2 f_{xx} + (\beta - \alpha)^2 f_{yy}],$$

где $S = (b - a)(\beta - \alpha)$.

Для повышения точности область можно разбить на прямоугольные ячейки, вычислить интеграл по каждой ячейке и затем эти значения сложить. В результате имеем составную формулу. Если для построения составной формулы, стороны исходного прямоугольника разбить на n и m равных частей, то для одной ячейки с номером i погрешность интегрирования ψ_i будет приблизительно равна

$$\psi_i \approx \frac{1}{24} S_i \left[\left(\frac{b - a}{n} \right)^2 f_{xx,i} + \left(\frac{\beta - \alpha}{m} \right)^2 f_{yy,i} \right],$$

где S_i — площадь i-ой ячейки, а индекс i у производных означает, что их значения вычисляются в середине i-ой ячейки. Суммируя эти выражения по всем ячейкам, получим погрешность ψ составной формулы:

$$\psi = \sum_i \psi_i \approx \frac{1}{24} \left[\left(\frac{b - a}{n} \right)^2 \sum_i S_i f_{xx,i} + \left(\frac{\beta - \alpha}{m} \right)^2 \sum_i S_i f_{yy,i} \right] \approx \\ \approx \frac{1}{24} \left[\left(\frac{b - a}{n} \right)^2 \iint_G f_{xx}\, dxdy + \left(\frac{\beta - \alpha}{m} \right)^2 \iint_G f_{yy}\, dxdy \right] = O(n^{-2} + m^{-2}).$$

В результате получилась формула второго порядка точности.

Для вычисления интеграла в области с криволинейной границей на область накладывают прямоугольную сетку. Те ячейки, которые целиком лежат в области называют **внутренними** и интеграл в них считают так, как было описано выше. Рассмотрим теперь ячейки прямоугольной сетки, часть точек которых принадлежит области, а часть нет. Для каждой из них оставим только ту часть ячейки, которая принадлежит области. Полученные ячейки называют **граничными**. Площадь граничной ячейки вычисляют приближенно, заменяя в пределах ячейки истинную границу хордой и получая в результате площадь многоугольника (треугольника, трапеции и т.п.). Точку $\bar{x}, \bar{y}$ выбирают в центре тяжести ячейки[5]. Тогда за значение интеграла по каждой ячейке принимается произведение площади ячейки на значение подынтегральной функции, вычисленной в центре тяжести ячейки.

Вычисление в граничных ячейках довольно трудоемко, так как требует отслеживания положения границы области внутри ячейки. Поэтому, для упрощения вычислений при достаточно мелкой сетке, граничные ячейки вообще можно отбросить. Погрешность при этом будет $O(n^{-1})$.

В ряде случаев непрямоугольную область целесообразно привести к прямоугольнику путем соответствующей замены переменных. Например, для случая криволинейного четырехугольника $a \le x \le b,\ \varphi_1(x) \le y \le \varphi_2(x)$ замена

$$t = \frac{y - \varphi_1(x)}{\varphi_2(x) - \varphi_1(x)}$$

приводит область к прямоугольнику $a \le x \le b,\ 0 \le t \le 1$.

Метод ячеек без труда переносится на интегралы любого числа измерений.

4.7.2 Метод последовательного интегрирования

Рассмотрим интеграл вида

$$I = \iint\limits_G f(x,y)\,dxdy = \int\limits_\alpha^\beta F(y)\,dy, \quad \text{где} \quad F(y) = \int\limits_{\varphi_1(y)}^{\varphi_2(y)} f(x,y)\,dx.$$

Следовательно, для нахождения I достаточно найти интеграл от функции $F(y)$, который вычисляется с помощью одной из квадратурных формул. При этом значения функции $F(y)$, которые вычисляются в узлах y_j этой квадратурной формулы находятся как интегралы от функции $f(x, y_j)$ в пределах от $\varphi_1(y_j)$ до $\varphi_2(y_j)$. Таким образом, для вычисления значений $F(y_j)$ могут в свою очередь применяться квадратурные формулы.

Например, при использовании формулы трапеций получаются следующие рас-

[5]Напомним, что координаты центра тяжести треугольника получаются как среднее арифметическое координат его вершин. В случае многоугольника его, для простоты, можно разбить на треугольники и вычислить интеграл по каждому треугольнику.

четные формулы для двойного интеграла:

$$F(y_j) \approx \frac{\varphi_2(y_j)-\varphi_1(y_j)}{n_1}\Big(\frac{f(x_0,y_j)}{2}+f(x_1,y_j)+\cdots+f(x_{n_1-1},y_j)+\frac{f(x_{n_1},y_j)}{2}\Big),$$

$$x_k=\varphi_1(y_j)+\frac{\varphi_2(y_j)-\varphi_1(y_j)}{n_1}k,\quad k=0,1,\ldots,n_1,$$

$$I \approx \frac{\beta-\alpha}{n_2}\Big(\frac{F(y_0)}{2}+F(y_1)+\cdots+F(y_{n_2-1})+\frac{F(y_{n_2})}{2}\Big),$$

$$y_j=\alpha+\frac{\beta-\alpha}{n_2}j,\quad j=0,1,\ldots,n_2.$$

Очевидно, что число арифметических операций пропорционально произведению $n_1 n_2$.

Переход от кратного интеграла к повторному, а затем вычисление последнего с помощью многократного использования квадратурных формул может быть перенесен на общий случай m-кратных интегралов. При этом число арифметических операций есть $O(n_1 n_2 \ldots n_m)$.

Для вычисления интеграла с точностью ε обычно задают числа $n_i,\ i=1,\ldots,m$ и производят расчеты с соответствующим числом узлов вдоль каждой переменной. Затем числа n_i удваивают и повторяют вычисления. Эти действия продолжают до тех пор, пока в пределах заданной точности результат перестает изменяться.

4.7.3 Метод статистических испытаний (метод Монте-Карло)

Методом статистических испытаний или **методом Монте-Карло** принято называть совокупность приемов, позволяющих получать решения задач при помощи многократных случайных испытаний. Оценки искомой величины выводятся статистическим путем и носят вероятностный характер.

Напомним, что величину ξ называют **случайной величиной с плотностью распределения** $\rho(x)$, если вероятность того, что величина примет значения между x_1 и x_2 равна

$$P(x_1<\xi<x_2)=\int\limits_{x_1}^{x_2}\rho(x)\,dx.$$

Из определения вероятности при этом следует, что плотность распределения неотрицательна и интеграл от нее в пределах от минус до плюс бесконечности равен 1. Очевидно, что если значения ξ всегда заключены между a и b, то $\rho(x)=0$, вне указанных пределов.

Случайную величину γ называют **равномерно распределенной на отрезке** $[a,b]$, если ее плотность распределения есть

$$\rho(x)=\begin{cases}\dfrac{1}{b-a}, & \text{при } x\in[a,b],\\ 0, & \text{при } x\notin[a,b].\end{cases}$$

Это значит, во-первых, что все ее значения лежат в отрезке $[a,b]$ и, во-вторых, вероятность попадания значений случайной величины γ в любой интервал из отрезка $[a,b]$ пропорциональна длине этого интервала и не зависит от его положения.

Значения случайной величины принято называть **случайными числами.** Последовательность чисел $\xi_1,\xi_2,\ldots$, являющихся значениями одной и той же случайной

величины ξ при независимых между собой испытаниях с повторяющимися условиями, называют **случайной последовательностью** с соответствующим законом распределения.

Идея метода Монте-Карло заключается в том, что рассматривается некоторая случайная величина ξ, математическое ожидание которой $M\xi$ равно искомой величине I. Проводится серия n независимых испытаний, в результате которых получаются случайные чисел $\xi_1, \xi_2, \ldots, \xi_n$. Среднее арифметическое этих чисел

$$\zeta_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

является случайной величиной с тем же математическим ожиданием, то есть $M\zeta_n = M\xi = I$. Согласно закону больших чисел в форме Хинчина, для любого $\varepsilon > 0$ вероятность $P(|\zeta_n - I| > \varepsilon) \to 0$ при $n \to \infty$. Таким образом, при больших n величина $I \approx \zeta_n$.

Если случайная величина ξ имеет конечную дисперсию

$$D\xi = M(\xi^2) - (M\xi)^2, \tag{4.40}$$

то доказывается, опираясь на независимость испытаний, что дисперсия $D\zeta_n = \dfrac{1}{n} D\xi$. В этом случае для оценки погрешности метода можно применить неравенство Чебышева, в соответствии с которым

$$P(|\zeta_n - M\zeta_n| \geq \varepsilon) \leq \frac{D\zeta_n}{\varepsilon^2}.$$

Так как $M\zeta_n = I$, а $D\zeta_n = \dfrac{1}{n} D\xi$, то для того чтобы с уровнем δ имели $P(|\zeta_n - I| \geq \varepsilon) \leq \delta$, достаточно выполнения неравенства $(D\xi)/(n\varepsilon^2) \leq \delta$. Отсюда следует, что

$$n \geq \frac{D\xi}{\varepsilon^2 \delta}.$$

Эта оценка означает, что сходимость ζ_n к I медленная. Уменьшение ε, например, в 10 раз проводит к увеличению числа испытаний n в 100 раз. Поэтому в расчетах число n обычно велико. Если, например, $D\xi = 1,\ \varepsilon = 0.01,\ \delta = 0.003$, то $n = 1/3 \cdot 10^7$.

Однако, оценка количества испытаний, основанная на применении неравенства Чебышева, является очень грубой. Более точную оценку можно получить опираясь на центральную предельную теорему для одинаково распределенных случайных величин. В ней утверждается, что для любых $x_1 < x_2$

$$\lim_{n\to\infty} P\left(x_1 < \frac{\zeta_n - I}{\sqrt{\frac{D\xi}{n}}} < x_2 \right) = \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-t^2/2}\, dt.$$

Если положить в этом неравенстве $-x_1 = x_2 = x > 0$, то

$$\lim_{n\to\infty} P\left(\left| \frac{\zeta_n - I}{\sqrt{\frac{D\xi}{n}}} \right| < x \right) = \frac{2}{\sqrt{2\pi}} \int_{0}^{x} e^{-t^2/2}\, dt = \Phi(x).$$

Функция $\Phi(x)$ называется функцией Лапласа. Таблица ее значений приводится в любом учебнике по теории вероятностей.

Из теоремы следует, что при больших значениях n можно считать, что

$$P\left(|\zeta_n - I| < x\sqrt{\frac{D\xi}{n}}\right) \approx \Phi(x) \tag{4.41}$$

Если теперь задать произвольное значение доверительной вероятности γ, то по таблице значений функции Лапласа можно найти такое значение $x = x_\gamma$, что $\Phi(x_\gamma) = \gamma$. Тогда из (4.41) вытекает, что вероятность неравенства

$$|\zeta_n - I| < x_\gamma\sqrt{\frac{D\xi}{n}} \tag{4.42}$$

приблизительно равна γ.

Заметим, что $\gamma = 0.95$ при $x_\gamma = 1.96$, $\gamma = 0.997$ при $x_\gamma = 3$ и $\gamma = 0.99999$ при $x_\gamma = 5$. Таким образом, неравенства

$$|\zeta_n - I| < 1.96\sqrt{\frac{D\xi}{n}}, \quad |\zeta_n - I| < 3\sqrt{\frac{D\xi}{n}} \quad \text{и} \quad |\zeta_n - I| < 5\sqrt{\frac{D\xi}{n}}$$

выполняются, соответственно, с вероятностями 0.95, 0.997 и 0.99999.

Вернемся к рассмотренному выше примеру, в котором предполагалось выполнение равенства $D\xi = 1$ и искалось значение n такое, что $P(|\zeta_n - I| > 0.01) < 0.003$. Легко заметить теперь, что это неравенство будет выполнено, если $\dfrac{3}{\sqrt{n}} \le 0,01$. Отсюда следует, что $n \ge 30000$. Это значение, конечно, тоже велико, но много меньше того, которое было получено с помощью неравенства Чебышева.

Для того, чтобы воспользоваться оценкой (4.42) необходимо знать величину $D\xi$. Ее достаточно просто получить в ходе вычислений. Достаточно одновременно с ζ_n вычислять $\dfrac{1}{n}\sum\limits_{i=1}^{n}(\xi_i)^2$, так как при больших значениях n эта сумма приблизительно равна $M(\xi^2)$. Тогда, согласно (4.40)

$$D\xi \approx \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - (\zeta_n)^2.$$

Таким образом, алгоритм нахождения неизвестной величины I с заданной точностью ε и значением доверительной вероятности γ выглядит следующим образом. Из уравнения $\Phi(x_\gamma) = \gamma$ находится x_γ. Затем последовательно определяются значения случайной величины $\xi_1, \xi_2, \ldots$ и вычисляются ζ_n, $D\xi$ до тех пор, пока не станет выполняться неравенство $x_\gamma\sqrt{\dfrac{D\xi}{n}} < \varepsilon$. Тогда полагают $I \approx \zeta_n$.

Отметим еще раз, что результат вычислений, полученный методом Монте-Карло, носит вероятностный характер и, в принципе, может сколь угодно сильно отличаться от точного значения величины I.

Рассмотрев общую идею метода, перейдем непосредственно к вычислению интегралов.

Пусть необходимо вычислить интеграл

$$I = \int_0^1 f(x)\,dx.$$

Если η - равномерно распределенная на $[0,1]$ случайная величина, то ее плотность распределения

$$\rho(x) = \begin{cases} 1, & \text{при } x \in [0,1], \\ 0, & \text{при } x \notin [0,1]. \end{cases}$$

Величина $\xi = f(\eta)$ также случайная и

$$M\xi = \int\limits_{-\infty}^{\infty} f(x)\rho(x)\,dx = \int\limits_{0}^{1} f(x)\,dx = I.$$

Здесь считается, что вне отрезка $[0,1]$ функция $f(x)$ доопределена, например, нулем. Поэтому, в соответствии с методом Монте-Карло,

$$I \approx \frac{1}{n}\sum_{i=1}^{n} \xi_i = \frac{1}{n}\sum_{i=1}^{n} f(\eta_i).$$

Отсюда следует, что имея равномерно распределенную на $[0,1]$ последовательность случайных чисел η_i, легко приближенно вычислить искомый интеграл.

Перейдем теперь к вычислению кратных интегралов.

Пусть функция $y = f(x_1, \ldots, x_m)$ непрерывна в ограниченной замкнутой области Ω и требуется вычислить m-кратный интеграл

$$I = \int\limits_{\Omega} \cdots \int f(x_1, \ldots, x_m)\,dx_1 \ldots dx_m.$$

Геометрически число I представляет собой объем $(m+1)-$мерного прямого цилиндра в пространстве $Ox_1 \ldots x_m y$, построенного на основании Ω и ограниченного сверху данной поверхностью $y = f(x_1, \ldots, x_m)$.[6]

Преобразуем интеграл так, чтобы новая область интегрирования σ целиком содержалась внутри единичного $m-$мерного куба. Пусть область Ω расположена в $m-$мерном параллелепипеде $a_i \le x_i \le b_i, \quad i = 1, \ldots, m$. Тогда достаточно положить $x_i = a_i + (b_i - a_i)z_i, \quad i = 1, \ldots, m$. Якобиан этого преобразования равен

$$J = \frac{D(x_1, \ldots, x_m)}{D(z_1, \ldots, z_m)} = \begin{vmatrix} b_1 - a_1 & 0 & \cdots & 0 \\ 0 & b_2 - a_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & b_m - a_m \end{vmatrix} = \prod_{i=1}^{m} (b_i - a_i). \tag{4.43}$$

Тогда

$$I = J \int\limits_{\sigma} \cdots \int F(z_1, \ldots, z_m)\,dz_1 \ldots dz_m = J\tilde{I}, \tag{4.44}$$

где $F(z_1, \ldots, z_m) = f(a_1 + (b_1 - a_1)z_1, \ldots, a_m + (b_m - a_m)z_m)$,

$$\tilde{I} = \int\limits_{\sigma} \cdots \int F(z_1, \ldots, z_m)\,dz_1 \ldots dz_m.$$

Таким образом, для вычисления интеграла I достаточно найти $\tilde{I}$.

[6] Если функция отрицательна, то для подобного геометрического толкования надо взять ее с противоположным знаком.

Если $\eta = (\eta^{(1)}, \ldots, \eta^{(m)})$ - равномерно распределенная в m-мерном единичном кубе случайная величина, то ее плотность распределения

$$\rho(z_1, \ldots, z_m) = \begin{cases} 1, & \text{при } z_i \in [0,1],\ i = 1, \ldots, m, \\ 0, & \text{во всех остальных случаях.} \end{cases}$$

Определим функцию $\tilde{F}$,

$$\tilde{F}(z_1, \ldots, z_m) = \begin{cases} F(z_1, \ldots, z_m), & \text{при } (z_1, \ldots, z_m) \in \sigma, \\ 0, & \text{во всех остальных случаях.} \end{cases}$$

Тогда величина $\xi = \tilde{F}(\eta)$ также случайная и

$$M\xi = \int\limits_{R^m} \cdots \int \tilde{F}(z_1, \ldots, z_m)\rho(z_1, \ldots, z_m)\, dz_1 \ldots dz_m =$$

$$= \int\limits_0^1 \ldots \int\limits_0^1 \tilde{F}(z_1, \ldots, z_m)\, dz_1 \ldots dz_m = \int\limits_\sigma \cdots \int F(z_1, \ldots, z_m)\, dz_1 \ldots dz_m = \tilde{I}.$$

Таким образом, интеграл $\tilde{I}$ совпадает с математическим ожиданием случайной величины ξ. Поэтому, в соответствии с методом Монте-Карло,

$$\tilde{I} \approx \frac{1}{n}\sum_{i=1}^{n} \xi_i = \frac{1}{n}\sum_{i=1}^{n} \tilde{F}(\eta_i) = \frac{1}{n}{\sum_{i=1}^{n}}' F(\eta_i).$$

Здесь штрих у суммы означает, что в нее включаются только те слагаемые, для которых $\eta_i \in \sigma$. Тогда в соответствии с (4.44)

$$I \approx J\frac{1}{n}{\sum_{i=1}^{n}}' F(\eta_i).$$

Для вычисления интеграла по приведенной формуле необходимо вырабатывать последовательность значений случайной величины η. Для этого можно выбрать m равномерно распределенных на отрезке $[0,1]$ последовательностей случайных чисел $(\eta_i^{(1)}, \ldots, \eta_i^{(m)})$, $i = 1, 2, \ldots$ или, что одно и то же, вырабатывать одну последовательность случайных чисел и брать в ней по m элементов, которые и будут служить значениями случайной величины η.

В настоящее время выработка случайных чисел на ЭВМ заключается в том, что по некоторому алгоритму происходит их вычисление (подобно тому, как вычисляются значения элементарных функций). Поскольку эти числа генерируются по наперед заданному алгоритму, то они не совсем случайны (их называют **псевдослучайными**), хотя и обладают свойственными случайным числам статистическими характеристиками. Алгоритмы вычисления псевдослучайных чисел входят практически во все распространенные языки программирования.

Возникает вопрос, какими методами удобнее вычислять интегралы — статистическими или традиционными сеточными? Точность метода статистических испытаний невелика, и для однократных интегралов они невыгодны. Для кратных интегралов положение меняется.

Пусть гладкая функция m переменных интегрируется по сеточным формулам p-го порядка точности, причем сетка имеет k шагов по каждой переменной. Тогда

полное число узлов есть $n = k^m$, а погрешность расчета $\varepsilon = O(k^{-p})$. Поэтому число узлов, требуемое для достижения данной точности ε, есть $n = O((1/\varepsilon)^{m/p})$. Это число экспоненциально растет с ростом числа измерений.

При интегрировании методом Монте-Карло погрешность $\varepsilon = O(n^{-1/2})$. Поэтому полное число узлов $n = O((1/\varepsilon)^2)$ независимо от числа измерений.

Таким образом, если число измерений $m < 2p$, то сеточные методы требуют меньшего числа узлов и более выгодны. При $m > 2p$ предпочтительнее статистические методы. В многомерном случае редко можно встретить метод с порядком точности выше 2, поэтому для трехмерного интеграла лучше сеточный метод, а для пятимерного — Монте-Карло.

4.8 ЗАДАЧИ К ГЛАВЕ 4

4.8.1 Примеры решения задач

1. Доказать, что для гладкой функции f справедлива формула

$$f'(x) \approx \frac{-25f(x) + 48f(x+h) - 36f(x+2h) + 16f(x+3h) - 3f(x+4h)}{12h}.$$

Найти погрешность этой формулы.

Решение. Согласно формуле Тейлора имеем

$$f(x+kh) = f(x) + khf'(x) + \frac{(kh)^2}{2!}f''(x) + \frac{(kh)^3}{3!}f'''(x) + \\ + \frac{(kh)^4}{4!}f^{IV}(x) + \cdots + \frac{(kh)^s}{s!}f^{(s)}(x) + \frac{(kh)^{s+1}}{(s+1)!}f^{(s+1)}(\xi),$$

где ξ — некоторая точка, лежащая между x и $x + kh$. Тогда

$$\frac{-25f(x) + 48f(x+h) - 36f(x+2h) + 16f(x+3h) - 3f(x+4h)}{12h} = \\ = \frac{1}{12h}\Bigg(-25f(x) + 48\sum_{i=0}^{s}\frac{h^i}{i!}f^{(i)}(x) + 48\frac{h^{s+1}}{s+1!}f^{(s+1)}(\xi_1) - \\ - 36\sum_{i=0}^{s}\frac{(2h)^i}{i!}f^{(i)}(x) - 36\frac{(2h)^{s+1}}{(s+1)!}f^{(s+1)}(\xi_2) + 16\sum_{i=0}^{s}\frac{(3h)^i}{i!}f^{(i)}(x) + 16\frac{(3h)^{s+1}}{(s+1)!}f^{(s+1)}(\xi_3) - \\ - 3\sum_{i=0}^{s}\frac{(4h)^i}{i!}f^{(i)}(x) - 3\frac{(4h)^{s+1}}{(s+1)!}f^{(s+1)}(\xi_4)\Bigg) = \\ = \frac{1}{12h}\Bigg((-25+48-36+16-3)f(x) + h(48 - 2\cdot 36 + 3\cdot 16 - 4\cdot 3)f'(x) + \\ + \frac{h^2}{2}(48 - 4\cdot 36 + 9\cdot 16 - 16\cdot 3)f''(x) + \frac{h^3}{6}(48 - 8\cdot 36 + 27\cdot 16 - 64\cdot 3)f'''(x) + \\ + \frac{h^4}{24}(48 - 16\cdot 36 + 81\cdot 16 - 256\cdot 3)f^{IV}(x)\Bigg) + \\ + \frac{1}{12h}\frac{h^5}{120}\Big(48f^V(\xi_1) - 32\cdot 36f^V(\xi_2) + 243\cdot 16f^V(\xi_3) - 1024\cdot 3f^V(\xi_4)\Big) = \\ = f'(x) + \frac{h^4}{30}\Big(f^V(\xi_1) - 24f^V(\xi_2) + 81f^V(\xi_3) - 64f^V(\xi_4)\Big).$$

Здесь $s = 4$. При малом значении h точки $\xi_j,\ j = 1, 2, 3, 4$ близки. Следовательно, можно считать, что значения функции f^V в этих точках приблизительно равны и $f^V(\xi_1) - 24f^V(\xi_2) + 81f^V(\xi_3) - 64f^V(\xi_4) \approx -6f^V(\xi)$ [7], поэтому

$$\frac{-25f(x) + 48f(x+h) - 36f(x+2h) + 16f(x+3h) - 3f(x+4h)}{12h} = f'(x) - \frac{h^4}{5}f^V(\xi).$$

2. Оценить погрешность, совершаемую при вычислении интеграла

$$I = \int_0^1 e^{-x^2}\,dx$$

по формуле трапеций, если шаг $h = 0.1$.

Решение. Для погрешности ψ формулы трапеций имеем:

$$|\psi| \le \frac{h^2(b-a)}{12}M_2.$$

Здесь $b - a$ — длина отрезка интегрирования, которая по условию задачи равна 1. M_2 — максимум модуля второй производной подынтегральной функции на отрезке, по которому интегрируется функция. Так как

$$\left(e^{-x^2}\right)'' = 2(2x^2 - 1)e^{-x^2},$$

легко получить, что наибольшее значение на отрезке $[0, 1]$ модуль этой функции принимает при $x = 0$, то есть $M_2 = 2$. Окончательно имеем:

$$|\psi| \le \frac{0.1^2}{12}2 \approx 1.667 \cdot 10^{-2}.$$

3. Показать, что квадратурная формула

$$\int_{-\pi}^{\pi} f(x)\,dx \approx \frac{2\pi}{n}\sum_{k=1}^{n} f\big(\alpha + (k-1)\frac{2\pi}{n}\big), \qquad \alpha \in \big[\pi, \pi + \frac{2\pi}{n}\big]$$

является точной для любого тригонометрического многочлена степени не выше $(n - 1)$. Напомним, что функции $f(x)$ называется **тригонометрическим многочленом степени** m, если

$$f(x) = a_0 + \sum_{j=1}^{m}(a_j \cos jx + b_j \sin jx).$$

Решение. Достаточно проверить, что квадратурная формула точна для всех функций

$$f(x) = \sin jx,\ \ f(x) = \cos jx,\ \ j = 0, 1, \ldots, n-1.$$

[7] Легко доказать, что если $g(x)$ — непрерывная функция и $\alpha + \beta = 1$, причем $\alpha \ge 0,\ \beta \ge 0$, то существует такая точка z, лежащая между точками x и y, что $\alpha g(x) + \beta g(y) = g(z)$. Действительно, если $g(x) = g(y)$, то утверждение справедливо. Если же $g(x) \ne g(y)$, то предположим для определенности, что $g(y) \le g(x)$. Тогда $g(y) \le \alpha g(x) + \beta g(y) = A \le g(x)$. Так как непрерывная функция принимает все промежуточные значения, лежащие на отрезке $[g(y), g(x)]$, между точками x и y есть точка, назовем ее z, в которой $g(z) = A$, что и требовалось доказать.

Из этого утверждения следует, что существуют такие точки ξ_5 и ξ_6, что $f^V(\xi_1) - 24f^V(\xi_2) + 81f^V(\xi_3) - 64f^V(\xi_4) = 82f^V(\xi_5) - 88f^V(\xi_6)$, при этом никакое предположение о малости шага не делается.

Интегралы от этих функций равны

$$\int_{-\pi}^{\pi} 1 \cdot dx = 2\pi, \quad \int_{-\pi}^{\pi} \cos jx\, dx = \int_{-\pi}^{\pi} \sin jx\, dx = 0, \quad j > 0.$$

Подсчитаем теперь значения сумм.

$$\frac{2\pi}{n} \sum_{k=1}^{n} 1 = 2\pi.$$

Пусть i — мнимая единица, то есть $i^2 = -1$. Тогда, используя формулу суммы геометрической прогрессии, получаем:

$$\frac{2\pi}{n} \sum_{k=1}^{n} \cos j\big(\alpha + (k-1)\frac{2\pi}{n}\big) + i\frac{2\pi}{n} \sum_{k=1}^{n} \sin j\big(\alpha + (k-1)\frac{2\pi}{n}\big) =$$
$$= \frac{2\pi}{n} \sum_{k=1}^{n} e^{ij\left(\alpha + (k-1)\frac{2\pi}{n}\right)} = \frac{2\pi}{n} e^{ij\alpha} \frac{1 - e^{ij2\pi}}{1 - e^{ij2\pi/n}} = 0, \quad j = 1, \ldots, n-1.$$

Таким образом, значения интегралов и сумм совпадают, что и требовалось доказать.

4. Найти коэффициенты и узлы квадратурной формулы Гаусса для вычисления интеграла

$$I = \int_{-1}^{1} \frac{f(x)\, dx}{\sqrt{1 - x^2}}.$$

Решение. В соответствии с первым утверждением теоремы из параграфа 4.5, узлы квадратурной формулы являются корнями многочлена, который ортогонален с весом $\rho(x) = (1 - x^2)^{-1/2}$ любому многочлену степени меньше n. Из задачи 19 предыдущей главы следует, что многочлен Чебышева $T_n(x) = \cos(n \cdot \arccos x))$ как раз и обладает указанным свойством. Таким образом, узлами квадратурной формулы являются корни многочлена Чебышева $T_n(x)$:

$$x_k = \cos \frac{(2k-1)\pi}{2n}, \quad k = 1, \ldots, n.$$

Коэффициенты формулы Гаусса находятся, согласно второму утверждению теоремы, по формуле

$$C_k = \int_{-1}^{1} \rho(x) \frac{\omega(x)}{(x - x_k)\omega'(x_k)}\, dx,$$

где $\omega(x) = \prod_{k=1}^{n} (x - x_k)$ и, следовательно, только числовым множителем отличается от $T_n(x)$. Поэтому формула для коэффициентов может быть переписана следующим образом:

$$C_k = \int_{-1}^{1} \frac{T_n(x)}{(x - x_k)T_n'(x_k)\sqrt{1 - x^2}}\, dx. \tag{4.45}$$

Вычислим коэффициенты. Сделаем в (4.45) замену переменной $x = \cos\varphi$. Получим:

$$C_k = \int_{0}^{\pi} \frac{\cos n\varphi}{(\cos\varphi - x_k)T_n'(x_k)}\, d\varphi = \frac{1}{2T_n'(x_k)} \int_{-\pi}^{\pi} \frac{\cos n\varphi}{(\cos\varphi - x_k)}\, d\varphi. \tag{4.46}$$

Отметим, что $P_{n-1}(x) = \dfrac{T_n(x)}{(x - x_k)}$ — многочлен степени $n-1$. Была проделана замена переменной x на $\cos\varphi$, в результате чего получился многочлен $P(\cos\varphi)$. Так как при замене многочлен $\dfrac{T_n(x)}{(x - x_k)}$ перешел в $\dfrac{\cos n\varphi}{(\cos\varphi - x_k)}$, значит это выражение является другой записью многочлена $P_{n-1}(\cos\varphi)$.

В [13] приведены формулы, в которых $\cos^r\varphi$ выражается в виде линейной комбинации $\cos r\varphi,\ \cos(r-2)\varphi, \ldots$. Учитывая этот факт, получаем, что $P_{n-1}(\cos\varphi) = \dfrac{\cos n\varphi}{(\cos\varphi - x_k)}$ является тригонометрическим многочленом степени $n-1$. Из предыдущей задачи следует тогда, что интеграл (4.46) может быть найден по квадратурной формуле с не менее чем n равномерно расположенными узлами и равными коэффициентами. Выбирая в качестве узлов точки

$$\pm\frac{\pi}{2n},\ \pm\frac{3\pi}{2n},\ \ldots,\ \pm\frac{(2n-1)\pi}{2n}$$

и учитывая, что

$$P_{n-1}\left(\cos\Big(\pm\frac{(2j-1)\pi}{2n}\Big)\right) = P_{n-1}(x_j) = \frac{T_n(x_j)}{(x_j - x_k)} = \begin{cases} 0, & j \neq k, \\ T_n'(x_k), & j = k, \end{cases}$$

получаем

$$\int\limits_{-\pi}^{\pi} \frac{\cos n\varphi}{(\cos\varphi - x_k)}\,d\varphi = \frac{2\pi}{2n}\sum_{j=1}^{n} P_{n-1}\left(\cos\Big(\pm\frac{(2j-1)\pi}{2n}\Big)\right) =$$

$$= \frac{\pi}{n}P_{n-1}\left(\cos\Big(\frac{(2k-1)\pi}{2n}\Big)\right) + \frac{\pi}{n}P_{n-1}\left(\cos\Big(-\frac{(2k-1)\pi}{2n}\Big)\right) = \frac{2\pi T_n'(x_k)}{n}.$$

Окончательно из этой формулы и (4.46) имеем выражение для коэффициентов квадратурной формулы Гаусса:

$$C_k = \frac{\pi}{n}, \qquad k = 1, \ldots, n.$$

Таким образом, получена формула численного интегрирования, которая является частным случаем формулы Гаусса и называется **формулой Эрмита**

$$\int\limits_{-1}^{1} \frac{f(x)\,dx}{\sqrt{1-x^2}} \approx \frac{\pi}{n}\sum_{k=1}^{n} f\Big(\cos\frac{2k-1}{2n}\pi\Big).$$

5. Предложить способ вычисления интеграла

$$\int\limits_{-1}^{1} \frac{dx}{\sqrt{1-x^4}}.$$

Решение. Подынтегральная функция обращается в бесконечность в точках ± 1. Представим ее в виде

$$\frac{1}{\sqrt{1-x^4}} = \frac{1}{\sqrt{1-x^2}} \cdot \frac{1}{\sqrt{1+x^2}}$$

и будем рассматривать функцию $(1-x^2)^{-1/2}$ как весовую. Тогда применима формула численного интегрирования Эрмита, полученная в предыдущей задаче:

$$\int\limits_{-1}^{1} \frac{dx}{\sqrt{1-x^4}} \approx \frac{\pi}{n} \sum_{k=1}^{n} \frac{1}{\sqrt{1+x_k^2}}, \quad x_k = \cos\frac{2k-1}{2n}\pi.$$

При $n = 6$ имеем

$$\int\limits_{-1}^{1} \frac{dx}{\sqrt{1-x^4}} \approx \frac{\pi}{6}\Big(\frac{2}{\sqrt{1+\cos^2\frac{\pi}{12}}} + \frac{2}{\sqrt{1+\cos^2\frac{3\pi}{12}}} + \frac{2}{\sqrt{1+\cos^2\frac{5\pi}{12}}}\Big) \approx 2.22133.$$

Для сравнения, значение интеграла с 5 верными знаками после запятой равно 2.22144.

4.8.2 Задачи

1. Пусть

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h},$$

причем функция $f(x)$ такова, что $|f'''(x)| \le 1$. Предположим, что значения функции $f(x)$ вычисляются с погрешностью не превосходящей величины δ. Какая наибольшая точность может быть достигнута при вычислении производной и как наилучшим способом выбрать шаг h?

2. Введем обозначения:

$$T_1 f(x) = f(x+h),\ T_{-1} f(x) = f(x-h),\ Ef(x) = f(x),\ \Delta_1 = T_1 - E,\ \Delta_{-1} = E - T_{-1}.$$

С помощью введенных операторов легко записать формулы численного дифференцирования:

$$\frac{\Delta_1}{h} f(x) = \frac{f(x+h) - f(x)}{h},\ \frac{\Delta_{-1}}{h} f(x) = \frac{f(x) - f(x-h)}{h}.$$

Доказать справедливость следующих равенств для численного дифференцирования произведения:

$$\frac{\Delta_1}{h}(fg) = \Big(\frac{\Delta_1}{h} f\Big)g + (T_1 f)\Big(\frac{\Delta_1}{h} g\Big) =$$
$$\Big(\frac{\Delta_1}{h} f\Big)(T_1 g) + f\Big(\frac{\Delta_1}{h} g\Big) = \Big(\frac{\Delta_1}{h} f\Big)g + f\Big(\frac{\Delta_1}{h} g\Big) + h\Big(\frac{\Delta_1}{h} f\Big)\Big(\frac{\Delta_1}{h} g\Big),$$
$$\frac{\Delta_{-1}}{h}(fg) = \Big(\frac{\Delta_{-1}}{h} f\Big)g + (T_{-1} f)\Big(\frac{\Delta_{-1}}{h} g\Big) =$$
$$\Big(\frac{\Delta_{-1}}{h} f\Big)(T_{-1} g) + f\Big(\frac{\Delta_{-1}}{h} g\Big) = \Big(\frac{\Delta_{-1}}{h} f\Big)g + f\Big(\frac{\Delta_{-1}}{h} g\Big) - h\Big(\frac{\Delta_{-1}}{h} f\Big)\Big(\frac{\Delta_{-1}}{h} g\Big).$$

3. Найти коэффициенты таким образом, чтобы полученные формулы численного дифференцирования имели максимальную точность:

$$f'(x) \approx (af(x) + bf(x+h) + cf(x-h))/h,$$
$$f'(x) \approx (af(x) + bf(x+h) + cf(x-2h))/h,$$
$$f''(x) \approx (af(x) + bf(x+h) + cf(x+2h))/h^2,$$
$$f''(x) \approx (af(x) + bf(x+h) + cf(x-h))/h^2,$$
$$f''(x) \approx (af(x) + bf(x-h) + cf(x-2h))/h^2.$$

4. На равномерной сетке получить формулы для численного нахождения третьей и четвертой производных.

5. Доказать следующие равенства:

$$f'(x+h) = \frac{f(x-h) - 4f(x) + 3f(x+h)}{2h} + \frac{h^2}{3} f'''(\xi),$$

$$f'(x-h) = \frac{-11f(x-h) + 18f(x) - 9f(x+h) + 2f(x+2h)}{6h} - \frac{h^3}{4} f^{IV}(\xi),$$

$$f'(x) = \frac{-2f(x-h) - 3f(x) + 6f(x+h) - f(x+2h)}{6h} + \frac{h^3}{12} f^{IV}(\xi),$$

$$f'(x) = \frac{f(x-2h) - 8f(x-h) + 8f(x+h) - f(x+2h)}{12h} + \frac{h^4}{30} f^{V}(\xi).$$

6. Для вычисления $\int\limits_0^1 f(x)\,dx$ применяется составная формула трапеций с равномерным распределением узлов. Оценить минимальное число разбиений, обеспечивающее точность $0.5 \cdot 10^{-3}$, если модуль второй производной функции $f(x)$ не превосходит 1.

7. Оценить минимальное число разбиений отрезка для вычисления интеграла $\int\limits_0^1 e^{x^2}\,dx$ по составной формуле центральных прямоугольников, обеспечивающее точность 10^{-4}.

8. Доказать, что для погрешности формулы трапеций справедливо равенство:

$$\int_x^{x+h} f(y)\,dy - \frac{h}{2}\big(f(x) + f(x+h)\big) = \frac{1}{2} \int_x^{x+h} (x-y)(x+h-y) f''(y)\,dy.$$

Указание. Дважды проинтегрировать по частям интеграл, стоящий в правой части равенства.

9. Используя результат предыдущей задачи, доказать оценку для погрешность составной формулы трапеций

$$\left| \int_a^b f(x)\,dx - h\Big(0.5f(a) + 0.5f(b) + \sum_{i=1}^{n-1} f(x_i)\Big) \right| \le \frac{h^2}{8} \int_a^b |f''(x)|\,dx.$$

10. Доказать, что ортогональные многочлены вида $\varphi_n(x) = x^n + \ldots$, определенные на симметричном относительно нуля отрезке с четным весом $\rho(x)$, удовлетворяют рекуррентному соотношению

$$\varphi_n(x) = x\varphi_{n-1}(x) + a_n\varphi_{n-2}(x).$$

Указание. Представить многочлен $x\varphi_{n-1}(x)$ в виде $x\varphi_{n-1}(x) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$ и показать, что в силу ортогональности $a_i = 0$ при $i < n-2$.

11. Доказать, что ортогональные многочлены на симметричном относительно нуля отрезке с четным весом $\rho(x)$ обладают свойством $\varphi_n(-x) = (-1)^n\varphi_n(x)$.

Указание. Воспользоваться методом математической индукции и соотношением предыдущей задачи.

12. Построить квадратурные формулы Гаусса с одним узлом для вычисления интегралов:

$$\int_0^1 x f(x)\,dx, \qquad \int_0^1 e^x f(x)\,dx.$$

13. Построить квадратурные формулы Гаусса с двумя узлами для вычисления интегралов:

$$\int_{-1}^1 x^2 f(x)\,dx, \qquad \int_{-\pi/2}^{\pi/2} \cos x f(x)\,dx.$$

14. Построить квадратурную формулу Гаусса с тремя узлами для вычисления интеграла:

$$\int_{-1}^1 f(x)\,dx.$$

15. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интеграла

$$\int_0^\infty e^{-x} f(x)\,dx.$$

16. Показать, что квадратурная формула

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)\,dx \approx \frac{\sqrt{\pi}}{3}\left(f\Big(-\frac{\sqrt{3}}{2}\Big) + f(0) + f\Big(\frac{\sqrt{3}}{2}\Big)\right)$$

точна для всех многочленов пятой степени.

17. Предложить алгоритмы для вычисления несобственных интегралов:

$$\text{а)}\ \int_0^1 \frac{\sin x}{x^{3/2}}\,dx, \quad \text{б)}\ \int_0^{\pi/2} \frac{\cos x}{\sqrt{\pi/2 - x}}\,dx, \text{в)}\ \int_0^{\pi/2} \ln(\sin x)\,dx.$$

Указание. а) Дважды воспользоваться формулой интегрирования по частям. б) Сделать замену переменной $\pi/2 - x = y^2$. в) Воспользоваться равенством $\ln(\sin x) = \ln x + \ln\Big(\frac{\sin x}{x}\Big)$.

18. Как вычислить интеграл $\int\limits_0^1 \dfrac{\ln x}{1+x^2}\,dx$ по составной квадратурной формуле с постоянным шагом h, чтобы погрешность имела порядок $O(h^2)$?

Указание. Применить результат задачи 9 к интегралу

$$\int_0^1 \Big(\frac{\ln x}{1+x^2} - \ln x\Big)\,dx.$$

19. Задана таблица значений функции $y = \cos x$:

$$y(\pi/12) = 0.966,\ y(\pi/8) = 0.924,\ y(\pi/6) = 0.866,\ y(5\pi/24) = 0.793,\ y(\pi/4) = 0.707.$$

Вычислить производную в точке $\pi/6$ и уточнить ее по значение по методу Рунге. Сравнить полученное значение с точным значением производной в этой точке.

20. Вывести формулу Симпсона, для вычисления интеграла, используя формулу трапеций и метод Рунге.

21. Квадратурной формулой Чебышева называется такая квадратурная формула, в которой все ее коэффициенты равны. Формула строится так, чтобы она была точна для многочленов как можно более высокой степени.

Вывести квадратурную формулу

$$\int_{-1}^{1} f(x)\,dx \approx C\sum_{i=1}^{n} f(x_i),$$

то есть найти коэффициент C и узлы x_i для случаев $n = 2, 3$.

4.8.3 Примеры тестовых вопросов к главе 4

1. Расположите следующие методы вычисления интегралов в порядке возрастания порядка точности методов. В случае равных порядков сначала укажите метод с большей погрешностью. В ответ запишите номера методов через пробел.

1) Левых прямоугольников. 2) Средних прямоугольников. 3) Симпсона. 4) Трапеций.

2. Ниже приведены формулы численного дифференцирования. Пусть k — порядок погрешности формулы, а n — порядок производной, которая вычисляется с помощью формулы. Для какой из формул величина $|n - k|$ максимальна?

$$a)\quad \frac{f(x+h) - f(x-h)}{2h}, \qquad b)\quad \frac{f(x) - f(x-h)}{h},$$
$$c)\quad \frac{f(x+h) - f(x)}{h}, \qquad d)\quad \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

3. Для вычисления интеграла $\int_0^3 (x^2 - 4x + 5)\,dx$ используется формула Симпсона с шагом 0.5. Какое число получится в результате вычислений?

4. Интеграл $\int_0^1 (x^8 + 7x^7 - 23)\,dx$ вычисляется по квадратурной формуле интерполяционного типа с узлами $x_i = 0.2i$ и квадратурной формуле Гаусса, содержащей 5 узлов. Пусть δ_I — погрешность квадратурной формулы интерполяционного типа, а δ_g — формулы Гаусса. Чему равно выражение $(\delta_I^2 + 1)\delta_g$?

5. Квадратурная формула

$$\int_a^b f(x)\,dx \approx A\sum_{i=1}^{n} f(x_i)$$

называется квадратурной формулой Чебышева, если ее узлы и вес A подобраны таким образом, чтобы она имела максимальный алгебраический порядок точности.

Чему должен быть равен вес A этой формулы, если $a = -1,\ b = 1,\ n = 5$? Ответ представить в виде десятичного числа.

6. Интеграл

$$\int_0^1 \frac{dx}{\sqrt[3]{x(1+\cos x)}}$$

вычисляется методом сведения к сумме двух интегралов, один из которых несобственный и находится аналитически, а второй не имеет особенностей и вычисляется численно. Какой должна быть подынтегральная функция при выделении несобственного интеграла?

$$a)\quad \frac{3}{x}, \qquad b)\quad \frac{1}{3x}, \qquad c)\quad \frac{2}{\sqrt[3]{x}},$$

$$d)\quad \frac{1}{2\sqrt[3]{x}} \qquad e)\quad (1+\cos x)^{-1/3}, \qquad f)\quad \frac{1}{\sqrt[3]{2x}}.$$

5 РЕШЕНИЕ НЕЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ И СИСТЕМ

Эта глава будет посвящена методам решения нелинейных алгебраических уравнений и систем. Нас будет интересовать вопрос о нахождении корней уравнения вида $f(x) = 0$ и системы уравнений

$$\begin{cases} f_1(x_1, \ldots, x_m) = 0, \\ \ldots\ldots\ldots\ldots\ldots \\ f_m(x_1, \ldots, x_m) = 0. \end{cases}$$

Если положить $\mathbf{x} = (x_1, \ldots, x_m)$ и $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, то можно представить систему в виде $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

Между линейными и нелинейными уравнениями существуют некоторые фундаментальные различия. Прежде всего, любая невырожденная система линейных уравнений имеет единственное решение. Для нелинейных уравнений это не так. Вещественных решение может не существовать, как в уравнении $e^x + 1 = 0$, или их может быть много, как в уравнении $\sin x = 0$.

Кроме того, любая невырожденная система линейных уравнений может быть решена с затратой конечного числа арифметических операций. Только некоторые нелинейные уравнения могут быть решены точно, например, квадратное уравнение. Даже для таких простых функций как многочлены Э. Галуа доказано, что если использовать обычные арифметические операции и операции извлечения корня, то невозможно записать формулу для нахождения корней произвольного многочлена степени 5 и выше.

Для систем нелинейных уравнений ситуация оказывается еще более сложной. Так как в общем случае не существует алгоритма нахождения решения за конечное число шагов, приходится ограничиваться приближенными решениями. Поэтому методы решения задачи будут итерационными.

Для одного уравнения зачастую можно будет довольно просто определить наличие решения уравнения на некотором промежутке и установить сходимость итераций. Однако гарантировать существование решения или сходимость к решению для системы уравнений намного сложнее. Поэтому нас, как правило, будет интересовать только поиск некоторого решения системы уравнений. Нахождение всех решений системы нелинейных алгебраических уравнений до сих пор остается предметом исследования математиков.

При оценке эффективности алгоритмов решения нелинейных уравнений обычно предполагают, что затраты на вычисление значения функции $f(x)$ велики. Поэтому делается все возможное, чтобы сократить количество требуемых вычислений значения функции.

5.1 РЕШЕНИЕ ОДНОГО АЛГЕБРАИЧЕСКОГО УРАВНЕНИЯ

В этом параграфе речь пойдет о решении уравнения

$$f(x) = 0, \tag{5.1}$$

где функция $f(x)$ определена и непрерывна на некотором конечном или бесконечном промежутке. Напомним, что точка x^* такая, что $f(x^*) = 0$, называется **корнем** уравнения (5.1) или **нулем** функции $f(x)$. Если в окрестности корня x^* функция $f(x)$ представима в виде $f(x) = (x - x^*)^k f_1(x)$, где k — целое положительное число, а $f_1(x)$ — ограниченная функция, причем $f_1(x^*) \neq 0$, то k называется **кратностью** корня[1]. Корень называется **простым**, если $k = 1$.

Нахождение корней этого уравнения происходит в два этапа:

- проводится отделение корней, то есть выделение областей содержащих только один корень[2];
- уточнение корня с помощью итерационного процесса.

Простой прием отделения корней состоит в том, что отрезок $[a, b]$, на котором определена непрерывная функция $f(x)$, разбивается на части. В точках деления x_n вычисляются значения функции. Если на концах отрезка $[x_n, x_{n+1}]$ функция принимает значения разных знаков, то это означает, что на нем есть нечетное число корней. Если же знаки функции на концах отрезка одинаковы, это означает, что на нем либо нет корней, либо корней четное число (с учетом кратности). Можно затем, разбив $[x_n, x_{n+1}]$ на части с помощью аналогичной процедуры уточнить количество корней на промежутке. Если на концах отрезка $[x_n, x_{n+1}]$ функция принимает значения разного знака и, кроме того, известно, что на этом отрезке она монотонная, то в этом случае можно гарантировать, что корень единственный, однако он может быть кратным. Например, у функции $f(x) = x^3$ ноль является корнем кратности 3. Как известно, в случае дифференцируемой функции о монотонности можно судить по первой производной. В случае, если производная не меняет знак на некотором промежутке, функция монотонна, причем, если $f'(x) > 0$ или же $f'(x) < 0$, то корень заведомо единственный и простой. Поэтому, в том случае, когда корни уравнения $f'(x) = 0$ легко находятся, их удобно взять в качестве точек деления x_n. Тогда на отрезке $[x_n, x_{n+1}]$ функция будет заведомо монотонна.

Сам же отрезок $[a, b]$ либо бывает задан, либо зачастую определяется из постановки задачи. Например, если x — концентрация вещества, то ее значение может лежать только в промежутке $[0, 1]$.

Другой метод определения корней — графический. Построение графиков зачастую позволяет выявить даже корни четной кратности.

Иногда удается заменить уравнение (5.1) эквивалентным ему уравнением $f_1(x) = f_2(x)$, в котором функции $y = f_1(x)$ и $y = f_2(x)$ имеют несложные графики. Абсциссы

[1] Для гладких функций используют другое, эквивалентное определение кратного корня: число x^* является корнем кратности k функции $f(x)$, если

$$f(x^*) = f'(x^*) = \cdots = f^{(k-1)}(x^*) = 0, \quad f^{(k)}(x^*) \neq 0.$$

[2] Предполагается, что функция такова, что для каждого корня есть интервал, содержащий только этот корень.

точек пересечения этих графиков будут корнями исходного уравнения. Например, уравнение $x \sin x - 1 = 0$ удобно преобразовать к виду $\sin x = 1/x$, не забыв при этом проанализировать случай $x = 0$.

Далее будем предполагать, что процедура отделения корней произведена и требуется только уточнить корень.

5.1.1 Метод деления отрезка пополам

Указанный метод зачастую называют **методом бисекции** или **методом дихотомии**. Он является одним из наиболее простых и надежных. От функции $y = f(x)$ достаточно потребовать только непрерывность.

Поскольку отделение корня произошло, будем считать, что на концах отрезка $[a, b]$ функция принимает значения разных знаков[3] и, значит, корень находится внутри отрезка. Ищется середина отрезка, то есть точка $c = a + (b - a)/2$ [4] и вычисляется $f(c)$. Из двух отрезков $[a, c]$ и $[c, b]$ оставляют тот, на котором функция принимает на концах значения разных знаков. После этого к выбранному отрезку применяют процесс деления пополам. Итерации заканчивают либо когда длина отрезка станет меньше заданной абсолютной погрешности ε_1, либо когда значение функции при вычислении его в середине отрезка станет меньше чем ε_2. В первом случае в качестве решения может быть выбрана любая точка отрезка, так как мы гарантированы, что расстояние от корня до любой точки отрезка меньше его длины. Во втором случае остановка означает, что середина отрезка случайно оказалась нулем функции, то есть решение найдено.

Каждая итерация уменьшает длину отрезка вдвое. Поэтому легко оценить максимальное число итераций m, которое необходимо совершить для достижения заданной точности ε_1. Должны иметь $2^{-m}(b - a) < \varepsilon_1$, откуда следует $m > \log_2(b - a) - \log_2 \varepsilon_1$.

Для сравнения различных методов введем понятие **скорости сходимости метода**

Пусть x^* — корень уравнения (5.1), а x_i — приближение к корню на i-ой итерации.

Определение 5.1.1 *Будем говорить, что метод сходится со скоростью r, если*

$$\lim_{i \to \infty} \frac{|x^* - x_{i+1}|}{|x^* - x_i|^r} = C,$$

где C — некоторая ненулевая константа.

Если $r = 1$, то скорость сходимости называют **линейной**, *если $r = 2$ —* **квадратичной**.

Для метода деления отрезка пополам $r = 1$, $C = 1/2$.

Зачастую скорость сходимости r имеет более важное значение, чем константа C, методы с большими значениями r привлекательнее. Тем не менее, константа C также может быть важна. Если скорость сходимости линейна и $C \geq 1$, то нет никакой гарантии, что $|x^* - x_i| \to 0$, то есть метод может не сойтись.

[3] Способ проверки изменения знака функции $f(x)$ на концах отрезка $[a, b]$, состоящий в проверке условия $f(a)f(b) < 0$ может оказаться неприемлемым. Если $f(a)$ и $f(b)$ малы, то $f(a)f(b)$ может оказаться машинным нулем. Поэтому проверку лучше выполнять следующим образом $(f(a)/|f(a)|)f(b) < 0$.

[4] В параграфе 1.4 объяснялось почему для нахождения середины отрезка лучше выбрать эту формулу.

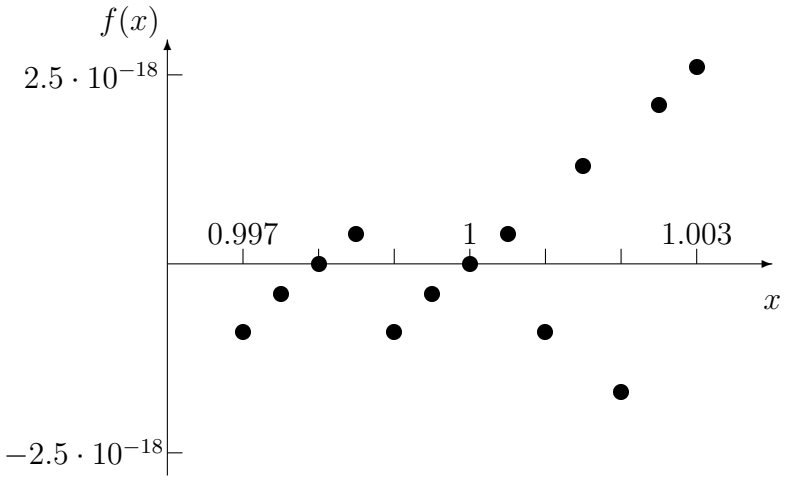


Рис. 5.1 Значения функции $f(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$

Метод деления отрезка пополам прост и надежен. Он применим к любой непрерывной и не обязательно дифференцируемой функции $f(x)$. Метод устойчив к ошибкам округления. Однако скорость сходимости метода невелика, за одну итерацию точность увеличивается примерно вдвое. Это означает, что для уточнение трех цифр потребует 10 итераций, так как $2^{10} = 1024$. Поэтому его можно рекомендовать в том случае, когда значений функции $f(x)$ вычисляется просто. Метод неприменим для вычисления корней четной кратности.

Все приведенные выше утверждения сделаны в предположении о том, что знак величины $f(x_i)$ определяется правильно. Однако, это выполняется не всегда. На рисунке 5.1 изображены значения функции

$$f(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1 = (x-1)^7,$$

полученные по программе, написанной на языке C, в которой переменные имеют тип long double. При таком задании типа переменных величина $\varepsilon_{\text{маш}} \approx 5.4 \cdot 10^{-20}$. Очевидно, что при $x < 1$ значения функции должны быть отрицательными, а при $x > 1$ — положительными.

До тех пор, пока ошибки округления не влияют на определение знака функции, метод нечувствителен к погрешностям вычислений. Если же знак $f(x_i)$ оказывается неверным, то следующий отрезок выбирается неправильно и оценка $|x^* - x_i| \leq (b-a)/2^i$ перестает быть справедливой. Ясно, что если максимум абсолютной ошибки, допускаемой при вычислении $f(x)$ в произвольной точке отрезка $[a, b]$ есть Δ, то знак f будет определяться правильно до тех пор, пока выполняется условие $|f(x)| > \Delta$. Так как вблизи корня x^* значения функции близки к нулю, вблизи корня существует **интервал неопределенности** $(x^* - \delta, x^* + \delta)$, в котором знак функции может определяться неправильно. Когда приближения попадают в этот интервал, дальнейшее их продвижение к корню становится весьма проблематичным. Установить интервал неопределенности заранее бывает крайне сложно. Этот интервал зависит от неизвестного значения корня, от того, насколько пологим является график функции в окрестности корня, от величины погрешности, допускаемой при вычислении значений функции. Попадание в этот интервал можно установить в ходе расчетов на основе анализа поведения итераций. Когда они начинают вести себя неустойчиво, нет никакого смысла продолжать вычисления дальше.

Тот факт, что знак функции вблизи корня может определяться неверно влияет

не только на метод деления отрезка пополам, но и на другие методы, которые будут изучаться ниже.

5.1.2 Метод простой итерации

Метод простой итерации состоит в том, что (5.1) заменяется равносильным уравнением

$$x = S(x) \tag{5.2}$$

задается начальное приближение x_0 и затем итерации образуются по правилу

$$x_{n+1} = S(x_n), \quad n = 0, 1, \ldots.$$

Рассмотрения вопрос о сходимости метода. Напомним, что функция $S(x)$ называется **липшиц-непрерывной** с постоянной q на множестве X, если для любых $x', x'' \in X$ выполняется неравенство

$$|S(x') - S(x'')| \le q|x' - x''|.$$

Теорема 5.1.1 *Пусть функция $S(x)$ липшиц-непрерывна с постоянной $q \in (0, 1)$ на отрезке $[a - \delta, a + \delta]$, причем*

$$|S(a) - a| \le (1 - q)\delta. \tag{5.3}$$

Тогда уравнение (5.2) имеет на этом отрезке единственное решение x^* и метод простой итерации сходится к решению x^* при любом начальном приближении $x_0 \in [a - \delta, a + \delta]$.

Доказательство. Покажем сначала, что если точка $x \in [a - \delta, a + \delta]$, то $S(x) \in [a-\delta, a+\delta]$. Действительно, учитывая условие липшиц-непрерывности и неравенство (5.3), имеем

$$|S(x) - a| \le |S(x) - S(a)| + |S(a) - a| \le q|x - a| + (1 - q)\delta \le q\delta + (1 - q)\delta = \delta,$$

то есть $S(x) \in [a - \delta, a + \delta]$.

Введем на множестве точек отрезка $[a - \delta, a + \delta]$ расстояние по формуле $\rho(x, y) = |x - y|$. В результате получим полное метрическое пространство. Если теперь определить в этом пространстве оператор $Ux = S(x)$, то из доказанного выше следует, что значения оператора лежат в этом же пространстве. Условие липшиц-непрерывности означает, что оператор сжимающий. Тогда утверждение теоремы следует из теоремы Банаха о неподвижной точке (см. [20]).

Замечание 1. Условие липшиц-непрерывности функции $S(x)$ с константой $q \in (0, 1)$ будет выполнено, если функция $S(x)$ дифференцируема на отрезке $[a-\delta, a+\delta]$ и

$$|S'(x)| \le q < 1 \ \text{ при } x \in [a - \delta, a + \delta].$$

Это утверждение следует из того, что согласно теореме Лагранжа о конечных приращениях $|S(x\,') - S(x\,'')| = |S'(\xi)||x\,' - x\,''|$, где ξ — некоторая точка из $[a-\delta, a+\delta]$.

Следствие 1. Пусть уравнение (5.2) имеет решение x^*, функция $S(x)$ непрерывно дифференцируема на отрезке $[x^* - \delta, x^* + \delta]$ и $|S'(x^*)| < 1$. Тогда существует $\epsilon > 0$ такое, что на отрезке $[x^* - \epsilon, x^* + \epsilon]$ уравнение (5.2) не имеет других решений и метод последовательных приближений сходится, если только $x_0 \in [x^* - \epsilon, x^* + \epsilon]$.

Доказательство. Так как $S(x)$ непрерывно дифференцируема и $|S'(x^*)| < 1$, найдутся числа $\epsilon \in (0, \delta)$ и $q \in (0, 1)$ такие, что

$$|S'(x)| \le q < 1 \quad \text{для всех} \quad x \in [x^* - \epsilon, x^* + \epsilon].$$

Так как на отрезке $[x^* - \epsilon, x^* + \epsilon]$ выполнены все условия теоремы, утверждение доказано.

Следствие 2. При выполнении условий теоремы

$$|x_n - x^*| \le q^n |x_0 - x^*|, \quad n = 0, 1, \ldots$$

Доказательство. Утверждение следует из соотношений

$$|x_{n+1} - x^*| = |S(x_n) - S(x^*)| \le q|x_n - x^*|.$$

Следствие 3. Пусть ε — относительная погрешность, с которой ищется корень уравнения (5.2), то есть $|x_n - x^*| \le \varepsilon |x_0 - x^*|$. Тогда в силу следствия 2 заключаем, что заданная точность будет достигнута, если $n > \dfrac{\ln \varepsilon}{\ln q}$.

Следствие 4. Если условия теоремы выполнены, то справедлива оценка $|x^* - x_n| \le \dfrac{q}{1-q}|x_n - x_{n-1}|$.

Доказательство. Имеем

$$|x^* - x_n| = |S(x^*) - S(x_{n-1})| \le q|x^* - x_{n-1}| \le q|x^* - x_n| + q|x_n - x_{n-1}|.$$

Отсюда следует требуемое неравенство.

Для определения скорости сходимости метода простой итерации заметим, что

$$C = \lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|} = \lim_{n\to\infty} \frac{|S(x_n) - S(x^*)|}{|x_n - x^*|} = |S'(x^*)|.$$

Следовательно, если $S'(x^*) \neq 0$, скорость сходимости линейная.

На рисунке 5.2 показано как ведут себя члены последовательности $x_0, x_1, \ldots$ при $|S'(x)| < 1$ (рисунки a), b)) и при $|S'(x)| > 1$ (рисунки c), d)). Заметим, что при $0 < S'(x) < 1$ последовательность сходится к корню монотонно и, следовательно, сходимость будет **односторонней**, то есть все приближения к корню расположены с одной стороны. При $-1 < S'(x) < 0$ сходимость к корню является **двусторонней**, то есть при всех значениях n корень x^* лежит между двумя соседними приближениями x_n и x_{n+1}.

Для сходимости метода большое значение имеет выбор функции $S(x)$. Из следствия 2 можно сделать вывод, что чем величина q меньше, тем быстрее сходится метод. При неудачном выборе функции $S(x)$ метод вообще может не сойтись. Например, поставим задачу найти наибольший положительный корень уравнения

$$x^3 + x = 1000.$$

Грубой прикидкой получаем, что приближенное значение корня $x_0 = 10$, причем корень $x^* < x_0$. Если переписать уравнение в виде $x = 1000 - x^3$, то метод не сойдется так как $S(x) = 1000 - x^3$, $S'(x) = -3x^2$ и в окрестности корня $|S'| \gg 1$. Записав уравнение в виде $x = \sqrt[3]{1000 - x}$ и взяв интервал $[9, 10]$, на котором находится корень, получим $S' = -1/3 \cdot (1000 - x)^{-2/3}$. Тогда на выбранном отрезке

$$|S'(x)| \le \frac{1}{3} 990^{-2/3} \approx \frac{1}{300}.$$

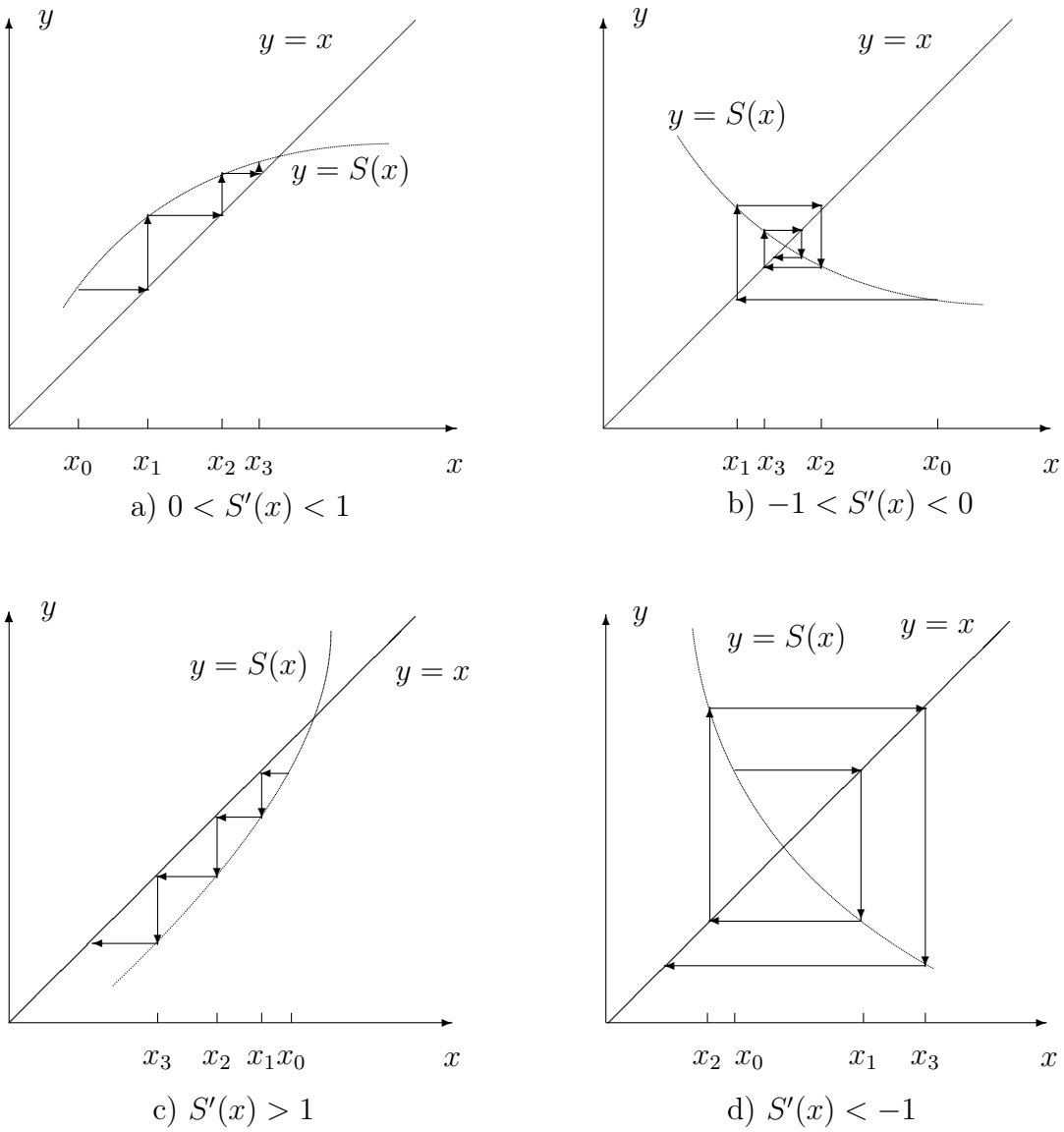


Рис. 5.2 Иллюстрация сходимости $a), b)$ и расходимости $c), d)$ метода простой итерации

Значит $q \ll 1$ и сходимость очень быстрая.

В общем случае не существует рекомендаций по преобразованию уравнения (5.1) в уравнение (5.2). В некоторых случаях получить необходимое представление (5.2) удается путем использования обратных функций. Поясним это на примере. Пусто необходимо получить корень уравнения $2x\cos x - \sin x = 0$, лежащий в интервале $(\pi/4, \pi/2)$. Если переписать уравнение в виде $x = 0.5\,\mathrm{tg}\,x$, получим $S(x) = 0.5\,\mathrm{tg}\,x$ и $S'(x) = 1/(2\cos^2 x) > 1$. Выбирая же обратную к тангенсу функцию арктангенс, уравнение перепишем в виде $x = \mathrm{arctg}\,2x$. В этом случае $S(x) = \mathrm{arctg}\,2x$ и $S'(x) = 2/(1+4x^2) < 1$.

Можно определить функцию $S(x)$ по формуле $S(x) = x + \tau(x)f(x)$, причем функция $\tau(x)$ такова, что не меняет знак на отрезке, где ищется корень. Если, например, $\tau(x) = \tau = const$, то получается метод **релаксации**

$$x_{n+1} = x_n + \tau f(x_n), \tag{5.4}$$

при этом параметр τ называют **параметром релаксации**.

Условие $|S'(x)| < 1$ означает, что

$$-1 < 1 + \tau f'(x) < 1$$

отсюда следует, что метод сходится, если $-2 < \tau f'(x) < 0$.

Если, например, в окрестности корня $f'(x) < 0,\ 0 < m_1 \le |f'(x)| \le M_1$, то метод релаксации сходится, если $0 < \tau < 2/M_1$.

Для определения оптимального значения параметра τ, обеспечивающего максимальную скорость сходимости, запишем уравнение для погрешности $y_n = x_n - x^*$ и подберем τ так, чтобы эта погрешность как можно быстрее сходилась к нулю. С этой целью подставляем $x_n = y_n + x^*$ в равенство (5.4)

$$y_{n+1} = y_n + \tau f(x^* + y_n) = y_n + \tau(f(x^*) + y_n f'(x^* + \theta y_n)), \quad |\theta| < 1.$$

Так как $f(x^*) = 0$,

$$y_{n+1} = (1 + \tau f'(x^* + \theta y_n))y_n.$$

Отсюда следует

$$|y_{n+1}| \le \max_x |1 + \tau f'(x)||y_n| \le \max(|1 - \tau m_1|, |1 - \tau M_1|)|y_n|.$$

Параметр τ следует выбрать таким образом, чтобы функция $g(\tau) = \max(|1 - \tau m_1|, |1 - \tau M_1|)$ принимала минимальное значение. Аналогично тому, как это было сделано в пункте 2.2.2 заключаем, что минимум достигается, если

$$|1 - \tau M_1| = |1 - \tau m_1|.$$

При этом $\tau_o = \dfrac{2}{M_1 + m_1}$ и

$$|y_{n+1}| \le \frac{M_1 - m_1}{M_1 + m_1}|y_n|.$$

Метод простых итераций обладает рядом важных преимуществ. Он позволяет искать комплексные корни, если задавать комплексное начальное приближение. Метод не является чувствительным к ошибкам округления, так как ошибочное значение может рассматриваться как новое начальное приближение. При удачном подборе уравнения (5.2) удается получить более быструю сходимость по сравнению с методом деления отрезка пополам. В отличие от рассматриваемого ниже метода Ньютона

не требуется вычислять значение производной функции $f(x)$. К недостаткам метода следует отнести, прежде всего, трудность определения момента достижения заданной точности в общем случае.

Вопрос о практическом определении количества итераций, необходимых для достижения заданной точности этого и других методов будет рассмотрен ниже.

5.1.3 Метод Ньютона

Рассматриваемый ниже метод часто называют **методом касательных** или **методом линеаризации**. Пусть x_n - приближенное значение корня x^* уравнения (5.1) и функция $f(x)$ непрерывно дифференцируема. Тогда можно записать

$$0 = f(x^*) = f(x_n) + (x^* - x_n)f'(\xi),$$

где ξ лежит между x^* и x_n. Заменяя $f'(\xi)$ на значение производной в известной точке x_n, получим приближение x_{n+1} корня x^*. Выражая тогда x_{n+1}, имеем следующий итерационный процесс:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \ldots. \tag{5.5}$$

Геометрически (см. рисунок 5.3), x_{n+1} — абсцисса точки пересечения касательной, проведенной в точке $(x_n, f(x_n))$, с осью Ox.

Сделаем ряд дополнительных предположений относительно функции $f(x)$. Будем считать, что

- x^* — простой корень, то есть $f'(x^*) \neq 0$;
- $f(x)$ — дважды непрерывно дифференцируемая в окрестности корня функция.

Метод Ньютона можно рассматривать как частный случай метода простой итерации с $S(x) = x - f(x)/f'(x)$. Так как

$$S'(x) = 1 - \frac{f'^2(x) - f(x)f''(x)}{f'^2(x)} = \frac{f(x)f''(x)}{f'^2(x)},$$

то $S'(x^*) = 0$. Поэтому в малой окрестности корня $|S'(x)| < 1$ и, значит, метод сходится.

Установим характер сходимости и правило выбора начального приближения.

Теорема 5.1.2 *Пусть функция $f(x)$ дважды непрерывно дифференцируемая на отрезке $[a, b]$, имеет внутри отрезка один корень x^*, монотонна и $f''(x)$ не меняет знак на $[a, b]$. Пусть кроме того, x_0 совпадает с тем концом отрезка, где знаки функции и второй производной совпадают. Тогда последовательность $x_0, x_1, \ldots$, определенная согласно (5.5), монотонно сходится к x^*.*

Доказательство. Предположим для определенности, что $f(b) > 0, f''(b) > 0$. Монотонность последовательных приближений докажем по индукции. Пусть для некоторого x_n выполняется неравенство $x^* < x_n < b$. Покажем тогда, что $x^* < x_{n+1} < x_n$.

Действительно, перепишем (5.5) в виде

$$x_n - x_{n+1} = \frac{f(x_n) - f(x^*)}{f'(x_n)} = \frac{(x_n - x^*)f'(\xi_n)}{f'(x_n)}, \tag{5.6}$$

где $\xi_n \in (x^*, x_n)$. Так как по предположению f'' не меняет знак и $f''(b) > 0$, то $f''(x) > 0$ всюду на $[a, b]$. Отсюда следует, что $f'(x)$ функция возрастающая и, значит, $f'(x_n) > f'(\xi_n)$. Кроме того, $f(x)$ — монотонна и имеет внутри отрезка один корень, значит $f(a) < 0, f(b) > 0$. Следовательно, $f'(x) > 0$ и

$$0 < \frac{f'(\xi_n)}{f'(x_n)} < 1.$$

Тогда из (5.6) получим $0 < x_n - x_{n+1} < x_n - x^*$, откуда следует требуемое утверждение.

Покажем теперь, что последовательные приближения сходятся к корню x^*. Последовательность $x_n,\ n = 0, 1, \ldots$ монотонно убывает и ограничена снизу x^*. Значит она имеет предел. Пусть $\bar{x} = \lim\limits_{n\to\infty} x_n$. Переходя к пределу в (5.5), получим

$$\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}.$$

Так как $f'(x) > 0$, отсюда, следует, что $f(\bar{x}) = 0$, то есть $\bar{x}$ - корень. Но по условию теоремы на $[a, b]$ только один корень, который равен x^*. Значит $\bar{x} = x^*$. Теорема доказана.

Замечание 1. Если выбрать не тот конец отрезка, то метод может не сойтись. На рисунке 5.3 касательная, проведенная через точку $(a, f(a))$, пересекает ось абсцисс вне отрезка $[a, b]$.

Замечание 2. Если на $[a, b]$ нет монотонности, либо $f''(x)$ меняет знак, то есть нет строгой выпуклости или вогнутости, то метод может не сойтись см. рисунок 5.4, из которого следует, что метод "зацикливается".

Замечание 3. В связи с тем, что последовательные приближения сходятся монотонно, можно предложить следующий критерий остановки, гарантирующий нахождение корня с заданной точностью ε. Вычисления следует продолжать до тех пор, пока $f(x_n)$ и $f(x_n \pm \varepsilon)$ (знак плюс выбирается если последовательность возрастает, а минус — если убывает) имеют одинаковые знаки. Однако такой подход требует дополнительных вычисление значений функции $f(x)$.

Замечание 3. В параграфе 5.1.7 будет показано, что при достаточно малом $\varepsilon > 0$ в качестве критерия остановки вычислений может быть выбрано условие выполнения неравенства $|x_{n-1} - x_n| < \varepsilon$.

Оценим теперь скорость сходимости вблизи простого корня. Перепишем (5.5) в виде

$$x_{n+1} - x^* = \Big(x_n - \frac{f(x_n)}{f'(x_n)}\Big) - \Big(x^* - \frac{f(x^*)}{f'(x^*)}\Big) = S(x_n) - S(x^*) =$$
$$= (x_n - x^*)S'(x^*) + \frac{1}{2}(x_n - x^*)^2 S''(\xi_n),$$

ξ_n лежит между x_k и x^*. Как было показано выше $S'(x^*) = 0$. Поэтому,

$$\lim_{k\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^2} = \frac{1}{2}|S''(x^*)|.$$

Таким образом, вблизи простого корня скорость сходимости квадратичная.

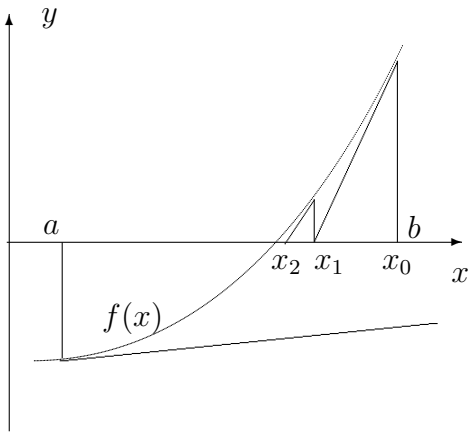


Рис. 5.3 Метод Ньютона

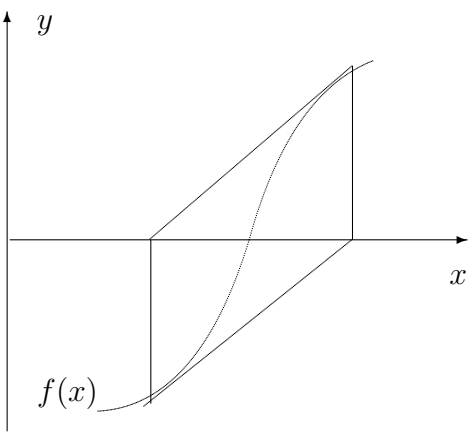


Рис. 5.4 "Зацикливание"метода Ньютона

Предположим теперь, что корень x^* имеет кратность $k > 1$. Тогда $f'(x^*) = 0$, поэтому знаменатель в формуле (5.5) стремится к нулю при $x_k \to x^*$. Однако

$$\lim_{x\to x^*} \frac{f(x)}{f'(x)} = \lim_{x\to x^*} \frac{(x-x^*)^k f_1(x)}{k(x-x^*)^{k-1} f_1(x) + (x-x^*)^k f_1{}'(x)} = 0.$$

Значит в окрестности корня поправка к приближения x_n в формуле (5.5) мала и формула имеет смысл. Можно показать (см. параграф 5.3.1 пример 2), что метод сходится, но, в отличие от простого корня, скорость сходимости линейная. Однако, легко модифицировать метод Ньютона, записав его в виде

$$x_{n+1} = x_n - k\frac{f(x_n)}{f'(x_n)},$$

чтобы скорость сходимости была квадратичной (см. параграф 5.3.1 пример 3).

Основным преимуществом метода Ньютона является его быстрая сходимость вблизи простого корня. Как и в методе итераций ошибка вычислений не накапливается, можно находить комплексные корни при задании комплексного начального приближения. Сходимость сохраняется в случае кратного корня, однако, не такая быстрая.

Одним из недостатков метода Ньютона является необходимость вычисления $f'(x)$. Это может потребовать много времени, оказаться трудным или даже невыполнимым делом, если, например, функция не задана аналитически. Преодолеть указанный недостаток можно, заменив производную конечными разностями, как это описано в предыдущей главе. Иной путь состоит в использовании описанного в следующем параграфе метода, который не требует вычисления производных.

5.1.4 Метод секущих

Заменим в методе Ньютона производную $f'(x_n)$ разделенной разностью, найденной по двум последним итерациям, то есть по формуле

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

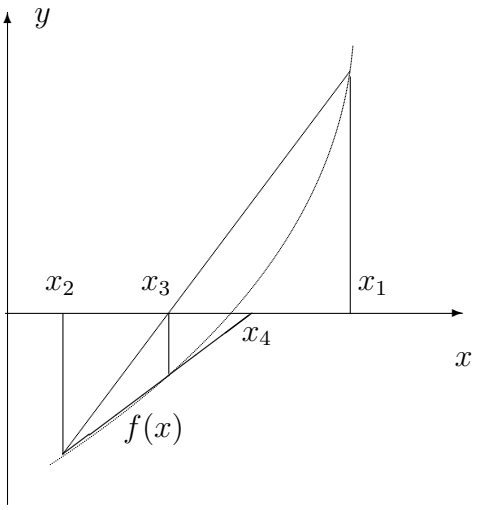


Рис. 5.5 Метод секущих

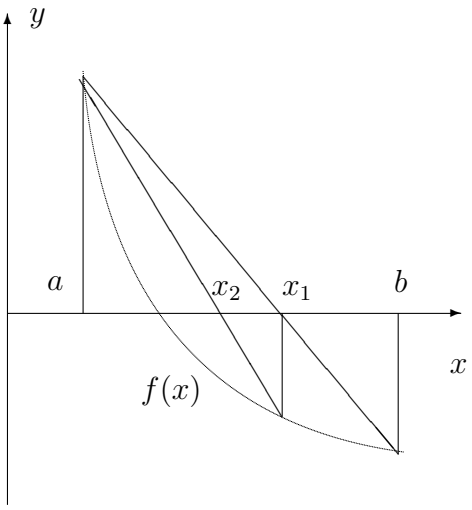


Рис. 5.6 Метод хорд

В результате получим следующий итерационный процесс, названный **методом секущих**:

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})f(x_n)}{f(x_n) - f(x_{n-1})}. \tag{5.7}$$

Геометрически (см. рисунок 5.5) формула (5.7) означает, что за очередное приближение выбирается абсцисса точки пересечения с осью Ox прямой, проходящей через точки $(x_{n-1}, f(x_{n-1}))$ и $(x_n, f(x_n))$. Иначе говоря, функция $f(x)$ интерполируется многочленом первой степени с узлами интерполяции x_n, x_{n-1} и за очередное приближение выбирается корень этого многочлена.

Для начала вычислений помимо начальной точки x_0 надо задать еще x_1. Такие процессы, где для вычисления очередного приближения надо знать два предыдущих называют **двухшаговыми**.

Проведенное изменение метода Ньютона влияет на характер сходимости. При определенном задании начальных точек x_0, x_1 метод может не сойтись, может быть не монотонным как это видно на рисунке. Доказано, что скорость сходимости метода секущих $r = 0.5(1 + \sqrt{5}) \approx 1.618$. При этом константа $C = |f''(x^*)/(2f'(x^*)|^{1/r}$. Одна итерация метода секущих требует одного вычисления функции f, а метода Ньютона двух вычислений (f и f'). Поэтому, в некоторых случаях трудоемкость одной итерации метода секущих может приблизительно оказать равной трудоемкости двух итераций метода Ньютона. Две итерации метода секущих дают скорость сходимости равную $r^2 \approx 2.618$. Таким образом, при трудоемком вычислении функций f и f' его можно рассматривать как более быстрый по сравнению с методом Ньютона.

В знаменателе формулы (5.7) стоит разность значений функции. Вдали от корня это несущественно, но вблизи корня значения функции малы и очень близки, поэтому может возникнуть потеря значащих цифр. Это особенно сказывается вблизи кратного корня. Заметим, что приводить к общему знаменателю в формуле (5.7) не следует, так как в этом случае возрастает потеря точности в расчетах.

5.1.5 Метод хорд

Пусть отрезок $[a, b]$ содержит единственный корень x^* уравнения (5.1). Заменим дугу $y = f(x)$ хордой, проходящей через точки $(a, f(a)), (b, f(b))$. В качестве очередного приближения берется точка пересечения хорды с осью абсцисс. Так как уравнение прямой, проходящей через заданные точки имеет вид

$$\frac{x-a}{b-a} = \frac{y-f(a)}{f(b)-f(a)},$$

получим, что $y = 0$ при

$$x = a - \frac{f(a)(b-a)}{f(b)-f(a)}. \tag{5.8}$$

Если записать уравнение прямой в виде

$$\frac{x-b}{a-b} = \frac{y-f(b)}{f(a)-f(b)},$$

то получим

$$x = b - \frac{f(b)(b-a)}{f(b)-f(a)}. \tag{5.9}$$

Выбирая из двух отрезков тот, где функция меняет знак, повторяем процесс.

Для доказательства сходимости будем считать, что $f''(x)$ не меняет знак на $[a, b]$. Пусть, например, $f''(x) > 0$ и $f(a) > 0$ (см. рисунок 5.6). Тогда кривая лежит под хордой, конец a будет неподвижен, а последовательность x_n будет монотонно убывать. В этом случае алгоритм удобно записать, воспользовавшись формулой (5.9),

$$x_{n+1} = x_n - \frac{f(x_n)(x_n-a)}{f(x_n)-f(a)}, \quad x_0 = b. \tag{5.10}$$

Если же $f''(x) > 0$ и $f(a) < 0$, то неподвижная точка — b и на основании (5.8) алгоритм запишем в виде

$$x_{n+1} = x_n - \frac{f(x_n)(x_n-b)}{f(x_n)-f(b)} \quad x_0 = a. \tag{5.11}$$

В результате получаем монотонно возрастающую последовательность приближений x_n. Аналогичные рассуждения можно провести, если $f''(x) < 0$.

В любом случае неподвижен тот конец, для которого совпадают знаки функций и второй производной (сравните с методом Ньютона). Последовательные приближения лежат по ту сторону корня, где $f(x)$ имеет знак противоположный знаку $f''(x)$. Так как последовательность приближений монотонна и ограничена сверху или снизу, то она сходится к некоторому $\bar{x}$. Переходя к пределу в соответствующем равенстве, например, в (5.11) получим

$$\bar{x} = \bar{x} - \frac{f(\bar{x})(\bar{x}-b)}{f(\bar{x})-f(b)}.$$

Отсюда следует, что $f(\bar{x}) = 0$, то есть $\bar{x}$ — корень. Так как по предположению на отрезке $[a, b]$ только один корень x^*, получаем, что $\bar{x} = x^*$.

Так как метод хорд и касательных сходится к корню с разных сторон, иногда используют так называемый **комбинированный метод**, когда применяется одновременно метод хорд и касательных: например, при $f'(x) > 0$, $f''(x) > 0$ на $[a, b]$,

расчетные формулы имеют вид

$$x_0 = a,\ \bar{x}_0 = b,\ x_{n+1} = x_n - \frac{f(x_n)}{f(\bar{x}_n) - f(x_n)}(\bar{x}_n - x_n),\ \bar{x}_{n+1} = \bar{x}_n - \frac{f(\bar{x}_n)}{f'(\bar{x}_n)}.$$

Тогда $x_n < x^* < \bar{x}_n$ и итерации проводятся до тех пор, пока $\bar{x}_k - x_n \geq \varepsilon$.

Все остальные случаи: $f'(x) > 0,\ f''(x) < 0;\ f'(x) < 0,\ f''(x) > 0;\ f'(x) < 0,\ f''(x) < 0$ сводятся к уже рассмотренному путем замены исходного уравнения на уравнение $-f(x) = 0$ или же $\pm f(-z) = 0$, где $z = -x$.

Основное преимущество комбинированного метода — двусторонняя сходимость как и у метода деления отрезка пополам. По сравнению сравнению с методом деления отрезка пополам комбинированный метод сходится быстрее.

5.1.6 Метод парабол (квадратичной интерполяции)

Пусть известны три приближения x_n, x_{n-1}, x_{n-2} к корню уравнения $f(x) = 0$. Построим интерполяционный многочлен второй степени с узлами в этих точках и в качестве нового приближения к корню выберем тот из корней этого многочлена, который ближе к точке x_n. Многочлен в форме Ньютона имеет вид:

$$P_2(x) = f(x_n) + (x - x_n)f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1})f(x_n, x_{n-1}, x_{n-2}).$$

Пусть $z = x - x_n,\ a = f(x_n, x_{n-1}, x_{n-2}),\ b = a(x_n - x_{n-1}) + f(x_n, x_{n-1}),\ c = f(x_n)$, Тогда уравнение запишется в виде $az^2 + bz + c = 0$. Тот из двух корней этого квадратного уравнения, который меньше по модулю, и определяет новое приближение, которое равно $x_{n+1} = x_n + z$.

Так как для начала расчетов надо задавать три точки x_0, x_1, x_2, то такой метод называется **трехшаговым.**

Можно показать, что для простого корня x^* выполняется равенство

$$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^{1.84}} = \left|\frac{f'''(x^*)}{6f'(x^*)}\right|^{0.42}.$$

Таким образом, метод сходится быстрее, чем метод секущих и медленнее метода Ньютона.

Достоинство метода состоит в том, что задавая даже действительные начальные приближения можно сойтись к комплексному корню. Поэтому программа, реализующая метод парабол должна быть написана с учетом комплексной арифметики.

Метод парабол очень эффективен при нахождении корней многочлена высокой степени. Если $f(x)$ — многочлен, то сходимость при любом начальном приближении не доказана, хотя на практике он всегда сходится, причем достаточно быстро. Если корень x^* найден, то $f(x)/(x - x^*)$ — многочлен, уже не содержащий этот корень либо содержит этот корень, но меньшей кратности. Находя и удаляя последовательно корни, сможем найти все корни многочлена. Надо только помнить, что из-за погрешности округлений при делении вносятся ошибки в коэффициенты, а корни многочлена чувствительны к изменению коэффициентов. Практика показывает, что выгоднее начинать исключение с меньших корней. Берут начальное приближение $x_0 = -1,\ x_1 = 1,\ x_2 = 0$. Тогда итерации обычно сходятся к наименьшему корню. Его исключают и затем берут то же приближение, ищут следующий корень и так далее.

В том случае, когда ищутся только действительные корни, можно воспользоваться методом **обратной квадратичной интерполяции**. Он заключается в том, что по трем точкам $(x_{n-2}, y_{n-2}), (x_{n-1}, y_{n-1}), (x_n, y_n)$, где $y_n = f(x_n)$ строится парабола $x = ay^2 + by + c$. Тогда точка x_{n+1} пересечения этой параболы с осью OX легко находится $x_{n+1} = c$.

Так как в этом методе используется только вещественная арифметика и нет нужды извлекать квадратный корень, он требует меньших затрат. Однако, для реализации метода необходимо задание хороших начальных приближений.

5.1.7 Критерии контроля точности и окончания счета

Одной из основных проблем, с которыми приходится сталкиваться при исследовании итерационных методов, является проблема определения момента окончания вычислений. Как уже отмечалось выше, она решается просто для комбинированного метода хорд и касательных, метода деления отрезка пополам. Не возникает проблем и при использовании двустороннего метода, каким является, например, метод простой итерации в случае $-1 < S'(x) < 0$. Так как для двустороннего метода корень x^* всегда лежит между двумя соседними приближениями, итерации проводят до тех пор, пока $|x_{n-1} - x_n| \geq \varepsilon$, где ε — заданная абсолютная погрешность, с которой ищется корень.

Применение критерия достижения заданной точности

$$|x_{n-1} - x_n| < \varepsilon \tag{5.12}$$

для метода, не являющегося двусторонним (Ньютона, хорд, секущих, парабол, простой итерации при $0 < S'(x) < 1$), вообще говоря, не гарантирует выполнения неравенства $|x^* - x_n| < \varepsilon$ и, следовательно, может привести к ошибочному заключению о необходимом количестве итераций. Рассмотрим, например, следующий итерационный процесс

$$x_{n+1} = \frac{x_n}{x_n + 1},\ x_0 = 1,\ n = 0, 1, \ldots.$$

Легко заметить, что в этом случае $x_n = 1/(n+1)$, а $x^* = 0$. Если $\varepsilon = 0.01$, то x_{100} приближает корень с заданной точностью, то есть число итераций равно 100. В то же время использование критерия (5.12) дает $|x_9 - x_{10}| = 1/110 < \varepsilon$, поэтому итерации прекратятся при $n = 10$. Таким образом, будет сделано в 10 раз меньше итераций, чем положено.

Для того, чтобы разобраться в сути дела, воспользуемся неравенством

$$|x^* - x_n| \leq \frac{q}{1-q}|x_n - x_{n-1}|, \tag{5.13}$$

которое было получено в следствии 4 параграфа 5.1.2 для приближений x_n, найденных по методу простой итерации. Из него видно, что оценка (5.12) применима, если дробь $\dfrac{q}{1-q} \leq 1$, то есть $0 \leq q \leq 1/2$. Действительно, в это случае имеем:

$$|x^* - x_n| \leq \frac{q}{1-q}|x_n - x_{n-1}| \leq |x_{n-1} - x_n| < \varepsilon.$$

Если же q близко к 1, то итерационный процесс сходится медленно и в этом случае расстояние между двумя последовательными приближениями не характеризует расстояние между x_n и корнем.

Напомним, что метод Ньютона можно трактовать, как метод простой итерации с функцией $S(x) = x - f(x)/f'(x)$, причем для дважды непрерывно дифференцируемой функции $f(x)$ и простого корня x^* выполняется равенство $S'(x^*) = 0$. Значит, в некоторой окрестности корня выполняется неравенство $|S'(x)| < q = 1/2$. Следовательно, для метода Ньютона критерий остановки вычислительного процесса (5.12) применим при достаточно малом значении ε.

Для прекращения вычислений иногда предлагают другой критерий: $|f(x_n)| < \varepsilon$, то есть проверяют насколько хорошо приближение x_n удовлетворяет уравнению (5.1). Однако и этот критерий следует признать плохим. Величина $|f(x_n)|$ может быть малой, в то время как модуль $|x^* - x_n|$ велик и наоборот, может быть большой величина $|f(x_n)|$, а модуль $|x^* - x_n|$ мал. Этот факт легко объяснить, если учесть, что при $C \neq 0$ уравнения $f(x) = 0$ и $Cf(x) = 0$ эквивалентны. При этом число $|Cf(x_n)|$ можно сделать сколь угодно малым или сколь угодно большим за счет выбора константы C.

Можно предложить следующий, универсальный метод определения момента прекращения итераций при нахождении простого корня или корня нечетной кратности. После нахождения x_n и $f(x_n)$ вычисляются $f(x_n + \varepsilon)$, $f(x_n - \varepsilon)$. Если знак $f(x_n)$ и одного из чисел $f(x_n + \varepsilon)$ или $f(x_n - \varepsilon)$ различны, то $|x^* - x_n| < \varepsilon$ и, следовательно, вычисления следует прекратить. Существенным недостатком этого метода является необходимость на каждой итерации дополнительного вычисления значений функции, что может существенно увеличить объем вычислений.

Покажем, как можно иначе строго оценить точность и, следовательно, определить момент прекращения вычислений.

Справедлива следующая теорема.

Теорема 5.1.3 *Пусть функция $f(x)$ дифференцируема на $[a, b]$, $|f'(x)| \geq m_1 > 0$ и x^* — корень уравнения $f(x) = 0$, лежащий на этом отрезке, а x_n — приближение к корню. Тогда* $|x^* - x_n| \leq \dfrac{|f(x_n)|}{m_1}$.

Доказательство. По теореме Лагранжа $f(x_n) - f(x^*) = f'(\xi)(x_n - x^*)$. Значит $|f(x_n)| \geq m_1|x_n - x^*|$, откуда следует утверждение.

Таким образом, критерием прекращения вычислений может служить выполнение неравенства $\dfrac{|f(x_n)|}{m_1} < \varepsilon$. Однако, применение этой теоремы на практике затруднено в связи с тем, что редко известно значение m_1.

В следствии 3 параграфа 5.1.2 была получена формула для определения числа итераций в методе простой итерации. Однако эта формула применяется редко, поскольку она требует знания величины q. Кроме того, эта оценка, как правило, дает завышенные результаты.

Для получения другого критерия пригодного для метода простых итераций, воспользуемся оценкой (5.13). Напомним, что в этой оценке величина q определена из условия выполнения в окрестности корня неравенства $|S'(x)| \leq q$. Вблизи корня можно считать, что

$$q \approx \frac{S(x_{n-2}) - S(x_{n-1})}{x_{n-2} - x_{n-1}} = \frac{x_{n-1} - x_n}{x_{n-2} - x_{n-1}}.$$

Поэтому, для того, чтобы $|x^* - x_n| < \varepsilon$, достаточно потребовать, чтобы

$$\frac{q}{1-q}|x_n - x_{n-1}| \approx \frac{(x_n - x_{n-1})^2}{2x_{n-1} - x_n - x_{n-2}} < \varepsilon.$$

При выполнении этого условия итерации можно прекращать. Применять этот метод надо с осторожностью, так как он справедлив только в вблизи корня.

Контроль абсолютной погрешности не всегда разумен и возможен, так как не учитывает величину порядка искомого корня. Абсолютная погрешность по существу фиксирует разряд приближенного значения корня. Если задать абсолютную погрешность равной 10^{-3}, а искомый корень равен 1010101.0, причем вычисления проводятся на машине с семью десятичными разрядами, то итерационный процесс просто зациклится.

Относительная погрешность определяет общее количество верных цифр приближенного значения корня. Если корень мал, то относительная оценка для двустороннего метода $|x_{n+1} - x_n| < \varepsilon|x_n|$ также может оказаться неприемлемой, так как в правой части может произойти образование машинного нуля или же правая часть может оказаться меньше расстояния между двумя соседними числами [5]. Поэтому разумным следует признать следующий критерий, который объединяет контроль абсолютной ε_a и относительной ε_o погрешностей

$$|x_{n+1} - x_n| < \varepsilon_a + \varepsilon_o|x_n|. \tag{5.14}$$

На практике для односторонних методов в качестве критерия, для определения окончания вычислений применяется одновременно пара критериев $|f(x_n)| < \varepsilon_a^1$ и (5.14) с достаточно малыми значениями $\varepsilon_a,\ \varepsilon_o,\ \varepsilon_a^1$, хотя, как отмечалось выше, они не гарантируют, что x_n — приближение к корню с требуемой точностью. Кроме того, в программе часто задают еще одно ограничение — максимальное число итераций, ориентировочно требуемое для обеспечения сходимости.

При применении некоторых методов (секущих, парабол) вычитаются два близких значения функции, что может привести к потере значащих цифр и "разболтке" счета. Существует прием, называемый **способом Гаврика**, который позволяет учесть возникающую "разболтку". Выбирается некоторое не очень маленькое число δ и проводят итерации до тех пор, пока $|x_{n+1} - x_n| < \delta$. Затем расчеты продолжают, пока $|x_{n+1} - x_n|$ убывает. Первое возрастание разности двух соседних приближений означает начало "разболтки" и в этом случае итерации прекращают. Следует пояснить, что итерации прежде чем начать сходиться могут в начале расходиться. Поэтому применять с самого начала условие $|x_{n+1} - x_n| > |x_n - x_{n-1}|$ в качестве критерия окончания счета нельзя. По этой причине вычисления до определенного момента проводятся без контроля поведения разности двух соседних итераций.

5.2 РЕШЕНИЕ СИСТЕМЫ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Рассмотрим методы решения системы нелинейных алгебраических уравнений:

$$\begin{cases} f_1(x_1, \ldots, x_m) = 0, \\ \cdots\cdots\cdots\cdots\cdots \\ f_m(x_1, \ldots, x_m) = 0. \end{cases} \tag{5.15}$$

[5] Напомним, что в машинной арифметике представимо только конечное число вещественных чисел.

Как уже отмечалось в начале главы, эту систему будем записывать в матричном виде

$$\mathbf{F}(\mathbf{x}) = \mathbf{0},$$

где $\mathbf{x} = (x_1, \ldots, x_m)^{\mathrm{T}}$ и $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^{\mathrm{T}}$.

Аналогично решению одного уравнения, поиск решения системы начинается с его локализации, то есть определения множества, содержащего одно решение и являющееся его достаточно малой окрестностью. Иногда такое выделение решения удается провести, опираясь на физические, геометрические или иные соображения. Однако, в общем случае это довольно сложная проблема и она считается решенной, если удается подобрать хорошее начальное приближение $\mathbf{x}^0$ к решению $\mathbf{x}^*$ системы.

Второй этап процесса нахождения решения — его уточнение путем применения какого-нибудь итерационного метода. В общем случае **одношаговые итерационные методы** для решения (5.15) имеют вид

$$\mathbf{B}_k \frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau_k} + \mathbf{F}(\mathbf{x}^k) = \mathbf{0}. \tag{5.16}$$

Здесь $k = 0, 1, \ldots$ — номер итерации, начальный вектор $\mathbf{x}^0$ задается произвольным образом, $\mathbf{B}_k$ — матрица, имеющая обратную и τ_k — числовой параметр. Из (5.16) следует, что нахождение $\mathbf{x}^{k+1}$ сводится к решению системы линейных алгебраических уравнений

$$\mathbf{B}_k \mathbf{x}^{k+1} = \mathbf{B}_k \mathbf{x}^k - \tau_k \mathbf{F}(\mathbf{x}^k). \tag{5.17}$$

Метод называется **явным**, если $\mathbf{B}_k$ — единичная матрица и **неявным** в противном случае.

Метод называется **стационарным**, если $\mathbf{B}_k, \tau_k$ не зависят от k.

Система (5.17) в свою очередь может решаться итерациями, которые называют **внутренними**, в отличие от итераций (5.16), называемых **внешними**.

Рассмотрим сначала стационарный метод. Перепишем (5.17) в виде

$$\mathbf{x}^{k+1} = \mathbf{B}^{-1}(\mathbf{B}\mathbf{x}^k - \tau \mathbf{F}(\mathbf{x}^k)). \tag{5.18}$$

Тогда вопрос о сходимости метода может быть решен с помощью теоремы о неподвижной точке оператора $\mathbf{S}$, действующего в пространстве m-мерных векторов. Взяв, например, $\mathbf{B} = \mathbf{E}$, где $\mathbf{E}$ — единичная матрица, получим **метод релаксации**

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau \mathbf{F}(\mathbf{x}^k).$$

Подобно тому, как это было сделано в предыдущем параграфе, можно доказать, что метод сходится, если в окрестности решения $\|\mathbf{S}'(\mathbf{x})\| < 1$, где $\mathbf{S}'(\mathbf{x}) = \mathbf{E} - \tau \mathbf{F}'(\mathbf{x})$, и

$$\mathbf{F}'(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \ldots & \frac{\partial f_1(\mathbf{x})}{\partial x_m} \\ \ldots\ldots\ldots\ldots\ldots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \ldots & \frac{\partial f_m(\mathbf{x})}{\partial x_m} \end{pmatrix}.$$

Метод Ньютона. Предположим, что функции $f_i(\mathbf{x})$ имеют вторые производные в окрестности решения системы уравнений (5.15). Пусть известно приближение $\mathbf{x}^k$ к решению $\mathbf{x}^*$. Тогда, согласно формуле Тейлора

$$f_i(\mathbf{x}^*) = f_i(\mathbf{x}^k) + \sum_{j=1}^{m} \frac{\partial f_i(\mathbf{x}^k)}{\partial x_j}(x_j^* - x_j^k) + O(\|\mathbf{x}^* - \mathbf{x}^k\|^2), \quad i = 1, \ldots, m. \tag{5.19}$$

Отбрасывая величину $O(\|\mathbf{x}^*-\mathbf{x}^k\|^2)$ и учитывая, что x^* — решение системы, получим:

$$\sum_{j=1}^{m} \frac{\partial f_i(\mathbf{x}^k)}{\partial x_j}(x_j^* - x_j^k) + f_i(\mathbf{x}^k) \approx 0, \quad i = 1, \ldots, m.$$

Заменим теперь приближенное равенство на точное, а $\mathbf{x}^*$ на $\mathbf{x}^{k+1}$. Тогда получим систему линейных алгебраических уравнений для нахождения очередного приближения $\mathbf{x}^{k+1}$:

$$\sum_{j=1}^{m} \frac{\partial f_i(\mathbf{x}^k)}{\partial x_j}(x_j^{k+1} - x_j^k) + f_i(\mathbf{x}^k) = 0, \quad i = 1, \ldots, m$$

или в векторном виде

$$\mathbf{F}'(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k) + \mathbf{F}(\mathbf{x}^k) = \mathbf{0}. \tag{5.20}$$

Таким образом, метод Ньютона может быть записан в виде (5.16) с

$$\mathbf{B}_k = \mathbf{F}'(\mathbf{x}^k), \ \tau_k = 1.$$

Для нахождения $\mathbf{x}^{k+1}$ надо решить систему (5.20) относительно вектора $\Delta^k = \mathbf{x}^{k+1} - \mathbf{x}^k$, после чего положить $\mathbf{x}^{k+1} = \mathbf{x}^k + \Delta^k$. Начальное приближение $\mathbf{x}^0$ выбирается произвольным образом.

Покажем, что если начальное приближение выбрано достаточно близко к решению системы, то метод сходится и имеет квадратичную скорость сходимости.

Обозначим через $S(\mathbf{x}^*, r) = \{\mathbf{x} : \ \|\mathbf{x} - \mathbf{x}^*\| < r\}$.

Теорема 5.2.1 *Предположим, что при некоторых положительных числах r, a_1, a_2 выполнены условия:*

$$\|(\mathbf{F}'(\mathbf{x}))^{-1}\| \le a_1, \quad \mathbf{x} \in S(\mathbf{x}^*, r); \tag{5.21}$$

$$\|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})(\mathbf{y} - \mathbf{x})\| \le a_2\|\mathbf{x} - \mathbf{y}\|^2, \quad \mathbf{x}, \mathbf{y} \in S(\mathbf{x}^*, r). \tag{5.22}$$

Пусть $C = a_1 a_2$, $b = \min(r, C^{-1})$. Тогда, если $\mathbf{x}^0 \in S(\mathbf{x}^*, b)$, итерационный процесс Ньютона (5.20) сходится с оценкой погрешности

$$\|\mathbf{x}^* - \mathbf{x}^k\| \le C^{-1}(C\|\mathbf{x}^* - \mathbf{x}^0\|)^{2^k}. \tag{5.23}$$

Замечание. Условие (5.22) выполнено, если в окрестности решения функции f_i имеют ограниченные вторые производные. Это следует из формулы (5.19), если в ней $\mathbf{x}^k$ заменить на $\mathbf{x}$, а $\mathbf{x}^*$ — на $\mathbf{y}$.

Доказательство теоремы. Покажем сначала, что если $\mathbf{x}^0 \in S(\mathbf{x}^*, b)$, то $\mathbf{x}^k \in S(\mathbf{x}^*, b)$, при всех k. Воспользуемся для этого методом математической индукции. Предположим, что это утверждение уже доказано для некоторого k. Тогда, в силу неравенства $b \le r$ имеем $\mathbf{x}^k \in S(\mathbf{x}^*, r)$. Поэтому выполняется неравенство (5.22), если положить в нем $\mathbf{x} = \mathbf{x}^k$, $\mathbf{y} = \mathbf{x}^*$. Запишем его с учетом того, что $\mathbf{F}(\mathbf{x}^*) = \mathbf{0}$ и $\mathbf{F}(\mathbf{x}^k) = -\mathbf{F}'(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k)$:

$$\|\mathbf{F}'(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^*)\| \le a_2\|\mathbf{x}^k - \mathbf{x}^*\|^2.$$

Тогда из (5.21) и полученного неравенства следует:

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| = \|(\mathbf{F}'(\mathbf{x}^k))^{-1}\mathbf{F}'(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^*)\| \le a_1\|\mathbf{F}'(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^*)\| \le a_1 a_2\|\mathbf{x}^k - \mathbf{x}^*\|^2. \tag{5.24}$$

По предположению $\mathbf{x}^k \in S(\mathbf{x}^*, b)$, поэтому из (5.24) следует, что

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| < a_1 a_2 b^2 = (Cb)b \leq b.$$

Полученное неравенство означает, что $\mathbf{x}^{k+1} \in S(\mathbf{x}^*, b)$. Таким образом, доказано, что все члены последовательности $\mathbf{x}^k$ лежат в $S(\mathbf{x}^*, b)$. Тогда для всех членов этой последовательности выполняется неравенство (5.24).

Пусть $d_k = C\|\mathbf{x}^k - \mathbf{x}^*\|$. Умножим (5.24) на C. В результате получим $d_{k+1} \leq d_k^2$. Воспользуемся снова методом математической индукции, чтобы доказать неравенство

$$d_k \leq d_0^{2^k},$$

которое, очевидно, означает выполнение неравенства (5.23). Ясно, что при $k = 0$ неравенство выполняется. Предположим, что при некотором k оно справедливо. Тогда

$$d_{k+1} \leq d_k^2 \leq \left(d_0^{2^k}\right)^2 = d_0^{2^{k+1}}.$$

Таким образом, при всех k выполняется неравенство (5.23).

Заметим, что по определению b выполняется неравенство

$$C\|\mathbf{x}^0 - \mathbf{x}^*\| < Cb \leq 1.$$

Тогда из (5.23) следует, что $\|\mathbf{x}^k - \mathbf{x}^*\| \to 0$ при $k \to \infty$. Теорема доказана.

Вычисления по формуле (5.20) требуют на каждой итерации решения системы линейных алгебраических уравнений, что весьма трудоемко. Поэтому существует модификация метода. **Модифицированный метод Ньютона** заключается в следующем. Выбирается последовательность $k_0 = 0 < k_1 < k_2 < \dots$ и при $k_n \leq k < k_{n+1}$ вычисления проводятся по формуле

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\mathbf{F}'(\mathbf{x}^{k_n}))^{-1}\mathbf{F}(\mathbf{x}^k) = \mathbf{0}.$$

Скорость сходимости при такой модификации становится линейной, но обращать матрицу приходится редко, за счет чего может произойти выигрыш в общем времени нахождения решения.

Нелинейный метод Якоби. На каждой итерации решается m независимых скалярных уравнений

$$f_i(x_1^k, x_2^k, \dots, x_{i-1}^k, x_i^{k+1}, x_{i+1}^k, \dots, x_m^k) = 0, \quad i = 1, \dots, m.$$

Для отыскания x_i^{k+1} применяется один из методов предыдущего параграфа (для различных i может быть свой метод).

Нелинейный метод Зейделя. Естественной модификацией метода Якоби является метод Зейделя, в котором уже найденные компоненты вектора на новой итерации участвуют в вычислении последующих компонент:

$$f_i(x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, \dots, x_m^k) = 0, \quad i = 1, \dots, m.$$

Возможно использование гибридных методов, когда, внешние итерации делаются одним методом, а внутренние — другим. При этом число внутренних итераций может

быть фиксированным и не очень большим. Например, если внешние итерации делать по Зейделю, а внутренние по Ньютону, получаются следующие расчетные формулы:

$$\frac{\partial f_i}{\partial x_i}(x_1^{k+1},\dots,x_{i-1}^{k+1},y_i^s,x_{i+1}^k,\dots,x_m^k)(y_i^{s+1}-y_i^s)+f(x_1^{k+1},\dots,x_{i-1}^{k+1},y^s,x_{i+1}^k,\dots,x_m^k)=0,$$
$$s=0,1,\dots,n,\quad y_i^0=x_i^k,\quad x_i^{k+1}=y_i^{n+1},\quad i=1,\dots,m.$$

Здесь индекс s соответствует внутренним итерациям.

Иногда делают только одну внутреннюю итерацию, то есть получают метод

$$\frac{\partial f_i}{\partial x_i}(x_1^{k+1},\dots,x_{i-1}^{k+1},x_i^k,\dots,x_m^k)(x_i^{k+1}-x_i^k)+$$
$$+f_i(x_1^{k+1},\dots,x_{i-1}^{k+1},x_i^k,\dots,x_m^k)=0,\ i=1,\dots,m,\quad k=0,1,\dots$$

5.3 ЗАДАЧИ К ГЛАВЕ 5

5.3.1 Примеры решения задач

1. Найти корни уравнения $x\sin x=1$, значения которых больше 30.

Решение. Перепишем уравнение в виде

$$\sin x=\frac{1}{x}.$$

Тогда ясно, что при больших значениях x величины $1/x$ малы и корни уравнения близки к нулям синуса, то есть $x_n\approx\pi n$. Эти значения легко уточнить. Положим $x_n=\pi n+\varepsilon_n$, где ε_n — небольшие добавки. Тогда

$$\frac{1}{\pi n+\varepsilon_n}-\sin(\pi n+\varepsilon_n)=\frac{1}{\pi n+\varepsilon_n}-(-1)^n\sin\varepsilon_n=0.$$

В силу малости ε_n можно положить $\sin\varepsilon_n\approx\varepsilon_n$ и $\dfrac{1}{\pi n+\varepsilon_n}\approx\dfrac{1}{\pi n}$. В результате имеем

$$\varepsilon_n\approx\frac{(-1)^n}{\pi n}.$$

Поэтому

$$x_n\approx\pi n+\frac{(-1)^n}{\pi n}. \tag{5.25}$$

Так как по условию корень должен быть больше 30, в (5.25) следует взять $n\geq 10$.

2. Показать, что если уравнение $f(x)=0$ имеет на отрезке $[a,b]$ корень x^* кратности $k>1$, причем на этом отрезке $f(x)$ дважды непрерывно дифференцируемая функция, то вблизи корня метод Ньютона сходится, однако скорость сходимости не является квадратичной.

Решение. Пусть $S(x)=x-\dfrac{f(x)}{f'(x)}$. Тогда метод Ньютона можно записать в виде $x_{n+1}=S(x_n)$. Так как для корня x^* выполняется равенство $x^*=S(x^*)$, получаем

$$x_{n+1}-x^*=S(x_n)-S(x^*)=(x_n-x^*)S'(x^*)+\frac{1}{2}(x_n-x^*)^2S''(\xi_n). \tag{5.26}$$

Здесь ξ_n некоторая точка, лежащая между x^* и x_n. Производная

$$S'(x) = \frac{f(x)f''(x)}{\big(f'(x)\big)^2}$$

имеет в точке x^* неопределенность типа "ноль на ноль", так как в силу условия кратности корня, знаменатель обращается в ноль. Оценим $S'(x)$ в малой окрестности корня x^*. Заметим, что в малой окрестности x^* функция $f(x) \approx a(x-x^*)^k$. Поэтому

$$S'(x) = \frac{f(x)f''(x)}{\big(f'(x)\big)^2} \approx \frac{a(x-x^*)^k \cdot ak(k-1)(x-x^*)^{k-2}}{a^2k^2(x-x^*)^{2k-2}} = \frac{k-1}{k}.$$

Таким образом, при x_n близком к x^* имеем из (5.26)

$$\frac{x_{n+1}-x^*}{x_n-x^*} = \frac{S(x_n)-S(x^*)}{x_n-x^*} \approx S'(x^*) = \frac{k-1}{k} < 1.$$

Отсюда следует, что

$$|x_{n+1}-x^*| \le \Big(\frac{k-1}{k}\Big)^n |x_1-x^*| \to 0 \ \text{ при } \ n \to \infty.$$

Заметим, что чем больше кратность корня, тем ближе $\dfrac{k-1}{k}$ к 1 и, значит, тем медленнее сходимость.

3. В условиях предыдущей задачи модернизировать метод Ньютона таким образом, чтобы получить квадратичную скорость сходимости. Предполагается, что кратность корня k известна заранее.

Решение. Пусть $S(x,\theta) = x - \theta\dfrac{f(x)}{f'(x)}$. Метод Ньютона соответствует параметру $\theta = 1$. Как и в предыдущей задаче справедливо равенство (5.26), в котором $S(x)$ заменяется на $S(x,\theta)$. Значит, для обеспечения квадратичной скорости сходимости надо чтобы $S'(x^*,\theta) = 0$. Имеем вблизи корня

$$S'(x,\theta) = 1-\theta+\theta\frac{f(x)f''(x)}{\big(f'(x)\big)^2} \approx 1-\theta+\theta\frac{k-1}{k} = \frac{k-\theta}{k}.$$

Следовательно, $S'(x^*,\theta) = 0$, если $\theta = k$, то есть метод имеет вид:

$$x_{n+1} = x_n - k\frac{f(x_n)}{f'(x_n)}.$$

4. Предложить итерационный процесс для нахождения всех корней уравнения $x^3+4x^2-1=0$ методом простой итерации.

Решение. Произведем отделение корней, определив таблично промежутки, на которых изменяются знаки функции $f(x) = x^3+4x^2-1$.

x	-4	-3	-2	-1	0	1
$\mathrm{sign} f(x)$	$-$	$+$	$+$	$+$	$-$	$+$

Таким образом, корни уравнения лежат на промежутках $[-4,-3]$, $[-1,0]$, $[0,1]$.

Для определения корня, лежащего на промежутке $[-4,-3]$ разделим уравнение на x^2 и запишем его в виде

$$x = S(x) = \frac{1}{x^2} - 4.$$

Имеем $S'(x) = -2x^{-3}$. Значит, при $x \in [-4,-3]$ выполняются неравенства $0 \le S'(x) < 2/27$. Отсюда следует, что на отрезке $|S'(x)| < 1$ и функция $S(x)$ возрастающая. Так как $S(-4) = -4 + 1/16 > -4$, а $S(-3) = -4 + 1/9 < -3$, все значения функции $S(x) \in [-4,-3]$ при $x \in [-4,-3]$. Таким образом, условия сходимости метода простой итерации выполнены и, значит, при любом начальном приближении $x_0 \in [-4,-3]$ метод

$$x_{k+1} = \frac{1}{x_k^2} - 4$$

сходится.

Для двух других отрезков уравнение переписывается в виде $x^2(x+4) = 1$ вследствие чего получаются два итерационных процесса

$$x_{k+1} = -\frac{1}{\sqrt{x_k + 4}}, \quad x_{k+1} = \frac{1}{\sqrt{x_k + 4}}.$$

Легко проверить, что $S_-(x) = -\frac{1}{\sqrt{x+4}}$ отображает отрезок $[-1,0]$ в себя и $|S_-'(x)| < 1$ на $[-1,0]$, а $S_+(x) = \frac{1}{\sqrt{x+4}}$ отображает отрезок $[0,1]$ в себя и $|S_+'(x)| < 1$ на $[0,1]$. Поэтому оба эти итерационных процесса сходятся.

5.3.2 Задачи

1. Указать алгоритм вычисления $\sqrt{a}$ с заданным числом верных десятичных знаков, рассматривая $\sqrt{a}$ как решение уравнения $x^2 - a = 0$.

Указание. Воспользоваться методом Ньютона. Доказать, что метод сходится при любом $x_0 > 0$.

2. Построить итерационный процесс Ньютона для нахождения $\sqrt[p]{a}$.

3. Показать,что обратную величину $y = 1/x$, где $x > 0$, можно вычислить приближенно, используя только операции умножения и вычитания по формуле $y_{n+1} = y_n(2 - xy_n)$.

Указание. Применить метод Ньютона для решения уравнения $x - 1/y = 0$.

4. Имеется программа для решения уравнения $x = \varphi(x)$ методом итераций. Требуется решить уравнение $f(x) = 0$. Определить промежуток, в котором находится положительный корень и преобразовать уравнение к виду, в котором его можно решить указанным методом.

a) $f(x) = \operatorname{tg} \frac{\pi}{2} x - 2.5x$, b) $f(x) = \ln 3x - 3.5x + 2.5$.

5. Исследовать сходимость метода простой итерации $x_{n+1} = x_n^2 - 2x_n + 2$ в зависимости от выбора начального приближения x_0.

Указание. Рассмотреть случаи: $x_0 = 0,\ x_0 = 1,\ x_0 = 2,\ x_0 < 0,\ 0 < x_0 < 1,\ 1 < x_0 < 2,\ x_0 > 2$.

6. Уравнение $x + \ln x = 0$, имеющее простой корень $x^* \approx 0.6$, решается одним из методов простой итерации:

$$x_{n+1} = -\ln x_n, \qquad x_{n+1} = e^{-x_n},$$
$$x_{n+1} = \frac{x_n + e^{-x_n}}{2}, \quad x_{n+1} = \frac{3x_n + 5e^{-x_n}}{8}.$$

Исследовать эти методы и выдать рекомендации по их использованию.

7. Пусть $f(x)$ дважды непрерывно дифференцируемая на отрезке $[a,b]$ функция, имеющая на этом отрезке ноль x^* неизвестной кратности $p \geq 1$. Предложить модификацию метода Ньютона, имеющую квадратичную скорость сходимости.

Указание. Воспользоваться тем фактом, что x^* — ноль функции $g(x) = \dfrac{f(x)}{f'(x)}$ кратности 1.

8. Написать расчетные формулы для решения системы нелинейных алгебраических уравнений, если внешние итерации считаются по методу Ньютона, а внутренние по методу Зейделя, причем выполняется только одна внутренняя итерация.

5.3.3 Примеры тестовых вопросов к главе 5

1. Методом простых итераций решается уравнение $x - \cos x = 0$. В каком виде надо представить уравнение, чтобы итерации сходились?

а) $x = \arccos x$;

б) $x = \cos x$;

в) $x = 2x - \cos x$;

г) $x = 2x - \arccos x$.

2. Корень уравнения $f(x) = 0$ ищется при помощи метода Ньютона. Какой из приведенных ниже критериев гарантирует нахождение корня с точностью ε?

а) $|x_n - x_{n-1}| < \varepsilon$;

б) $|f(x_n)| < \varepsilon$;

в) $|f(x_n)| < \min\limits_x |f'(x)| \cdot \varepsilon$;

г) $|f(x_n)| < \max\limits_x |f'(x)| \cdot \varepsilon$;

д) Никакой из критериев 1-4 не гарантирует нахождение корня с точностью ε.

3. Уравнение $f(x) = 0$ решается методом деления отрезка пополам. Установлено, что корень находится в промежутке $[2,3]$. Какое наименьшее количество итераций надо сделать, чтобы с гарантией получить корень с точностью до 10^{-6}? Предполагается, что после окончания итераций в качестве корня берется середина отрезка.

4. Корень уравнения $f(x) = 0$, лежащий на отрезке $[a,b]$, ищется с помощью итерационного процесса

$$x_{n+1} = x_n - \tau f(x_n), \quad x_0 \in [a,b], \quad n = 0,1,\ldots$$

Выберите значение τ, при котором метод сходится, если в качестве начального приближения выбирается произвольная точка отрезка, $a = 0.5,\ b = 1.5,\ f(x) = xe^{1-x} - 0.9$.

а) $\tau = 1$;

б) $\tau = 2$;

в) $\tau = -1$;

г) $\tau = -2$;

д) При любом значении τ метод расходится.

5. Среди приведенных ниже методов решения системы двух нелинейных уравнений с двумя неизвестными $f(x,y) = 0,\ g(x,y) = 0$ укажите тот, в котором внешние итерации совершаются методом Зейделя, а внутренние — простых итераций.

а)

$$\chi_{k+1} = \chi_k - \tau f(\chi_k, y_n),\ \chi_0 = x_n,\ k = 0, \dots, K,\ x_{n+1} = \chi_{K+1},$$
$$\zeta_{k+1} = \zeta_k - \tau g(x_{n+1}, \zeta_k),\ \zeta_0 = y_n,\ k = 0 \dots, K,\ y_{n+1} = \zeta_{K+1},\ n = 0, 1 \dots;$$

б)

$$\chi_{k+1} = \chi_k - \tau f(\chi_k, y_n),\ \chi_0 = x_n,\ k = 0, \dots, K,\ x_{n+1} = \chi_{K+1},$$
$$\zeta_{k+1} = \zeta_k - \tau g(x_n, \zeta_k),\ \zeta_0 = y_n,\ k = 0 \dots, K,\ y_{n+1} = \zeta_{K+1},\ n = 0, 1 \dots;$$

в)

$$\chi_{k+1} = \chi_k - \tau f(\chi_k, y_n),\ \chi_0 = x_n,\ k = 0, \dots, K,\ x_{n+1} = \chi_{K+1},$$
$$\zeta_{k+1} = \zeta_k - \frac{g(x_{n+1}, \zeta_k)}{\frac{\partial g(x_{n+1}, \zeta_k)}{\partial y}},\ \zeta_0 = y_n,\ k = 0 \dots, K,\ y_{n+1} = \zeta_{K+1},\ n = 0, 1 \dots;$$

г)

$$\chi_{k+1} = \chi_k - \tau f(\chi_k, y_n),\ \chi_0 = x_n,\ k = 0, \dots, K,\ x_{n+1} = \chi_{K+1},$$
$$\zeta_{k+1} = \zeta_k - \frac{g(x_n, \zeta_k)}{\frac{\partial g(x_{n+1}, \zeta_k)}{\partial y}},\ \zeta_0 = y_n,\ k = 0 \dots, K,\ y_{n+1} = \zeta_{K+1},\ n = 0, 1 \dots$$

6. Для какой функции $f(x)$ уравнение $f(x) = 0$ не может быть решено методом касательных?

а) $f(x) = \begin{cases} x^4, & x > 0, \\ -x^4 & x \le 0; \end{cases}$

б) $f(x) = x^3$;

в) $f(x) = \begin{cases} x^2, & x > 0, \\ -x^2 & x \le 0; \end{cases}$

г) $f(x) = \begin{cases} \sqrt{x}, & x > 0, \\ -\sqrt{-x} & x \le 0; \end{cases}$

д) Для всех приведенных в п.1-4 функций метод касательных применим.

6 ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Многие задачи механики, физики, химии, биологии и других наук при их математическом описании сводятся к дифференциальным уравнениям. Конкретная прикладная задача может приводиться к дифференциальному уравнению любого порядка, или к системе уравнений любого порядка. Известно, что обыкновенное дифференциальное уравнение n-го порядка

$$u^{(n)} = f(x, u, u', \ldots, u^{(n-1)})$$

путем замены $u^{(k)}(x) = u_{k+1}(x)$ можно свести к эквивалентной системе уравнений n уравнений первого порядка

$$u_k'(x) = u_{k+1}(x), \quad k = 1, \ldots, n-1,$$
$$u_n' = f(x, u_1, u_2, \ldots, u_n).$$

Аналогично, произвольную систему дифференциальных уравнений любого порядка можно заменить некоторой эквивалентной системой уравнений первого порядка. Поэтому в дальнейшем, как правило, будут рассматриваться системы уравнений первого порядка, причем такие, которые разрешены относительно производной

$$u_k' = f_k(x, u_1, u_2, \ldots, u_n), \quad k = 1, \ldots, n. \tag{6.1}$$

Для краткости будем записывать их в векторной форме

$$\mathbf{u}'(x) = \mathbf{f}(x, \mathbf{u}),$$
$$\mathbf{u} = (u_1, \ldots, u_n), \quad \mathbf{f} = (f_1, \ldots, f_n).$$

Известно, что (6.1) имеет множество решений. В общем случае все решения могут быть записаны в форме n параметрического семейства функций $\mathbf{u} = \mathbf{u}(x, \mathbf{c})$, где $\mathbf{c} = (c_1, \ldots, c_n)$. Для определения значений этих параметров, то есть для выделения единственного решения, нужно наложить n дополнительных условий на функции $u_k(x)$. **Задача Коши** или **задача с начальными условиями** имеет дополнительные условия вида

$$\mathbf{u}(x_0) = \mathbf{u}_0, \tag{6.2}$$

то есть заданы значения всех функций в одной и той же точке x_0. Решение при этом обычно требуется найти на некотором отрезке $x_0 \le x \le x_0 + X$ или $x_0 - X \le x \le x_0$, так что точку x_0 можно считать начальной точкой этого отрезка.

Если правые части системы (6.1) непрерывны и ограничены в окрестности точки $(x_0, \mathbf{u}_0)$, то задача 6.1-6.2 имеет решение, но, вообще говоря не единственное. Если

же правые части удовлетворяют условию Липшица по переменным $\mathbf{u}$, то решение задачи Коши единственно и непрерывно зависит от координат начальной точки.

Методы решения задачи Коши для обыкновенных дифференциальных уравнений делятся на четыре группы: графические, аналитические, приближенные и численные.

Графические методы используют геометрические построения. В частности, одним из них является метод изоклин для решения дифференциального уравнения первого порядка. Он основан на геометрическом определении интегральных кривых по заранее построенному поля направлений.

Аналитические методы изучаются в курсе (разделе) "Обыкновенные дифференциальные уравнения"и позволяют найти точные решения поставленных задач. Классы уравнений, для которых разработаны методы получения точных решений сравнительно узки и охватывают только малую часть возникающих на практике задач.

Приближенными называют методы, в которых решение получается как предел некоторой последовательности функций, которые могут быть выражены через элементарные функции или при помощи квадратур. Ограничиваясь конечным числом членов последовательности, получают приближенное выражение для решения.

Численные методы - это алгоритмы вычисления приближенных значений решения на некоторой выбранной сетке значений аргумента x_i. Решение при этом получается в виде таблицы. Численные методы не позволяют найти общего решения системы. Они могут дать какое-то частное решение. С появлением ЭВМ численные методы решения стали одним из основных способов решения практических задач для обыкновенных дифференциальных уравнений.

Численные методы можно применять только к корректно поставленным задачам. Однако формального выполнения условия корректности может оказаться недостаточным. Надо, чтобы задача была **хорошо обусловленной**, то есть малым изменениям начальных условий соответствовало малое изменение решения. В противном случае, небольшое изменение начальных условий или эквивалентные этим изменениям небольшие погрешности численного метода могут существенно изменить решение.

В качестве примера плохой обусловленности рассмотрим задачу

$$u' = u - \pi x, \quad 0 \le x \le 100, \tag{6.3}$$

$$u(0) = \pi. \tag{6.4}$$

Общее решение уравнения (6.3) имеет вид

$$u(x, c) = \pi(1 + x) + ce^x.$$

Таким образом, для того, чтобы удовлетворить условию (6.4) достаточно положить $c = 0$. Тогда $u(100) \approx 317.3009$. Если при определении начального условия вместо числа π задать $\widetilde{u}(0) = 3.1416$, то есть изменение начального условия $\delta = 3.1416 - \pi \approx 7.4 \cdot 10^{-6}$, то $c = \delta$. Тогда

$$\widetilde{u}(100) - u(100) = \delta e^{100} \approx 10^{38},$$

то есть решение изменилось сильно.

В этой главе будут рассмотрены методы решения задачи Коши. Для простоты записи почти всюду ограничимся случаем одного уравнения первого порядка. Алгоритмы для случая системы n уравнений легко получаются из алгоритмов для одного уравнения путем формальной замены $u(x)$ и $f(x, u)$ на $\mathbf{u}(x)$ и $\mathbf{f}(x, \mathbf{u})$.

Всюду в дальнейшем будем предполагать, что $f(x,u)$ гладкая функция, то есть имеет столько непрерывных производных, сколько требуется при соответствующих рассуждениях.

6.1 ПРИБЛИЖЕННЫЕ МЕТОДЫ

6.1.1 Метод Пикара

Рассмотрим задачу Коши для уравнения первого порядка

$$u'(x) = f(x, u(x)), \quad x_0 \le x \le x_0 + X \quad u(x_0) = u_0. \tag{6.5}$$

Проинтегрировав это уравнение в промежутке (x_0, x), заменим задачу (6.5) эквивалентным интегральным уравнением

$$u(x) = u_0 + \int\limits_{x_0}^{x} f(\xi, u(\xi))d\xi. \tag{6.6}$$

Решая это уравнение методом последовательных приближений, получим **итерационный процесс Пикара**:

$$y_m(x) = u_0 + \int\limits_{x_0}^{x} f(\xi, y_{m-1}(\xi))d\xi, \quad y_0(x) \equiv u_0. \tag{6.7}$$

Сходимость функций $y_m(x)$ к $u(x)$ доказана в [20] п.1.3.

В качестве примера применим метод Пикара к решению задачи

$$u' = u^2 + x^2, \quad u(0) = 0.$$

Доказано, что решение этого уравнения не может быть выражено через элементарные функции.

Имеем

$$y_0(x) = 0, \quad y_1(x) = \int\limits_0^x \xi^2 d\xi = \frac{1}{3}x^3,$$

$$y_2(x) = \int\limits_0^x (\xi^2 + y_1^2(\xi))d\xi = \frac{1}{3}x^3\Big(1 + \frac{1}{21}x^4\Big),$$

$$y_3(x) = \int\limits_0^x (\xi^2 + y_2^2(\xi))d\xi = \frac{1}{3}x^3\Big(1 + \frac{1}{21}x^4 + \frac{2}{693}x^8 + \frac{1}{19845}x^{12}\Big).$$

Легко заметить, что при $|x| < 1$ приближения быстро сходятся, что позволяет найти решение с высокой точностью.

Метод Пикара целесообразно применять, если интегралы, стоящие в правой части соотношений (6.6) выражаются через элементарные функции. Если же правая часть сложная и интегралы приходится находить численными методами, то метод Пикара становится не слишком удобным.

6.1.2 Метод степенных рядов

Пусть требуется решить задачу Коши (6.5). Предположим, что правая часть дифференциального уравнения является аналитической функцией. В теории обыкновенных дифференциальных уравнений доказывается, что в этом случае в окрестности точки x_0 существует решение задачи Коши, которое может быть представлено в виде ряда Тейлора

$$u(x) = \sum_{i=0}^{\infty} \frac{u^{(i)}(x_0)}{i!}(x - x_0)^i.$$

Таким образом, для построения решения достаточно найти значения производных решения в точке x_0. Из уравнения находим, что первая производная $u'(x_0) = f(x_0, u_0)$. Для нахождения второй производной продифференцируем уравнение по x. В результате получим

$$u''(x) = \frac{d}{dx}f(x, u(x)) = \frac{\partial f(x,u)}{\partial x} + \frac{\partial f(x,u)}{\partial u}u' = \frac{\partial f(x,u)}{\partial x} + \frac{\partial f(x,u)}{\partial u}f(x,u). \quad (6.8)$$

Подставляя $x = x_0,\ u = u_0$ в полученное соотношение, найдем значение $u''(x_0)$.

$$u'''(x) = \frac{du''(x)}{dx} = \frac{d}{dx}\left(\frac{\partial f(x,u)}{\partial x} + \frac{\partial f(x,u)}{\partial u}f(x,u)\right) =$$
$$= \frac{\partial^2 f(x,u)}{\partial x^2} + 2\frac{\partial^2 f(x,u)}{\partial x \partial u}f(x,u) + \frac{\partial^2 f(x,u)}{\partial u^2}f^2(x,u) +$$
$$+ \frac{\partial f(x,u)}{\partial u}\left(\frac{\partial f(x,u)}{\partial x} + \frac{\partial f(x,u)}{\partial u}f(x,u)\right). \quad (6.9)$$

Отсюда вычисляем $u'''(x_0)$ и так далее. В результате находим $u(x)$ из приближенного равенства

$$u(x) \approx \sum_{i=0}^{n} \frac{u^{(i)}(x_0)}{i!}(x - x_0)^i. \quad (6.10)$$

При малых значениях $|x - x_0|$ полученная формула может дать хорошее приближение к точному решению даже для небольших значений n. В то же время, если величина $|x - x_0|$ больше радиуса сходимости ряда погрешность формулы (6.10) не стремится к нулю при росте числа n. В этом случае можно поступить следующим образом.

Разобьем отрезок $[x_0, x_0 + X]$ на отрезки $[x_{m-1}, x_m],\ m = 1, \ldots, M$. Будем последовательно строить приближения к решению на каждом отрезке. Для этого предположим, что на отрезке $[x_{m-1}, x_m]$ приближение найдено и равно $u_m(x)$. Возьмем тогда $u_m(x_m)$ в качестве начального значения решения в точке x_m и по описанному выше алгоритму построим приближение решения на отрезке $[x_m, x_{m+1}]$.

Метод степенных рядов применяется редко, так как требует выведения формул для вычисления большого числа производных. Однако его применение может оказаться целесообразным в том случае, когда приходится решать большое число раз одно и то же уравнение при различных начальных данных.

6.2 МЕТОДЫ РУНГЕ–КУТТА

6.2.1 Вывод простейших расчетных формул

Все изучаемые далее методы относятся к классу численных методов решения обыкновенных дифференциальных уравнений. Выведем сначала простейшие расчетные формулы, которые получаются из наглядных соображений.

Поставим задачу выразить значение решения в точке $x+h$, если известно решение в точке x. Проинтегрируем с этой целью уравнение (6.5) на промежутке $(x, x+h)$. В результате получим

$$u(x+h) = u(x) + \int_x^{x+h} u'(\xi)\,d\xi = u(x) + \int_x^{x+h} f(\xi, u(\xi))\,d\xi. \tag{6.11}$$

Таким образом, поставленная задача была бы решена, если бы было известно значение интеграла в правой части полученного равенства. Если использовать для вычисления интеграла формулу левых прямоугольников, получим

$$u(x+h) = u(x) + hf(x, u(x)) + O(h^2). \tag{6.12}$$

При малых значениях величины h можно пренебречь в равенстве (6.12) величиной порядка $O(h^2)$. В результате получим

$$u(x+h) \approx u(x) + hf(x, u(x)).$$

Из этого соотношения получается расчетная формула

$$y_{n+1} = y_n + hf(x_n, y_n), \tag{6.13}$$

где $x_n = x_0 + nh,\ n = 0, 1, \ldots,$ y_n — приближенное значение решения в точке x_n. Метод нахождения приближенного решения по формуле (6.13) называется **методом Эйлера**.

Если при вычислении интеграла в правой части равенства (6.11) воспользоваться формулой правых прямоугольников, получится другая расчетная формула (**неявный метод Эйлера**)

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}). \tag{6.14}$$

В этом соотношении y_{n+1} присутствует как в левой, так и в правой части, то есть задано неявно. Поэтому для нахождения y_{n+1} придется решить, вообще говоря, нелинейное уравнение (6.14).

Повышение точности методов, очевидно, связано с более точным вычислением интеграла в равенстве (6.11). Если воспользоваться с этой целью формулой трапеций, получим

$$u(x+h) = u(x) + \frac{h}{2}\Big(f(x, u(x)) + f(x+h, u(x+h))\Big) + O(h^3). \tag{6.15}$$

Отсюда следует расчетная формула

$$y_{n+1} = y_n + \frac{h}{2}\Big(f(x_n, y_n) + f(x_{n+1}, y_{n+1})\Big). \tag{6.16}$$

Полученная формула снова определяет y_{n+1} неявно. Для вывода явных формул будем рассуждать следующим образом. Предположим, что $\widetilde{u}$ таково, что

$$\widetilde{u} = u(x+h) + O(h^2). \tag{6.17}$$

Тогда по теореме Лагранжа о конечных приращениях имеем

$$f(x+h,\widetilde{u}) - f(x+h,u(x+h)) = \left.\frac{\partial f(x+h,u)}{\partial u}\right|_{u=\overline{u}} \cdot (\widetilde{u} - u(x+h)) = O(h^2),$$

где $\overline{u}$ — некоторая величина, лежащая между $\widetilde{u}$ и $u(x+h)$. Поэтому в силу (6.15) имеем

$$\begin{aligned} u(x+h) = u(x) + \frac{h}{2}\Big(f(x,u(x)) + f(x+h,\widetilde{u})\Big) + \\ + \frac{h}{2}\Big(f(x+h,u(x+h)) - f(x+h,\widetilde{u})\Big) + O(h^3) = \\ = u(x) + \frac{h}{2}\Big(f(x,u(x)) + f(x+h,\widetilde{u})\Big) + O(h^3). \end{aligned} \tag{6.18}$$

Заметим, что согласно формуле (6.12) условию (6.17) удовлетворяет функция $\widetilde{u} = u(x) + hf(x,u(x))$. Отсюда и из (6.18) получаем расчетные формулы

$$\begin{aligned} \widetilde{y}_{n+1} &= y_n + hf(x_n,y_n), \\ y_{n+1} &= y_n + \frac{h}{2}\Big(f(x_n,y_n) + f(x_{n+1},\widetilde{y}_{n+1})\Big). \end{aligned} \tag{6.19}$$

Используя аналогичные рассуждения и применяя для вычисления интеграла в правой части (6.11) формулу средних прямоугольников

$$\int_x^{x+h} f(\xi,u(\xi))d\,\xi = hf\Big(x+\frac{h}{2},u(x+\frac{h}{2})\Big) + O(h^3),$$

получим другую пару расчетных формул

$$\begin{aligned} \widetilde{y}_{n+1/2} &= y_n + \frac{h}{2}f(x_n,y_n), \\ y_{n+1} &= y_n + hf\Big(x_n+\frac{h}{2},\widetilde{y}_{n+1/2}\Big). \end{aligned} \tag{6.20}$$

Полученные выше расчетные формулы (6.13),(6.19),(6.20) относятся к семейству методов Рунге–Кутта.

6.2.2 Общая формулировка методов Рунге–Кутта

Будем считать, что в точке x_n известно приближенное значение y_n решения $u(x_n)$ дифференциального уравнения (6.5). Покажем как приближенно можно вычислить $u(x_n+h)$.

Пусть q — целое положительное число. Предположим, что заданы следующие наборы чисел

$$\alpha_2,\ldots,\alpha_q; \quad p_1,\ldots,p_q; \quad \beta_{ij},\ 0<j<i\le q.$$

Тогда сначала вычисляем

$$\begin{aligned} k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + \alpha_2 h, y_n + \beta_{21} k_1), \\ &\dots \\ k_q &= hf(x_n + \alpha_q h, y_n + \beta_{q\,1} k_1 + \dots + \beta_{q\,q-1} k_{q-1}), \end{aligned} \tag{6.21}$$

после чего полагаем

$$y_{n+1} = y_n + p_1 k_1 + p_2 k_2 + \dots + p_q k_q. \tag{6.22}$$

Описанный алгоритм называется **q - этапным явным методом Рунге — Кутта** для задачи (6.5).

Для выбора параметров α_i, β_{ij}, p_i введем величину $\psi_n(h)$ и назовем ее **погрешностью метода на шаге** или **локальной погрешностью**. Для получения выражения для величины $\psi_n(h)$ поступим следующим образом. Всюду в (6.21), (6.22) заменим y_n на $u(x_n)$, а y_{n+1} на $u(x_n + h)$. Если приближенное решение не совпадает с точным, то равенство (6.22) перестанет выполняться. Разность между левой $u(x_n + h)$ и правой частью (обозначим ее значение u_{approx}) преобразованного таким образом соотношения (6.22) и есть $\psi_n(h)$.

Название "погрешностью метода на шаге"связано со следующими соображениями. Если исходя из начальной точки $(x_n, u(x_n))$ посчитать по численному алгоритму приближенное решение, получится u_{approx}. Взяв же эту точку в качестве начального данного и решив задачу Коши для уравнения $u' = f(x, u)$ на промежутке $x_n, x_n + h$, получим $u(x_n + h)$. Таким образом, исходные данные для $u(x_n + h)$ и u_{approx} одинаковые и, значит, их разность — та ошибка, которую совершают, сделав только один шаг вычислений. [1]

При подборе параметров естественно потребовать, чтобы локальная погрешность была как можно меньше. С этой целью наложим следующие ограничения на производные функции ψ_n. Пусть параметры выбраны так, что для всех гладких функций $f(x, u)$ выполняются равенства $\psi_n^{(k)}(0) = 0,\ k = 0, 1, \dots, s$ и существует такая гладкая функция $f(x, u)$, что $\psi_n^{(s+1)}(0) \neq 0$. Тогда по формуле Тейлора

$$\psi_n(h) = \sum_{k=0}^{s} \frac{\psi_n^{(k)}(0)}{k!} h^k + \frac{\psi_n^{(s+1)}(\xi)}{(s+1)!} h^{s+1} = \frac{\psi_n^{(s+1)}(\xi)}{(s+1)!} h^{s+1} = O(h^{s+1}), \quad 0 < \xi < h. \tag{6.23}$$

Число s называется **порядком погрешности** метода или просто **порядком метода**.

Для дальнейшего нам будет важно, что, в соответствии с формулой Тейлора, погрешность метода на шаге $\psi(h)$ может быть представлена в виде (индекс "n" опускаем)

$$\psi(h) = \frac{\psi^{(s+1)}(0)}{(s+1)!} h^{s+1} + O(h^{s+2}) = C(x, u) h^{s+1} + O(h^{s+2}), \tag{6.24}$$

Здесь $C(x, u) = \dfrac{\psi^{(s+1)}(0)}{(s+1)!} \neq 0$. Из этого равенства и определения $\psi(h)$ получаем, что $C(x, u)$ явно выражается через значения в точке (x, u) функции f и ее производных до порядка не выше $s+1$. Следовательно, $C(x, u)$ непрерывно зависит от своих аргументов для гладкой функции f. Величина $C(x, u)h^{s+1}$ называется **главным членом погрешности**.

---

[1] Зачастую вместо локальной погрешности вводится величина $\dfrac{\psi_n}{h}$, которую называют **погрешностью аппроксимации уравнения (6.22) на решении исходного уравнения**.

Построим методы для различных значений q. Возьмем сначала $q = 1$. Тогда

$$\psi(h) = u(x+h) - [u(x) + p_1 h f(x, u(x))], \quad \psi(0) = 0,$$
$$\psi'(h) = u'(x+h) - p_1 f(x, u(x)), \quad \psi'(0) = u'(x) - p_1 f(x, u(x)) = (1-p_1) f(x, u(x)),$$
$$\psi''(h) = u''(x+h), \quad \psi''(0) = u''(x).$$

Отсюда видно, что $\psi'(0) = 0$ для любой функции f тогда и только тогда, когда $p_1 = 1$. Заметим, что при этом значении параметра получается уже выведенный ранее другим путем метод Эйлера. Если взять уравнение $u' = u$, то есть выбрать функцию $f(x, u) = u$, то $\psi''(0) = u''(x) = u'(x) = u(x) \neq 0$. Это означает, что у метода Эйлера погрешность имеет первый порядок.

Возьмем теперь $q = 2$. Для сокращения объема выкладок введем обозначения

$$\alpha = \alpha_2, \quad \beta = \beta_{21}, \quad f_x = \frac{\partial f}{\partial x}, \quad f_u = \frac{\partial f}{\partial u}, \quad f_{xx} = \frac{\partial^2 f}{\partial x^2} \quad \text{и т. д.}$$

Кроме того, будем опускать аргументы у функции f и ее производных, если они вычислены в точке $(x, u(x))$ и писать черту над функцией f и ее производными, если они вычислены в точке $(x + \alpha h, u(x) + \beta h f(x, u(x)))$. Тогда имеем

$$\psi(h) = u(x+h) - u(x) - p_1 h f - p_2 h \overline{f},$$
$$\psi'(h) = u'(x+h) - p_1 f - p_2 \overline{f} - p_2 h(\alpha \overline{f}_x + \beta \overline{f}_u f),$$
$$\psi''(h) = u''(x+h) - 2p_2(\alpha \overline{f}_x + \beta \overline{f}_u f) -$$
$$-p_2 h(\alpha^2 \overline{f}_{xx} + 2\alpha\beta \overline{f}_{xu} f + \beta^2 \overline{f}_{uu} f^2),$$
$$\psi'''(h) = u'''(x+h) - 3p_2(\alpha^2 \overline{f}_{xx} + 2\alpha\beta \overline{f}_{xu} f + \beta^2 \overline{f}_{uu} f^2) + O(h).$$

Если теперь в полученных выражениях для функции ψ и ее производных положить $h = 0$, а для производных функции u использовать соотношения (6.5),(6.8), то после приведения подобных получим

$$\psi(0) = u(x) - u(x) = 0, \quad \psi'(0) = (1 - p_1 - p_2) f,$$
$$\psi''(0) = (1 - 2p_2\alpha) f_x + (1 - 2p_2\beta) f_u f,$$
$$\psi'''(0) = u'''(x) - 3p_2(\alpha^2 f_{xx} + 2\alpha\beta f_{xu} f + \beta^2 f^2 f_{uu}).$$

Так как строится метод решения произвольного дифференциального уравнения с гладкой правой частью, функция f — произвольна. Поэтому для того, чтобы $\psi'(0) = \psi''(0) = 0$, необходимо выполнение равенств

$$1 - p_1 - p_2 = 0, \quad 1 - 2p_2\alpha = 0, \quad 1 - 2p_2\beta = 0. \tag{6.25}$$

Как и в случае $q = 1$ достаточно взять $f(x, u) = u$, чтобы убедиться, что $\psi'''(0) \neq 0$. Значит нельзя построить формул Рунге–Кутта со значениями $q = 2$ и $s = 3$.

Для определения четырех параметров получено три уравнения (6.25). Следовательно, можно построить бесконечное число методов, погрешность которых имеет второй порядок. В частности, при $p_1 = p_2 = 1/2$ и $\alpha = \beta = 1$ получается метод (6.19). Если же взять $p_1 = 0,\ p_2 = 1,\ \alpha = \beta = 1/2$, получится метод (6.20).

Очевидно, что увеличение порядка точности метода возможно только за счет увеличения числа q. При этом с ростом q увеличивается объем выкладок, которые необходимо проделать для получения расчетных формул. В следующей таблице указано значение s и соответствующее ему минимальное значение q:

s	1	2	3	4	5	6	7
q	1	2	3	4	6	7	9

Из таблицы следует, например, что при $q = 5$, нет расчетных формул с $s = 5$.

К числу наиболее употребительных, "классических" относится метод Рунге–Кутта, порядок погрешности которого равен 4:

$$k_1 = hf(x_n, y_n), \quad k_2 = hf(x_n + h/2, y_n + k_1/2), \quad k_3 = hf(x_n + h/2, y_n + k_2/2),$$
$$k_4 = hf(x_n + h, y_n + k_3), \quad y_{n+1} = y_n + \frac{1}{6}\Big(k_1 + 2k_2 + 2k_3 + k_4\Big). \tag{6.26}$$

Другой часто применяемый метод четвертого порядка имеет вид

$$k_1 = hf(x_n, y_n), \quad k_2 = hf(x_n + h/3, y_n + k_1/3), \quad k_3 = hf(x_n + 2h/3, y_n - k_1/3 + k_2),$$
$$k_4 = hf(x_n + h, y_n + k_1 - k_2 + k_3), \quad y_{n+1} = y_n + \frac{1}{8}\Big(k_1 + 3k_2 + 3k_3 + k_4\Big). \tag{6.27}$$

Следует отметить, что методы высокого порядка точности дают хороший результат для дифференциальных уравнений (6.5) с гладкой функцией $f(x, u)$. Если же эта функция не имеет достаточного числа непрерывных производных, расчеты по формулам более низкого порядка точности могут дать меньшую погрешность.

Важным достоинством всех методов Рунге–Кутта является тот факт, что они позволяют изменять величину шага h в процессе вычислений, то есть при различных значениях индекса n шаг может быть различным.

6.3 ГЛОБАЛЬНАЯ ОЦЕНКА ПОГРЕШНОСТИ ОДНОШАГОВЫХ МЕТОДОВ

Глобальной называется погрешность численного решения после выполнения нескольких шагов. Оценка величины такой погрешности представляет наибольший интерес.

Ограничимся рассмотрением **одношаговых методов**. Это такие алгоритмы, в которых для вычисления значения решения точке $k + 1$ необходимо знать решение только в одной k-ой точке. Все рассмотренные ранее методы (степенных рядов, Рунге-Кутта) являются одношаговыми. В общем виде одношаговые методы можно записать следующим образом

$$y_{k+1} = F(f, x_k, x_{k+1}, y_k). \tag{6.28}$$

Здесь F — некоторый функционал, свой для каждого метода, а его аргументы, это те данные, которые необходимы для вычисления значения решения в точке x_{k+1}. Поскольку в начальной точке x_0 значение u_0 решения задано, положим $y_0 = u_0$. Тогда формула (6.28) позволяет найти приближение y_k к решению во всех узлах x_k, $k = 0, 1, \ldots, N,\ x_N = x_0 + X$.

Реальное значение y_{k+1} получается не по формуле (6.28). Вычисления содержат погрешности, связанные с округлением чисел, приближениями при задании правой части дифференциального уравнения $f(x, u)$. Поэтому, учитывая эти погрешности, можно записать, что

$$y_{k+1} = F(f, x_k, x_{k+1}, y_k) + \delta_k. \tag{6.29}$$

Величина δ_k называется **вычислительной погрешностью на шаге.** Её точное значение найти практически невозможно, однако для нее существует оценка типа $|\delta_k| \le C(|y_k|+1)2^{-t}$. Здесь C — некоторая константа, а t — число разрядов, отводимых под мантиссу, при представлении чисел в ЭВМ.

В этом параграфе нас будет интересовать оценка глобальной погрешности, то есть оценка величины $u(x_n) - y_n$, где n — произвольное положительное целое число, не превосходящее N.

Для дальнейшего понадобится следующее утверждение.

Лемма 6.3.1 *Пусть на отрезке $[a,b]$ известны два различных решения $U_1(x)$, $U_2(x)$ дифференциального уравнения $u' = f(x,u)$. Предположим, что функция $f(x,u)$ непрерывна и непрерывно дифференцируема по переменной u. Тогда, существует такая функция $\overline{u}(x)$, заключенная между $U_1(x)$ и $U_2(x)$, что*

$$U_1(b) - U_2(b) = (U_1(a) - U_2(a)) \exp\left(\int_a^b \frac{\partial f(x,\overline{u}(x))}{\partial u}\, dx \right). \tag{6.30}$$

Доказательство. Прежде всего заметим, что из теоремы существования и единственности решения задачи Коши для обыкновенных дифференциальных уравнений следует, что ни в одной точке отрезка $[a,b]$ значения функций $U_1(x)$ и $U_2(x)$ совпадать не могут (в противном случае они совпадали бы на всем отрезке).

Запишем тождества ${U_1}' = f(x,U_1)$, ${U_2}' = f(x,U_2)$. Вычитая из первого второе, и, разделив на $U_1 - U_2$, получим

$$\frac{(U_1 - U_2)'}{U_1 - U_2} = \frac{f(x,U_1) - f(x,U_2)}{U_1 - U_2}.$$

Левая часть этого равенства может быть заменена выражением $(\ln|U_1 - U_2|)'$. Правая же часть является непрерывной функцией и согласно теореме Лагранжа о конечных приращениях её можно представить в виде $\dfrac{\partial f(x,\overline{u}(x))}{\partial u}$, где $\overline{u}(x)$ заключено между $U_1(x)$ и $U_2(x)$. Таким образом, имеем

$$(\ln|U_1 - U_2|)' = \frac{\partial f(x,\overline{u}(x))}{\partial u}.$$

Интегрируя полученное равенство по отрезку $[a,b]$, и, применяя затем операцию потенцирования, получим требуемый результат.

Перейдем теперь к выводу оценки для глобальной погрешности. С этой целью введем функции $u_k(x)$, $k = 0,1,\ldots,n$, которые являются решениями задач Коши

$$\begin{aligned} u_k' &= f(x,u_k), \\ u_k(x_k) &= y_k. \end{aligned}$$

Заметим, что, вообще говоря, решение $u(x)$ задачи (6.5) не совпадает с $u_0(x)$. Это связано с тем, что из-за погрешности округления реальное значение y_0 может отличаться от u_0. В дальнейшем разность $u_0 - y_0 = u(x_0) - u_0(x_0)$ будет обозначаться ψ_0.

Определим теперь величины (см. рисунок 6.1)

$$\begin{aligned} \psi_{k+1} = u_k(x_{k+1}) - u_{k+1}(x_{k+1}) = u_k(x_{k+1}) - y_{k+1} = \\ = u_k(x_{k+1}) - F(f,x_k,x_{k+1},y_k) - \delta_k = \widetilde{\psi}_{k+1} - \delta_k. \end{aligned} \tag{6.31}$$

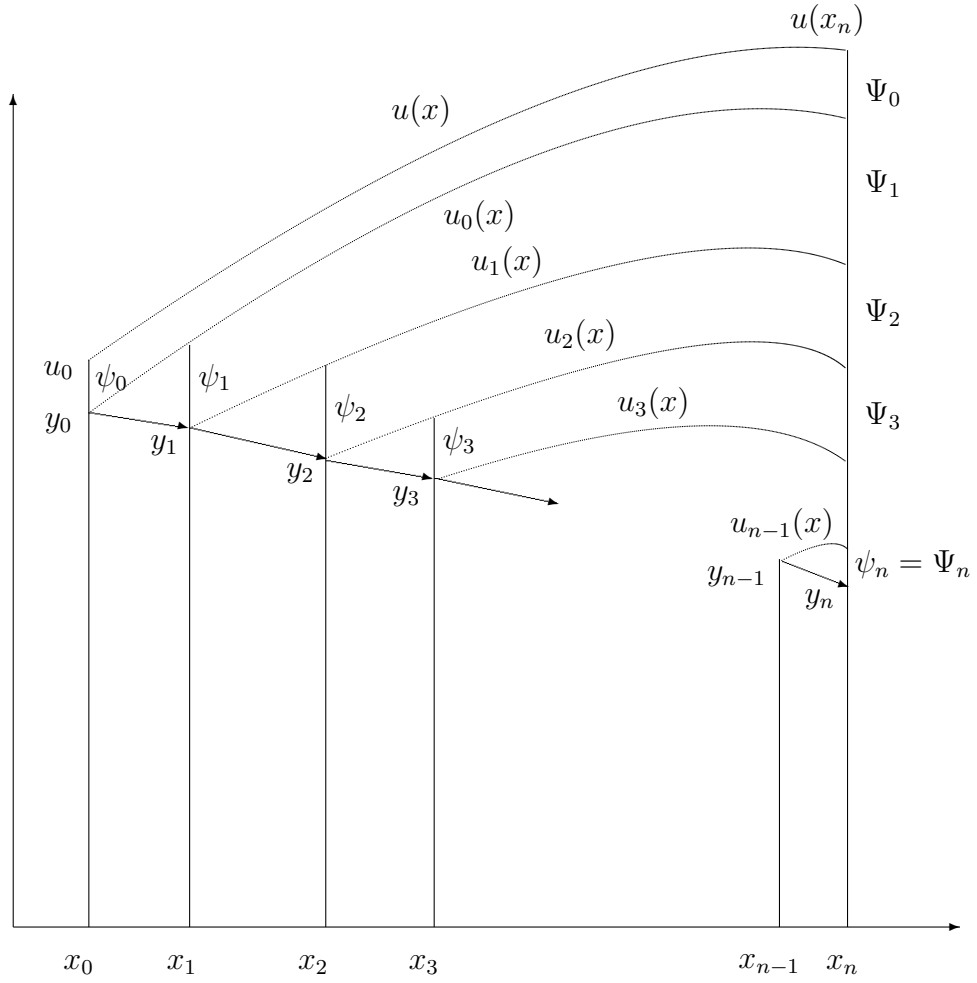


Рис. 6.1 Оценивание глобальной погрешности одношагового метода

Здесь введено обозначение $\widetilde{\psi}_{k+1} = u_k(x_{k+1}) - F(f, x_k, x_{k+1}, y_k)$. Рассмотрим, какой смысл имеет величина $\widetilde{\psi}_k$. $u_k(x_{k+1})$ — значение в точке x_{k+1} точного решения дифференциального уравнения, удовлетворяющего условию $u_k(x_k) = y_k$. $F(f, x_k, x_{k+1}, y_k)$ — число, полученное по расчетной формуле (6.28). Следовательно, $\widetilde{\psi}_{k+1}$ — погрешность одного шага рассматриваемого метода, если вычисления начинать из точки (x_k, y_k) и проводить без округлений. При этом величина шага равна $h_k = x_{k+1} - x_k$. $\widetilde{\psi}_k$ называется **погрешностью метода на шаге** или **локальной погрешностью**. Будем считать, что порядок погрешности метода равен s, то есть при всех k выполнено неравенство

$$|\widetilde{\psi}_k| \le C_1 h_{k-1}^{s+1}. \tag{6.32}$$

Глобальную погрешность можно представить в виде

$$u(x_n) - y_n = u(x_n) - u_n(x_n) = (u(x_n) - u_0(x_n)) + \sum_{i=1}^{n} (u_{i-1}(x_n) - u_i(x_n)) = \\ = \Psi_0 + \sum_{i=1}^{n} \Psi_i, \tag{6.33}$$

где $\Psi_0 = u(x_n) - u_0(x_n)$, $\Psi_i = u_{i-1}(x_n) - u_i(x_n)$.

Из леммы 6.3.1 следует, что

$$\Psi_0 = u(x_n) - u_0(x_n) = (u(x_0) - u_0(x_0)) \exp\left(\int_{x_0}^{x_n} \frac{\partial f(x, \overline{u}_0(x))}{\partial u}\, dx \right) = \\ = \psi_0 \exp\left(\int_{x_0}^{x_n} \frac{\partial f(x, \overline{u}_0(x))}{\partial u}\, dx \right),$$

$$\Psi_i = u_{i-1}(x_n) - u_i(x_n) = (u_{i-1}(x_i) - u_i(x_i)) \exp\left(\int_{x_i}^{x_n} \frac{\partial f(x, \overline{u}_i(x))}{\partial u}\, dx \right) = \\ = \psi_i \exp\left(\int_{x_i}^{x_n} \frac{\partial f(x, \overline{u}_i(x))}{\partial u}\, dx \right),$$

где $\overline{u}_0(x)$ заключено между $u_0(x)$ и $u(x)$, а $\overline{u}_i(x)$ — между $u_i(x)$ и $u_{i-1}(x)$. Подставляя эти соотношения в (6.33), имеем

$$u(x_n) - y_n = \psi_0 \exp\left(\int_{x_0}^{x_n} \frac{\partial f(x, \overline{u}_0(x))}{\partial u}\, dx \right) + \sum_{i=1}^{n} \psi_i \exp\left(\int_{x_i}^{x_n} \frac{\partial f(x, \overline{u}_i(x))}{\partial u}\, dx \right). \tag{6.34}$$

Сделаем следующее предположение относительно правой части дифференциального уравнения. Пусть существует константа L такая, что при всех $x \in [x_0, x_0 + X]$ в некоторой окрестности решения $u(x)$ выполняется неравенство

$$\left| \frac{\partial f(x, u)}{\partial u} \right| \le L.$$

Учитывая, что $x_0 \le x_i \le x_n \le x_0 + X$, при $i = 0, 1, \ldots, n$, имеем $x_n - x_i \le X$. Тогда

$$\exp\left(\int_{x_i}^{x_n} \frac{\partial f(x, \overline{u}_i(x))}{\partial u}\, dx \right) \le e^{LX}.$$

Из (6.34), (6.31), (6.32) следует теперь, что

$$|u(x_n)-y_n| \le e^{LX}\Big(|\psi_0|+\sum_{i=1}^{n}|\psi_i|\Big) \le e^{LX}\Big(|\psi_0|+\sum_{i=1}^{n}(|\widetilde{\psi}_i|+|\delta_{i-1}|)\Big) \le$$

$$\le e^{LX}\Big(|\psi_0|+\sum_{i=1}^{n}(C_1h_{i-1}^{s+1}+|\delta_{i-1}|)\Big). \quad (6.35)$$

Если $h=\max\limits_{k=0,\dots,n-1}h_k$, $\delta=\max\limits_{k=0,\dots,n-1}|\delta_k|$, то

$$\sum_{i=1}^{n}C_1h_{i-1}^{s+1}=\sum_{i=1}^{n}C_1h_{i-1}h^s \le C_1Xh^s,$$

$$\sum_{i=1}^{n}\delta_{i-1} \le n\delta \le N\delta.$$

Подставляя полученные неравенства в (6.35), имеем окончательную оценку

$$|u(x_n)-y_n| \le e^{LX}(|\psi_0|+C_1Xh^s+N\delta). \quad (6.36)$$

Неравенство (6.36) позволяет утверждать, что

$$|u(x_n)-y_n| \to 0 \ \text{ для } \ n=0,1,\dots,N,$$

если одновременно $h\to 0,\ N\delta\to 0,\ \psi_0\to 0$. В связи с присутствием в расчетах вычислительной погрешности, одного уменьшения шага зачастую недостаточно для повышения точности результатов. Это связано с тем, что $N=O(1/h)$ и уменьшение шага проводит к росту величины $N\delta$. Поэтому уменьшать шаг можно до тех пор, пока $N\delta \le C_1Xh^s$.

В оценку (6.36) входит множитель e^{LX}, который очень быстро растет с ростом LX. Поэтому необходимо с большой осторожностью относиться к результатам расчетов на промежутках, длина которых велика.

Оценка (6.36) может быть существенно улучшена, если

$$\frac{\partial f}{\partial u} \le -l < 0.$$

В этом случае

$$\exp\left(\int_{x_i}^{x_n}\frac{\partial f(x,\overline{u_i}(x))}{\partial u}\,dx\right) \le e^{-l(x_n-x_i)}.$$

Из (6.34), (6.31), (6.32) следует

$$|u(x_n)-y_n| \le |\psi_0|e^{-l(x_n-x_0)}+\sum_{i=1}^{n}|\psi_i|e^{-l(x_n-x_i)} \le$$

$$\le |\psi_0|e^{-l(x_n-x_0)}+\sum_{i=1}^{n}(C_1h_{i-1}^{s+1}+|\delta_{i-1}|)e^{-l(x_n-x_i)} \le$$

$$\le |\psi_0|e^{-l(x_n-x_0)}+\sum_{i=1}^{n}(C_1h_{i-1}^{s}+\frac{|\delta_{i-1}|}{h_{i-1}})e^{-l(x_n-x_i)}h_{i-1} \le$$

$$\le |\psi_0|e^{-l(x_n-x_0)}+\Big(C_1h^s+\max_{i=0,\dots,n-1}\frac{|\delta_i|}{h_i}\Big)\sum_{i=1}^{n}e^{-l(x_n-x_i)}h_{i-1}. \quad (6.37)$$

Заметим теперь, что при $x \in [x_{i-1}, x_i]$ справедливо неравенство $e^{-l(x_n-x_i)} \le e^{-l(x_n-x-h)}$, поэтому

$$\sum_{i=1}^{n} e^{-l(x_n-x_i)} h_{i-1} \;\le\; \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} e^{-l(x_n-x-h)}\,dx \;=\; \int_{x_0}^{x_n} e^{-l(x_n-x-h)}\,dx \;=\; e^{lh}\frac{1-e^{-l(x_n-x_0)}}{l}$$

Отсюда и из (6.37) имеем

$$|u(x_n) - y_n| \le |\psi_0| e^{-l(x_n-x_0)} + \Big(C_1 h^s + \max_{i=0,\dots,n-1} \frac{|\delta_i|}{h_i}\Big) e^{lh} \frac{1-e^{-l(x_n-x_0)}}{l}. \tag{6.38}$$

В этом случае оценка не ухудшается при увеличении промежутка интегрирования. Более того, погрешность, связанная с неточностью задания начальных данных довольно быстро убывает при увеличении $x_n - x_0$.

6.4 ПРАКТИЧЕСКАЯ ОЦЕНКА ПОГРЕШНОСТИ И ВЫБОР ДЛИНЫ ШАГА

Для проведения вычислений необходимо научиться выбирать величину шага h таким образом, чтобы, с одной стороны, шаг был мал и обеспечивал требуемую точность вычислительных результатов, с другой стороны — достаточно велик и не приводил к бесполезной вычислительной работе. Решение вопроса о величине шага h, с которым проводятся вычисления основано на определении величины погрешности. Представления (6.23),(6.24) для погрешности метода на шаге имеют только теоретическое значение, так как получение на их основе оценки погрешности весьма затруднительно. Это связано с тем, что выражение погрешности через правую часть уравнения и ее производные весьма громоздко, и, кроме того, оценки самих производных функции f во многих случаях тоже не простая задача.

Рассмотрим методы апостериорной оценки погрешности вычислений.

6.4.1 Метод Рунге контроля погрешности на шаге

Предположим, что, пользуясь некоторым определенным методом, например, методом Рунге–Кутта порядка s и исходя из заданной начальной точки (x_0, y_0), произведен расчет для двух шагов при длине шага h и получены при этом численные результаты y_1 и y_2. Исходя из той же начальной точки (x_0, y_0), сделаем теперь расчет для одного большого шага длины $2h$ и обозначим полученное численное решение $\overline{y}_2$. Как известно из формулы (6.24), погрешность приближенного решения y_1 представима в виде [2]

$$\psi_1 = u(x_0 + h) - y_1 = C(x_0, y_0) h^{s+1} + O(h^{s+2}), \tag{6.39}$$

где C выражается через производные правой части дифференциального уравнения порядка $s+1$ и непрерывно дифференцируема по своим аргументам, если правая часть дифференциального уравнения гладкая функция. Погрешность ψ_2 величины y_2 состоит из двух частей: из перенесенной погрешности первого шага $\psi_2^{(1)}$ и локальной погрешности второго шага $\psi_2^{(2)}$. Обозначим как и в предыдущем параграфе при

[2] В приводимых рассуждениях учитывается только погрешность метода. Считается, что вычислительные погрешности пренебрежимо малы.

получении глобальной оценки через $u_1(x)$ решение дифференциального уравнения такое, что $u_1(x_1) = y_1$. Тогда, используя лемму 6.3.1 и равенство (6.39), имеем

$$\psi_2^{(1)} = u(x_0+2h) - u_1(x_0+2h) = (u(x_0+h) - u_1(x_0+h)) \exp\left(\int\limits_{x_0+h}^{x_0+2h} \frac{\partial f}{\partial u}\, dx \right) =$$

$$= \psi_1 \exp\left(\int\limits_{x_0+h}^{x_0+2h} \frac{\partial f}{\partial u}\, dx \right) = \psi_1 e^{O(h)} = \psi_1(1+O(h)) = C(x_0,y_0)h^{s+1} + O(h^{s+2}). \quad (6.40)$$

По аналогии с (6.39) имеем

$$\psi_2^{(2)} = u_1(x_0+2h) - y_2 = C(x_1,y_1)h^{s+1} + O(h^{s+2}), \quad (6.41)$$

Так как $x_1 = x_0 + h$, $y_1 = y_0 + O(h)$ и $C(x,y)$ — непрерывно дифференцируемая функция, $C(x_1,y_1) = C(x_0,y_0) + O(h)$. Отсюда и из (6.40), (6.41) получим

$$u(x_0+2h) - y_2 = \psi_2 = \psi_2^1 + \psi_2^2 =$$
$$= C(x_0,y_0)h^{s+1} + O(h^{s+2}) + (C(x_0,y_0) + O(h))h^{s+1} + O(h^{s+2}) = 2C(x_0,y_0)h^{s+1} + O(h^{s+2}). \quad (6.42)$$

Для большого шага $2h$ по аналогии с (6.39)

$$u(x_0+2h) - \overline{y}_2 = C(x_0,y_0)(2h)^{s+1} + O(h^{s+2}). \quad (6.43)$$

Если пренебречь членами $O(h^{s+2})$, формулы (6.42) (6.43) позволяют исключить $u(x_0+2h)$ и вычислить величину погрешности (точнее главного члена погрешности). Действительно, вычитая (6.43) из (6.42) и отбрасывая члены порядка $O(h^{s+2})$ имеем

$$\overline{y_2} - y_2 \approx 2C(x_0,y_0)h^{s+1}(1-2^s).$$

Отсюда и из (6.42)

$$u(x_0+2h) - y_2 = \psi_2 \approx 2C(x_0,y_0)h^{s+1} \approx \frac{y_2 - \overline{y}_2}{2^s - 1}. \quad (6.44)$$

Итак получен следующий результат.

Теорема 6.4.1 *Пусть некоторым методом Рунге-Кутта порядка s в результате выполнения двух шагов длины h найдено численное значение, а в результате выполнения одного шага длины $2h$ получено значение $\overline{y}_2$. Тогда погрешность численного значения y_2 может быть приближенно найдена по формуле (6.44) и*

$$u(x_0+2h) = y_2 + \frac{y_2 - \overline{y}_2}{2^s - 1} + O(h^{s+2}). \quad (6.45)$$

Изложенный выше метод нахождения погрешности иногда называют экстраполяцией по Ричардсону.

6.4.2 Вложенные методы

Недостатком изложенного выше метода приближенного определения локальной погрешности является большое число обращений к процедуре вычисления правой части дифференциального уравнения. Так, например, при использовании "классического" метода 4-го порядка, для определения погрешности придется 11 раз вычислять правую часть. [3]

Возникает проблема нахождения более экономного метода определения локальной погрешности. Если взять два каких-нибудь метода, порядок которых равны s_1 и s_2, причем $s_1 < s_2$, и обозначить найденные по ним численные решения после первого шага y_1 и $\widetilde{y}_1$, то

$$\begin{aligned} u(x_0+h) - y_1 &= C_1 h^{s_1+1} + O(h^{s_1+2}), \\ u(x_0+h) - \widetilde{y}_1 &= C_2 h^{s_2+1} + O(h^{s_2+2}). \end{aligned}$$

Вычитая из первого равенства второе и учитывая, что $s_2 + 1 \geq s_1 + 2$, имеем

$$\widetilde{y}_1 - y_1 = (C_1 h^{s_1+1} + O(h^{s_1+2})) - (C_2 h^{s_2+1} + O(h^{s_2+2})) = C_1 h^{s_1+1} + O(h^{s_1+2}).$$

Таким образом, погрешность величины y_1 приближенно равна $\widetilde{y}_1 - y_1$.

Посмотрим на примере "классического" метода 4-го порядка удалось ли добиться желаемого сокращения вычислений правой части дифференциального уравнения? Так как $s_1 = 4$, минимально возможное значение s_2 — 5. Ранее отмечалось, что погрешность 5-го порядка может быть только у шести или более этапного метода. Поэтому, придется $4 + 6 - 1$ раз вычислять значения функции f. Таким образом, существенного выигрыша не произошло. Однако выход из сложившейся ситуации есть. Дело в том, что методов более высокого порядка есть бесконечно много. В приведенных выше рассуждениях брался любой, первый попавшийся метод. Можно было бы попытаться подобрать такой, чтобы для него подходили уже вычисленные для метода порядка s_1 значения функции f.

Идея **вложенных методов** [4] как раз и состоит в том, чтобы построить такие методы Рунге–Кутта, которые сами содержали бы кроме численного приближенного значения y_1 некоторое приближение $\widetilde{y}_1$ более высокого порядка, то есть более точное, чем y_1. Тогда $\widetilde{y}_1$ могло бы служить для определения погрешности на каждом шаге. Это смогло бы "удешевить"процедуру оценки погрешности и, как будет показано ниже, выбраковку шага.

Итак, надо найти такие числа

$$\alpha_2, \ldots, \alpha_q; \quad p_1, \ldots, p_q; \quad \widetilde{p}_1, \ldots, \widetilde{p}_q; \quad \beta_{i\,j},\ 0 < j < i \leq q,$$

чтобы погрешность приближенного решения

$$y_1 = y_0 + p_1 k_1 + \cdots + p_q k_q$$

имела порядок s_1, а

$$\widetilde{y}_1 = y_0 + \widetilde{p}_1 k_1 + \cdots + \widetilde{p}_q k_q$$

[3] Количество вычислений значений функции f равно 11, так как величина k_1, используемая при нахождении y_1, только множителем $1/2$ отличается от значения k_1, применяемого для вычисления $\overline{y}_2$.

[4] Если метод Рунге появился в начале прошлого века, то подход с использованием вложенных методов относительно нов. Он был предложен в работах Ингенда, Сантини и Фельдберга во второй половине 20-го века.

— порядок s_2. Обычно $s_2 = s_1 + 1$ или $s_2 = s_1 - 1$[5]. Здесь формально записано, что оба метода являются q-этапными. На самом деле, например, p_q может быть равным нулю. Тогда, если $p_{q-1} \neq 0$, то метод для вычисления y_1 фактически является $(q-1)$-этапным.

При $s_2 > s_1$ имеем:

$$u(x_0 + h) - y_1 \approx \widetilde{y}_1 - y_1 = \sum_{i=1}^{q} (\widetilde{p}_i - p_i) k_i.$$

Величину

$$E = \sum_{i=1}^{q} (\widetilde{p}_i - p_i) k_i$$

принято называть **контрольным членом**. Заметим, что специально вычислять $\widetilde{y}_1$ нет необходимости. Можно сразу вычислять контрольный член, задавая в программе в качестве исходных данных не $\widetilde{p}_i$, а $\widetilde{p}_i - p_i$.

Так "классический" метод Рунге-Кутта четвертого порядка (6.26) может использоваться совместно с методом второго порядка

$$y_1 = y_0 + \frac{1}{2}(-k_1 + 2k_2 + 2k_3 - k_4).$$

Контрольный член записывается в данном случае в виде

$$E = \frac{2}{3}(k_1 - k_2 - k_3 + k_4)$$

и известен как **контрольный член Егорова**.

Процедура построения вложенных методов очень громоздка. Поэтому ниже будут представлены результаты, взятые из работ, ставших уже классическими.

Прежде чем приводить примеры вложенных методов, опишем используемые ниже обозначения. В 1964 г. Бутчером предложено использовать для символического описания метода следующую таблицу

0					
α_2	$\beta_{2\,1}$				
α_3	$\beta_{3\,1}$	$\beta_{3\,2}$			
.	.	.			
α_q	$\beta_{q\,1}$	$\beta_{q\,2}$	.	$\beta_{q\,q-1}$	
	p_1	p_2	.	p_{q-1}	p_q

Так как одна таблица будет по существу использоваться для описания двух методов, придется добавить в таблицу еще одну строку с коэффициентами $\widetilde{p}_1, \widetilde{p}_2, \ldots, \widetilde{p}_q$, которые необходимы для вычисления $\widetilde{y}_1$. Заголовок будет иметь вид "$name\ s_1(s_2)$ где name — название метода или фамилия автора пары вложенных методов, s_1 — порядок погрешности y_1, s_2 — $\widetilde{y}_1$.

RKF 2(3)

0			
1	1		
1/2	1/4	1/4	
y_1	1/2	1/2	0
$\widetilde{y}_1$	1/6	1/6	4/6

RKF 2(3)B

0				
1/4	1/4			
27/40	-189/800	729/800		
1	214/891	1/33	650/891	
y_1	214/891	1/33	650/891	0
$\widetilde{y}_1$	533/2106	0	800/1053	-1/78

[5]В этом случае в отличие от сделанного в начале параграфа предположения, выполняется неравенство $s_1 > s_2$.

Обе пары вложенных методов принадлежат Е. Фельбергу. Фельбергом показано, что величина погрешности второго метода меньше. Следует отметить, что RKF 2(3) — трех этапные методы, то есть в них число вычислений правой части дифференциального уравнения $q = 3$, а RKF 2(3)B — четырех этапные. Таким образом, может показаться, что метод RKF 2(3)B не только более точный, но и более трудоемкий. На самом деле это не так. Дело в том, что коэффициенты $\beta_{4\,i}$ совпадают с p_i при всех i. Поэтому вычисленное на текущем шаге на последнем этапе значение функции f может быть снова использовано, так как оно совпадает со значением f, которое надо вычислить на первом этапе следующего шага.

Одним из наиболее используемых в настоящее время является метод Фельберга 4(5). Например, в популярный пакет программ MatLab включены методы RKF 2(3) и Фельберга 4(5).

Фельберг 4(5)

0						
1/4	1/4					
3/8	3/32	9/32				
12/13	1932/2197	-7200/2197	7296/2197			
1	439/216	-8	3680/513	-845/4104		
1/2	-8/27	2	-3544/2565	1859/4104	-11/40	
y_1	25/216	0	1408/2565	2197/4104	-1/5	0
$\widetilde{y}_1$	16/135	0	6656/12825	28561/56430	-9/50	2/55

Возникает вопрос, а почему бы не использовать лучшее из полученных значений, то есть $\widetilde{y}_1$, для численного результата и начального значения на следующем шаге? В пакете MatLab именно так и сделано. В качестве ответа на это вопрос можно было привести рассуждение о том, что тогда $\widetilde{y}_1 - y_1$ уже не оценивает погрешность этого приближения. Кроме того, Фельберг при подборе методов старался минимизировать коэффициенты погрешности приближенного решения младшего порядка, то есть y_1. Поэтому, если уж использовать для численного результата метод старшего порядка, то у него желательно минимизировать коэффициенты погрешности, решение же младшего порядка использовать только для оценивания погрешности. Такой метод, в котором к тому же $\beta_{q\,i} = p_i$ при всех i, так что результат последнего вычисления функции f может снова использоваться на следующем шаге, построили Дорман и Принс в 1980 г. Численные эксперименты показывают его высокую эффективность.

Дорман—Принс 5(4)

0							
1/5	1/5						
3/10	3/40	9/40					
4/5	44/45	-56/15	32/9				
8/9	19372/6561	-25360/2187	64448/6561	-212/729			
1	9017/3168	-355/33	46732/5247	49/176	-5103/18656		
1	35/384	0	500/1113	125/192	-2187/6784	11/84	
y_1	35/384	0	500/1113	125/192	-2187/6784	11/84	0
$\widetilde{y}_1$	5179/57600	0	7571/16695	393/640	-92097/339200	187/2100	1/40

6.4.3 Автоматическое управление длиной шага

Одним из основных преимуществ одношаговых методов является тот факт, что расстояние между точками x_k может быть произвольным, то есть длина $h_k = x_{k+1} -$

x_k может быть любой. Естественно поставить вопрос: как можно автоматизировать процедуру выбора шага, чтобы локальная погрешность при этом не превосходила заданную величину.

Рассмотрим сначала простейший прием, основанный на применении метода Рунге.

Задаются меры погрешности на шаге ε [6] и ε_1, причем $0 < \varepsilon_1 \leq \varepsilon$. Часто берут $\varepsilon/\varepsilon_1 = 2^{s+1}$, где s — порядок погрешности метода.

Пусть начальная длина шага равна h и x_0, y_0 — начальная точка. Выполняются вычисления двух шагов длины h и одного шага длины $2h$ как это было описано в пункте 6.4.1. Затем вычисляется приближенное значение локальной погрешности в точке $x_0 + 2h$

$$\psi = \frac{y_2 - \overline{y}_2}{2^s - 1}. \tag{6.46}$$

Здесь использованы те же обозначения, что и в 6.4.1. Вычисляется величина

$$err = \left| \frac{\psi}{M} \right|, \tag{6.47}$$

где M — масштабирующий множитель. Для вычисления абсолютной погрешности полагают $M = 1$, а для относительной — $M = |y_2|$. Часто используют масштабирование типа $M = \max(|y_2|, |y_0|, 1)$, или $M = \max(|y_2|, |y_0|, 10^{-6})$ или что-нибудь еще в этом роде. Таким образом величина err это модуль абсолютной или относительной локальной погрешности.[7]

Если окажется, что $\varepsilon < err$, то шаг h считается слишком большим. Делается попытка провести расчеты заново, взяв новый шаг вдвое меньше прежнего. Заметим, что y_1, вычисленное со старым шагом, совпадает с $\overline{y}_2$ для нового шага. Поэтому эту величину пересчитывать не требуется. Расчеты проводятся до тех пор, пока не окажется выполненным неравенство $err \leq \varepsilon$. Может случиться, что это неравенство выполняется только при очень маленьком значении шага h. Зачастую задают еще минимально допустимое значение шага $h_{\min}$ и если оказывается, что $h < h_{\min}$, то вычисления прекращаются.

Если при некотором шаге h выполняется неравенство $err \leq \varepsilon$, то считается, что значения $y_1,\ y_2$ решения в точках $x_0 + h,\ x_0 + 2h$ найдены верно. В этом случае $(x_0 + 2h, y_2)$ определяют как новую начальную точку и по описанному алгоритму продолжают вычисления далее для нахождения решения в следующей точке. Если перед этим выполнялось неравенство $\varepsilon_1 \leq err \leq \varepsilon$, то продолжить вычисления пытаются с шагом h. Если же выполнялось неравенство $err < \varepsilon_1$, то это означало, что локальная погрешность слишком маленькая и можно попытаться продолжить вычисления с шагом $2h$. [8]

В современных коммерческих программах для нахождения решения обыкновенных дифференциальных уравнений алгоритм автоматического управления длиной шага несколько более сложен.

---

[6] В описаниях программ и в зарубежной литературе часто вместо ε используется обозначение *tol* от слова tolerance — допуск.

[7] Как отмечалось ранее все описанные методы применимы для решения систем уравнений. Тогда $\psi,\ y_2, y_0, M$ — вектора, причем приведенные выше формулы нахождения скалярной величины M заменяют по аналогии формулами для нахождения компонент вектора M. Вычисляется отношение соответствующих компонент векторов ψ и M, получается вектор абсолютной или относительной локальной погрешности и err определяется как норма этого вектора.

[8] Разумеется необходимо следить чтобы при таком выборе шага не выйти за пределы отрезка, на котором ищется решение.

Величина err вычисляется также по формуле (6.47). Только локальная погрешность, входящая в эту формулу определяется либо по формуле (6.46), если применяется метод Рунге определения локальной погрешности, либо по формуле $\psi = y_1 - \widetilde{y_1} = E$, если используются вложенные формулы. Разумеется, что в последнем случае выполняется для нахождения ψ один, а не два шага. Далее подбирается оптимальная длина шага $h_{\text{опт}}$.

Оптимальной считается длина шага, при которой локальная погрешность (абсолютная или относительная) равна ε. Поскольку локальная погрешность приблизительно равна $\approx Ch^{s+1}$, имеем

$$Ch_{\text{опт}}^{s+1} = \varepsilon,$$

откуда

$$h_{\text{опт}} = \Big(\frac{\varepsilon}{C}\Big)^{1/(s+1)}. \tag{6.48}$$

С другой стороны при том шаге h, с которым проводились вычисления имеем $err \approx Ch^{s+1}$. Выражая отсюда C, и, подставляя его в (6.48), получим

$$h_{\text{опт}} = h\Big(\frac{\varepsilon}{err}\Big)^{1/(s+1)}. \tag{6.49}$$

Для получения хорошей программы требуется известная осторожность. Поэтому (6.49) умножают на гарантийный фактор fac. fac обычно выбирают 0.8, 0.9. Кроме того, нельзя допускать, чтобы длина шага возрастала или убывала слишком быстро. Для этого можно выбрать новую длину шага, например, следующим образом

$$h_{new} = h \cdot \min(facmax, \max(facmin, fac\Big(\frac{\varepsilon}{err}\Big)^{1/(s+1)})). \tag{6.50}$$

Максимальный коэффициент увеличения шага $facmax$ обычно выбирается между 1.5 и 5, а минимальный $facmin$ между 0.2 и 0.7.

Если при выбранном шаге $err \le \varepsilon$, вычисленное решение считается принятым, и можно продолжать расчеты далее, исходя из уже подсчитанного значения, причем в качестве длины нового шага берется h_{new}. В противном случае результаты вычисления на текущем шаге отбрасываются и вычисления повторяются с новым шагом h_{new}. Для шагов, выполненных непосредственно после отбрасывания забракованных шагов, рекомендуется положить $facmax = 1$.

Как уже отмечалось, в пакете MatLab реализован метод Фельберга 4(5). При этом подбор шага организован следующим образом. Задана точность $\varepsilon = 10^{-6}$. Вычисляются $h_{\max} = X/5$, $h_{\min} = X/20000$ и начальное значение шага $h = X/100$, где X — длина промежутка, на котором ищется решение. Полагается

$$h_{new} = \max(h_{\min}, \min(h_{\max}, 0.8h(\varepsilon/err)^{1/5})).$$

При этом для вычисления масштабирующего множителя M используется формула $M = \max(\|y\|, 1)$, где y — вычисленное значение решения в соответствующей точке.

6.5 УСТОЙЧИВОСТЬ ЧИСЛЕННЫХ МЕТОДОВ, ЖЕСТКИЕ ЗАДАЧИ

6.5.1 Устойчивые и неустойчивые уравнения и системы. Жесткие дифференциальные уравнения

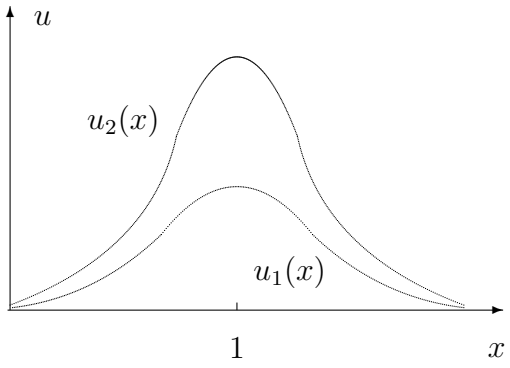


Рис. 6.2 Решения уравнения (6.51)

В параграфе 6.3 была доказана лемма, из которой следует, что если $\dfrac{\partial f}{\partial u} < 0$, то интегральные кривые с ростом x сближаются, если же $\dfrac{\partial f}{\partial u} > 0$, то кривые расходятся. В первом случае говорят об **устойчивом уравнении**, а во втором уравнение называют **неустойчивым**. Предположим, что при численном решении неустойчивого уравнения на каком-то шаге с номером k приближенное решение ненамного отличается от точного, а на всех последующих шагах локальные погрешности равны нулю, то есть на k-ом шаге перешли на другую, пусть даже близкую интегральную кривую и больше с новой интегральной кривой не сходим. Все равно из-за расходимости кривых с ростом x значения приближенного и точного решений будут расходиться. Поэтому не следует рассчитывать на высокую точность при решении неустойчивого уравнения.

Уравнение общего вида может демонстрировать на разных частях отрезка интегрирования оба типа поведения - устойчивость и неустойчивость. Например, уравнение

$$u'(x) = -2a(x-1)u \tag{6.51}$$

имеет однопараметрическое семейство решений

$$u(x) = Ce^{-a(x-1)^2}. \tag{6.52}$$

Если рассмотреть эти решения при $a > 0$ на промежутке $[0, 2]$, то сначала при $x < 1$ производная $\dfrac{\partial f}{\partial u} = -2a(x-1) > 0$ и интегральные кривые расходятся (см. рисунок 6.2). Пусть $a = 10$ и взяты два близких при $x = 0$ решения: $u_1(0) = 0.001$, $u_2(0) = 0.002$. Тогда

$$u_2(x) = 2u_1(x) = 0.002e^{10(1-(x-1)^2)}.$$

Отсюда следует, что $u_2(1) - u_1(1) = 0.001e^{10} > 200$. При $x > 1$ производная $\dfrac{\partial f}{\partial u} = -2a(x-1) < 0$ и интегральные кривые сходятся. При $x = 2$ расстояние между ними вновь мало.

Когда речь идет о системах уравнений, вместо производной $\dfrac{\partial f}{\partial u}$ рассматривается матрица Якоби, то есть матрица $\mathbf{J} = (J_{i\,j})$, где $J_{i\,j} = \dfrac{\partial f_i}{\partial u_j}$. Устойчивость системы непосредственно связана с собственными числами матрицы Якоби. Положительные

вещественные части собственных чисел обычно соответствуют областям с расходящимися, неустойчивыми решениями. Наличие у всех собственных чисел отрицательных вещественных частей влечет за собой устойчивость решений. При каждом фиксированном x могут встречаться собственные числа как с положительными, так и с отрицательными вещественными частями.

Введем теперь понятие жесткости. Рассмотрим сначала физический смысл этого понятия.

Пусть на отрезке $[0, 1]$ задана задача Коши

$$u'(t) = -a(u(t) - \sin t) + \cos t, \quad u(0) = 1. \tag{6.53}$$

Ее решение имеет вид $u(t) = e^{-at} + \sin t$. Если $a > 0$, то уравнение устойчиво. Пусть $a \gg 0$. Тогда, на промежутке близком к 0 решение быстро уменьшается от 1 практически до 0. Потом оно почти не отличается от функции $\sin t$. Решение вблизи $t = 0$ качественно отличается от решения при больших t. Начальный отрезок решения называется **пограничным слоем**, остальная часть отрезка интегрирования — **режим медленного или плавного изменения**. Таким образом, если переменная t имеет смысл времени, то решение обладает двумя несоизмеримыми характерными временами: изменение решения в пограничном слое — это процесс с малым характерным временем и режим медленного изменения, который определяет большой, по сравнению с первым процессом, отрезок интегрирования. Уравнения и системы уравнений подобного типа, то есть уравнения и системы, моделирующие физический процесс, компоненты которого обладают несоизмеримыми характерными временами, или же процесс, характерное время которого много меньше отрезка интегрирования, называют **жесткими.**

Дадим теперь строгое определение жесткой системы уравнений.

Определение 6.5.1 *Система уравнений*

$$\mathbf{u}'(x) = \mathbf{f}(x, \mathbf{u}),$$
$$\mathbf{u} = (u_1, \ldots, u_n), \quad \mathbf{f} = (f_1, \ldots, f_n)$$

называется ***жесткой на решении*** $\mathbf{v}(x)$ ***и на данном интервале*** $(x_0, x_0 + X)$, *если для собственных чисел* $\lambda_k(x), \quad k = 1, \ldots, n$ *матрицы Якоби*

$$\mathbf{J}(t, \mathbf{v}(t)) = \left(\frac{\partial f_i(x, \mathbf{v}(x))}{\partial u_j} \right)$$

выполняются условия

- $\mathrm{Re}\lambda_k(x) < 0, \quad k = 1, \ldots, n$, *для всех* $x \in (x_0, x_0 + X)$,
- $\sup\limits_{x \in (x_0, x_0 + X)} s(x) \gg 1,$

где

$$s(x) = \frac{\max\limits_{k=1,\ldots,n} |\mathrm{Re}\lambda_k(x)|}{\min\limits_{k=1,\ldots,n} |\mathrm{Re}\lambda_k(x)|}.$$

Число s называют **числом жесткости** системы. Заметим, что для линейных систем уравнений матрица Якоби, а, следовательно, и жесткость не зависят от выбранного решения. Для системы с постоянными коэффициентами жесткость не зависит от промежутка интегрирования.

Пояснить это определение можно на простых примерах.

Рассмотрим систему

$$u_1' = -a_1 u_1, \quad u_2' = -a_2 u_2, \quad a_1 \gg a_2 > 0, \quad x > 0. \tag{6.54}$$

Модуль её решения $u_1(x) = u_1(0)e^{-a_1 x}$, $u_2(x) = u_2(0)e^{-a_2 x}$ с ростом x монотонно убывает, причем компонента u_1 затухает много быстрее, чем u_2. Поэтому начиная с некоторого момента поведение вектора решения полностью определяется второй компонентой. Таким образом, налицо два процесса — быстрый для первой компоненты и медленный для второй. Этот пример может показаться надуманным так как ясно, что каждое уравнение можно решать отдельно, независимо друг от друга. Поэтому рассмотрим другую систему

$$u_1' = 998u_1 + 1998u_2, \quad u_2' = -999u_1 - 1999u_2, \quad u_1(0) = u_2(0) = 1, \quad x > 0. \tag{6.55}$$

Собственные числа матрицы коэффициентов

$$\begin{pmatrix} 998 & 1998 \\ -999 & -1999 \end{pmatrix}$$

равны -1 и -1000. Решение системы $u_1 = 4e^{-x} - 3e^{-1000x}$, $u_2 = -2e^{-x} + 3e^{-1000x}$. Очень быстро решение практически полностью начинает совпадать с функциями $u_1 = 4e^{-x}$, $u_2 = -2e^{-x}$, то есть после тонкого пограничного слоя решение начинает меняться медленно.

Замечание. Некоторые авторы [2],[15] при определении жесткости не требуют, чтобы все собственные числа имели отрицательную действительную часть. Жесткими они называют системы, у которых среди собственных чисел матрицы Якоби имеются большие по абсолютной величине, обязательно обладающие большой по модулю отрицательной действительной частью, а собственные числа с положительной вещественной частью имеют малую величину.

6.5.2 Устойчивые численные методы

Предположим, что заранее известна та или иная характерная особенность в поведении решения исходной задачи. Естественно потребовать чтобы численное решение обладало этой особенностью. Поэтому, если применять численный метод для нахождения решения устойчивых уравнений или систем, то как и для точного решения для численного решения погрешность в задании начальных данных должна затухать с ростом x. Если у численного метода это свойство выполнено, то будем говорить, что он **устойчив**.

Для исследования методов на устойчивость обычно рассматривают модельное уравнение

$$u'(x) = -\lambda u(x), \quad x > 0, \quad u(0) = u_0, \quad \lambda > 0. \tag{6.56}$$

У этого уравнения $f(x,u) = -\lambda u$, $\dfrac{\partial f}{\partial u} = -\lambda < 0$, а решение равно $u = u_0 e^{-\lambda x}$. Убывание погрешности при задании начальных данных для этого уравнения эквивалентно убыванию модуля решения, то есть $|u(x+h)| < |u(x)|$. Это следует из того, что в силу линейности уравнения, погрешность удовлетворяет тому же дифференциальному уравнению, что и функция u. Действительно, пусть $\widetilde{u} = u + \delta u$ — решение уравнения (6.56), полученное при измененных начальных данных. Тогда для погрешности δu имеем

$$\delta u' = \widetilde{u}' - u' = -\lambda\widetilde{u} + \lambda u = -\lambda\delta u.$$

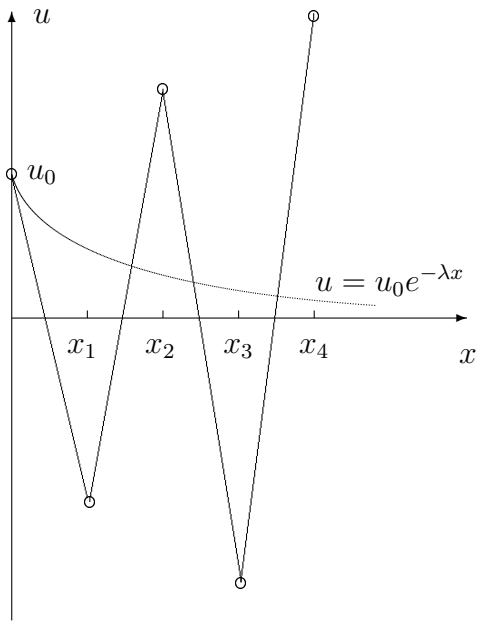


Рис. 6.3 Решение уравнения (6.56) методом Эйлера. $\lambda h > 2$.

Проанализируем, например, метод Эйлера применительно к уравнению (6.56). Имеем с учетом определения функции f

$$y_{n+1} = y_n + h(-\lambda)y_n = (1 - \lambda h)y_n. \tag{6.57}$$

Убывание модуля решения означает, что $|y_{n+1}| < |y_n|$, что эквивалентно условию $|1 - \lambda h| < 1$. Отсюда следует ограничение на шаг h:

$$0 < h < \frac{2}{\lambda}. \tag{6.58}$$

Таким образом, численный метод может быть устойчивым или неустойчивым для конкретной задачи в зависимости от выбора шага.

Численный метод называется **абсолютно устойчивым**, если он устойчив при любом шаге $h > 0$ и **условно устойчивым**, если он устойчив при некоторых ограничениях на шаг h. Из (6.58) следует, что метод Эйлера условно устойчив.

Посмотрим что будет в случае невыполнения условия устойчивости (6.58) то есть при $h\lambda \geq 2$. Тогда множитель $1 - \lambda h \leq -1$, поэтому y_n и y_{n+1} будут иметь разные знаки и $|y_{n+1}| \geq |y_n|$ (см. рисунок 6.3, на котором точки, полученные в результате численного решения соединены прямыми). Заметим, что подобный эффект растущих осцилляций практически всегда связан с неустойчивостью численного метода.

Применяя аналогичные рассуждения к другим численным методам можно получить для них условия устойчивость. Например, классический метод Рунге–Кутта четвертого порядка устойчив, если $0 < \lambda h < 2,785$.

Приведем пример абсолютно устойчивого метода. Рассмотрим неявный метод Эйлера (6.14). Решая им модельное уравнение (6.56), имеем

$$y_{n+1} = y_n - h\lambda y_{n+1}.$$

Отсюда

$$y_{n+1} = y_n(1 + h\lambda)^{-1}.$$

Таблица 6.5.2

	Явный метод			Неявный метод			Точное
Шаг	$h = 0.1$	$h = 0.01$	$h = 0.001$	$h = 0.1$	$h = 0.01$	$h = 0.001$	решение
y_1	$-2.713 \cdot 10^{20}$	$-7.968 \cdot 10^{95}$	1.471	1.542	1.479	1.472	1.472
y_2	$2.713 \cdot 10^{20}$	$7.968 \cdot 10^{95}$	-0.735	-0.771	-0.739	-0.736	-0.736

Следовательно, при всех $h > 0$ выполнено неравенство $|y_{n+1}| < |y_n|$, то есть метод устойчив.

Приведенные примеры отображают ситуацию характерную и в более общих случаях — явные методы являются условно устойчивыми, а среди неявных методов существуют абсолютно устойчивые. Условная устойчивость является недостатком явных методов, так как иногда вынуждает брать слишком мелкий шаг h. Абсолютно устойчивые неявные методы лишены этого недостатка, однако их применение к системам уравнений приводит к необходимости решения на каждом шаге, в общем случае, нелинейных алгебраических уравнений.

Рассмотрим теперь проблемы, возникающие при решении жестких задач. Начнем с задачи (6.53). Если $a = 1000$, то для метода Эйлера ограничение устойчивости на шаг сетки имеет вид $h < 2 \cdot 10^{-3}$. Пока нас интересует решение в пограничном слое, такой малый шаг интегрирования имеет физический смысл. Но в области плавного изменения решения, а эта область занимает основную часть промежутка интегрирования, выбор такого шага смысла не имеет.

Аналогичная ситуация имеет место для систем уравнений. При решении системы (6.54) выбор шага h диктует наибольшее по модулю собственное число матрицы, в данном случае a_1, а поведение решения системы начиная с некоторого момента почти полностью определяется другой компонентой.

То же самое можно сказать и про решение задачи (6.55). Взяв, например, $h = 0.01$, и применив метод Эйлера, получим в точке $x_1 = h$

$$y_1 = 1 + 0.01(998 \cdot 1 + 1998 \cdot 1) = 30.96, \quad y_2 = 1 + 0.01(-999 \cdot 1 - 1999 \cdot 1) = -28.98.$$

Легко видеть, что эти значения очень далеки от истинного значения решения. Это связано с тем, что в силу соображений устойчивости численного метода шаг должен определяться наибольшим по модулю собственным числом матрицы. В данном случае собственные числа равны -1 и -1000. Поэтому для обеспечения устойчивости согласно (6.58) шаг должен быть меньше 0.002.

В таблице 6.5.2 приведены результаты вычисления решения задачи (6.55) в точке $x = 1$, полученные с применением явного и неявного методов Эйлера, а также значение точного решения. Вычисления проводились с помощью пакета Mathcad 14. Из таблицы видно, что явный метод при шагах 0.1 и 0.01 расходится, то есть дает результаты, которые ничего общего не имеют с истинными значениями решения. Приближенные значения решения, полученные неявным методом Эйлера, даже при шаге 0.1 вполне приемлемы. Напомним, что это метод первого порядка точности и, значит, величина погрешности соизмерима с шагом.

Большинство стандартных методов не приспособлены для решения жестких задач. В настоящее время существуют весьма эффективные методы, позволяющие справляться с описанными проблемами, однако их обсуждение выходит за рамки этой книги.

6.6 МНОГОШАГОВЫЕ МЕТОДЫ

6.6.1 Общая форма линейных многошаговых методов

Общая теория многошаговых методов, предназначенных для решения задачи Коши (6.5), начала развиваться во второй половине 20-го века, хотя первые методы были сформулированы еще в 19 веке. Записываются эти методы следующим образом:

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = h(\beta_k f_{n+k} + \beta_{k-1} f_{n+k-1} + \cdots + \beta_0 f_n). \qquad (6.59)$$

В этой формуле $\alpha_i,\ \beta_i$ — вещественные числа, h обозначает величину шага, а $f_i = f(x_i, y_i),\ x_i = x_0 + ih$. Предполагается, что выполнены следующие условия:

$$\alpha_k \neq 0, \quad |\alpha_0| + |\beta_0| > 0. \qquad (6.60)$$

Формулу (6.59) называют **линейным многошаговым**, а точнее **линейным k-шаговым методом**. При этом различают **явные** ($\beta_k = 0$) и **неявные** ($\beta_k \neq 0$) многошаговые методы.

Для того, чтобы вычислить последовательность приближенных значений y_n, необходимо сначала получить k начальных (стартовых) значений $y_0, \ldots, y_{k-1}$. Затем процесс вычислений может следовать по одному из двух возможных путей. Если $\beta_k = 0$, то есть метод явный, то y_{n+k} легко вычислить. Если же метод неявный, то правая часть содержит $f(x_{n+k}, y_{n+k})$ и в общем случае необходимо решать нелинейное уравнение относительно y_{n+k}.

Для рассмотрения вопроса о нахождении решения этого уравнения, перепишем его в виде

$$y_{n+k} = \frac{h\beta_k}{\alpha_k} f(x_{n+k}, y_{n+k}) + g_n, \qquad (6.61)$$

где g_n содержит известные величины $y_{n+j}, f_{n+j},\ j = 0, 1, \ldots, k-1$. Тогда, используя теорему Банаха о неподвижной точке [20], можно доказать, что если L — константа Липшица для функции f по переменной y [9], то при

$$h < \left| \frac{\alpha_k}{\beta_k L} \right|$$

нелинейное уравнение (6.61) имеет единственное решение. Это решение может быть найдено с помощью итерационного процесса

$$y_{n+k}^{m+1} = \frac{h\beta_k}{\alpha_k} f(x_{n+k}, y_{n+k}^{m}) + g_n, \qquad (6.62)$$

где $m = 0, 1, \ldots$ — номер итерации, а y_{n+k}^0 можно выбрать произвольно. Однако, желательно подобрать y_{n+k}^0 по возможности ближе к y_{n+k}. Надлежащий выбор обеспечивается при помощи явной многошаговой формулы. Тогда явный метод называется **предсказывающим** или **предиктором**, неявный **исправляющим** или **корректором**, а весь комбинированный процесс **предсказывающе-исправляющим** или **предиктор-корректором**.

[9] Напомним, что константа Липшица это такая константа L, что для любого $x \in [x_0, x_0 + X]$ и любых y_1, y_2 выполняется неравенство

$$|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2|.$$

Существует два способа реализации метода предиктор-корректор. В первом случае итерации (6.62) проводят до тех пор, пока не будет достигнута сходимость. Во втором случае на каждом шаге корректор применяется фиксированное число раз, скажем t, после чего полученное значение y_{n+k}^t принимается за приближение к $u(x_{n+k})$.

Может показаться, что явный многошаговый метод является вообще самым простым и экономным с точки зрения вычислений. Однако на практике явные методы используются редко. Причина этого будет пояснена ниже при изучении методов Адамса.

Введем теперь понятие локальной погрешности и порядка многошагового метода. Поскольку численное решение, полученное многошаговым методом, зависит не только от начального условия задачи Коши, но и от выбора стартовых значений, определение локальной погрешности в этом случае будет не таким простым, как для одношаговых методов.

Определение 6.6.1 ***Локальной погрешностью*** *многошагового метода (6.59) называется величина $u(x_k) - y_k$, где $u(x)$ — точное решение задачи (6.5), а y_k — численное решение, полученное по формуле (6.59) при точных стартовых значениях $y_i = u(x_i),\ i = 0, 1, \ldots, k-1$.*

Определение 6.6.2 *Говорят, что многошаговый метод (6.59) имеет* ***порядок*** *p, если его локальная погрешность равна $O(h^{p+1})$ для всех дифференциальных уравнений (6.5) с достаточно гладкой правой частью.*

Для того, чтобы сформулировать соотношения между порядком метода и свободными параметрами метода $\alpha_i,\ \beta_i$, рассмотрим оператор

$$L(u, x, h) = \sum_{i=0}^{k} [\alpha_i u(x + ih) - h\beta_i u'(x + ih)]. \tag{6.63}$$

Здесь $u(x)$ — какая-либо дифференцируемая функция, которая задана на интервале, включающем значения $x + ih,\ i = 0, 1, ..., k$.

Лемма 6.6.1 *Рассмотрим дифференциальное уравнение (6.5) с непрерывно дифференцируемой функцией $f(x, u)$ и решением $u(x)$. Тогда для локальной погрешности выполняется равенство*

$$u(x_k) - y_k = \left(\alpha_k - h\beta_k \frac{\partial f(x_k, \eta)}{\partial u} \right)^{-1} L(u, x_0, h), \tag{6.64}$$

где η некоторое промежуточное значение между $u(x_k)$ и y_k.

Доказательство. Из определения локальной погрешности следует, что y_k может быть найдено из уравнения

$$\sum_{i=0}^{k-1} [\alpha_i u(x_i) - h\beta_i f(x_i, u(x_i))] + \alpha_k y_k - h\beta_k f(x_k, y_k) = 0.$$

Подставляя сюда (6.63), имеем

$$L(u, x_0, h) = \alpha_k (u(x_k) - y_k) - h\beta_k (f(x_k, u(x_k)) - f(x_k, y_k)).$$

Отсюда теперь следует утверждение леммы, если учесть теорему о среднем, согласно которой

$$f(x_k, u(x_k)) - f(x_k, y_k) = \frac{\partial f(x_k, \eta)}{\partial u}(u(x_k) - y_k).$$

Следствие. Многошаговый метод (6.59) имеет порядок p тогда и только тогда, когда $L(u, x, h) = O(h^{p+1})$ для всех достаточно гладких функций $u(x)$.

Теорема 6.6.1 *Многошаговый метод (6.59) имеет порядок p тогда и только тогда, когда*

$$\sum_{i=0}^{k} \alpha_i = 0 \quad u \quad \sum_{i=0}^{k} \alpha_i i^q = q \sum_{i=0}^{k} \beta_i i^{q-1} \quad при \quad q = 1, \ldots, p.$$

Доказательство. Подставим в (6.63) разложение по формуле Тейлора для функций $u(x + ih)$ и $u'(x + ih)$. В результате получим

$$L(u, x, h) = \sum_{i=0}^{k} \left[\alpha_i \sum_{q \geq 0} \frac{i^q h^q}{q!} u^{(q)}(x) - h\beta_i \sum_{r \geq 0} \frac{i^r h^r}{r!} u^{(r+1)}(x) \right] =$$

$$= u(x) \sum_{i=0}^{k} \alpha_i + \sum_{q \geq 0} \frac{i^q h^q}{q!} u^{(q)}(x) \left[\sum_{i=0}^{k} \alpha_i i^q - q \sum_{i=0}^{k} \beta_i i^{q-1} \right].$$

В силу произвольности $u(x)$ отсюда следует, что $L(u, x, h) = O(h^{p+1})$ тогда и только тогда, когда выполнены условия теоремы.

В теории многошаговых методов фундаментальную роль играют многочлены

$$\rho(\chi) = \alpha_k \chi^k + \alpha_{k-1} \chi^{k-1} + \cdots + \alpha_0, \quad \sigma(\chi) = \beta_k \chi^k + \beta_{k-1} \chi^{k-1} + \cdots + \beta_0, \tag{6.65}$$

введенные Далквистом. Для многошагового метода условия первого порядка ($p = 1$) принято называть условиями **согласованности.** Очевидно, что с помощью этих многочленов их можно записать в виде

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1). \tag{6.66}$$

Воспользуемся теоремой 6.6.1 для того, чтобы построить явный ($\beta_k = 0$) 2-шаговый метод максимально возможного порядка. Для нахождения коэффициентов α_i, β_i имеем равенства

$$\begin{gathered} \alpha_0 + \alpha_1 + \alpha_2 = 0, \\ \alpha_1 + 2\alpha_2 = \beta_0 + \beta_1, \quad (q = 1) \\ \alpha_1 + 4\alpha_2 = 2\beta_1, \quad (q = 2) \\ \alpha_1 + 8\alpha_2 = 3\beta_1, \quad (q = 3). \end{gathered}$$

Выбирая один из коэффициентов произвольным образом, например, положив $\alpha_2 = 1$, получим $\alpha_0 = -5$, $\alpha_1 = 4$, $\beta_0 = 2$, $\beta_1 = 4$. Если же к приведенным выше соотношениям для коэффициентов метода добавить еще одно равенство, то полученная система будет иметь только тривиальное решение. Таким образом, максимальный порядок явного 2-шагового метода равен трем и метод имеет вид

$$y_{n+2} + 4y_{n+1} - 5y_n = h(4f_{n+1} + 2f_n). \tag{6.67}$$

Исследуем этот метод. Попытаемся решить с его помощью задачу Коши

$$u' = u, \quad u(0) = 1 \tag{6.68}$$

Так как $f(x,u) = u$, уравнение (6.67) примет вид

$$y_{n+2} + 4(1-h)y_{n+1} - (5+2h)y_n = 0 \tag{6.69}$$

Если в качестве стартовых взять значения точного решения $y_0 = 1,\ y_1 = e^h$, получим в сопоставимых точках[10]

x_i	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$u(x_i)$	1.00	1.10	1.22	1.35	1.49	1.65	1.82	2.01	2.23	2.46	2.72
$y_i^{(0.1)}$	1.00	1.10	1.22	1.35	1.49	1.66	1.82	2.04	2.09	3.06	-0.13
$y_i^{(0.05)}$	1.00	1.10	1.22	1.35	1.48	1.42	-3.47	-122	-2927	-6.9*10^4	-1.6*10^6

Несмотря на малую локальную погрешность, результаты являются неудовлетворительными. Более того, можно проверить, что с последующим уменьшением шага они все ухудшаются. Для объяснения этого факта найдем общее решение уравнения (6.69). Для этого подставим в него $y_i = \zeta^i$. В результате после сокращения на ζ^n получим характеристическое уравнение

$$\zeta^2 + 4(1-h)\zeta - (5+2h) = 0. \tag{6.70}$$

В соответствии с теоремой 6.6.2 (см. ниже), общее решение уравнения (6.69) вычисляется по формуле

$$y_n = C_1\zeta_1^n(h) + C_2\zeta_2^n(h), \tag{6.71}$$

где C_1, C_2 — произвольные коэффициенты и для конкретного решения определяются по стартовым значениям, а $\zeta_1(h), \zeta_2(h)$ — корни уравнения (6.70)

$$\zeta_{1,2} = -2(1-h) \pm \sqrt{4(1-h)^2 + (5+2h)} = \\ = -2(1-h) \pm \sqrt{(3-h)^2 + 3h^2} = -2(1-h) \pm (3-h) + O(h^2).$$

Следовательно, $\zeta_1 = 1 + h + O(h^2)$ и приближает e^h, а ζ_1^n аппроксимирует e^x в точке $x = nh$. Второй корень $\zeta_2 = -5 + O(h)$. Он по модулю больше 1 и с ростом n становится очень большим. Поэтому даже при малом значении коэффициента C_2, второе слагаемое в правой части (6.71)[11] начинает со временем преобладать в решении y_n.

Поскольку нас интересует вопрос о сходимости метода (6.59), то есть о стремлении y_n к $u(x)$ при $h \to 0$ и $nh = x$, необходимо изучить поведение решения y_n при $h \to 0$ $(n \to \infty)$. Очевидно, что (6.59) при $h \to 0$ сводится к формуле

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = 0. \tag{6.72}$$

Ее можно рассматривать как численный метод (6.59), примененный к решению дифференциального уравнения $u' = 0$. Подставив в формулу (6.72) $y_j = \chi^j$ и разделив ее на χ^n, получим, что χ должно быть корнем уравнения

$$\rho(\chi) = \alpha_k\chi^k + \alpha_{k-1}\chi^{k-1} + \cdots + \alpha_0 = 0. \tag{6.73}$$

Справедливо следующее утверждение, которое приведем без доказательства

[10] Верхний индекс у y_i обозначает величину шага h.

[11] Это слагаемое часто называют **паразитным решением.**

Теорема 6.6.2 *Пусть многочлен $\rho(\chi)$ имеет корни $\chi_1, \ldots, \chi_r$ кратности $m_1, \ldots, m_r$ соответственно. Тогда общее решение уравнения (6.72) задается формулой*

$$y_n = p_1(n)\chi_1^n + \cdots + p_r(n)\chi_r^n, \tag{6.74}$$

где $p_j(n)$ — многочлены степеней $m_j - 1$

Из формулы (6.74) следует, что для ограниченности y_n при $n \to \infty$ требуется, чтобы корни (6.73) лежали в круге единичного радиуса, а корни, принадлежащие единичной окружности, были простыми.

Введем в связи с этим следующее определение.

Определение 6.6.3 *Многошаговый метод (6.59) называется* ***ноль-устойчивым*** *или* ***D-устойчивым*** *если корни многочлена $\rho(\chi)$ лежат на или внутри единичной окружности, причем корни, принадлежащие единичной окружности, являются простыми.*

В рассмотренном выше примере метод является неустойчивым.

Определение 6.6.4 *Многошаговый метод (6.59) называется* ***сходящимся*** *если для всех задач Коши $u' = f(x,u),\ u(x_0) = u_0$, удовлетворяющих требованиям теоремы существования и единственности, выполнено условие $y_n \to u(x)$ при $nh = x,\ h \to 0$, если для стартовых значений выполняются соотношения*

$$u(x_0 + ih) - y_i \to 0 \quad при \quad h \to 0, \quad i = 0, 1, \ldots, k-1.$$

Доказывается, что устойчивость и согласованность являются необходимыми и достаточными условиями сходимости многошагового метода. Более того, если метод устойчив и имеет порядок p, то $u(x_n) - y_n = O(h^p)$, если этому условию удовлетворяют стартовые значения.

6.6.2 Методы Адамса

Перейдем теперь к построению конкретных расчетных формул.

Предположим, что $x_n = x_0 + nh$ и известны приближенные значения

$$y_n,\ y_{n-1}, \ldots, y_{n-k+1}$$

точного решения

$$u(x_n), u(x_{n-1}), \ldots, u(x_{n-k+1})$$

задачи (6.5). Выведем формулу для нахождения y_{n+1}. Для этого проинтегрируем уравнение (6.5) на промежутке (x_n, x_{n+1}). Получим

$$u(x_{n+1}) = u(x_n) + \int\limits_{x_n}^{x_{n+1}} F(x)\,dx, \tag{6.75}$$

где $F(x) = f(x, u(x))$. В правую часть (6.75) входит искомое решение $u(x)$. Но поскольку известны его приближенные значения $y_n,\ y_{n-1}, \ldots, y_{n-k+1}$, можно найти величины $F_i = f(x_i, y_i)$ при $i = n-k+1, \ldots, n$. Поэтому естественно заменить функцию

$F(x)$, входящую в правую часть (6.75), интерполяционным многочленом, проходящим через точки $(x_i, F_i),\ i = n-k+1, \dots, n$. В результате, используя интерполяционный многочлен Ньютона $P_{k-1}(x)$, получим формулы **явных методов Адамса**

$$y_{n+1} = y_n + \int\limits_{x_n}^{x_{n+1}} P_{k-1}(x)\,dx = y_n + \int\limits_{x_n}^{x_{n+1}} \Big(F_n + (x-x_n)F(x_n, x_{n-1}) + \dots + \\ + (x-x_n)(x-x_{n-1})\dots(x-x_{n-k+2})F(x_n, \dots, x_{n-k+1})\Big)\,dx. \quad (6.76)$$

Выполняя интегрирование и выражая разделенные разности через значения функции F, находим расчетные формулы **явных методов Адамса** для различных значений k

$$k=1,\ \ y_{n+1} = y_n + hF_n,$$
$$k=2,\ \ y_{n+1} = y_n + h\Big(\frac{3}{2}F_n - \frac{1}{2}F_{n-1}\Big),$$
$$k=3,\ \ y_{n+1} = y_n + h\Big(\frac{23}{12}F_n - \frac{16}{12}F_{n-1} + \frac{5}{12}F_{n-2}\Big),$$
$$k=4,\ \ y_{n+1} = y_n + h\Big(\frac{55}{24}F_n - \frac{59}{24}F_{n-1} + \frac{37}{24}F_{n-2} - \frac{9}{24}F_{n-3}\Big).$$

Приведенные выше формулы получены при интегрировании интерполяционного многочлена Ньютона от x_n до x_{n+1}, то есть вне интервала интерполяции (x_{n-k+1}, x_n). Поэтому эти формулы иногда называют **экстраполяционными**. Известно, что при экстраполяции обычно получается приближение, которое существенно хуже, чем при интерполяции. Поэтому, для увеличения точности формул, заменим многочлен $P_{k-1}(x)$, на интерполяционный многочлен $P^*_{k-1}(x)$, использующий значения решения в точках $x_{n-k+2}, \dots, x_{n+1}$. В результате получим:

$$y_{n+1} = y_n + \int\limits_{x_n}^{x_{n+1}} P^*_{k-1}(x)\,dx = y_n + \int\limits_{x_n}^{x_{n+1}} \Big(F_{n+1} + (x-x_{n+1})F(x_{n+1}, x_n) + \dots + \\ + (x-x_{n+1})(x-x_n)\dots(x-x_{n-k+2})F(x_{n+1}, \dots, x_{n-k+1})\Big)\,dx. \quad (6.77)$$

Избавляясь как и ранее от разделенных разностей, получаем **неявные** или **интерполяционные формулы Адамса**.

$$k=1,\ \ y_{n+1} = y_n + hF_{n+1},$$
$$k=2,\ \ y_{n+1} = y_n + h\Big(\frac{1}{2}F_{n+1} + \frac{1}{2}F_n\Big),$$
$$k=3,\ \ y_{n+1} = y_n + h\Big(\frac{5}{12}F_{n+1} + \frac{8}{12}F_n - \frac{1}{12}F_{n-1}\Big),$$
$$k=4,\ \ y_{n+1} = y_n + h\Big(\frac{9}{24}F_{n+1} + \frac{19}{24}F_n - \frac{5}{24}F_{n-1} + \frac{1}{24}F_{n-2}\Big).$$

Определим локальную погрешность методов Адамса. Рассмотрим явные методы. В соответствии с определением локальной погрешности, необходимо оценить величину $u(x_k) - y_k$, при условии, что $y_i = u(x_i),\ i = 0, \dots, k-1$. Имеем с учетом (6.75),

(6.76)

$$u(x_k) - y_k = (u(x_k) - u(x_{k-1})) - (y_k - y_{k-1}) = \int\limits_{x_{k-1}}^{x_k} F(x)\,dx - \int\limits_{x_{k-1}}^{x_k} P_{k-1}(x)\,dx =$$

$$= \int\limits_{x_{k-1}}^{x_k} (F(x) - P_{k-1}(x))\,dx.$$

Следовательно,

$$|u(x_k) - y_k| \le h \max_{x\in[x_{k-1},\,x_k]} |F(x) - P_{k-1}(x)|. \tag{6.78}$$

Учитывая, что $P_{k-1}(x)$ — интерполяционный многочлен для функции $F(x)$, опираясь на оценку для погрешности интерполяции (см. пункт 3.1.2) имеем

$$\max_{x\in[x_{k-1},\,x_k]} |F(x) - P_{k-1}(x)| = O(h^k).$$

Отсюда и из (6.78) следует, что $|u(x_k) - y_k| = O(h^{k+1})$, то есть k-шаговый явный метод Адамса имеет порядок k. Аналогично доказывается, что k-шаговый неявный метод Адамса имеет порядок k.

Вопрос ноль-устойчивости решается для методов Адамса просто. Для явного и неявного методов имеем $\rho(\chi) = \chi^k - \chi^{k-1}$. Таким образом, кроме простого корня, равного 1, этот многочлен имеет еще нулевой корень кратности $k-1$. Поэтому методы Адамса ноль-устойчивы.

В заключение этого параграфа проведем сравнение методов Рунге-Кутта и Адамса. Рассмотрим, например, явный метод 4-го порядка Адамса и метод Рунге-Кутта 4-го порядка (6.26). Недостатком метода Рунге-Кутта следует считать тот факт, что для вычисления y_{n+1} по известному значению y_n потребуется 4 раза вычислить значение правой части $f(x,u)$ дифференциального уравнения. В том случае, когда эта функция имеет сложную структуру, указанная процедура может оказаться весьма трудоемкой. Метод же Адамса требует только однократного вычисления функции f на каждом шаге. В то же время метод Адамса требует нестандартное начало расчетов, так как для рассматриваемого примера требуется знать 4 стартовых значения, в то время как из начального условия известно только одно. Приходится находить недостающие условия, например, применяя метод Рунге-Кутта того же порядка или метод Эйлера с малым шагом. Это, а также тот факт, что метод построен для точек, расположенных с постоянным шагом, являются недостатком метода Адамса.

Подход Адамса для построения расчетных формул может быть обобщен. Было получено и исследовано много методов, основанных на соотношении

$$u(x_{n+1}) = u(x_n) + \int\limits_{x_{n-s}}^{x_{n+1}} f(x, u(x))\,dx. \tag{6.79}$$

В нем, как и выше, $f(x,u(x))$ заменяется интерполяционным многочленом $P(x)$ или $P^*(x)$.

6.7 ЗАДАЧИ К ГЛАВЕ 6

6.7.1 Примеры решения задач

1. Найти первые семь членов разложения в степенной ряд решения $u = u(x)$ задачи Коши

$$u'' + 0.1(u')^2 + (1 + 0.1x)u = 0, \quad u(0) = 1, \quad u'(0) = 2.$$

Решение. Решение ищем в виде степенного ряда

$$u(x) = u(0) + \frac{u'(0)}{1!}x + \frac{u''(0)}{2!}x^2 + \cdots + \frac{u^{(n)}(0)}{n!}x^n + \ldots$$

Непосредственно из начальных условий имеем $u(0) = 1,\ u'(0) = 2$. Для определения $u''(0)$ разрешим данное уравнение относительно u'':

$$u'' = -0.1(u')^2 - (1 + 0.1x)u. \tag{6.80}$$

Используя начальные условия, получим

$$u''(0) = -0.1 \cdot 4 - 1 \cdot 1 = -1.4.$$

Дифференцируем теперь последовательно по x левую и правую части равенства (6.80):

$$\begin{aligned} u''' &= -0.2u'u'' - 0.1(xu' + u) - u', \\ u^{IV} &= -0.2(u'u''' + u''^2) - 0.1(xu'' + 2u') - u'', \\ u^{V} &= -0.2(u'u^{IV} + 3u''u''') - 0.1(xu''' + 3u'') - u''', \\ u^{VI} &= -0.2(u'u^{V} + 4u''u^{IV} + 3u'''^2) - 0.1(xu^{IV} + 4u''') - u^{IV}. \end{aligned}$$

Подставляя начальные условия и значение $u''(0)$ находим:

$$u'''(0) = -1.54, \quad u^{IV}(0) = 1.224, \quad u^{V}(0) = 0.1768, \quad u^{VI}(0) = -0.7308.$$

Таким образом, искомое приближенное решение записывается в виде

$$u(x) \approx 1 + 2x - 0.7x^2 - 0.2567x^3 + 0.051x^4 + 0.00147x^5 - 0.000101x^6.$$

2. Вывести формулы для локальной погрешности явных методов Адамса.

Решение. Формулы Адамса для решения уравнения $u' = f(x, u)$ получаются при замене в равенстве

$$u(x_{n+1}) = u(x_n) + \int\limits_{x_n}^{x_{n+1}} F(x)\,dx,$$

где $F(x) = f(x, u(x)) = u'(x)$ функции $F(x)$ интерполяционным многочленом Ньютона $P_{k-1}(x)$. Следовательно, погрешность методов Адамса получается из-за погрешности интерполяции. Таким образом, учитывая формулу для погрешности интерполяции (3.15), локальная погрешность формул Адамса R_k равна

$$R_k = \int\limits_{x_n}^{x_{n+1}} (F(x) - P_{k-1}(x)\,dx = \int\limits_{x_n}^{x_{n+1}} \frac{F^{(k)}(\xi)}{k!}\omega_{k-1}(x)\,dx.$$

Здесь ξ — некоторая, зависящая от x точка, а функция $\omega_{k-1}(x) = (x - x_n)\ldots(x - x_{n-k+1})$ и, значит, положительна на интервале (x_n, x_{n+1}), так как $x_{n-j} = x_n - jh,\ h > 0$.

В математическом анализе доказывается теорема о среднем, в соответствии с которой существует такая точка $\eta \in (a,b)$, что

$$\int_a^b g_1(x)g_2(x)\,dx = g_1(\eta)\int_a^b g_2(x)\,dx$$

если $g_1(x)$ — непрерывная функция, а функция $g_2(x)$ не меняет знак на $[a,b]$. Поэтому для локальной погрешности формул Адамса R_k имеем

$$R_k = \frac{u^{(k+1)}(\eta)}{k!}\int_{x_n}^{x_{n+1}} (x-x_n)\dots(x-x_{n-k+1})\,dx.$$

В частности,

$$R_1 = u''(\eta)\int_{x_n}^{x_{n+1}} (x-x_n)\,dx = \frac{1}{2}u''(\eta)h^2 \quad \text{при} \quad k=1,$$

$$R_2 = \frac{u'''(\eta)}{2}\int_{x_n}^{x_{n+1}} (x-x_n)(x-x_{n-1})\,dx = \frac{5}{12}u'''(\eta)h^3 \quad \text{при} \quad k=2,$$

$$R_3 = \frac{u^{IV}(\eta)}{6}\int_{x_n}^{x_{n+1}} (x-x_n)(x-x_{n-1})(x-x_{n-2})\,dx = \frac{3}{8}u^{IV}(\eta)h^4 \quad \text{при} \quad k=3,$$

$$R_4 = \frac{u^{V}(\eta)}{24}\int_{x_n}^{x_{n+1}} (x-x_n)(x-x_{n-1})(x-x_{n-3})(x-x_{n-4})\,dx = \frac{251}{720}u^{V}(\eta)h^5 \quad \text{при} \quad k=4.$$

3. Получить методы, аналогичные явным методам Адамса для нахождения решения задачи Коши для уравнения второго порядка $u'' = f(x,u)$.

Решение. Пусть $x_n = x_0 + nh$. Интегрируя дифференциальное уравнение на промежутке $[x_n, x]$, имеем

$$u'(x) = u'(x_n) + \int_{x_n}^{x} f(t,u(t))\,dt.$$

Заменяя функцию $F(t) = f(t,u(t)) = u''(t)$ интерполяционным многочленом Ньютона

$$P_{k-1}(t) = F_n + (t-x_n)F(x_n,x_{n-1}) + \dots + (t-x_n)(t-x_{n-1})\dots(t-x_{n-k+2})F(x_n,\dots,x_{n-k+1}),$$

проходящем через точки (t_i, F_i), $i = n, n-1\dots, n-k+1$, $F_i = f(t_i)$, получим

$$u'(x) = u'(x_n) + \int_{x_n}^{x} P_{k-1}(t)\,dt + \int_{x_n}^{x} \big(F(t) - P_{k-1}(t)\big)\,dt. \tag{6.81}$$

Выражение, стоящее под вторым интегралом является погрешностью интерполяции, которую обозначим $r_k(t)$. Интегрируя равенство (6.81) по промежутку $[x_n, x_{n+1}]$, получим

$$u(x_{n+1}) = u(x_n) + hu'(x_n) + \int_{x_n}^{x_{n+1}}\int_{x_n}^{x} P_{k-1}(t)\,dtdx + \rho_k^{(1)}, \tag{6.82}$$

где

$$\rho_k^{(1)} = \int\limits_{x_n}^{x_{n+1}} \int\limits_{x_n}^{x} \big(F(t) - P_{k-1}(t)\big)\, dtdx$$

— остаточный член.

Проинтегрируем теперь равенство (6.81) по промежутку $[x_{n-1}, x_n]$. Тогда

$$u(x_n) = u(x_{n-1}) + hu'(x_n) + \int\limits_{x_{n-1}}^{x_n} \int\limits_{x_n}^{x} P_{k-1}(t)\, dtdx + \rho_k^{(2)}, \tag{6.83}$$

где

$$\rho_k^{(2)} = \int\limits_{x_{n-1}}^{x_n} \int\limits_{x_n}^{x} \big(F(t) - P_{k-1}(t)\big)\, dtdx$$

Вычитая (6.83) из (6.82), исключим слагаемое, содержащее u':

$$u(x_{n+1}) = 2u(x_n) - u(x_{n-1}) + \int\limits_{x_n}^{x_{n+1}} \int\limits_{x_n}^{x} P_{k-1}(t)\, dtdx - \int\limits_{x_{n-1}}^{x_n} \int\limits_{x_n}^{x} P_{k-1}(t)\, dtdx. \tag{6.84}$$

Если ввести новые переменные $t = x_n + h\tau,\ x = x_n + h\zeta$, то (6.84) перепишется в виде

$$u(x_{n+1}) = 2u(x_n) - u(x_{n-1}) + h^2 \int\limits_0^1 \int\limits_0^{\zeta} P_{k-1}(x_n + h\tau)\, d\tau d\zeta - h^2 \int\limits_{-1}^0 \int\limits_0^{\zeta} P_{k-1}(x_n + h\tau)\, d\tau d\zeta.$$

Заменяя теперь во втором интеграле ζ на $-\zeta$, имеем

$$\begin{gathered} u(x_{n+1}) - 2u(x_n) + u(x_{n-1}) = h^2 \int\limits_0^1 \int\limits_{-\zeta}^{\zeta} P_{k-1}(x_n + h\tau)\, d\tau d\zeta = \\ = h^2 \int\limits_0^1 \int\limits_{-\zeta}^{\zeta} \Big(F_n + h\tau F(x_n, x_{n-1}) + h^2\tau(\tau+1)F(x_n, x_{n-1}, x_{n-2}) + \cdots + \\ + h^{k-1}\tau(\tau+1)\ldots(\tau+k-2)F(x_n, \ldots, x_{n-k+1})\Big) d\tau d\zeta. \end{gathered} \tag{6.85}$$

Отсюда получаем формулы, которые называются **явными формулами Штермера**:

$$u(x_{n+1}) - 2u(x_n) + u(x_{n-1}) = h^2 F_n, \quad \text{при} \quad k = 2,$$

$$u(x_{n+1}) - 2u(x_n) + u(x_{n-1}) = h^2 F_n + \frac{h^4}{6} F(x_n, x_{n-1}, x_{n-2}) =$$

$$= h^2 \Big(\frac{13}{12} F_n - \frac{1}{6} F_{n-1} + \frac{1}{12} F_{n-2}\Big), \quad \text{при} \quad k = 3,$$

$$u(x_{n+1}) - 2u(x_n) + u(x_{n-1}) = h^2 F_n + \frac{h^4}{6} F(x_n, x_{n-1}, x_{n-2}) + \frac{h^5}{2} F(x_n, x_{n-1}, x_{n-2}, x_{n-3}) =$$

$$= h^2 \Big(\frac{7}{6} F_n - \frac{5}{12} F_{n-1} + \frac{1}{3} F_{n-2} - \frac{1}{12} F_{n-3}\Big), \quad \text{при} \quad k = 4.$$

6.7.2 Задачи

1. Вывести формулу, позволяющую вычислить третью и четвертую производную решения задачи Коши

$$u' = f(x,u), \quad u(x_0) = u_0,$$

которое ищется методом степенных рядов.

2. Найти приближенное решение задачи Коши на промежутке $0 \le x \le 0.5$

$$u'(x) = u^3(x) + x, \quad u(0) = 0$$

методом степенных рядов. Найти решение этой же задачи методами Пикара, Рунге-Кутта и сравнить полученные решения.

Ответ: $u(x) = x^2/2 + ...$

3. Исследовать на устойчивость метод Рунге-Кутта (6.19).

4. Покажите, что применение метода Рунге-Кутта (6.26) для решения задачи Коши $u'(x) = f(x), \quad x \in [a,b], \quad u(a) = 0$ приводит к формуле Симпсона для интегрирования функции $f(x)$.

5. Используя соотношение (6.79) с $s = 1$ и интерполяционный многочлен $P(x)$, построенный по узлам интерполяции $x_{n-k+1}, \ldots, x_n$, получить методы Нюстрема

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n), \quad \text{при } k = 1,$$
$$y_{n+1} = y_{n-1} + h\left(\frac{7}{3}f(x_n, y_n) - \frac{2}{3}f(x_{n-1}, y_{n-1}) + \frac{1}{3}f(x_{n-2}, y_{n-2})\right), \quad \text{при } k = 3.$$

6. Используя соотношение (6.79) с $s = 1$ и интерполяционный многочлен $P^*(x)$, построенный по узлам интерполяции $x_{n-k+1}, \ldots, x_{n+1}$, получить методы Милна-Симпсона

$$y_{n+1} = y_{n-1} + 2hf(x_{n+1}, y_{n+1}), \quad \text{при } k = 0,$$
$$y_{n+1} = y_{n-1} + h\left(\frac{1}{3}f(x_{n+1}, y_{n+1}) + \frac{4}{3}f(x_n, y_n) + \frac{1}{3}f(x_{n-1}, y_{n-1})\right), \quad \text{при } k = 2.$$

7. Используя соотношение (6.79) с $s = 4$ и интерполяционный многочлен $P(x)$, построенный по узлам интерполяции x_{n-2}, x_{n-1}, x_n, получить метод Милна

$$y_{n+1} = y_{n-3} + h\left(\frac{8}{3}f(x_n, y_n) - \frac{4}{3}f(x_{n-1}, y_{n-1}) + \frac{8}{3}f(x_{n-2}, y_{n-2})\right).$$

8. Доказать, что применение к линейному дифференциальному уравнения с постоянными коэффициентами $u'(x) + pu(x) + q = 0$ любого 2-х этапного метода Рунге-Кутта второго порядка дает один и тот же результат.

9. Вывести формулы для локальной погрешности неявных методов Адамса.

10. Вывести неявные формулы Штермера для нахождения решения задачи Коши для уравнения второго порядка $u'' = f(x,u)$.

6.7.3 Примеры тестовых вопросов к главе 6

1. Среди систем линейных дифференциальных уравнений $\mathbf{u}' + \mathbf{A}\mathbf{u} = \mathbf{0}$, где $\mathbf{A}$ — заданная матрица, выбрать ту, которая обладает наибольшей жесткостью.

а) $\mathbf{A} = \begin{pmatrix} 1000 & 1 & 1 & 1 \\ 0 & 200 & 1 & 1 \\ 0 & 0 & 30 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$;

б) $\mathbf{A} = \begin{pmatrix} -1000 & 1 & 1 & 1 \\ 0 & 200 & 1 & 1 \\ 0 & 0 & -30 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$;

в) $\mathbf{A} = \begin{pmatrix} -1 & 1 & 1 & 1 \\ 0 & 20 & 1 & 1 \\ 0 & 0 & -3500 & 1 \\ 0 & 0 & 0 & 4000 \end{pmatrix}$;

г) $\mathbf{A} = \begin{pmatrix} -1 & 1 & 1 & 1 \\ 0 & -20 & 1 & 1 \\ 0 & 0 & -300 & 1 \\ 0 & 0 & 0 & -4000 \end{pmatrix}$.

2. Задача $u' + 10u = e^{\sin x},\ u(0) = 1$ решается неявным методом Эйлера. Выберите, какое ограничение накладывает на шаг h условие устойчивости.

а) $h < 1/100$;

б) $h < 1/50$;

в) $h < 1$;

г) $h < 50$;

д) $h < 100$;

е) $h < \infty$.

3. Задача $u' = u + x,\ u(0) = 0$ решается методом степенных рядов. Ищется разложение решения в ряд вплоть до членов x^3. Чему равно это приближение?

а) $1 + x + x^2/2$;

б) $x + x^3$;

в) $x^2/2! + x^3/3!$;

г) $1 + x + x^2/2! + x^3/3!$;

д) $x^2 + x^3$;

е) x.

4. Пусть

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = (h\beta_k f_{n+k} + \beta_{k-1} f_{n+k-1} + \cdots + \beta_0 f_n)$$

— многошаговый метод для решения задачи Коши $u' = f(x,u),\ u(x_0) = u_0$. Какое из приведенных ниже условий является условием согласованности, если $k = 2$?

а) $\alpha_0 + \alpha_1 + \alpha_2 = 0,\ \alpha_1 + 2\alpha_2 = \beta_1 + \beta_2$;

б) $\alpha_0 + \alpha_1 + \alpha_2 = 1,\ \alpha_1 + 2\alpha_2 = \beta_0 + \beta_1 + \beta_2$;

в) $\alpha_0 + \alpha_1 + \alpha_2 = 0,\ \alpha_1 + 2\alpha_2 = \beta_0 + \beta_1 + \beta_2$;

г) $\alpha_0 + \alpha_1 + 2\alpha_2 = 0,\ \beta_0 + \beta_1 + \beta_2 = 0$;

д) $\alpha_0 + \alpha_1 + 2\alpha_2 = 0,\ \alpha_0 + \alpha_1 + \alpha_2 = \beta_1 + \beta_2$.

5. Рассматриваются два метода 4-го порядка для решения задачи Коши $u' = f(x,u),\ u(0) = u_0$ — Рунге-Кутта и Адамса. Какие из перечисленных ниже свойств присущи методу Адамса, а какие Рунге-Кутта? Ответ запишите в виде: буква А и далее без пробела перечень номеров свойств в порядке возрастания, присущих методу Адамса, пробел, буквы РК и далее без пробела перечень номеров свойств в порядке возрастания, присущих методу Рунге-Кутта (например, А123 РК45678).

1) Многократное вычисление функции f при совершении одного шага метода;

2) однократное вычисление функции f при совершении одного шага метода;

3) специальные методы вычисления начальных точек;

4) вычисление только с постоянным шагом;

5) возможность менять шаг в процессе вычисления не меняя алгоритма вычисления.

6. Пусть s — порядок точности метода Рунге-Кутта, которым ищется приближенное решение задачи Коши. Пусть ε — точность, с которой требуется найти решение. Выбран шаг h и получена при этом оценка погрешности метода, которая равна err. По какой формуле в практических расчетах вычисляют оптимальное значение шага, обеспечивающее заданную точность?

а) $h_{opt} = h(\varepsilon/err)^{s+1}$;

б) $h_{opt} = h(\varepsilon/err)^{1/(s+1)}$;

в) $h_{opt} = 0.9h(\varepsilon/err)^{s+1}$;

г) $h_{opt} = 0.8h(\varepsilon/err)^{1/(s+1)}$;

д) $h_{opt} = h(\varepsilon \cdot err)^{s+1}$.

7 РЕШЕНИЕ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ. ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ

Этот раздел посвящен в основном методам решения краевых задач для обыкновенных дифференциальных уравнений. В заключительном параграфе раздела рассматриваются интегральные уравнения. Между краевыми задачами и интегральными уравнениями существует тесная связь. Существуют методы сведения некоторых краевых задач к интегральным уравнениям и наоборот. Здесь будет рассмотрен метод, не связанный с редукцией интегрального уравнения к краевой задаче.

Напомним сначала как формулируются краевые задачи.

Пусть имеется уравнение m-го порядка ($m \geq 2$) относительно неизвестной функции $u(x)$

$$F(x, u, u', \ldots, u^{(m)}) = 0, \quad a \leq x \leq b. \tag{7.1}$$

Для однозначного определения решения этого уравнения нужны дополнительный условия. В случае задачи Коши такими условиями являются заданные в некоторой точке отрезка $[a, b]$ значения функции u и ее производных до порядка $m-1$ включительно. Однако на практике часто встречаются задачи, когда дополнительные условия известны не в одной, а в нескольких точках $a \leq x_1 \leq \cdots \leq x_k \leq b$, то есть заданы условия вида

$$\begin{gathered}\varphi_i\big(u(x_1), u'(x_1), \ldots, u^{(m-1)}(x_1), u(x_2), u'(x_2), \ldots, u^{(m-1)}(x_2), \\ \ldots, u(x_k), u'(x_k), \ldots, u^{(m-1)}(x_k)\big) = 0, \quad i = 1, \ldots, m.\end{gathered} \tag{7.2}$$

Их называют **краевыми** или **граничными условиями**. Решить краевую задачу (7.1),(7.2), значит найти на отрезке $[a, b]$ такую функцию $u(x)$, которая обращает (7.1),(7.2) в тождество.

На практике чаще всего встречаются задачи, в которых условия (7.2) заданы в двух, как правило, крайних точках отрезка $[a, b]$, то есть $k = 2$. Отсюда и пошло название краевые или граничные условия. Краевые задачи называют **линейными**, если функции F, φ_i — линейные относительно искомой функции u и ее производных.

По аналогии легко сформулировать постановку краевой задачи для системы дифференциальных уравнений. Для системы m уравнений первого порядка ($m \geq 2$) задаются m соотношений в k точках ($k \geq 2$), связывающие значения неизвестных функций в этих точках.

Как и для задачи Коши существует большое число различных методов численного решения краевых задач. В этом параграфе на примере решения простых уравнений и систем будут рассмотрены лишь некоторые, наиболее употребительные методы.

7.1 МЕТОД СТРЕЛЬБЫ

Метод стрельбы заключается в сведении решения краевой задачи к решению последовательности задач Коши. Изложим идею метода на примере задачи

$$u_1' = f_1(x, u_1, u_2), \quad u_2' = f_2(x, u_1, u_2), \quad x \in [a, b], \tag{7.3}$$

$$\varphi_1(u_1(a), u_2(a)) = 0, \tag{7.4}$$

$$\varphi_2(u_1(b), u_2(b)) = 0. \tag{7.5}$$

Предположим, что функция φ_1 такова, что позволяет выразить $u_2(a)$ через значение $u_1(a)$, то есть $u_2(a) = \beta(u_1(a))$. Зададим пока произвольным образом $u_1(a)$, положив $u_1(a) = \alpha$, где α — произвольное число. Тогда из (7.4) имеем $u_2(a) = \beta(\alpha)$. Выбирая α и β в качестве начальных значений в точке a для функций u_1, u_2 соответственно, решаем задачу Коши для системы (7.3). Полученные решения зависят от переменной x и параметра α. Этот параметр надо выбрать так, чтобы выполнялось второе граничное условие (7.5)

$$\Phi(\alpha) = \varphi_2(u_1(b, \alpha), u_2(b, \alpha)) = 0. \tag{7.6}$$

Уравнение $\Phi(\alpha) = 0$ можно решать любым пригодным для этого способом. Например, в случае применения метода деления отрезка пополам, надо подобрать такие значения α_1, α_2, чтобы знаки величин $\Phi(\alpha_1)$ и $\Phi(\alpha_2)$ были противоположными. Затем вычисляется $\alpha_3 = (\alpha_1 + \alpha_2)/2$ и $\Phi(\alpha_3)$. Из двух полученных отрезков на оси $O\alpha$ выбирается тот, на концах которого функция $\Phi(\alpha)$ принимает значения разных знаков и т.д. до достижения заданной точности $|\Phi(\alpha_n)| < \varepsilon$.

При численной реализации метода надо только учесть, что функция $\Phi(\alpha)$ задана не аналитически, а алгоритмически. В качестве алгоритма выступает метод решения задачи Коши, например, Рунге-Кутта.

Решение существенно упрощается в случае линейной краевой задачи. Рассмотрим для примера задачу

$$u_1' = a_{11}u_1 + a_{12}u_2 + b_1, \quad u_2' = a_{21}u_1 + a_{22}u_2 + b_2, \quad x \in [a, b], \tag{7.7}$$

$$c_1u_1(a) + c_2u_2(a) = c_3, \tag{7.8}$$

$$d_1u_1(b) + d_2u_2(b) = d_3, \tag{7.9}$$

где a_{ij}, b_i — функции переменной x, а c_i, d_i — числа, причем $c_1^2 + c_2^2 \neq 0,\ d_1^2 + d_2^2 \neq 0$. Пусть для определенности $c_2 \neq 0$.

Вместо задачи (7.7)-(7.9) рассмотрим вспомогательные задачи Коши

$$u_{11}' = a_{11}u_{11} + a_{12}u_{21}, \quad u_{21}' = a_{21}u_{11} + a_{22}u_{21}, \quad x \in [a, b], \tag{7.10}$$

$$u_{11}(a) = 1, \qquad u_{21}(a) = -\frac{c_1}{c_2}. \tag{7.11}$$

Здесь (7.10) — однородная система уравнений, соответствующая системе (7.7), а величины $u_{11}(a),\ u_{21}(a)$ удовлетворяют однородному условию, соответствующему (7.8).

$$u_{12}' = a_{11}u_{12} + a_{12}u_{22} + b_1, \quad u_{22}' = a_{21}u_{12} + a_{22}u_{22} + b_2, \quad x \in [a, b], \tag{7.12}$$

$$u_{12}(a) = 0, \quad u_{22}(a) = \frac{c_3}{c_2}, \tag{7.13}$$

Заметим, что $u_{12}(a),\ u_{22}(a)$ удовлетворяют условию (7.8). Легко проверить, что функции

$$u_1(x) = \alpha u_{11}(x) + u_{12}(x),\ u_2(x) = \alpha u_{21}(x) + u_{22}(x) \tag{7.14}$$

удовлетворяют при любом значении α системе уравнений (7.7) и краевому условию (7.8)[1]. Осталось подобрать α так, чтобы выполнялось второе краевое условие (7.9). Для этого достаточно положить

$$\alpha = \frac{d_3 - d_1 u_{12}(b) - d_2 u_{22}(b)}{d_1 u_{11}(b) + d_2 u_{21}(b)}.$$

Подставляя теперь это значение α в (7.14), получим решение задачи (7.7)-(7.9). Таким образом, в линейном случае вместо решения последовательности задач Коши достаточно ограничиться решением всего двух задач Коши.

Следует однако заметить, что к результатам, полученным методам стрельбы необходимо относиться с определенной осторожностью. Проиллюстрируем проблемы, которые могут возникнуть при применении метода стрельбы. Рассмотрим краевую задачу

$$u_1' = au_2, \quad u_2' = au_1, \quad x \in [0,1], \quad a = const > 0, \tag{7.15}$$

$$u_2(0) = c, \quad u_2(1) = d, \tag{7.16}$$

решение которой имеет вид

$$u_1(x) = \frac{-e^{-ax} - e^{-a(2-x)}}{1 - e^{-2a}} c + \frac{e^{-a(1-x} + e^{-a(1+x)}}{1 - e^{-2a}} d,$$
$$u_2(x) = \frac{e^{-ax} - e^{-a(2-x)}}{1 - e^{-2a}} c + \frac{e^{-a(1-x} - e^{-a(1+x)}}{1 - e^{-2a}} d.$$

Получим теперь это решение методом стрельбы. В соответствии с методом запишем имеем две задачи

$$u_{11}' = au_{21}, \quad u_{21}' = au_{11}, \quad u_{11}(0) = 1, \quad u_{21}(0) = 0,$$
$$u_{12}' = au_{22}, \quad u_{22}' = au_{12}, \quad u_{12}(0) = 0, \quad u_{22}(0) = c.$$

Их решение

$$u_{11}(x) = \frac{1}{2}(e^{ax} + e^{-ax}), \quad u_{21}(x) = \frac{1}{2}(e^{ax} - e^{-ax}),$$
$$u_{12}(x) = \frac{c}{2}(e^{ax} - e^{-ax}), \quad u_{22}(x) = \frac{c}{2}(e^{ax} + e^{-ax}).$$

Значит

$$u_1(x) = \frac{\alpha + c}{2} e^{ax} + \frac{\alpha - c}{2} e^{-ax}, \quad u_2(x) = \frac{\alpha + c}{2} e^{ax} + \frac{c - \alpha}{2} e^{-ax}.$$

Предположим, что при нахождении α допущена ошибка $\delta\alpha$, то есть вместо величины α получена величина $\alpha + \delta\alpha$. Тогда решение будет получено с ошибкой. Например, ошибка $\delta u_1(x)$ вычисления функции $u_1(x)$ равна

$$\delta u_1(x) = \frac{\delta\alpha}{2} e^{ax} + \frac{\delta\alpha}{2} e^{-ax}.$$

[1]Отметим, что по построению $u_1(a) = \alpha$.

В частности, при $x = 1$ и большом значении a имеем

$$\delta u_1(1) = \frac{\delta\alpha}{2}(e^a + e^{-a}) \approx \frac{\delta\alpha}{2}e^a.$$

Таким образом, даже при малом значении $\delta\alpha$ ошибка при нахождении решения будет большой, так как величина e^a велика. Поэтому метод стрельбы при решении задачи (7.15),(7.16), будучи формально приемлемой процедурой, при больших a становится практически непригодным.

7.2 РАЗНОСТНЫЙ МЕТОД

Рассмотрим сначала метод на примере решения линейной краевой задачи для уравнения второго порядка

$$u''(x) - p(x)u(x) = f(x), \quad x \in [a, b], \tag{7.17}$$

$$u(a) = \alpha, \quad u(b) = \beta. \tag{7.18}$$

Разобьем отрезок $[a, b]$ на n частей, которые для простоты будем считать равными. Определим шаг $h = (b-a)/n$ и точки $x_i = a + ih,\ i = 0, \ldots, n$. Совокупность точек x_i называется **сеткой**, а каждая из точек — **узлом сетки.** При этом, так как расстояние между соседними узлами одинаково, сетку называют **равномерной**.

В параграфе 4.1 была получена формула (4.5), из которой следует, что для четыре раза непрерывно дифференцируемой функции $u(x)$ выполняется равенство

$$u''(x_i) = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} - \frac{h^2}{12}u^{IV}(\xi_i), \quad \xi_i \in [x_{i-1}, x_{i+1}]. \tag{7.19}$$

Возьмем в (7.17) $x = x_i,\ i = 1, \ldots, n-1$ и заменим затем вторую производную решения в соответствии с формулой (7.19). Вводя обозначения $p_i = p(x_i),\ f_i = f(x_i)$, получаем

$$\frac{u(x_{i-1}) - (2 + h^2p_i)u(x_i) + u(x_{i+1})}{h^2} = f_i + \frac{h^2}{12}u^{IV}(\xi_i). \tag{7.20}$$

Обозначим через y_i приближенное значение решения в точке x_i. Для построения уравнений для y_i отбросим в (7.20) слагаемое порядка h^2. В результате получим систему, состоящую из $(n-1)$-го уравнения

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = f_i \quad i = 1, \ldots, n-1. \tag{7.21}$$

Таким образом, каждое из уравнений системы получилось путем замены производной **конечной разностью**, то есть выражением $(y_{i-1} - 2y_i + y_{i+1})/(h^2)$. Отсюда и произошло название метода. Если воспользоваться граничными условиями (7.18), можно положить $y_0 = \alpha,\ y_n = \beta$. В результате, для нахождения величин y_i имеем

$$\begin{cases} y_0 = \alpha, \\ y_{i-1} - (2 + h^2p_i)y_i + y_{i+1} = f_ih^2, \quad i = 1, \ldots, n-1, \\ y_n = \beta. \end{cases} \tag{7.22}$$

Матрица системы (7.22) является трехдиагональной, поэтому система может быть решена методом прогонки. В связи с этим, рассматриваемый метод решения краевой задачи иногда называют **методом прогонки.** Заметим, что если $p(x) \geq 0$, то

выполнено условие диагонального преобладания и, как было показано ранее, система (7.22) однозначно разрешима, причем прогонка устойчива. Будем в дальнейшем считать, что выполнено более жесткое условие $p(x) \geq p^{(0)} > 0$ на $[a, b]$.

Итак, нахождение приближенного решения краевой задачи в точках x_i свелось к решения системы линейных алгебраических уравнений. Естественно при этом возникает вопрос о том, насколько близки значения y_i и $u(x_i)$? Для ответа на этот вопрос введем величины $v_i = u(x_i) - y_i$. Получим теперь систему уравнений, которой удовлетворяют введенные величины. С этой целью вычтем из уравнений (7.20) уравнения (7.21). После простых преобразований получим

$$\begin{cases} v_0 = 0, \\ (2 + h^2 p_i) v_i = v_{i-1} + v_{i+1} - \dfrac{h^4}{12} u^{IV}(\xi_i), \ i = 1, \ldots, n-1, \\ v_n = 0. \end{cases} \tag{7.23}$$

Пусть $\|v\| = \max\limits_{i=0,\ldots,n} |v_i| = |v_m|$ Последнее равенство означает, что m то значение номера, при котором $|v_i|$ принимает максимальное значение. Поскольку $v_0 = v_n = 0$, можно считать, что $0 < m < n$. Выбирая в (7.23) $i = m$ и учитывая предположение о том, что $p(x) \geq p^{(0)} > 0$ имеем

$$(2 + h^2 p_m)\|v\| = \left| v_{m-1} + v_{m+1} - \frac{h^4}{12} u^{IV}(\xi_m) \right| \leq$$
$$\leq |v_{m-1}| + |v_{m+1}| + \frac{h^4}{12} |u^{IV}(\xi_m)| \leq 2\|v\| + \frac{h^4}{12} \max_{x \in [a,b]} |u^{IV}(x)|.$$

Отсюда следует, что

$$\|v\| \leq \frac{\max\limits_{x \in [a,b]} |u^{IV}(x)|}{12 p^{(0)}} h^2. \tag{7.24}$$

Таким образом показано, что если решение задачи (7.17),(7.18) существует, имеет непрерывную четвертую производную и $p(x) \geq p^{(0)} > 0$, то $|y_i - u(x_i)| = O(h^2)$.

Замечание 1. Из оценки (7.24) следует, что в отличии от метода стрельбы, погрешность при нахождении решения тем меньше, чем значения функции $p(x)$ больше.

Замечание 2. Если $p(x) < 0$, то достаточное условие устойчивости прогонки не выполнено. Однако, обычно в практике численных расчетов нарушение этого условия не вызывает заметного ухудшения устойчивости. Исключение составляют те случаи, когда определитель системы близок к нулю. Для опознания таких ситуаций проводят расчеты при нескольких разбиениях отрезка $[a, b]$ на части с уменьшающимся шагом h. Если при этом все решения близки, то можно сделать вывод, что решение краевой задачи приближенно найдено.

Замечание 3. Даже в том случае, когда $p(x) > 0$ и, значит, прогонка устойчива, очень маленькое значение шага может привести к существенным ошибкам. Для того, чтобы пояснить это достаточно заметить, что если $h^2 < \varepsilon_{\text{маш}}$ (здесь $\varepsilon_{\text{маш}}$ — машинный эпсилон см. параграф 1.2), а $0 < p(x) \leq 1$, то при вычислении на ЭВМ коэффициентов системы (7.22) из-за округления окажется, что $2 + h^2 p_i$ равно 2. Поэтому, если даже все остальные вычисления будут выполнены точно, фактически, будет найдено решение системы

$$\begin{cases} \widetilde{y}_0 = \alpha, \\ \widetilde{y}_{i-1} - 2\widetilde{y}_i + \widetilde{y}_{i+1} = f_i h^2, \quad i = 1, \ldots, n-1, \\ \widetilde{y}_n = \beta, \end{cases}$$

которая соответствует задаче

$$u''(x) = f(x), \quad x \in [a,b],$$
$$u(a) = \alpha, \quad u(b) = \beta.$$

Таким образом, будет найдено приближенное решение другой задачи.

Перейдем теперь к некоторым обобщениям. Рассмотрим на отрезке $[a,b]$ задачу

$$u'' + A(x)u' + B(x)u = C(x), \tag{7.25}$$
$$F_1 u(a) + D_1 u'(a) = E_1, \tag{7.26}$$
$$F_2 u(b) + D_2 u'(b) = E_2. \tag{7.27}$$

Как и ранее введем сетку $x_i = a + ih,\ i = 0, \ldots, n,\ h = (b-a)/n$ и попытаемся найти приближенные значения решения в узлах сетки. Приближенные значения решения задачи (7.25)-(7.27) в точке x_i обозначим y_i. Помимо соотношения (7.19) понадобятся равенства

$$u'(x_i) = \frac{u(x_{i+1}) - u(x_i)}{h} - \frac{h}{2}u''(x_i) + O(h^2), \tag{7.28}$$
$$u'(x_i) = \frac{u(x_i) - u(x_{i-1})}{h} + \frac{h}{2}u''(x_i) + O(h^2), \tag{7.29}$$
$$u'(x_i) = \frac{u(x_{i+1}) - u(x_{i-1})}{2h} + O(h^2). \tag{7.30}$$

Используя (7.19),(7.28)–(7.30), заменим производные в (7.25)–(7.27) соответствующими конечными разностями. В результате получим систему линейных алгебраических уравнений для нахождения приближенного решения в узлах сетки

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + A(x_i)\frac{y_{i+1} - y_{i-1}}{2h} + B(x_i)y_i = C(x_i), \quad i = 1, \ldots, n-1, \tag{7.31}$$
$$F_1 y_0 + D_1 \frac{y_1 - y_0}{h} = E_1, \tag{7.32}$$
$$F_2 y_n + D_2 \frac{y_n - y_{n-1}}{h} = E_2. \tag{7.33}$$

Система уравнений (7.31)–(7.33) называется **разностной схемой**. Если в (7.31) перенести $C(x_i)$ в левую часть и вместо y подставить u, то левая часть не будет равна нулю. Ее порядок в силу (7.19),(7.30),(7.25) будет равен $O(h^2)$. В этом случае говорят, что **разностные уравнения (7.31) аппроксимируют дифференциальное уравнение (7.25) со вторым порядком.**

Если в (7.32) перенести E_1 в левую часть и вместо y подставить u, то порядок левой части в силу (7.28),(7.26) будет равен $O(h)$. Аналогичное утверждение справедливо для (7.33). Это означает, что **граничные условия аппроксимируются с первым порядком.** Говоря о приближении в целом, то есть всей разностной схемы, естественно ориентироваться на те соотношения, которые приближают хуже всего. Поэтому разностная схема (7.31)–(7.33) аппроксимирует краевую задачу (7.25)–(7.27) с первым порядком.

В связи с этим встает вопрос, можно ли так изменить разностные граничные условия (7.32),(7.33) чтобы получить второй порядок аппроксимации. Рассмотрим граничное условие (7.26). Перепишем его с учетом (7.28) в виде

$$F_1 u(a) + D_1\left(\frac{u(x_1) - u(a)}{h} - \frac{h}{2}u''(a) + O(h^2)\right) = E_1. \tag{7.34}$$

Выразим теперь $u''(a)$ из (7.25) и подставим в (7.34)

$$F_1u(a) + D_1\left(\frac{u(x_1)-u(a)}{h} - \frac{h}{2}\Big(C(a) - B(a)u(a) - A(a)u'(a)\Big) + O(h^2)\right) = E_1$$

или с учетом (7.28)

$$\begin{aligned} &F_1u(a)+\\ &+D_1\left(\frac{u(x_1)-u(a)}{h} - \frac{h}{2}\Big(C(a) - B(a)u(a) - A(a)\frac{u(x_1)-u(a)}{h} + O(h)\Big) + O(h^2)\right) = E_1 \end{aligned} \tag{7.35}$$

Если теперь отбросить в (7.35) слагаемые порядка $O(h^2)$, получим искомое разностное граничное условие

$$F_1y_0 + D_1\frac{y_1-y_0}{h} - \frac{h}{2}D_1\Big(C(a) - B(a)y_0 - A(a)\frac{y_1-y_0}{h}\Big) = E_1. \tag{7.36}$$

Аналогично можно получить второе разностное граничное условие

$$F_2y_n + D_2\frac{y_n-y_{n-1}}{h} + \frac{h}{2}D_2\Big(C(b) - B(b)y_n - A(b)\frac{y_n-y_{n-1}}{h}\Big) = E_2. \tag{7.37}$$

В результате получается разностная схема (7.31), (7.36),(7.37), которая аппроксимирует краевую задачу со вторым порядком.

Рассмотрим теперь вопрос об организации расчетов по предложенной разностной схеме. Для этого в уравнениях (7.36), (7.31),(7.37) приведем подобные и перепишем их в виде

$$\begin{cases} -b_0y_0 + c_0y_1 &= d_0, \\ a_iy_{i-1} - b_iy_i + c_iy_{i+1} &= d_i, \quad i = 1,\dots,n-1, \\ a_ny_{n-1} - b_ny_n &= d_n, \end{cases} \tag{7.38}$$

где

$$b_0 = -F_1h + D_1 + D_1\big(A(a) - B(a)h\big)\frac{h}{2}, \quad c_0 = A(a)D_1\frac{h}{2} + D_1, \quad d_0 = E_1h + C(a)D_1\frac{h^2}{2},$$

$$a_i = 1 - A(x_i)\frac{h}{2}, \quad b_i = 2 - B(x_i)h^2, \quad c_i = 1 + A(x_i)\frac{h}{2}, \quad d_i = C(x_i)h^2, \quad i = 1,\dots,n-1,$$

$$a_n = A(b)D_2\frac{h}{2} - D_2, \quad b_n = -F_2h - D_2 + D_2\big(A(b) + B(b)h\big)\frac{h}{2}, \quad d_n = E_2h - C(b)D_2\frac{h^2}{2}.$$

Система уравнений (7.38) решается методом прогонки.

При практической оценке погрешности приближенного решения можно воспользоваться принципом Рунге. В соответствии с этим принципом, находятся приближенные решения задачи при разбиении отрезка интегрирования с шагом h и $h/2$. Обозначим полученные значения в узле x соответственно через $y(x)$ и $\widetilde{y}(x)$. Тогда $|\widetilde{y}(x) - u(x)| \approx |y(x) - \widetilde{y}(x)|/3$ для случая второго порядка аппроксимации и $|\widetilde{y}(x) - u(x)| \approx |y(x) - \widetilde{y}(x)|$ для первого порядка аппроксимации.

Выше был рассмотрен случай линейной краевой задачи. Значительно труднее решать нелинейные задачи. Рассмотрим, например, краевую задачу

$$u'' = f(x,u), \quad x \in [a,b], \tag{7.39}$$

$$u(a) = \alpha, \quad u(b) = \beta. \tag{7.40}$$

Будем предполагать, что функция $f(x,u)$ ограничена и непрерывна вместе со своими вторыми производными.

По аналогии с линейной задачей введем равномерную сетку с шагом h, заменим в уравнении (7.39) вторую производную по формуле (7.19) и отбросим слагаемые порядка $O(h^2)$. В результате получим систему нелинейных алгебраических уравнений

$$\begin{cases} y_0 = \alpha, \\ y_{i-1} - 2y_i + y_{i+1} = f(x_i, y_i)h^2 \quad i = 1, \ldots, n-1, \\ y_n = \beta. \end{cases} \tag{7.41}$$

Предполагая, что

$$\Big|\frac{\partial f}{\partial u}\Big| \geq m_1 > 0,$$

можно доказать неравенство

$$|u(x_i) - y_i| \leq \frac{h^2}{12m_1} \max_{x\in[a,b]} |u^{IV}(x)|.$$

Это означает, что при $h \to 0$ приближенное решение сходится равномерно со вторым порядком к точному решению.

Как уже отмечалось, система уравнений (7.41) нелинейная. Поэтому следует обсудить метод, которым она может быть решена. Одним из возможных подходов является применение метода последовательных приближений

$$\begin{cases} y_0^{k+1} = \alpha, \\ y_{i-1}^{k+1} - 2y_i^{k+1} + y_{i+1}^{k+1} = f(x_i, y_i^k)h^2, \quad i = 1, \ldots, n-1, \\ y_n^{k+1} = \beta. \end{cases} \tag{7.42}$$

Здесь верхний индекс означает номер итерации. Для определения y_i^{k+1} на каждой итерации получается система линейных алгебраических уравнений, решаемая методом прогонки. Доказывается, что итерации сходятся при выполнении условия

$$\frac{(b-a)^2}{8} \max\Big|\frac{\partial f}{\partial u}\Big| < 1.$$

Другим подходом для решения (7.41) является применение метода Ньютона. Для этого будем считать, что приближение y_i^k известно. Подставим в (7.41) $y_i = y_i^k + \delta_i$ и воспользуемся тем, что

$$f(x_i, y_i) = f(x_i, y_i^k + \delta_i) = f(x_i, y_i^k) + \frac{\partial f(x_i, y_i^k)}{\partial u}\delta_i + O(\delta_i^2).$$

В результате, если отбросим слагаемое порядка $O(\delta_i^2)$, получим систему для нахождения приближения δ_i^k к величине δ_i, с помощью которого находится очередное значение $y_i^{k+1} = y_i^k + \delta_i^k$. Эта система имеет вид

$$\begin{cases} \delta_0^k = 0, \\ \delta_{i-1}^k - \Big(2 + h^2\dfrac{\partial f(x_i, y_i^k)}{\partial u}\Big)\delta_i^k + \delta_{i+1}^k = f(x_i, y_i^k)h^2 - y_{i-1}^k + 2y_i^k - y_{i+1}^k, \quad i = 1, \ldots, n-1, \\ \delta_n^k = 0 \end{cases}$$

и решается методом прогонки.

7.3 МЕТОД ГАЛЕРКИНА

Рассмотрим в общих чертах еще один метод решения краевых задач, который носит название **метода Галеркина**[2].

Пусть в гильбертовом пространстве $\mathcal{H}$ требуется решить уравнение

$$\mathcal{L}u = f. \tag{7.43}$$

Здесь $\mathcal{L}$ — оператор с областью определения $D(\mathcal{L}) \in \mathcal{H}$, причем замыкание множества $D(\mathcal{L})$ совпадает с $\mathcal{H}$, а f — заданный элемент из $\mathcal{H}$.

Напомним [20], что если элементы $\varphi_1, \varphi_2, \ldots$ гильбертова пространства образуют базис, то любой элемент, ортогональный всем векторам базиса равен нулевому элементу. Поэтому если бы нашелся такой элемент $u \in D(\mathcal{L})$, что при всех k для скалярного произведения выполняется равенство

$$(\mathcal{L}u - f, \varphi_k) = 0, \tag{7.44}$$

то это означало бы, что u удовлетворяет уравнению (7.43) и, таким образом, является искомым решением.

Введем последовательность конечномерных подпространств $\mathcal{H}_n \subset \mathcal{H}$. Будем считать, что выполнено **свойство полноты**, которое в данном случае означает следующее: для любых $u \in \mathcal{H}$ и $\varepsilon > 0$ существует число $\widetilde{n} = \widetilde{n}(u, \varepsilon)$ такое, что

$$\inf_{v \in \mathcal{H}_n} \|u - v\| < \varepsilon \tag{7.45}$$

для всех $n > \widetilde{n}$. Иначе говоря, полнота последовательности $\mathcal{H}_n$ означает, что любой элемент $u \in \mathcal{H}$ может быть с любой степенью точности приближен элементами из последовательности пространств $\mathcal{H}_n$.

Пусть $\varphi_i^{(n)},\ i = 1, \ldots, N_n$ — базис пространства $\mathcal{H}_n$ и все элементы базиса принадлежат области определения оператора $\mathcal{L}$. Тогда приближение по Галеркину ищется в виде

$$u^{(n)} = \sum_{i=1}^{N_n} \alpha_i \varphi_i^{(n)}. \tag{7.46}$$

Коэффициенты α_i выбираются так, чтобы невязка $\mathcal{L}u^{(n)} - f$ была ортогональна всем базисным элементам пространства $\mathcal{H}_n$, то есть

$$(\mathcal{L}u^{(n)} - f, \varphi_k^{(n)}) = 0, \quad k = 1, \ldots, N_n. \tag{7.47}$$

В том случае, когда $\mathcal{L}$ — линейный оператор, уравнения (7.47) принимают вид

$$\begin{aligned}(\mathcal{L}u^{(n)} - f, \varphi_k^{(n)}) &= \left(\mathcal{L}\Big(\sum_{i=1}^{N_n} \alpha_i \varphi_i^{(n)}\Big) - f, \varphi_k^{(n)}\right) = \\ &= \sum_{i=1}^{N_n} \alpha_i (\mathcal{L}\varphi_i^{(n)}, \varphi_k^{(n)}) - (f, \varphi_k^{(n)}) = 0, \quad k = 1, \ldots, N_n. \end{aligned} \tag{7.48}$$

Таким образом, получена система линейных алгебраических уравнений для коэффициентов α_i, после нахождения которых, по формуле (7.46) определяется приближенное решение $u^{(n)}$.

[2]Метод иногда называют методом Бубнова-Галеркина.

Если оператор $\mathcal{L}$ нелинейный, то система (7.47) тоже будет нелинейной и ее решение при больших N_n затруднено.

Вопрос о сходимости $u^{(n)}$ при $n \to \infty$ к точному решению и о скорости сходимости здесь не рассматривается.

Рассмотрим пример. Пусть речь идет о решении некоторой краевой задачи вида

$$\mathcal{L}u(x) = f(x), \quad a \le x \le b, \quad u(a) = A, \; u(b) = B,$$

где $\mathcal{L}$ — линейный дифференциальный оператор. Если ввести какую-нибудь гладкую функцию $\varphi_0(x)$ такую, что $\varphi_0(a) = A$ и $\varphi_0(b) = B$, то функция $v = u - \varphi_0$ удовлетворяет однородным граничным условиям и является решением уравнения $\mathcal{L}v = f - \mathcal{L}\varphi_0$. Будем искать v методом Галеркина. Для этого в качестве пространства $\mathcal{H}$ выберем $L_2(a, b)$. Областью определения оператора $\mathcal{L}$ назовем гладкие функции, обращающиеся в ноль на концах отрезка. Выберем полную систему линейно-независимых функций $\varphi_i, \; i = 1, 2, \ldots$, принадлежащих области определения оператора $\mathcal{L}$. По теореме Вейерштрасса любую непрерывную функцию можно приблизить со сколь угодно высокой точностью алгебраическими или тригонометрическими многочленами. Поэтому можно положить

$$\varphi_i(x) = (x - a)^i(b - x) \quad \text{или} \quad \varphi_i = \sin \frac{\pi i(x - a)}{b - a}, \quad i = 1, 2, \ldots$$

Разумеется приведенные примеры не исчерпывают всего множества возможных вариантов базисных функций. Пространством $\mathcal{H}_n$ назовем подпространство пространства $L_2(a, b)$, в котором функции $\varphi_i, \; i = 1, 2, \ldots, n$ образуют базис.

Конкретизируем теперь наш пример. Пусть требуется решить краевую задачу

$$u'' + u + x = 0, \quad u(0) = u(1) = 0.$$

Таким образом, $\mathcal{L}u = u'' + u, \; f = -x$. Очевидно, что можно взять функцию $\varphi_0 = 0$. Возьмем $n = 2$, а $\varphi_1(x) = x(1 - x), \; \varphi_2(x) = x^2(1 - x)$. Тогда приближение решения

$$u^{(2)}(x) = \alpha_1\varphi_1(x) + \alpha_2\varphi_2(x) = \alpha_1 x(1 - x) + \alpha_2 x^2(1 - x) = x(1 - x)(\alpha_1 + \alpha_2 x)$$

и

$$\mathcal{L}u^{(2)} - f = -2\alpha_1 + \alpha_2(2 - 6x) + x(1 - x)(\alpha_1 + \alpha_2 x) + x.$$

Система (7.47) принимает вид

$$\int_0^1 (\mathcal{L}u^{(2)} - f)x(1 - x)\, dx = \frac{3}{10}\alpha_1 + \frac{3}{20}\alpha_2 - \frac{1}{12} = 0,$$

$$\int_0^1 (\mathcal{L}u^{(2)} - f)x^2(1 - x)\, dx = \frac{3}{20}\alpha_1 + \frac{13}{105}\alpha_2 - \frac{1}{20} = 0.$$

Решая эту систему, получим

$$\alpha_1 = \frac{71}{369}, \quad \alpha_2 = \frac{7}{41}.$$

Таким образом,

$$u^{(2)}(x) = x(1 - x)\Big(\frac{71}{369} + \frac{7}{41}x\Big).$$

Легко найти точное решение $u(x)$ краевой задачи

$$u(x) = \frac{\sin x}{\sin 1} - x.$$

Для сравнения приведем значения точного и приближенного решения в некоторых точках

x	0.20000	0.40000	0.60000	0.80000
$u(x)$	0.03610	0.06278	0.07102	0.05220
$u^{(2)}(x)$	0.03625	0.06257	0.07076	0.05264

Обычно при небольшом числе n метод Галеркина дает неважные результаты. Разумеется величина погрешности очень чувствительна к тому, насколько удачно выбрана система функций φ_i для данной задачи. Заметим также, что, вообще говоря, матрица системы (7.48) не имеет никаких особенностей, которые позволяли бы использовать упрощенные методы решения системы уравнений. В следующем пункте будет рассмотрен метод выбора базисных функций, позволяющий упростить вид системы (7.48).

Заметим, что метод прогонки определяет значения решения только в узлах сетки. Метод же Галеркина позволяет найти значение решения в произвольной точке отрезка $[a, b]$.

7.4 МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

В этом параграфе будут рассмотрены идеи метода на примере уравнения диффузии для одномерной области, которое имеет вид

$$-\frac{d}{dx}\Big(p(x)\frac{du}{dx}\Big) + q(x)u = f(x), \tag{7.49}$$

где $p(x) \geq p_0 > 0,\ q(x) \geq 0$. Будем считать, что p, q и f — кусочно непрерывные функции на отрезке $[0, 1]$ с возможными точками разрыва первого рода. Сформулируем краевые условия

$$u(0) = 0, \quad u(1) = 0 \tag{7.50}$$

и поставим задачу о нахождении решения уравнения (7.49), удовлетворяющего условию (7.50). При этом под решением понимается непрерывная, кусочно-дифференцируемая функция u, для которой функция pu' дифференцируема и выполняются равенства (7.49),(7.50).

В соответствии с методом Галеркина для любой базисной функции φ должно выполняться равенство

$$\int_0^1 \left(-\frac{d}{dx}\Big(p\frac{du}{dx}\Big) + qu - f\right)\varphi\, dx = 0. \tag{7.51}$$

Так как функция φ выбирается из области определения оператора, она должна быть непрерывной, кусочно-дифференцируемой и равной нулю на концах отрезка. Тогда, используя формулу интегрирования по частям и учитывая, что $\varphi(0) = \varphi(1) = 0$, получим

$$\int_0^1 \Big(p\frac{du}{dx}\frac{d\varphi}{dx} + (qu - f)\varphi\Big)\, dx = 0. \tag{7.52}$$

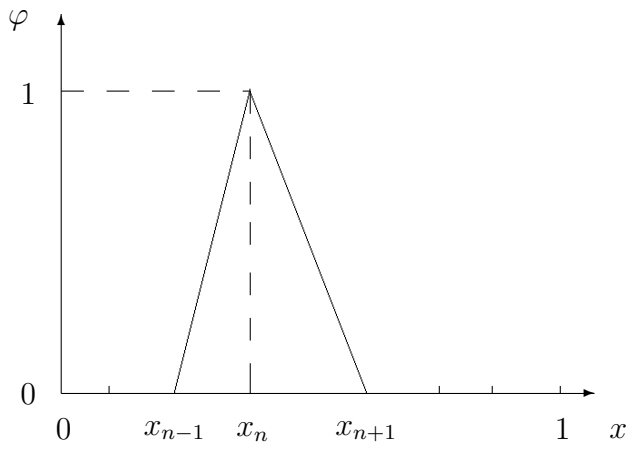


Рис. 7.1 График базисной функции

Перейдем теперь к выбору базисных функций. Разобьем отрезок $[0,1]$ на части точками x_n так, что $0 = x_0 < x_1 < \cdots < x_N = 1$. Пусть $h_{n-0.5} = x_n - x_{n-1},\ n = 1,\ldots,N$. Введем для каждого значения n функцию

$$\varphi_n(x) = \begin{cases} 0, & 0 \le x \le x_{n-1}, \\ \dfrac{x - x_{n-1}}{h_{n-0.5}}, & x_{n-1} \le x \le x_n, \\ \dfrac{x_{n+1} - x}{h_{n+0.5}}, & x_n \le x \le x_{n+1}, \\ 0, & x_{n+1} \le x \le 1. \end{cases} \tag{7.53}$$

График этой функции изображен на рисунке 7.1. Каждая из функций φ_n непрерывна, отлична от нуля только в промежутке (x_{n-1}, x_{n+1}) и принимает максимальное значение равное 1 в точке x_n. Эти функции обладают свойством полноты в том смысле, что любую непрерывную кусочно-линейную функцию с возможными изломами в точках $x_i,\ i = 1,\ldots,N-1$ и равную нуля на концах отрезка можно представить в виде линейной комбинации таких функций. В свою очередь любая непрерывная функция может быть приближена кусочно-линейными непрерывными функциями.

Пусть $y(x)$ — непрерывная кусочно-линейная функция с возможными изломами в точках $x_i,\ i = 1,\ldots,n-1,\ y(0) = y(1) = 0$. Обозначим коэффициенты разложения ее по функциям φ_i через y_i, то есть

$$y(x) = \sum_{i=1}^{N-1} y_i \varphi_i(x). \tag{7.54}$$

Выбирая в этом равенстве $x = x_k$, с учетом того, что

$$\varphi_i(x_k) = \begin{cases} 0, & \text{при } i \ne k, \\ 1, & \text{при } i = k, \end{cases}$$

получаем $y(x_k) = y_k$. Таким образом, коэффициентами разложения являются значения функции $y(x)$ в точках x_k.

Анализируя графики функций φ_k, легко заметить, что они "почти" ортогональны в $L_2(0,1)$, то есть

$$(\varphi_k, \varphi_n) = \int_0^1 \varphi_k(x)\varphi_n(x)\,dx = 0 \ \text{ при } n \ne k-1, k, k+1.$$

В первую очередь этим определяется специфика выбранного базиса.

Будем искать теперь приближенное решение в виде (7.54). Подставим его в (7.52) вместо $u(x)$ и $\varphi_k(x)$ вместо $\varphi(x)$. Заметим при этом, что так как функция $\varphi_k(x)$ равна нулю вне отрезка $[x_{k-1}, x_{k+1}]$ интегрировать достаточно не по всему отрезку $[0,1]$, а только по отрезку $[x_{k-1}, x_{k+1}]$.

$$0 = \int\limits_{x_{k-1}}^{x_{k+1}} \Big(p\frac{dy}{dx}\frac{d\varphi_k}{dx} + (qy - f)\varphi_k \Big)\, dx =$$

$$= \sum_{i=1}^{N-1} y_i \int\limits_{x_{k-1}}^{x_{k+1}} p\frac{d\varphi_i}{dx}\frac{d\varphi_k}{dx}\, dx + \sum_{i=1}^{N-1} y_i \int\limits_{x_{k-1}}^{x_{k+1}} \varphi_i\varphi_k\, dx - \int\limits_{x_{k-1}}^{x_{k+1}} f\varphi_k\, dx =$$

$$= \sum_{i=k-1}^{k+1} y_i \int\limits_{x_{k-1}}^{x_{k+1}} p\frac{d\varphi_i}{dx}\frac{d\varphi_k}{dx}\, dx + \sum_{i=k-1}^{k+1} y_i \int\limits_{x_{k-1}}^{x_{k+1}} \varphi_i\varphi_k\, dx - \int\limits_{x_{k-1}}^{x_{k+1}} f\varphi_k\, dx. \quad (7.55)$$

Последнее равенство здесь написано на основании свойства "почти ортогональности" базисных функций.

Преобразуем слагаемые, стоящие в левой части равенства (7.55). Интегралы, входящие в первое и второе слагаемые разобьем на два: интегралы по промежуткам $[x_{k-1}, x_k]$ и $[x_k, x_{k+1}]$. В результате с учетом (7.53) получим

$$\sum_{i=k-1}^{k+1} y_i \int\limits_{x_{k-1}}^{x_k} p\frac{d\varphi_i}{dx}\frac{d\varphi_k}{dx}\, dx = y_{k-1} \int\limits_{x_{k-1}}^{x_k} p\frac{-1}{h_{k-0,5}}\frac{1}{h_{k-0,5}}\, dx +$$

$$+ y_k \int\limits_{x_{k-1}}^{x_k} p\frac{1}{h_{k-0,5}}\frac{1}{h_{k-0,5}}\, dx = \frac{y_k - y_{k-1}}{h_{k-0.5}} p_{k-0.5}, \quad (7.56)$$

где

$$p_{k-0.5} = \frac{1}{h_{k-0.5}} \int\limits_{x_{k-1}}^{x_k} p\, dx.$$

Аналогично

$$\sum_{i=k-1}^{k+1} y_i \int\limits_{x_k}^{x_{k+1}} p\frac{d\varphi_i}{dx}\frac{d\varphi_k}{dx}\, dx = -\frac{y_{k+1} - y_k}{h_{k+0.5}} p_{k+0.5}. \quad (7.57)$$

Имеем далее

$$\sum_{i=k-1}^{k+1} y_i \int\limits_{x_{k-1}}^{x_k} q\varphi_i\varphi_k\, dx = y_{k-1}\frac{1}{h_{k-0.5}^2} \int\limits_{x_{k-1}}^{x_k} q(x_k - x)(x - x_{k-1})\, dx +$$

$$+ y_k \frac{1}{h_{k-0.5}^2} \int\limits_{x_{k-1}}^{x_k} q(x - x_{k-1})(x - x_{k-1})\, dx = y_{k-1} q_{k-0.5}^{(1)} + y_k q_{k-0.5}^{(2)}, \quad (7.58)$$

где

$$q_{k-0.5}^{(1)} = \frac{1}{h_{k-0.5}^2} \int\limits_{x_{k-1}}^{x_k} q(x_k - x)(x - x_{k-1})\, dx, \qquad q_{k-0.5}^{(2)} = \frac{1}{h_{k-0.5}^2} \int\limits_{x_{k-1}}^{x_k} q(x - x_{k-1})^2\, dx,$$

$$\sum_{i=k-1}^{k+1} y_i \int_{x_k}^{x_{k+1}} q\varphi_i\varphi_k\,dx = y_k \frac{1}{h_{k+0.5}^2} \int_{x_k}^{x_{k+1}} q(x_{k+1}-x)^2\,dx +$$

$$+ y_{k+1}\frac{1}{h_{k+0.5}^2} \int_{x_k}^{x_{k+1}} q(x-x_k)(x_{k+1}-x)\,dx = y_k q_{k+0.5}^{(3)} + y_{k+1} q_{k+0.5}^{(1)}, \quad (7.59)$$

$$q_{k+0.5}^{(3)} = \frac{1}{h_{k+0.5}^2} \int_{x_k}^{x_{k+1}} q(x_{k+1}-x)^2\,dx.$$

Подставляя в (7.55) равенства (7.56)-(7.59) и вводя обозначение

$$F_k = \int_{x_{k-1}}^{x_{k+1}} f\varphi_k\,dx,$$

получим

$$p_{k-0.5}\frac{y_k - y_{k-1}}{h_{k-0.5}} - p_{k+0.5}\frac{y_{k+1}-y_k}{h_{k+0.5}} +$$
$$+ q_{k-0.5}^{(1)} y_{k-1} + (q_{k-0.5}^{(2)} + q_{k+0.5}^{(3)}) y_k + q_{k+0.5}^{(1)} y_{k+1} = F_k, \quad k = 1, \ldots, N-1. \quad (7.60)$$

Добавляя к системе уравнений (7.60) граничные условия $y_0 = y_N = 0$, получаем систему линейных алгебраических уравнение относительно y_i, которая решается методом прогонки.

7.5 ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ

Рассмотрим один из методов решения на примере интегрального уравнения Фредгольма второго рода [20]. Пусть требуется найти решение $u(x)$ уравнения

$$u(x) - \lambda \int_a^b K(x,t)u(t)\,dt = f(x),\ a \le x \le b. \quad (7.61)$$

Будем считать функции $K(x,t),\ f(x)$ гладкими. Выберем какую-нибудь квадратурную формулу и заменим с ее помощью интеграл в левой части равенства (7.61). В результате получим

$$u(x) - \lambda \sum_{j=1}^{n} c_j K(x,t_j)u(t_j) - \lambda r_n = f(x). \quad (7.62)$$

Здесь c_j — коэффициенты, t_j — узлы, а r_n — остаточный член квадратурной формулы.

Полагая в (7.61) $x = t_i,\ i = 1, \ldots, n$, получим

$$u(t_i) - \lambda \sum_{j=1}^{n} c_j K(t_i,t_j)u(t_j) - \lambda r_n = f(t_i). \quad (7.63)$$

Отбрасывая остаточный член квадратурнрой формулы и обозначая приближенное значение величины $u(t_i)$ через y_i, получим систему линейных алгебраических уравнений

$$y_i - \lambda \sum_{j=1}^{n} c_j K(t_i, t_j) y_j = f(t_i),\ i = 1, \ldots, n. \tag{7.64}$$

Для решения этой системы могут быть применены стандартные методы решения систем линейных алгебраических уравнений.

Матрица $\mathbf{A}$ системы (7.64) имеет вид

$$\mathbf{A} = \begin{pmatrix} 1 - \lambda c_1 K(t_1, t_1) & -\lambda c_2 K(t_1, t_2) & -\lambda c_3 K(t_1, t_3) & \cdots & -\lambda c_n K(t_1, t_n) \\ -\lambda c_1 K(t_2, t_1) & 1 - \lambda c_2 K(t_2, t_2) & -\lambda c_3 K(t_2, t_3) & \cdots & -\lambda c_n K(t_2, t_n) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -\lambda c_1 K(t_n, t_1) & -\lambda c_2 K(t_n, t_2) & -\lambda c_3 K(t_n, t_3) & \cdots & 1 - \lambda c_n K(t_n, t_n) \end{pmatrix}.$$

Легко заметить теперь, что даже в том случае, когда ядро $K(x, t)$ является симметричным, то есть $K(x, t) = K(t, x)$, матрица $\mathbf{A}$ может оказаться не симметричной из-за коэффициентов квадратурной формулы. Системы с симметричной матрицей предпочтительнее, так как можно использовать специальные методы нахождения решения. Поэтому рассмотрим вопрос о том, как в случае симметричного ядра преобразовать систему (7.64)для получения симметричной матрицы.

Можно предложить два подхода. В первом достаточно умножить i-ое уравнение на c_i. В результате получим систему

$$c_i y_i - \lambda \sum_{j=1}^{n} c_i c_j K(t_i, t_j) y_j = c_i f(t_i),\ i = 1, \ldots, n, \tag{7.65}$$

матрица которой симметрична.

Во втором подходе, который применим при положительных коэффициентах c_i, i-ое уравнение умножается на $\sqrt{c_i}$ и вводятся новые величины $v_i = \sqrt{c_i} y_i$. Тогда для v_i получаются уравнения

$$v_i - \lambda \sum_{j=1}^{n} \sqrt{c_i}\sqrt{c_j} K(t_i, t_j) v_j = \sqrt{c_i} f(t_i),\ i = 1, \ldots, n. \tag{7.66}$$

Описанный выше метод нахождения таблицы приближенных значений решения интегрального уравнения иногда называют **квадратурным методом**.

Замечание. В том случае, когда требуется найти приближенное значение решения в точке x, не являющейся узлом квадратурной формулы, можно применить интерполяцию. Однако, есть и другой подход. Достаточно воспользоваться соотношением

$$u(x) \approx \lambda \sum_{j=1}^{n} c_j K(x, t_j) y_j + f(x),$$

которое получается из (7.62) путем отбрасывания остаточного члена квадратурной формулы и замены $u(t_j)$ на y_j.

В случае, когда необходимо решить уравнение Вольтерра второго рода

$$u(x) - \lambda \int_a^x K(x, t) u(t)\, dt = f(x),\ a \le x \le b, \tag{7.67}$$

целесообразно поступить следующим способом. Задаться набором точек $a \le x_1 < x_2 < \cdots < x_n \le b$. Положить в (7.67) $x = x_i$

$$u(x_i) - \lambda \int_a^{x_i} K(x_i, t)u(t)\,dt = f(x_i)$$

и для вычисления интеграла использовать квадратурную формулу с узлами в точках $x_1, \ldots, x_i$. Легко заметить, что в этом случае получится система уравнений с треугольной матрицей. Поэтому отыскание решения такой системы не вызывает труда.

Остановимся вкратце на решении уравнения Вольтерра первого рода:

$$\int_a^x K(x, t)u(t)\,dt = f(x),\ a \le x \le b. \tag{7.68}$$

Выбирая точки $a = x_0 < x_1 < x_2 < \cdots < x_n \le b$, имеем равенства

$$\int_a^{x_i} K(x_i, t)u(t)\,dt = f(x_i),\ i = 1, \ldots, n. \tag{7.69}$$

В отличие от уравнения Вольтерра второго рода, нет необходимости считать узлы t_i квадратурной формулы совпадающими с узлами x_i. Воспользовавшись, например, формулой средних прямоугольников, и полагая $t_i = 0.5(x_{i-1} + x_i)$, получим из (7.69):

$$\begin{aligned}
&(x_1 - x_0)K(x_1, t_1)y_1 = f(x_1),\\
&(x_1 - x_0)K(x_2, t_1)y_1 + (x_2 - x_1)K(x_2, t_2)y_2 = f(x_2),\\
&\cdots,
\end{aligned}$$

откуда найдем $y_i \approx u(t_i)$.

Если узлы квадратурной формулы t_i выбрать равными x_i и такими, что в них входит точка t_0, то нахождение $u(a) = u(x_0) = u(t_0) \approx y_0$ непосредственно из (7.69) невозможно, так как при подстановке в него $x = x_0$ пропадает интеграл. В то же время без значения y_0 последующие значения $y_1, y_2, \ldots$ найти не удастся. Для нахождения решения в этом случае продифференцируем уравнение (7.68). В результате получим

$$K(x, x)u(x) + \int_a^x \frac{\partial K(x, t)}{\partial x} u(t)\,dt = f'(x).$$

Полагая $x = x_0 = a$, получим $K(x_0, x_0)u(x_0) = f'(x_0)$. Тогда можно положить

$$y_0 = u(x_0) = \frac{f'(x_0)}{K(x_0, x_0)}.$$

Применяя далее, например, формулу трапеций, и полагая $x = x_1$, имеем

$$\frac{x_1 - x_0}{2}(K(x_1, x_0)y_0 + K(x_1, x_1)y_1) = f(x_1).$$

Отсюда находим y_1 и так далее.

7.6 ЗАДАЧИ К ГЛАВЕ 7

7.6.1 Примеры решения задач

1. При каких коэффициентах $a,\ b,\ c$ разностная схема

$$-\frac{y_{k-1}-2y_k+y_{k+1}}{h^2}+(ay_{k-1}+by_k+cy_{k+1})=f(x_k)+\frac{h^2}{12}f''(x_k),\quad k=1\ldots,K-1,$$
$$x_k=kh,\quad Kh=1,\quad y_0=y_K=0$$

аппроксимирует задачу

$$-u''(x)+u(x)=f(x),\quad 0\le x\le 1,\quad u(0)=u(1)=0$$

с четвертым порядком, если известно, что решение дифференциальной задачи имеет непрерывные шестые производные?

Решение. В соответствии с определением, порядок величины

$$\psi_k=-\frac{u_{k-1}-2u_k+u_{k+1}}{h^2}+(au_{k-1}+bu_k+cu_{k+1})-f(x_k)-\frac{h^2}{12}f''(x_k),$$

где $u_k=u(x_k)$ и есть порядок аппроксимации, так как граничные условия аппроксимируются точно.

Для того, чтобы определить порядок величины ψ_k воспользуемся формулой Тейлора

$$u(x_k\pm h)=u(x_k)\pm\frac{u'(x_k)}{1!}h+\frac{u''(x_k)}{2!}h^2\pm\frac{u'''(x_k)}{3!}h^3+\frac{u^{IV}(x_k)}{4!}h^4\pm\frac{u^V(x_k)}{5!}h^5+O(h^6).\tag{7.70}$$

Подставим эту формулу в выражение для ψ_k. Учитывая, что в выражения для u_{k-1} и u_{k+1} слагаемые с нечетными степенями h входят с разными знаками, а с четными — с одинаковыми, получим

$$\begin{aligned}\psi_k&=-u''(x_k)-2\frac{u^{IV}(x_k)}{4!}h^2+O(h^4)+\\&+\Bigg[a\Big(u(x_k)-\frac{u'(x_k)}{1!}h+\frac{u''(x_k)}{2!}h^2-\frac{u'''(x_k)}{3!}h^3+O(h^4)\Big)+bu(x_k)+\\&+c\Big(u(x_k)+\frac{u'(x_k)}{1!}h+\frac{u''(x_k)}{2!}h^2+\frac{u'''(x_k)}{3!}h^3+O(h^4)\Big)\Bigg]-f(x_k)-\frac{h^2}{12}f''(x_k)=\\&=\big[-u''(x_k)+u(x_k)-f(x_k)\big]+\frac{h^2}{12}\big[-u^{IV}(x_k)+u''(x_k)-f''(x_k)\big]+\\&+(a+b+c-1)u(x_k)+(c-a)\Big(u'(x_k)h+\frac{u'''(x_k)}{3!}h^3\Big)+\Big(a+c-\frac{1}{6}\Big)\frac{u''(x_k)}{2!}h^2+O(h^4).\end{aligned}$$

Выражения, стоящие в квадратных скобках в правой части равенства равны нулю. Действительно, равенство нулю первой скобки означает, что u — решение задачи. Если же продифференцировать дважды дифференциальное уравнение, получим равенство нулю второй скобки. Для того, чтобы $\psi_k=O(h^4)$ достаточно теперь потребовать, чтобы

$$a+b+c=1,\quad c-a=0,\quad a+c=\frac{1}{6}.$$

Отсюда $a = c = 1/12,\ b = 5/6$.

2. Для задачи

$$u'' - 3u = \sin x, \quad x \in [0, \pi], \quad u(0) = 0, \quad u(\pi) - u'(\pi) = 1$$

предложить разностную схему второго порядка аппроксимации.

Решение. Пусти $x_k = kh,\ k = 0, \ldots, K,\ Kh = \pi$. Обозначим через y_k значение приближенного решения в точке x_k. Тогда для точек $k = 1, \ldots, K-1$ можно записать

$$\frac{y_{k-1} - 2y_k + y_{k+1}}{h^2} - 3y_k = \sin x_k.$$

Заменяя здесь y_k на $u_k = u(x_k)$ и воспользовавшись формулой Тейлора (7.70) и тем фактом, что u — решение дифференциального уравнения, имеем

$$\psi_k = \frac{1}{h^2}\Big(u(x_k) - \frac{u'(x_k)}{1!}h + \frac{u''(x_k)}{2!}h^2 - \frac{u'''(x_k)}{3!}h^3 + O(h^4) -$$
$$- 2u(x_k) + u(x_k) + \frac{u'(x_k)}{1!}h + \frac{u''(x_k)}{2!}h^2 + \frac{u'''(x_k)}{3!}h^3 + O(h^4)\Big) - 3u(x_k) - \sin x_k =$$
$$= u''(x_k) - 3u(x_k) - \sin x_k + O(h^2) = O(h^2).$$

Первое граничное условие задаем точно: $y_0 = 0$. Для получения приближения для второго условия запишем:

$$\frac{u(x_K) - u(x_{K-1})}{h} = \frac{1}{h}\left(u(x_K) - \left(u(x_K) - \frac{u'(x_K)}{1!}h + \frac{u''(x_K)}{2!}h^2 + O(h^3)\right)\right) =$$
$$= u'(x_K) - \frac{u''(x_K)}{2}h + O(h^2).$$

Отсюда следует, что второе граничное условие переписывается в виде

$$u(x_K) - \frac{u(x_K) - u(x_{K-1})}{h} - \frac{u''(x_K)}{2}h + O(h^2) = 1. \tag{7.71}$$

Из дифференциального уравнения имеем $u''(x_K) = 3u(x_K) + \sin x_K = 3u(x_K)$, так как $x_K = \pi$. Подставляя в (7.71) выражение для $u''(x_K)$, получим

$$u(x_K) - \frac{u(x_K) - u(x_{K-1})}{h} - \frac{3u(x_K)}{2}h + O(h^2) = 1.$$

Отбрасывая члены порядка h^2, получим аппроксимацию второго граничного условия:

$$y_K - \frac{y_K - y_{K-1}}{h} - \frac{3y_K}{2}h = 1.$$

После приведения подобных это соотношение можно переписать в виде

$$y_K(2h - 2 - 3h^2) + 2y_{K-1} = 2h.$$

3. Пусть α_j, где $j = 0, 1 \ldots, m$ и $f_n,\ n = 0, 1, \ldots$ — заданные числа. Тогда уравнение

$$\alpha_m y_{n+m} + \alpha_{m-1} y_{n+m-1} + \cdots + \alpha_0 y_n = f_n \tag{7.72}$$

относительно неизвестных y_n называется **неоднородным разностным уравнением**. Индекс n может пробегать и другое множество значений, например, $0, \pm 1, \pm 2, \ldots$ или $0, 1, \ldots, N$.

Если для всех значений n выполняются равенства $f_n = 0$, то уравнение называется **однородным разностным уравнением**. Каждое отдельное решение уравнения (7.72) называется **частным решением**, а m параметрическое семейство решений, содержащее любое частное решение, называется **общим решением.**

Доказать, что общее решение неоднородного уравнения равно сумме общего решения однородного уравнения и частного решения неоднородного уравнения.

Решение. Пусть

$$\mathcal{L}y_n = \alpha_m y_{n+m} + \alpha_{m-1} y_{n+m-1} + \cdots + \alpha_0 y_n.$$

Легко проверить, что для произвольных чисел $\mu,\ \nu$ и любых $y_n,\ z_n$ выполняется равенство

$$\mathcal{L}(\mu y_n + \nu z_n) = \mu \mathcal{L} y_n + \nu \mathcal{L} z_n. \tag{7.73}$$

Тогда, если $y_n^o = y_n(c_1, \ldots, c_m)$ — общее решение однородного уравнения, а z_n — частное решение неоднородного, то для их суммы имеем

$$\mathcal{L}(y_n^o + z_n) = \mathcal{L}y_n^o + \mathcal{L}z_n = 0 + f_n = f_n.$$

Следовательно доказано, что сумма общего решения однородного уравнения и частного решения неоднородного уравнения является решением неоднородного уравнения.

Покажем теперь, что если Y_n — решение неоднородного уравнения, то можно подобрать такие параметры $c_1^*, \ldots, c_m^*$, что $Y_n = y_n(c_1^*, \ldots, c_m^*) + z_n$. Это будет означать, что любое частное решение неоднородного уравнения получается из решения $y_n^o + z_n$ путем соответствующего подбора параметров $c_1, \ldots, c_m$, то есть $y_n^o + z_n$ — общее решение неоднородного уравнения. Заметим, что

$$\mathcal{L}(Y_n - z_n) = \mathcal{L}Y_n - \mathcal{L}z_n = f_n - f_n = 0,$$

то есть $Y_n - z_n$ — частное решение однородного уравнения. Так как $y_n^o = y_n(c_1, \ldots, c_m)$ — общее решение однородного уравнения, существуют такие значения $c_1^*, \ldots, c_m^*$ параметров общего решения, что $y_n(c_1^*, \ldots, c_m^*) = Y_n - z_n$. Значит, $Y_n = y_n(c_1^*, \ldots, c_m^*) + z_n$, что и требовалось доказать.

4. Получить формулу для общего члена последовательности чисел Фибоначчи.

Решение. Числа Фибоначчи f_n определяются соотношениями

$$f_0 = 0, \quad f_1 = 1, \quad f_{n+2} = f_{n+1} + f_n,\ n = 0, 1, \ldots$$

Это означает, что числа Фибоначчи являются решением однородного разностного уравнения

$$f_{n+2} - f_{n+1} - f_n = 0,\ n = 0, 1, \ldots$$

Для того, чтобы найти общее решение этого уравнения надо воспользоваться утверждением теоремы 6.6.2. В соответствии с этим утверждением, составляется многочлен $\chi^2 - \chi - 1$ и находятся его корни $\chi_{1\,2} = \dfrac{1 \pm \sqrt{5}}{2}$. Тогда общее решение

$$f_n = c_1 \chi_1^n + c_2 \chi_2^n = c_1 \Big(\frac{1+\sqrt{5}}{2}\Big)^n + c_2 \Big(\frac{1-\sqrt{5}}{2}\Big)^n,$$

где c_1, c_2 — произвольные константы. Их следует подобрать так, чтобы $f_0 = 0,\ f_1 = 1$. Имеем

$$0 = f_0 = c_1 + c_2, \quad 1 = f_1 = c_1 \frac{1+\sqrt{5}}{2} + c_2 \frac{1-\sqrt{5}}{2}.$$

Отсюда следует, что $c_1 = -c_2 = 1/\sqrt{5}$. Окончательно имеем

$$f_n = \frac{1}{\sqrt{5}}\left[\Big(\frac{1+\sqrt{5}}{2}\Big)^n - \Big(\frac{1-\sqrt{5}}{2}\Big)^n\right].$$

7.6.2 Задачи

1. Для задачи

$$u'' + 2u = e^{x+1}, \quad -1 \le x \le 1, \quad u'(-1) = u(-1), \quad u'(1) = 0$$

предложить разностную схему второго порядка аппроксимации.

2. Для задачи

$$u'' + u = e^x, \quad 0 \le x \le 1, \quad u(0) = u(1) = 0$$

построить разностную схему 6-го порядка аппроксимации.

3. Методом конечных элементов построить разностную схему для дифференциальной задачи

$$-\big(a(x)u'(x)\big)' = 1, \quad 0 \le x \le 1, \quad u(0) = u(1) = 0, \quad a(x) = \begin{cases} 1, & 0 \le x \le \frac{\sqrt{2}}{2}, \\ 2, & \frac{\sqrt{2}}{2} < x \le 1 \end{cases}.$$

В качестве базисных выбрать функции (7.53).

4. Найти решение задачи

$$\frac{y_{k-1} - 2y_k + y_{k+1}}{h^2} = 1, \quad k = 1, \ldots, K-1, \quad y_0 = 0, \quad y_K = 1.$$

Указание. Воспользоваться результатом задачи 3 пункта 7.6.1 и утверждением теоремы 6.6.2. Частное решение неоднородного разностного уравнения искать в виде $y_k^{(\text{ч})} = ak^2$, где a — константа.

5. Решите задачу $u'' - u + x = 0, \quad u(0) = u(1) = 0$ методом Галеркина, ограничившись в разложении решения по базису двумя слагаемыми. Сравнить полученное приближенное решение с точным решением

$$u(x) = \frac{e^x - e^{-x}}{e^{-1} - e} + x.$$

7.6.3 Примеры тестовых вопросов к главе 7

1. Выберите метод решения задачи

$$y'' + \sin(xy - 1) = 0, \ 0 \le x \le \pi \quad y(0) = 0, \quad y(\pi) = 0.$$

а) Адамса;

б) Рунге-Кутта;

в) прогонки;

г) стрельбы.

2. Какие задачи рациональнее решать методом прогонки, а какие — стрельбы? Ответ запишите в виде: С<перечень номеров задач в порядке возрастания, соответствующих методу стрельбы> пробел П<перечень номеров задач в порядке возрастания, соответствующих методу прогонки>. Например С135 П24

- 1) $u'' + \sin u = x, \quad x \in [0,1], \quad u(0) = u(1) = 0;$
- 2) $u'' + \cos u = u, \quad x \in [0,1], \quad u(0) = u(1) = 0;$
- 3) $u'' - u = \cos x, \quad x \in [0,1], \quad u(0) = u(1) = 0;$
- 4) $u'' + u' - u = e^x, \quad x \in [0,1], \quad u(0) = u(1) = 0;$
- 5) $u'' + e^x u' - u = 1, \quad x \in [0,1], \quad u(0) = u(1) = 0;$
- 6) $u'' - 4u = 2x^3 - 9x^2, \quad x \in [0,1], \quad u(0) = u(1) = 0.$

3. Какая из перечисленных ниже схем соответствует методу прогонки для решения задачи

$$\frac{d^2u}{dx^2} - \frac{du}{dx} - u = 0, \quad u(0) = \frac{du(0)}{dx}, \quad u(1) = 0$$

на сетке $x_k = kh,\ k = 0, \ldots, K,\ Kh = 1$?

а) $\dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - \dfrac{y_k - y_{k-1}}{h} = y_k,\ k = 1, \ldots, K-1, \quad hy_0 = y_1 - y_0,\ y_1 = 0;$

б) $\dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - \dfrac{y_k - y_{k-1}}{h} = y_k,\ k = 1, \ldots, K-1, \quad y_1 = y_0,\ y_K = 0;$

в) $\dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - \dfrac{y_k - y_{k-1}}{h} = y_k,\ k = 1, \ldots, K-1, \quad (1+h)y_0 = y_1,\ y_K = 0;$

г) $\dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - \dfrac{y_k - y_{k-1}}{h} = y_k,\ k = 1, \ldots, K-1, \quad (1-h)y_0 = y_1,\ y_1 = 0;$

д) $\dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - \dfrac{y_k - y_{k-1}}{h} = y_k,\ k = 1, \ldots, K-1, \quad hy_0 = y_1 - y_0,\ y_K = 1;$

е) $\dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - \dfrac{y_{k+1} - y_{k-1}}{h} = y_k,\ k = 1, \ldots, K-1, \quad hy_0 = y_1 - y_0,\ y_K = 0.$

4. Пусть задача $u'' + u - x^2 = 0,\ u(0) = 0,\ u(1) = 0$ решается методом Галеркина. Какие из приведенных ниже функций могут быть выбраны в качестве базисных?

а) $\varphi_n(x) = x^n,\ n = 0, 1, \ldots;$

б) $\varphi_n(x) = \sin nx,\ n = 0, 1, \ldots;$

в) $\varphi_{2n}(x) = \cos nx,\ \varphi_{2n+1}(x) = \sin nx\ n = 0, 1, \ldots;$

г) $\varphi_{2n}(x) = \cos 2nx,\ \varphi_{2n+1}(x) = \sin(2n+1)x\ n = 0, 1, \ldots;$

д) $\varphi_n(x) = \sin \pi nx,\ n = 1, \ldots;$

е) $\varphi_n(x) = (1-x)^n,\ n = 0, 1, \ldots.$

5. Какие из приведенных ниже функций могут рассматриваться как конечные элементы на отрезке $[0,1]$?

а) $\varphi_n(x) = x^n, \quad n = 0, 1, \ldots;$

б) $\varphi_n(x) = \sin nx, \quad n = 0, 1, \ldots;$

в) $\varphi_n(x) = \begin{cases} \sin nx, & x < 0.5, \\ x^n & x \geq 0,5, \end{cases} \qquad n = 0, 1, \ldots;$

г) $\varphi_n(x) = \begin{cases} -1, & 0 \leq x < 1/n, \\ -1 + 2n(x - 1/n), & 1/n \leq x < 1/(2n), \\ 1, & 1/(2n) \leq x \leq 1, \end{cases} \qquad n = 0, 1, \ldots;$

д) среди перечисленных нет функций, которые могут быть выбраны в качестве конечных элементов.

6. Какие из приведенных ниже систем могут использоваться для приближенного решения уравнения

$$u(x) - 2\int_0^1 x\frac{\sin t}{t}u(t)\,dt = x^2?$$

В приведенных формулах $x_i = t_i = ih, \ i = 0, 1, \ldots, n, \ nh = 1.$

а) $y_i - 2\sum\limits_{j=1}^{n} x_i\frac{\sin t_i}{t_i}y_j = x_i^2, \quad i = 1, \ldots, n;$

б) $y_i - 2\sum\limits_{j=1}^{n} x_i\frac{\sin t_j}{t_j}y_j = x_j^2, \quad i, j = 1, \ldots, n;$

в) $y_i - 2\sum\limits_{j=1}^{n} x_i\frac{\sin t_j}{t_j}y_j = x_i^2, \quad i = 1, \ldots, n;$

г) $y_j - 2\sum\limits_{i=1}^{n} x_i\frac{\sin t_i}{t_i}y_j = x_i^2, \quad i, j = 1, \ldots, n;$

д) $y_j - 2\sum\limits_{i=1}^{n} x_j\frac{\sin t_i}{t_i}y_i = x_j^2, \quad j = 1, \ldots, n;$

е) $y_i - 2x_iy_0 - 2\sum\limits_{j=1}^{n-1} x_i\frac{\sin t_j}{t_j}y_j = x_i^2, \quad i = 0, \ldots, n-1.$

8 РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Значительное число задач физики и техники приводится к линейным и нелинейным уравнениям в частных производных. Универсальным методом решения таких задач является метод конечных разностей (разностный метод) или, как его еще называют, метод сеток. Он позволяет сводить приближенное решение задачи к нахождению решения систем алгебраических уравнений. Для такого сведения, то есть построения разностной схемы необходимо проделать следующие шаги: заменить область непрерывного изменения аргумента дискретной областью, дифференциальный оператор заменить разностным и сформулировать аналог начальных и (или) граничных условий. Этим проблемам, а также вопросу выбора методов решения полученных систем уравнений будет посвящена эта глава. Следует отметить, что часть понятий уже вводилась и использовалась в предыдущей главе. Здесь будут сделаны некоторые обобщения, поэтому в качестве примеров могут рассматриваться и методы решения задач для обыкновенных дифференциальных уравнений.

8.1 ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ РАЗНОСТНЫХ СХЕМ

8.1.1 Сетки и сеточные функции

При численном решении той или иной задачи мы, естественно, не можем воспроизвести решение для всех значений аргумента, изменяющегося в некоторой области Ω евклидового пространства. Естественно поэтому выбрать в этой области некоторое конечное множество точек ω и приближенное решение искать только в этих точках. Такое множество точек называется **сеткой**, а сами точки — **узлами сетки**. Функция, определенная в узлах сетки, называется **сеточной функцией**.

Рассмотрим некоторые примеры сеток.

Равномерная сетка на отрезке. Пусть $\Omega = [a, b]$. Разобьем этот отрезок на N равных частей. Точки деления $x_i = a + ih,\ i = 0, 1, \ldots, N$ — узлы сетки. Величина $h = (b - a)/N$ называется **шагом сетки**. ω_h — множество внутренних узлов, то есть $\omega_h = \{x_i\ :\ i = 1, \ldots, N - 1\}$. Если граничные точки x_0 и x_N включены в сетку, то ее будем обозначать $\overline{\omega}_h$.

Неравномерная сетка на отрезке. Отрезок разбивается на N частей произвольными точками такими, что $a = x_0 < x_1 < \cdots < x_N = b$.

Равномерная сетка в прямоугольнике на плоскости. Пусть $\overline{\Omega} = \{(t, x)\ :\ a \le x \le b,\ 0 \le t \le T\}$. Разобьем отрезки $[a, b]$ и $[0, T]$ на N_1 и N_2 частей соответственно. Пусть $h = (b - a)/N_1,\ \tau = T/N_2$ — шаги по каждой переменной. Через точки деления проведем прямые параллельные соответствующим осям. В результате пересечения этих

прямых получим узлы (t_j, x_i), которые образуют сетку $\overline{\omega}_{\tau,h} = \{(t_j, x_i) \ : \ (t_j, x_i) \in \overline{\Omega}\}$. **Соседними узлами** называют узлы, лежащие на одной и той же горизонтальной или вертикальной прямой, расстояние между которыми равно величине шага сетки по соответствующей переменной.

Равномерная сетка в двумерной области Пусть на плоскости $(x^{(1)}, x^{(2)})$ задана область Ω с границей Γ. Пусть $\overline{\Omega} = \Omega \bigcup \Gamma$. Проведем прямые $x^{(k)}_{i_k} = i_k h_k, \ k = 1, 2, \ i = 0, \pm 1, \pm 2, \ldots$. Точки пересечения этих прямых, которые попадут внутрь области Ω назовем внутренними узлами сетки и их совокупность обозначим ω_h. Точки пересечения прямых с границей Γ назовем граничными узлами, их множество обозначим γ_h, а $\overline{\omega}_h = \omega_h \bigcup \gamma_h$.

В зависимости от решаемой задачи и от метода ее решения, разбиение множества точек сетки на внутренние и граничные может осуществляться по-разному. Соответствующие примеры разбиения встретятся в параграфе 8.7.

Аналогично можно провести построения для области из n-мерного пространства R^n.

Вместо функции непрерывного аргумента $u(x), \ x \in \overline{\Omega} \in R^n$ будем рассматривать **сеточные функции** $y(x)$, то есть функции точки $x \in \overline{\omega}_h$. Сеточную функцию $y(x)$ можно представить в виде вектора. Для этого достаточно пронумеровать все узлы сетки в некотором порядке и рассматривать значения функции в узлах сетки как компоненты вектора. При этом размерность вектора будет равна числу узлов сетки. Обычно рассматривается некоторое множество сеток $\{\omega_h\}$, зависящих от h как от параметра. Поэтому и сеточные функции зависят от параметра h при этом если по каждому аргументу шаг сетки постоянный, то $h = (h_1, \ldots, h_n)$.

Функции $u(x)$ являются элементами некоторого линейного нормированного пространства $\mathcal{H}_0$. Множество сеточных функций при этом образуют пространство $\mathcal{H}_h$. Рассматривая множество сеток $\{\omega_h\}$, получаем множество пространств $\{\mathcal{H}_h\}$ сеточных функций, зависящих от параметра h. В пространстве $\mathcal{H}_h$ вводится норма, являющаяся сеточным аналогом нормы в исходном пространстве $\mathcal{H}_0$. Например, если в $\mathcal{H}_0$ введена норма $\|u\| = \max\limits_{x \in \overline{\Omega}} |u(x)|$, то ее сеточным аналогом будет $\|y\| = \max\limits_{x \in \overline{\omega}_h} |y(x)|$.

В случае, когда в $\mathcal{H}_0$ вводится норма $\|u\|^2 = \int\limits_a^b u(x)\,dx$,[1] норма в $\mathcal{H}_h$ выглядит следующим образом

$$\|y\| = \Big(\sum_{i=0}^{N} y^2(x_i) h\Big)^{1/2}.$$

Иногда в сумму не включают одну или обе крайние точки.

Поскольку y_h будет решением разностной схемы и должно приближать решение исходной задачи, то есть функцию $u(x)$, необходимо решить каким образом оценивать их близость. Проблема состоит в том, что эти функции находятся в разных функциональных пространствах. Для решения этой проблемы сначала поставим в соответствие функции $u(x)$ из $\mathcal{H}_0$ сеточную функцию $[u]_h(x)$ из $\mathcal{H}_h$. Это соответствие можно осуществить по разному. Например, если $u(x)$ — непрерывная функция, то $[u]_h(x)$ — сеточная функция такая, что в узлах сетки ее значения совпадают со значениями функции $u(x)$[2]. Для интегрируемой функции $u(x)$ можно в точке $x \in \omega_h \in R^1$

[1] Здесь $\overline{\Omega} = [a, b]$

[2] Будем в дальнейшем считать, что именно такое соответствие установлено.

положить

$$[u]_h(x) = \frac{1}{h} \int\limits_{x-h/2}^{x+h/2} u(x)\,dx.$$

Мерой близость $u(x)$ и y_h может служить теперь значение величины $\|[u]_h - y_h\|_h$, где $\|\cdot\|_h$ — норма пространства $\mathcal{H}_h$.

8.1.2 Разностные операторы и разностные схемы

Дифференциальное уравнение, которое необходимо решить, будем записывать в виде $Lu = f$, где L — дифференциальный оператор. Например, для уравнения $u''(x) = f(x),\ x \in [a,b]$ оператор $Lu = u''$.

Для однозначного определения решения дифференциального уравнения задаются дополнительные условия. Это могут быть начальные, краевые, начальные и краевые условия. Будем записывать эти дополнительные условия в виде $lu = g$. Таким образом, оператор l определяется дополнительными условиями. Например, если для приведенного выше уравнения заданы краевые условия $u'(a) = 0,\ u(b) = 1$, то

$$lu = \begin{cases} u', & x = a, \\ u, & x = b, \end{cases} \qquad g = \begin{cases} 0, & x = a, \\ 1, & x = b. \end{cases}$$

Таким образом, под исходной дифференциальной задачей будем понимать задачу нахождения функции $u = u(x)$, удовлетворяющей условиям

$$Lu = f, \quad lu = g. \tag{8.1}$$

При переходе к сеточным функциям необходимо будет заменить операторы, фигурирующие в постановке дифференциальной задачи, некоторыми операторами, определенными на этих функциях.

Оператор будем называть **разностным**, если он задан на множестве сеточных функций и каждой сеточной функции ставит в соответствие сеточную функцию, определенную, быть может, на другой сетке. Рассмотрим примеры разностных операторов:

1. Оператор L_h^0 задан на функциях y, определенных на равномерной сетке $\overline{\omega}_h$ на отрезке $[a,b]$ (см. предыдущий пункт), по формуле

   $$(L_h^0 y)_i = \frac{y_{i+1} - y_{i-1}}{2h}, \quad i = 1, 2, \ldots, N-1.$$

   Здесь индекс i в левой части равенства определяет номер узла, в котором оператор принимает указанное справа значение. Таким образом, значение оператора — сеточная функция, определенная на ω_h;

2. $(L_h^+ y)_i = \dfrac{y_{i+1} - y_i}{h}, \quad i = 0, 1, \ldots, N-1;$

3. $(L_h^- y)_i = \dfrac{y_i - y_{i-1}}{h}, \quad i = 1, 2, \ldots, N;$

4. $(\Lambda_h y)_i = \dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \quad i = 1, 2, \ldots, N-1;$

5. $(l_h y)_i = \begin{cases} \dfrac{y_1 - y_0}{h}, & i = 0, \\ y_N, & i = N. \end{cases}$ Значение этого оператора — сеточная функция, определенная в двух крайних точках сетки.

Множество узлов сетки, в которых вычисляется значение функции, фигурирующей в запись разностного оператора, называют **шаблоном**. Например, в предыдущем примере для оператора L_h^0 шаблон состоит из двух точек x_{i-1}, x_{i+1}, а для оператора Λ_h из трех: x_{i-1}, x_i, x_{i+1}.

Замена оператора L на L_h называется **аппроксимацией на сетке** оператора L оператором L_h, или **разностной аппроксимацией** оператора L. Изучение разностных аппроксимаций L_h оператора L обычно производят сначала локально, то есть в любой фиксированной точке сетки.

Погрешностью аппроксимации оператора L оператором L_h называют разность

$$\psi = L_h u - Lu.$$

При исследовании погрешности аппроксимации обычно считают, что функции, на которые действует оператор L достаточно гладкие, то есть имеют столько производных сколько необходимо для проведения всех выкладок.

Говорят, что L_h **имеет k-ый порядок аппроксимации в точке x на классе функций $\mathcal{U}$**, если

$$\psi(x) = L_h u(x) - Lu(x) = O(h^k)$$

для любой функции $u \in \mathcal{U}$, то есть $|\psi(x)| \le Mh^k$, где M положительная константа, не зависящая от h и $k > 0$.

Как отмечалось выше, том случае, когда рассматриваются функции многих переменных, h — вектор, поэтому в этом и последующих определениях либо вместо h^k понимается $|h|^k$, либо по каждой из переменных рассматривается свой порядок аппроксимации. Например, при двух переменных можно говорить о порядке k_1 по переменной $x^{(1)}$ и k_2 по переменной $x^{(2)}$, то есть $\phi = O(h_1^{k1}) + O(h_2^{k2})$.

Используя формулу Тейлора

$$u(x \pm h) = u(x) \pm u'(x)h + u''(x)\frac{h^2}{2!} \pm u'''(x)\frac{h^3}{3!} + u^{IV}(x)\frac{h^4}{4!} + O(h^5), \tag{8.2}$$

легко получить

$$u'(x) - L_h^0 u(x) = u'(x) - \frac{u(x+h) - u(x-h)}{2h} = u'(x) -$$
$$- \frac{u(x) + u'(x)h + u''(x)\frac{h^2}{2!} + O(h^3) - u(x) + u'(x)h - u''(x)\frac{h^2}{2!} + O(h^3)}{2h} = O(h^2).$$

Таким образом, оператор L_h^0 аппроксимирует оператор дифференцирования со вторым порядком на классе трижды непрерывно дифференцируемых функций. Аналогично проверяется, что операторы L_h^+, L_h^- аппроксимируют оператор дифференцирования с первым порядком на классе дважды непрерывно дифференцируемых функций. Для оператора же Λ_h имеем

$$u''(x) - \Lambda_h u(x) = -\frac{h^2}{12}u^{IV} + O(h^4) = O(h^2). \tag{8.3}$$

Обычно требуется оценка погрешности аппроксимации в некоторой сеточной норме. Говорят, что L_h имеет **k-ый порядок аппроксимации** или **аппроксимирует оператор L с порядком** k на классе функций $\mathcal{U}$, если

$$\|L_h[u]_h - [Lu]_h\|_h = O(h^k)$$

для любой функции $u \in \mathcal{U}$.

Вернемся к задаче, поставленной для дифференциального уравнения. Пусть требуется решить уравнение $Lu = f$ с дополнительным условием $lu = g$. Заменяя функции u, f, g сеточными функциями y_h, f_h, g_h, а операторы L и l разностными операторами L_h и l_h соответственно, получим систему алгебраических уравнений

$$L_h y_h = f_h, \quad l_h y_h = g_h \tag{8.4}$$

относительно значений сеточной функции y_h. Эту систему алгебраических уравнение принято называть **разностной задачей** или **разностной схемой**. Заметим, что сетка и соответственно сеточные функции y_h, f_h, g_h, а также разностные операторы L_h и l_h зависят от шага h, поэтому точнее говорить не об одной задаче (8.4), а о семействе задач, зависящих от параметра h.

Разностная схема для краевой задачи для обыкновенного дифференциального уравнения уже встречалась ранее в параграфе (7.2).

Определим что следует понимать под приближением дифференциальной задачи (8.1) разностной задачей (8.4).

Пусть $u = u(x)$ — решение задачи (8.1). **Невязкой разностного уравнения** или **погрешностью аппроксимации разностного уравнения на решении** $u = u(x)$ назовем величину $\psi_h^{(1)} = L_h[u]_h - f_h$. **Невязкой** или **погрешностью аппроксимации дополнительных условий (начальных, граничных) на решении** $u = u(x)$ назовем величину $\psi_h^{(2)} = l_h[u]_h - g_h$.

Будем говорить, что **разностная схема (8.4) аппроксимирует задачу (8.1) с порядком** k, если

$$\|\psi_h^{(1)}\|_h^{(1)} = O(h^k), \quad \|\psi_h^{(2)}\|_h^{(2)} = O(h^k), \quad k > 0.$$

В приведенном определении фигурируют две нормы $\|\cdot\|_h^{(1)}$ и $\|\cdot\|_h^{(2)}$. Это связано с тем, что указанные сеточные функции определены на различном множестве узлов сетки и уже поэтому оцениваются в различных нормах. Например, если в качестве нормы использовании максимальное по модулю значение функции, то для функции $\psi_h^{(1)}$ максимум берется по внутренним узлам сетки, а для $\psi_h^{(2)}$ — по той части границы, где задано дополнительное условие. В общем же случае и принцип выбора нормы для $\psi_h^{(1)}$ и $\psi_h^{(2)}$ может быть различным, то есть для одной из функций в качестве нормы может, например, быть выбран максимум модуля функции, а для другой — сумма модулей значений функции.

Оценка невязок $\psi_h^{(1)}$, $\psi_h^{(2)}$ проводится в предположении, что решение исходной задачи (8.1) существует и имеет столько производных, сколько требуется при получении k-го порядка аппроксимации.

Заметим, что

$$\|\psi_h^{(1)}\|_h^{(1)} = \|L_h[u]_h - f_h\|_h^{(1)} = \|L_h[u]_h - [Lu]_h + [f]_h - f_h\|_h^{(1)} \leq$$
$$\leq \|L_h[u]_h - [Lu]_h\|_h^{(1)} + \|[f]_h - f_h\|_h^{(1)}.$$

Аналогичные неравенства можно записать для дополнительных условий. Отсюда следует, что для того чтобы разностная схема (8.4) аппроксимировала задачу (8.1) с порядком k, достаточно чтобы операторы L_h и l_h аппроксимировали операторы L и l соответственно с порядком k на решении задачи (8.1), а $\|[f]_h - f_h\|_h^{(1)} = O(h^k)$ и $\|[g]_h - g_h\|_h^{(2)} = O(h^k)$.

При решении задачи (8.1) разностным методом основным является вопрос о погрешности, получаемой от замены решения $u(x)$ этой задачи, решением y_h разностной схемы.

Будем говорить, что **разностная схема сходится** если $\|[u]_h - y_h\|_h^{(3)} \to 0$ при $h \to 0$.

Будем говорить, что **разностная схема сходится с порядком** k если $\|[u]_h - y_h\|_h^{(3)} = O(h^k),\ k > 0$.

Естественно возникает вопрос есть ли связь между аппроксимацией разностной схемой дифференциальной задачи и сходимостью. Следует ли из аппроксимации сходимость? Приведем пример, иллюстрирующий введенные выше понятия и показывающий, что из аппроксимации не следует сходимость.

Рассмотрим задачу Коши для уравнения переноса

$$\frac{\partial u(t,x)}{\partial t} + a\frac{\partial u(t,x)}{\partial x} = 0 \text{ при } t \geq 0, \quad u(0,x) = g(x). \tag{8.5}$$

Здесь $Lu = \partial u/\partial t + a\partial u/\partial x$, $f = 0$, $a = const > 0$. Дополнительным является начальное условие, заданное при $t = 0$, то есть вдоль оси Ox. При этом $lu = u$. Легко проверить, что решением задачи (8.5) является функция $u(t,x) = g(x - at)$. Из вида решения следует, что вдоль любой прямой $x - at = C = const$ решение не меняется и его значение равно $g(C)$. Отметим также, что прямая $x - at = C$ пересекает ось Ox в точке C. Таким образом, для того чтобы найти значение решения в точке (t,x) достаточно провести через нее прямую, тангенс угла наклона которой к оси Ox равен $1/a$, найти точку пересечения этой прямой с осью Ox и вычислить значение функции g в этой точке.

Построим теперь разностную схему. Для этого введем сетку

$$\overline{\omega}_{\tau,h} = \{(t_n, x_i)\ :\ t_n = n\tau,\ n = 0,1,\dots,\ x_i = ih,\ i = 0,\pm1,\pm2\dots\}.$$

Здесь τ — шаг сетки по переменой t, а h — по x. Для сокращения записи обозначим $y_{\tau h}(t_n, x_i) = y_i^n$. Здесь и всюду в дальнейшем верхний индекс будет соответствовать временной переменной, а нижний — пространственным. Запишем теперь разностную схему

$$\frac{y_i^{n+1} - y_i^n}{\tau} + a\frac{y_{i+1}^n - y_i^n}{h} = 0, \quad y_i^0 = g(x_i), \quad i = 0,\pm1,\pm2\dots,\ n = 0,1,\dots\ . \tag{8.6}$$

В данном случае разностная схема получилась путем замены операторов дифференцирования $\partial/\partial t$ и $\partial/\partial x$ разностными операторами L_τ^+ и L_h^+ примененными по переменным t и x соответственно.

Разностный оператор

$$(L_{\tau h} y_{\tau h})_i^n = \frac{y_i^{n+1} - y_i^n}{\tau} + a\frac{y_{i+1}^n - y_i^n}{h}.$$

Его шаблон содержит три точки (см. рисунок 8.1) $(t_{n+1}, x_i),\ (t_n, x_i),\ (t_n, x_{i+1})$. Правая часть разностного уравнения $f_{\tau h} = 0$.

Зная шаблон, легко описать теперь алгоритм нахождения решения разностной схемы. Поскольку разностное уравнение связывает значения решения в трех точках, для определения значения решения в точке (t_{n+1}, x_i) достаточно знать его в двух других точках $(t_n, x_i), (t_n, x_{i+1})$. Так как разностное начальное условие определяет значение решения при $n = 0$ можно, положив в разностной схеме $n = 0$, найти

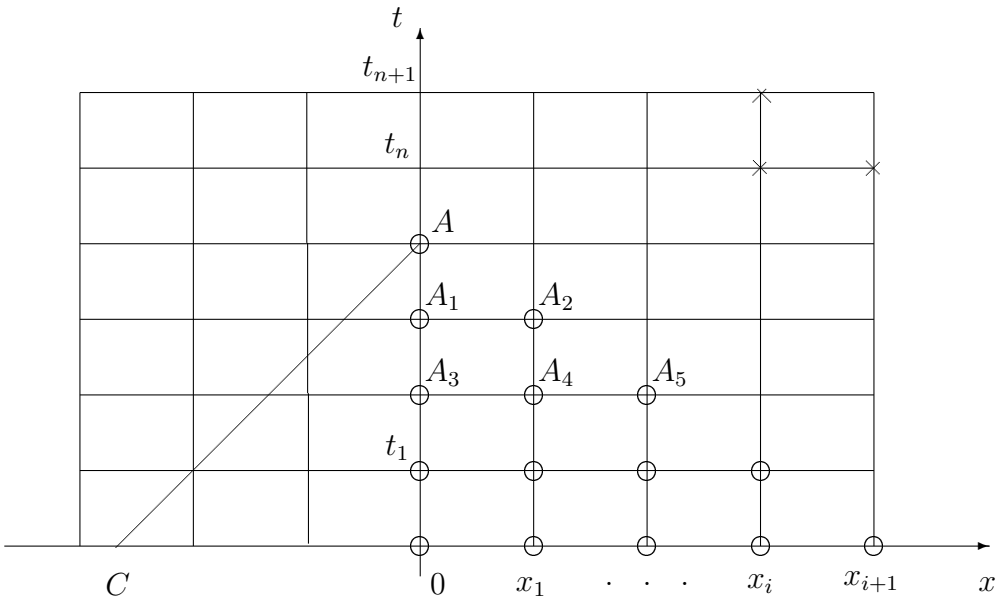


Рис. 8.1 Сетка для уравнения переноса.
Знаками "x" отмечены точки шаблона разностной схемы.

значения решения при $n = 1$ (говорят найти решение на первом слое[3]). Затем, взяв в разностном уравнении $n = 1$, найти решение на втором слое и так далее.

Исследуем порядок аппроксимации. Очевидно, что погрешность аппроксимации начального условия равна нулю (иногда в этом случае говорят о бесконечном порядке аппроксимации). Для разностного уравнения имеем в предположении, что решение дважды непрерывно дифференцируемая функция

$$\psi^{(1)}_{\tau h} = L_{\tau h}[u]_{\tau h} - f_{\tau h} = \frac{u(t+\tau,x)-u(t,x)}{\tau} + a\frac{u(t,x+h)-u(t,x)}{h} =$$
$$= \frac{u(t,x)+\tau\dfrac{\partial u(t,x)}{\partial t}+O(\tau^2)-u(t,x)}{\tau} + a\frac{u(t,x)+h\dfrac{\partial u(t,x)}{\partial x}+O(h^2)-u(t,x)}{h} =$$
$$= \frac{\partial u(t,x)}{\partial t} + a\frac{\partial u(t,x)}{\partial x} + O(\tau+h) = O(\tau+h).$$

Здесь в последнем равенстве воспользовались тем, что u — решение уравнения переноса, поэтому $\dfrac{\partial u(t,x)}{\partial t} + a\dfrac{\partial u(t,x)}{\partial x} = 0$. Таким образом, разностная схема имеет первый порядок аппроксимации по каждой из переменных.

Рассмотрим теперь, сходится ли решение разностной схемы к решению задачи (8.5). Проанализируем для этого значения точного и приближенного решения в точке A (см. рисунок 8.1). Как отмечалось выше, значения точного решения в точках A и C совпадают и равны $g(C)$. Решение разностной схемы в точке A вычисляется через значения в точках A_1 и A_2, в них решение вычисляется, в свою очередь, через значения в точках A_3, A_4, A_5 и так далее. Таким образом, решение разностной схемы в точке A определяется значениями функции g при положительных значениях аргумента. Будем теперь изменять начальные данные, то есть функцию $g(x)$, но таким образом, чтобы ее значение изменялось только для отрицательных значений аргумента. Тогда в точке A решение разностной схемы изменяться не будет, в то время

[3]**Слоем** называют множество узлов сетки, имеющих одну и ту же временную координату.

как решения задачи (8.5) в этой точке при различных функциях g будет различным. Значит, $y_{\tau,h}$ не может приближать $u(t,x)$. Это означает, что сходимости нет.

8.1.3 Устойчивость, теорема сходимости

Как было показано выше, из аппроксимации сходимость не следует, нужно какое-то дополнительное условие. Таким условием является устойчивость.

Определение 8.1.1 *Разностная схема (8.4) называется* **устойчивой**, *если ее решение непрерывно зависит от входных данных f_h, g_h и эта зависимость равномерна относительно шага сетки.*

Более точно сказанное выше означает, что если помимо схемы (8.4) имеется разностная схема

$$L_h\widetilde{y}_h = \widetilde{f}_h, \quad l_h\widetilde{y}_h = \widetilde{g}_h \tag{8.7}$$

то при достаточно малом h для любого $\varepsilon > 0$ найдется такое $\delta(\varepsilon)$, не зависящее от шага, что $\|y_h - \widetilde{y}_h\|_h^{(3)} \le \varepsilon$, если $\|f_h - \widetilde{f}_h\|_h^{(1)} \le \delta$ и $\|g_h - \widetilde{g}_h\|_h^{(2)} \le \delta$.

В том случае, когда операторы L_h и l_h линейны, условие устойчивости означает, что при достаточно малом h существуют константы C_1 и C_2, которые не зависят от h, такие, что для решения разностной схемы (8.4) выполняется оценка

$$\|y_h\|_h^{(3)} \le C_1\|f_h\|_h^{(1)} + C_2\|g_h\|_h^{(2)}. \tag{8.8}$$

Покажем, что если выполнена оценка (8.8), то схема устойчива. Действительно, из линейности операторов следует, что разность $y_h - \widetilde{y}_h$ является решением разностной схемы

$$L_h(y_h - \widetilde{y}_h) = f_h - \widetilde{f}_h, \quad l_h(y_h - \widetilde{y}_h) = g_h - \widetilde{g}_h.$$

Для этого решения справедлива оценка (8.8), согласно которой

$$\|y_h - \widetilde{y}_h\| \le C_1\|f_h - \widetilde{f}_h\|_h^{(1)} + C_2\|g_h - \widetilde{g}_h\|_h^{(2)} \le C_1\delta + C_2\delta \le \varepsilon.$$

Отсюда следует, что $\delta(\varepsilon)$ существует и $\delta(\varepsilon) = \varepsilon/(C_1 + C_2)$.

В дальнейшем, говоря о линейных разностных схемах, то есть о таких у которых операторы L_h, l_h линейны, под устойчивостью будем понимать выполнение неравенства (8.8).

В том случае, когда h — вектор, вводятся понятия безусловной или абсолютной и условной устойчивости. Схему называют **безусловно** или **абсолютно** устойчивой, если (8.8) выполняется при любых достаточно малых шагах сетки. Схема называется **условно** устойчивой, если для выполнения (8.8) необходимо, чтобы шаги сетки удовлетворяли некоторым дополнительным условиям. Соответствующие примеры будут приведены ниже.

Основной теоремой теории разностных схем является теорема сходимости. Сформулируем ее для линейных разностных схем.

Теорема 8.1.1 (Теорема сходимости) *Если разностная схема(8.4) аппроксимирует задачу (8.1) с порядком k и устойчива, то она сходится с порядком k.*

Доказательство Пусть $z_h = [u]_h - y_h$. Тогда

$$L_h z_h = L_h[u]_h - L_h y_h = L_h[u]_h - f_h = \psi_h^{(1)}. \tag{8.9}$$

Здесь в соответствии с определением $\psi_h^{(1)}$ — погрешность аппроксимации разностного уравнения на решении. Аналогично,

$$l_h z_h = \psi_h^{(2)}, \tag{8.10}$$

где $\psi_h^{(2)}$ — погрешность аппроксимации дополнительных условий на решении. В силу условий теоремы

$$\|\psi_h^{(1)}\|_h^{(1)} = O(h^k), \quad \|\psi_h^{(2)}\|_h^{(2)} = O(h^k)$$

Соотношения (8.9),(8.10) можно рассматривать как разностную схему относительно сеточной функции z_h. В соответствии с условием устойчивости отсюда имеем

$$\|[u]_h - y_h\|_h^{(1)} = \|z_h\|_h^{(1)} \le C_1\|\psi_h^{(1)}\|_h^{(1)} + C_2\|\psi_h^{(2)}\|_h^{(2)} = O(h^k),$$

что и требовалось доказать.

Замечание 1. При доказательстве теоремы условие устойчивости нужно было для того, чтобы применить неравенство (8.8) к разностной схеме (8.9),(8.10), записанной относительно сеточной функции z_h. Если, например, дополнительные условия аппроксимируются точно, то есть $\psi_h^{(2)} = 0$, то условие устойчивости достаточно проверить не для общего случая, а для разностной схемы вида $L_h y_h = f_h, \quad l_h y_h = 0$. Иногда это немного облегчает получение неравенства (8.8).

Замечание 2. Легко показать, что и для нелинейных разностных схем из аппроксимации и устойчивости следует сходимость.

Из теоремы сходимости следует, что для нахождения приближения решения необходимо научиться строить устойчивые разностные схемы, которые обладают свойствами аппроксимации. В следующих двух параграфах будут рассмотрены методы построения разностных схем и способы определения их устойчивости.

8.2 МЕТОДЫ ПОСТРОЕНИЯ РАЗНОСТНЫХ СХЕМ

В этом параграфе будут рассмотрены некоторые приемы, с помощью которых можно построить аппроксимирующие разностные схемы.

8.2.1 Замена производных разностными отношениями

Простейший прием для построения разностной схемы заключается в замене в дифференциальном операторе каждой производной соответствующим разностным отношением. С помощью этого приема была построена разностная схема (8.6) в предыдущем параграфе. Для той же задачи Коши для уравнения переноса (8.5) можно было бы построить еще несколько разностных схем. Например, выбрав для замены производной по временной переменной приближение

$$\frac{\partial u(t,x)}{\partial t} \approx \frac{u(t+\tau,x) - u(t,x)}{\tau}, \tag{8.11}$$

а для производной по пространственной переменной выражение

$$\frac{\partial u(t,x)}{\partial x} \approx \frac{u(t,x) - u(t,x-h)}{h}, \tag{8.12}$$

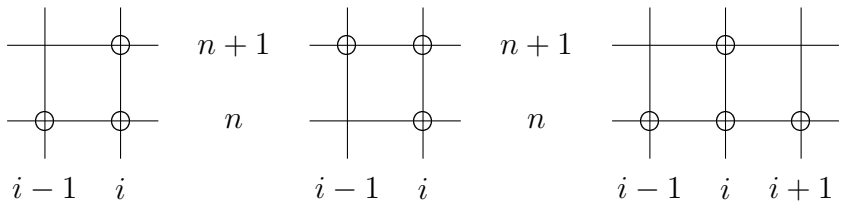


Рис. 8.2 Шаблоны схем: (8.13) слева, (8.15) в центре и (8.16) справа.

получим разностную схему

$$\frac{y_i^{n+1}-y_i^n}{\tau}+a\frac{y_i^n-y_{i-1}^n}{h}=0,\quad y_i^0=g(x_i),\quad i=0,\pm1,\pm2\ldots,\ n=0,1,\ldots\ . \tag{8.13}$$

Если же вместо (8.12) взять приближение

$$\frac{\partial u(t,x)}{\partial x}\approx\frac{u(t+\tau,x)-u(t+\tau,x-h)}{h}, \tag{8.14}$$

оставив для производной по t то же самое приближение (8.11), то схема будет иметь вид

$$\frac{y_i^{n+1}-y_i^n}{\tau}+a\frac{y_i^{n+1}-y_{i-1}^{n+1}}{h}=0,\quad y_i^0=g(x_i),\quad i=0,\pm1,\pm2\ldots,\ n=0,1,\ldots\ . \tag{8.15}$$

Шаблоны полученных схем представлены на рисунке 8.2. Если сравнить построенные разностные схемы, то можно заметить, что шаблон схемы (8.13) содержит на $(n+1)$-ом слое только одну точку и, поэтому, значение решения на $(n+1)$-ом слое выражается по явной формуле через значения решения на предыдущем слое:

$$y_i^{n+1}=\Big(1-\frac{a\tau}{h}\Big)y_i^n+\frac{a\tau}{h}y_{i-1}^n.$$

Такая схема называется **явной**. Для схемы (8.15) шаблон содержит на $(n+1)$-ом слое более одной точки и, следовательно, найти простое выражение для значения решения на этом слой через значения на предыдущем слое не удастся. Такая схема называется **неявной**.

Так же, как и в предыдущем параграфе проверяется, что обе схемы (8.13) и (8.15) имеют первый порядок аппроксимации по τ и h.

Для производной по x можно было бы воспользоваться приближением

$$\frac{\partial u(t,x)}{\partial x}\approx\frac{u(t,x+h)-u(t,x-h)}{2h},$$

Тогда получили бы схему

$$\frac{y_i^{n+1}-y_i^n}{\tau}+a\frac{y_{i+1}^n-y_{i-1}^n}{2h}=0,\quad y_i^0=g(x_i),\quad i=0,\pm1,\pm2\ldots,\ n=0,1,\ldots\ . \tag{8.16}$$

Это явная схема, так как шаблон (см. рисунок 8.2) содержит только одну точку на $(n+1)$-ом слое. Она аппроксимирует задачу (8.5) с порядком $O(\tau+h^2)$.

8.2.2 Метод неопределенных коэффициентов

Идея метода заключается в том, что выбирается шаблон разностной схемы. Затем схема представляется в виде некоторой комбинации с неопределенными пока коэффициентами значений сеточной функции в точках, входящих в шаблон. После этого коэффициенты подбираются таким образом, чтобы выполнялось условие аппроксимации.

Поясним теперь метод на примере разностной схемы для той же задачи (8.5). Выберем шаблон для разностного уравнения, состоящий из точек

$$(n\tau,(i-1)h),(n\tau,ih),(n\tau,(i+1)h),((n+1)\tau,ih).$$

Запишем теперь разностную схему в виде

$$y_i^{n+1}=\alpha_{-1}y_{i-1}^n+\alpha_0 y_i^n+\alpha_1 y_{i+1}^n. \tag{8.17}$$

Здесь $\alpha_j,\ j=-1,0,1$ — неопределенные пока коэффициенты. Коэффициент при y_i^{n+1} взят равным 1, так как если при этой величине стоит какой-нибудь ненулевой коэффициент, то на него всегда можно разделить обе части уравнения (8.17). Подберем коэффициенты так, чтобы порядок малости невязки разностного уравнения на решении задачи (8.5) был как можно больше. С этой целью выпишем невязку, опуская для краткости индекс n при переменной t, а при переменной x индекс i. При преобразованиях воспользуемся формулой Тейлора.

$$\begin{aligned}
u(t+\tau,x)-\sum_{j=-1}^{1}\alpha_j u(t,x+jh)=\\
=u(t,x)+\frac{\partial u(t,x)}{\partial t}\tau+\frac{\partial^2 u(t,x)}{\partial t^2}\frac{\tau^2}{2}+\frac{\partial^2 u(t,x)}{\partial t^2}\frac{\tau^3}{3!}+\dots-\\
-\sum_{j=-1}^{1}\alpha_j\Big(u(t,x)+\frac{\partial u(t,x)}{\partial x}jh+\frac{\partial^2 u(t,x)}{\partial x^2}\frac{(jh)^2}{2}+\frac{\partial^3 u(t,x)}{\partial x^3}\frac{(jh)^3}{3!}+\dots\Big)=\\
=u(t,x)\Big(1-\sum_{j=-1}^{1}\alpha_j\Big)-\frac{\partial u(t,x)}{\partial x}h\Big(\frac{a\tau}{h}+\sum_{j=-1}^{1}j\alpha_j\Big)+\frac{\partial^2 u(t,x)}{\partial x^2}\frac{h^2}{2}\Big(\frac{(a\tau)^2}{h^2}-\sum_{j=-1}^{1}j^2\alpha_j\Big)-\\
-\frac{\partial^3 u(t,x)}{\partial x^3}\frac{h^3}{3!}\Big(\frac{(a\tau)^3}{h^3}+\sum_{j=-1}^{1}j^3\alpha_j\Big)+\dots
\end{aligned}\tag{8.18}$$

При приведении подобных здесь воспользовались тем, что из уравнения (8.5) следуют равенства

$$\begin{gathered}
\frac{\partial u}{\partial t}=-a\frac{\partial u}{\partial x},\\
\frac{\partial^2 u}{\partial t^2}=\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)=\frac{\partial}{\partial t}\left(-a\frac{\partial u}{\partial x}\right)=-a\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right)=a^2\frac{\partial^2 u}{\partial x^2},\\
\frac{\partial^3 u}{\partial t^3}=\frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial t^2}\right)=\frac{\partial}{\partial t}\left(a^2\frac{\partial^2 u}{\partial x^2}\right)=a^2\frac{\partial^2}{\partial x^2}\left(\frac{\partial u}{\partial t}\right)=-a^3\frac{\partial^3 u}{\partial x^3}.
\end{gathered}$$

Будем считать, что шаги разностной схемы связаны соотношением $a\tau/h=r$, где r — некоторая константа. Поскольку разностная схема должна аппроксимировать

задачу Коши при любых начальных данных, в фиксированной точке (t, x) значения функции u и ее производных по x могут быть любыми и не зависеть друг от друга. Поэтому для того, чтобы получить как можно более высокий порядок аппроксимации желательно обратить в ноль как можно больше выражений, находящихся в скобках в правой части соотношения (8.18). Поскольку в нашем распоряжении всего три коэффициента, можно надеяться обратить в ноль первые три слагаемые. Таким образом, получаем:

$$\begin{cases} \alpha_{-1} + \alpha_0 + \alpha_1 & = 1, \\ -\alpha_{-1} + \alpha_1 & = -r, \\ \alpha_{-1} + \alpha_1 & = r^2. \end{cases} \tag{8.19}$$

Отсюда следует, что

$$\alpha_{-1} = \frac{r^2 + r}{2}, \quad \alpha_1 = \frac{r^2 - r}{2}, \quad \alpha_0 = 1 - r^2.$$

Заметим, что попытка обратить в ноль еще одно слагаемое приводит к равенству

$$-\alpha_{-1} + \alpha_1 = -r^3. \tag{8.20}$$

Левая часть этого равенства совпадает с левой частью второго из уравнений (8.19). Значит уравнение (8.20) будет выполнено, если совпадут правые части. Тогда $r = r^3$, то есть $r^2 = 1$. В этом случае для бесконечно дифференцируемой функции u получим, что и все последующие слагаемые в правой части равенства (8.18) обратятся в ноль, то есть порядок аппроксимации равен бесконечности.

Если же $r^2 \neq 1$, правая часть равенства (8.18) имеет порядок $O(h^3)$. Следует отметить, что это еще не означает, что порядок аппроксимации равен 3. Дело в том, что порядок оставшихся членов зависит от вида, в котором записана разностная схема. Например, умножение (8.17) на h приведет к тому, что порядок оставшихся членов станет $O(h^4)$. Для преодоления этой неоднозначности принято считать, что при определении порядка аппроксимации схема должна быть записана в таком виде, что при стремлении шагов сетки к нулю получается дифференциальное уравнение. Если перейти к пределу при τ и h стремящихся к нулю в равенстве (8.17), заменив y на u, получим с учетом первого уравнения системы (8.19) $u(t, x) = u(t, x)$. Таким образом, дифференциальное уравнение не получено. Подставим теперь в (8.17) найденные коэффициенты, разделим обе части на на τ и перегруппируем слагаемые. В результате получим

$$\frac{y_i^{n+1} - y_i^n}{\tau} + a\frac{y_{i+1}^n - y_{i-1}^n}{2h} = \frac{a^2\tau}{2}\frac{y_{i-1}^n - 2y_i^n + y_{i+1}^n}{h^2}. \tag{8.21}$$

Легко заметить, что это приближение уравнения

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = \frac{a^2\tau}{2}\frac{\partial^2 u}{\partial x^2}$$

и предельный переход даст исходное дифференциальное уравнение. Так как при получении (8.21) применялась операция деления на τ, получим что при определении порядка аппроксимации остаток правой части равенства (8.18) надо разделить на τ. Учитывая, что $\tau = O(h)$, имеем второй порядок аппроксимации.

Следует отметить, что для получения второго порядка аппроксимации при выборе коэффициентов α_j есть некоторый произвол. В первое из уравнений (8.19) можно добавить слагаемые порядка $O(h^3)$, второе уравнение — $O(h^2)$, третье — $O(h)$.

Если только $r^2 \neq 1$, на выбранном шаблоне получить разностную схему более высокого порядка аппроксимации невозможно. Повышение порядка аппроксимации требует увеличения количества точек, входящих в шаблон схемы.

8.2.3 Интегро-интерполяционный метод

Различные физические процессы (теплопроводности, диффузии, колебаний и т.д.) характеризуются некоторыми интегральными законами сохранения (энергии, массы, количества движения и т.д.). В основе вывода дифференциальных уравнений математической физики обычно лежат некоторые интегральные соотношения, выражающие законы сохранения для малого объема. Дифференциальные уравнения получаются путем предельного перехода в уравнениях баланса при стягивании объема к нулю. Метод конечных разностей физически означает переход от непрерывной среды к дискретной. При таком переходе естественно потребовать, чтобы все основные свойства физического процесса сохранились. Такими свойствами, прежде всего, являются законы сохранения. Разностные схемы, выражающие на сетке законы сохранения называют **консервативными** или **дивергентными**. Законы сохранения для всей сеточной области для консервативных разностных схем должны быть алгебраическим следствием разностных уравнений.

Интегро-интерполяционный метод позволяет строить консервативные разностные схемы. В основе метода лежит следующий подход. Выбирается шаблон разностной схемы, после чего область разбивается на ячейки, определенным образом связанные с шаблоном. Дифференциальное уравнение интегрируется по выбранным ячейкам. Затем, в зависимости от размерности задачи, применяются формулы Грина или Гаусса-Остроградского. В результате интеграл сводится к контурному или поверхностному, что соответствует физическому закону сохранения и затем приближенно вычисляется этот интеграл.

Проиллюстрируем сказанное на примере краевой задачи для одномерного стационарного уравнения теплопроводности

$$\frac{d}{dx}\Big(k(x)\frac{du(x)}{dx}\Big) = q(x)u(x) - f(x),\ x \in (0,1),\quad u(0) = U_0,\ u(1) = U_1. \tag{8.22}$$

Эта задача описывает распределение температуры $u(x)$ в стержне длины 1, концы которого имеют заданную температуру. Функция $f(x)$ описывает плотность распределения внешних источников тепла, $-q(x)u(x)$ — мощность стоков тепла пропорциональных температуре, а $w = -k(x)u'(x)$ — поток тепла.

Разобьем отрезок $[0,1]$ на I равных частей точками $x_i,\ i = 0,\dots,I,\ Ih = 1$. Выберем шаблон, состоящий из трех точек x_{i-1}, x_i, x_{i+1} и свяжем с этим шаблоном отрезок $[x_{i-1/2}, x_{i+1/2}]$. Здесь $x_{i+1/2} = (i+1/2)h$. Проинтегрируем теперь уравнение (8.22) по этому отрезку и воспользуемся формулой Ньютона-Лейбница. В результате получим

$$w(x_{i-1/2}) - w(x_{i+1/2}) + \int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx = \int\limits_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)\,dx. \tag{8.23}$$

Это уравнение описывает баланс тепла в выбранном отрезке:

- $w(x_{i-1/2})$ — количество тепла, втекающего в отрезок через сечение $x = x_{i-1/2}$;
- $-w(x_{i+1/2})$ — количество тепла, втекающего из отрезка через сечение $x = x_{i+1/2}$;
- $\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx$ — количество тепла, выделяющегося на отрезке за счет источников, распределенных с плотностью f;

- $\int\limits_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)\,dx$ — количество тепла, которое "вытекло"за счет теплообмена боковой поверхности стержня с внешней средой.

Для получения разностной схемы заменим w и интеграл, содержащий u и входящий в соотношение (8.23), линейными комбинациями значений функции u в узлах сетки. Для этого будем считать, что $u(x) \approx u(x_i) = u_i$ при $x \in (x_{i-1/2}, x_{i+1/2})$. Тогда

$$\int\limits_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)\,dx \approx hd_iu_i, \text{ где } d_i = \frac{1}{h}\int\limits_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx. \tag{8.24}$$

Для того, чтобы выразить величину w воспользуемся ее определением, которое запишем в виде $u' = -w/k$. Проинтегрируем это равенство по промежутку (x_{i-1}, x_i) и будем считать, что на этом промежутке $w(x) \approx w(x_{i-1/2}) = w_{i-1/2}$. В результате получим

$$u_{i-1} - u_i = \int\limits_{x_{i-1}}^{x_i} \frac{w(x)}{k(x)}\,dx \approx w_{i-1/2}\int\limits_{x_{i-1}}^{x_i} \frac{dx}{k(x)}.$$

Отсюда следует, что

$$w_{i-1/2} \approx -a_i\frac{u_i - u_{i-1}}{h}, \quad \text{где } a_i^{-1} = \frac{1}{h}\int\limits_{x_{i-1}}^{x_i} \frac{dx}{k(x)}. \tag{8.25}$$

Подставим полученные соотношения (8.24),(8.25) в (8.23) ($w_{i+1/2}$ выражается аналогично), заменим приближенное равенство точным, а u_i на y_i. В результате получим разностную схему

$$\frac{1}{h}\left(a_{i+1}\frac{y_{i+1} - y_i}{h} - a_i\frac{y_i - y_{i-1}}{h}\right) - d_iy_i + f_i = 0, \quad i = 1, \ldots, I-1, \quad y_0 = U_0,\ y_I = U_1, \tag{8.26}$$

где

$$f_i = \frac{1}{h}\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx.$$

Разностная схема представляет собой систему линейных алгебраических уравнений относительно значений сеточной функции y_i, которая легко решается методом прогонки.

Замечание. В том случае, когда функции $k(x), q(x), f(x)$ таковы, что интегралы от них аналитически не вычисляются или вычисляются сложно, можно использовать методы численного интегрирования для вычисления a_i, d_i, f_i. Например, в случае гладких функций, применение формулы средних прямоугольников дает для этих величин следующие выражения:

$$a_i = k(x_{i-1/2}), \quad d_i = q(x_i), \quad f_i = f(x_i). \tag{8.27}$$

Если же воспользоваться формулой трапеций, то

$$a_i = \frac{2k(x_{i-1})k(x_i)}{k(x_{i-1}) + k(x_i)}, \quad d_i = \frac{q(x_{i-1/2}) + q(x_{i+1/2})}{2}, \quad f_i = \frac{f(x_{i-1/2}) + f(x_{i+1/2})}{2}. \tag{8.28}$$

8.2.4 Аппроксимация начальных и граничных условий

Рассмотрим подходы, позволяющие аппроксимировать краевые или начальные условия, на примере условия второго рода для одномерного стационарного уравнения теплопроводности

$$u''(x) = q(x)u(x) - f(x),\ x \in (0,1),\quad u'(0) = U_0,\ u(1) = U_1. \tag{8.29}$$

Вводя равномерную сетку на отрезке $[a,b]$ и используя метод замены производных соответствующими разностями, легко получить аппроксимацию второго порядка для дифференциального уравнения

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = q(x_i)y_i - f(x_i),\quad i = 1,\dots,I-1,\quad x_i = ih,\ h = \frac{1}{I}. \tag{8.30}$$

Краевое условие в точке $x = 1$ аппроксимируется точно, то есть $y_I = U_1$. Главное наше внимание будет уделено сейчас условию в точке $x = 0$. Если воспользоваться тем же методом замены производных соответствующими разностями и записать

$$\frac{y_1 - y_0}{h} = U_0, \tag{8.31}$$

то невязка в этом случае равна

$$\frac{u(h) - u(0)}{h} - U_0 = \frac{u(0) + hu'(0) + O(h^2) - u(0)}{h} - U_0 = u'(0) - U_0 + O(h) = O(h).$$

Таким образом, для этого краевого условия имеем первый порядок аппроксимации и в целом разностная схема аппроксимирует задачу (8.29) с первым порядком. Желательно, поэтому более точно приблизить краевое условие в точке $x = 0$. Для этого можно предложить два способа рассуждений. При первом способе введем дополнительную точку сетки $x_{-1} = -h$ и дополнительное значение сеточной функции y_{-1}. Тогда уравнение (8.30) можно будет записать и для $i = 0$, то есть записать

$$\frac{y_{-1} - 2y_0 + y_1}{h^2} = q(0)y_0 - f(0). \tag{8.32}$$

Для граничного условия воспользуемся выражением

$$\frac{y_1 - y_{-1}}{2h} = U_0,$$

которое имеет второй порядок аппроксимации. Из этого соотношения и из (8.32) исключаем теперь дополнительное значение y_{-1}, получим искомое разностное граничное условие

$$y_1 = y_0\Big(1 + \frac{h^2}{2}q(0)\Big) + hU_0 - \frac{h^2}{2}f(0). \tag{8.33}$$

Второй способ заключается в том, что записывается формула Тейлора

$$u(h) = u(0) + hu'(0) + \frac{h^2}{2}u''(0) + O(h^3).$$

Заменим теперь $u'(0)$, воспользовавшись граничным условием, а $u''(0)$ — уравнением (8.29). Получим

$$u(h) = u(0) + hU_0 + \frac{h^2}{2}(q(0)u(0) - f(0)) + O(h^3).$$

Отбрасывая слагаемые порядка $O(h^3)$, получим соотношение (8.33).

8.3 МЕТОДЫ ИССЛЕДОВАНИЯ УСТОЙЧИВОСТИ

В настоящее время существует несколько подходов к исследованию разностных схем на устойчивость. В этом параграфе будут рассмотрены некоторые из таких подходов предназначенных для разностных схем, построенных для нестационарных задач. Пример доказательства устойчивости разностной схемы для стационарной задачи будет приведен ниже.

8.3.1 Спектральный критерий устойчивости

Пусть разностное уравнения связывает значения сеточной функции на двух соседних временных слоях, то есть в моменты времени t_n и $t_{n+1} = t_n + \tau$. Такую разностную схему будем называть **двухслойной.** Обозначим решение разностной схемы в момент времени t_n через y^n. Предположим, что разностная схема может быть записана в виде

$$y^{n+1} = Sy^n + \tau f^n, \quad n = 0, 1, \ldots, N, \quad N\tau = T, \quad y^0 = g. \tag{8.34}$$

Здесь f^n при каждом фиксированном n и g — заданные сеточные функции на слое, S — линейный разностный оператор, называемый **оператором перехода от слоя к слою**. Он действует на сеточную функцию на слое, то есть связывает значения сеточной функции в точках, соответствующих фиксированному значению времени. Будем предполагать всюду в этом пункте, что S линейный ограниченный оператор, который не зависит от n.

Например, рассмотрим задачу Коши для уравнения теплопроводности

$$\frac{\partial u(t,x)}{\partial t} = a\frac{\partial^2 u(t,x)}{\partial x^2} + f(t,x), \quad |x| < \infty, \quad 0 < t \le T, \quad a = const > 0, \quad u(0,x) = g(x). \tag{8.35}$$

Введем сетку $x_j = jh,\ j = 0, \pm 1, \pm 2, \ldots;\ t_n = n\tau,\ n = 0, 1, \ldots, N,\ N\tau = T$. Обозначим $y(t_n, x_j) = y_j^n,\ f(t_n, x_j) = f_j^n,\ g(x_j) = g_j$. Тогда для задачи (8.35) можно предложить разностную схему

$$\frac{y_j^{n+1} - y_j^n}{\tau} = a\frac{y_{j-1}^n - 2y_j^n + y_{j+1}^n}{h^2} + f_j^n, \quad y_j^0 = g_j. \tag{8.36}$$

Введем разностный оператор,

$$(\Lambda y^n)_j = \frac{y_{j-1}^n - 2y_j^n + y_{j+1}^n}{h^2}.$$

Этот оператор действует на сеточную функцию на слое, то есть является линейной комбинацией значений сеточной функции в точках, расположенных на одном временном слое t_n. Пусть E — единичный оператор, то есть $Ey_j^n = y_j^n$, а $S = E + \tau\Lambda$. Тогда разностная схема (8.36) может быть представлена в виде (8.34), если переписать ее, выразив y_j^{n+1} и всюду опустив нижний индекс j.

Для получения оценки решения y^n разностной схемы (8.34) выразим его через входные данные f и g. Пользуясь многократно уравнением (8.34), записанным для различных значений n, получим

$$y^n = Sy^{n-1} + \tau f^{n-1} = S(Sy^{n-2} + \tau f^{n-2}) + \tau f^{n-1} = S^2 y^{n-2} + \tau S f^{n-2} + \tau f^{n-1} =$$

$$= S^2(Sy^{n-3} + \tau f^{n-3}) + \tau S f^{n-2} + \tau f^{n-1} = \cdots = S^n g + \tau\sum_{k=0}^{n-1} S^k f^{n-1-k}, \tag{8.37}$$

где $S^0 = E,\ S^k = S(S^{k-1})$ [4]. Из (8.37) следует, что

$$\|y^n\| = \|S^n g + \tau \sum_{k=0}^{n-1} S^k f^{n-1-k}\| \le \|S^n g\| + \tau \sum_{k=0}^{n-1} \|S^k f^{n-1-k}\| \le$$

$$\le \|S^n\|\|g\| + \tau \sum_{k=0}^{n-1} \|S^k\|\|f^{n-1-k}\| \le \|S^n\|\|g\| + \tau n \max_{0\le k\le N} \|f^k\| \max_{0\le k\le N} \|S^k\|. \quad (8.38)$$

Заметим, что $\tau n = t_n \le T$. Потребуем, чтобы при достаточно малых, положительных τ и h существовала такая константа C, что при любых $0 \le k \le N$ выполняется неравенство

$$\|S^k\| \le C. \quad (8.39)$$

Тогда, в силу (8.38) имеем

$$\|y^n\| \le C(\|g\| + T \max_{0\le k\le N} \|f^k\|).$$

Это неравенство означает, что разностная схема (8.34) устойчива. Таким образом, вопрос об устойчивости свелся к проверке условия $\|S^k\| \le C$, которое должно выполняться равномерно по τ и h при любых целых неотрицательных k не больших N. Можно показать, что это условие не только достаточно, но и необходимо для устойчивости сходящейся разностной схемы (8.34) (см.[28]).

Получение оценки нормы степеней операторов является довольно сложной задачей. Поэтому упростим задачу и попытаемся получить собственные числа оператора S. Известно, что модуль любого собственного числа оператора не превосходит норму оператора, и, если λ — собственное число оператора S, то λ^k собственное число оператора S^k. Поэтому, по величине $|\lambda^k|$ можно будет судить о величине $\|S^k\|$. По крайней мере, условие $|\lambda^k| \le C$ будет являться необходимым для выполнения неравенство $\|S^k\| \le C$.

Пусть $v = \{v_j\}$ — ограниченная сеточная функция заданная на слое и определенная для всех целых значений индекса j. Так как предполагалось, что S — линейный оператор, его значение на функции v в j-ой точке имеет вид

$$(Sv)_j = \sum_p a_p v_{j+p}.$$

Здесь a_p — коэффициенты, зависящие, вообще говоря, от τ, h, а p пробегает некоторое, не обязательно конечное, множество значений, определяемое шаблоном оператора. Наложим ограничение на коэффициенты оператора. Потребуем, чтобы коэффициенты a_p не зависели от j. С учетом сделанного выше предположения о том, что оператор S не зависит от номера слоя, получается, что **коэффициенты оператора S являются постоянными**.

Собственные функции оператора должны удовлетворять равенству

$$(Sv)_j = \lambda v_j, \quad j = 0, \pm 1, \pm 2 \dots. \quad (8.40)$$

Будем искать решение уравнения (8.40) в виде $v_j = v_0 q^j$, где v_0 — нормировочный множитель, а q — некоторое число. Подставляя это выражение в (8.40), получим

$$\sum_p a_p v_0 q^{j+p} = \lambda v_0 q^j. \quad (8.41)$$

[4] Надеемся, что читатель различает степень у оператора S и верхний индекс у сеточных функций, означающий номер временного слоя.

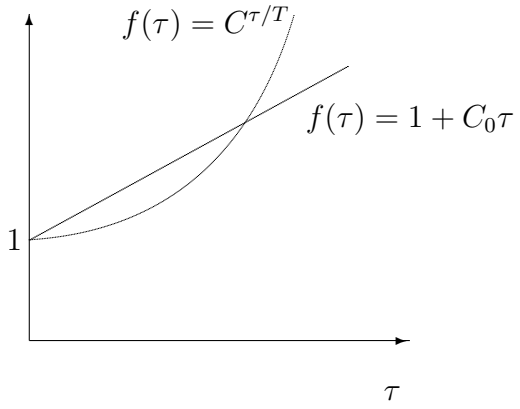


Рис. 8.3 Взаимное расположение кривых $f = C^{\tau/T}$ и $f = 1 + C_0\tau$ при малом τ и достаточно большом C_0.

Отсюда следует, что

$$\lambda = \sum_p a_p q^p. \tag{8.42}$$

Равенства (8.41),(8.42) показывают, что при любом $q \neq 0$ сеточная функция v со значениями $v_j = v_0 q^j$ является собственной функцией оператора S. Однако, если $|q| \neq 1$, сеточная функция v неограничена, так как $|v_j| \to \infty$ либо при $j \to \infty$, либо при $j \to -\infty$. Поэтому ограниченные собственные функции получаются только при $|q| = 1$. При рассмотрении действительных чисел множество собственных функций будет состоять тогда только из двух функций, соответствующих $q = \pm 1$. Это слишком мало, чтобы судить о норме оператора. Поэтому перейдем в поле комплексных чисел, то есть возьмем $q = e^{i\varphi},\ i^2 = -1,\ \varphi \in [0, 2\pi]$[5]. В результате имеем

$$v_j = v_0 e^{ij\varphi}, \quad \lambda(\varphi) = \sum_p a_p e^{ip\varphi}. \tag{8.43}$$

Как уже отмечалось, при достаточно малых положительных $\tau,\ h$, должно выполняться неравенство

$$|\lambda^n(\varphi)| \leq C, \quad \text{при } 0 \leq n \leq N, \quad \varphi \in [0, 2\pi]. \tag{8.44}$$

Можно считать, что в этом неравенстве константа C не меньше 1. Отсюда следует, что

$$|\lambda| \leq C^{1/N} = C^{\tau/T}.$$

Но тогда, см. рисунок 8.3 существует такая константа $C_0 \geq 0$, что при достаточно малых $\tau > 0$ выполняется неравенство $C^{\tau/T} \leq 1 + C_0\tau$.

Из приведенных рассуждений следует, что для ограниченности норм степеней оператора S и, значит, для устойчивости разностной схемы **необходимо** выполнение следующего условия, которое принято называть **спектральным критерием устойчивости Неймана**:

для устойчивости разностной схемы (8.34) должна существовать такая константа $C_0 \geq 0$, что при достаточно малых $\tau > 0,\ h$ и любом $\varphi \in [0, 2\pi]$ выполняется неравенство

$$|\lambda(\varphi)| \leq 1 + C_0\tau. \tag{8.45}$$

[5]Напомним, что $e^{i\varphi} = \cos\varphi + i\sin\varphi$.

Если условие устойчивости Неймана не выполнено, то ни при каком разумном выборе норм нельзя ожидать устойчивости, а в случае его выполнения можно ожидать, что при некотором выборе норм устойчивость имеет место.

Заметим, что если выполнено неравенство (8.45) с некоторой константой $C_0 > 0$, то

$$|\lambda(\varphi)|^n \le (1 + C_0\tau)^n \le e^{C_0 n\tau} \le e^{C_0 T}.$$

Величина $e^{C_0 T}$ хоть и конечна, не зависит от шагов сетки, но может быть очень большой. Поэтому иногда на числа $\lambda(\varphi)$ накладывают условие $|\lambda(\varphi)| \le 1$, которое является более жестким, чем условие (8.45).

Выше было показано, что спектральный критерий является необходимым условием устойчивости разностной задачи Коши. Однако в некоторых случаях это условие является и достаточным. Вывод о том, что спектральный критерий — необходимое условие устойчивости был основан на том факте, что при всех $\varphi \in [0, 2\pi]$ выполняется неравенство $|\lambda(\varphi)| \le \|S\|$. Если же будет выполнено равенство

$$\max_{\varphi\in[0,2\pi]} |\lambda(\varphi)| = \|S\|, \tag{8.46}$$

то это будет означать, что при любых $0 \le k \le N$ справедливо неравенство

$$\|S^k\| = \max_{\varphi\in[0,2\pi]} |\lambda(\varphi)|^k \le C.$$

Таким образом будет выполнено условие (8.39), обеспечивающее устойчивость.

Рассмотрим, когда можно гарантировать выполнение условия (8.46). Пусть сеточные функции y таковы, что сходится ряд $\sum\limits_{j=\infty}^{\infty} |y_j|^2$. Введем норму во множестве таких функций по формуле

$$\|y\| = h\sqrt{\sum_{j=-\infty}^{\infty} |y_j|^2},$$

и обозначим полученное пространство L_2^h.

Каждая сеточная функция $y \in L_2^h$ порождает периодическую функцию $\omega(\varphi)$ по формуле

$$\frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} y_j e^{-ij\varphi} = \omega(\varphi). \tag{8.47}$$

Выражение, стоящее в левой части равенства (8.47) — ряд Фурье, а y_j — коэффициенты Фурье функции $\omega(\varphi)$. Как известно, они находятся из равенства

$$y_j = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \omega(\varphi) e^{ij\varphi}\, d\varphi \tag{8.48}$$

и, кроме того, выполняется равенство Парсеваля [20]

$$\sum_{j=\infty}^{\infty} |y_j|^2 = \int_0^{2\pi} |\omega(\varphi)|^2\, d\varphi = \|\omega(\varphi)\|^2_{L_2(0,2\pi)}. \tag{8.49}$$

Применим к $y \in L_2^h$ оператор S и воспользуемся равенствами (8.48),(8.43)

$$(Sy)_j = \sum_p a_p y_{j+p} = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \omega(\varphi) e^{ij\varphi} \sum_p a_p e^{ip\varphi}\, d\varphi = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \omega(\varphi)\lambda(\varphi) e^{ij\varphi}\, d\varphi. \quad (8.50)$$

Из этого равенства следует, что $(Sy)_j$ являются коэффициентами ряда Фурье функции $\omega(\varphi)\lambda(\varphi)$. Поэтому для них справедливо равенство Парсеваля. С учетом (8.49) имеем

$$\sum_{j=-\infty}^{\infty} |(Sy)_j|^2 = \int_0^{2\pi} |\omega(\varphi)\lambda(\varphi)|^2\, d\varphi \le \max_{\varphi\in[0,2\pi]} |\lambda(\varphi)|^2 \int_0^{2\pi} |\omega(\varphi)|^2\, d\varphi = \\ = \max_{\varphi\in[0,2\pi]} |\lambda(\varphi)|^2 \sum_{j=-\infty}^{\infty} |y_j|^2. \quad (8.51)$$

Согласно определению нормы, неравенство (8.51) принимает вид

$$\|Sy\| \le \max_{\varphi\in[0,2\pi]} |\lambda(\varphi)| \|y\|. \quad (8.52)$$

Из неравенства (8.52) следует, что $\|S\| \le \max\limits_{\varphi\in[0,2\pi]} |\lambda(\varphi)|$ и, так как при любом задании нормы выполняется противоположное неравенство, имеем $\|S\| = \max\limits_{\varphi\in[0,2\pi]} |\lambda(\varphi)|$.

Из неравенства (8.51) следует также, что если $f^n \in L_2^h$ при всех n и $g \in L_2^h$, то решение разностной схемы (8.34) принадлежит L_2^h при всех n. Таким образом, доказана теорема

Теорема 8.3.1 *Если $f^n \in L_2^h$ при всех n и $g \in L_2^h$, то решение y^n разностной схемы (8.34) при всех n принадлежит L_2^h и спектральный критерий является необходимым и достаточным условием устойчивости разностной схемы в пространстве с нормой $\|y\| = \max\limits_{n=0,\dots,N} \|y^n\|_{L_2^h}$.*

Следствие 1. Покажем, что спектральный критерий легко обобщается на случай неявных разностных схем. Пусть разностная схема имеет вид

$$S_1 y^{n+1} = S_2 y^n + \tau f^n, \quad n = 0, 1, \dots, N, \quad N\tau = T, \quad y^0 = g, \quad (8.53)$$

где $(S_1 v)_j = \sum\limits_p b_p v_{j+p}$, $(S_2 v)_j = \sum\limits_p a_p v_{j+p}$ и существует оператор S_1^{-1}, обратный к оператору S_1. Последнее предположение означает, что из разностного уравнения можно однозначно выразить y^{n+1}. Тогда разностную схему можно привести к уже изученному виду, если подействовать на его обе части оператором S_1^{-1}. В результате получим разностное уравнение (8.34), в котором f^n заменена на $S_1^{-1} f^n$ и $S = S_1^{-1} S_2$. Собственные числа оператора S находятся из равенства $S_1^{-1} S_2 v = \lambda v$ или

$$\lambda S_1 v = S_2 v. \quad (8.54)$$

Выбирая у сеточной функции v значения $v_j = v_0 e^{ij\varphi}$, и подставляя их в соотношение (8.54) получим

$$\lambda \sum_p b_p v_0 e^{i(j+p)\varphi} = \sum_p a_p v_0 e^{i(j+p)\varphi}. \quad (8.55)$$

Отсюда следует, что

$$\lambda = \frac{\sum\limits_p a_p e^{ip\varphi}}{\sum\limits_p b_p e^{ip\varphi}}.$$

Таким образом, как для случая $S_1 = E$, так и для случая $S_1 \neq E$, для нахождения λ надо записать и решить уравнение (8.55). Заметим, что формально уравнение для λ получается, если в однородной (f^n=0) разностной схеме y_j^n заменить на $v_0 e^{ij\varphi}$, а y_j^{n+1} на $\lambda v_0 e^{ij\varphi}$.

Применим спектральный критерий к разностной схеме (8.36). Совершив формальную замену, описанную в следствии 1, имеем

$$\frac{\lambda v_0 e^{ij\varphi} - v_0 e^{ij\varphi}}{\tau} = a\frac{v_0 e^{i(j-1)\varphi} - 2v_0 e^{ij\varphi} + v_0 e^{i(j+1)\varphi}}{h^2}.$$

Если разделить обе части равенства на $v_0 e^{ij\varphi}$, умножить на τ и ввести обозначение $r = a\tau/h^2$, получим

$$\lambda = 1 + r(e^{-i\varphi} - 2 + e^{i\varphi}) = 1 + r(\cos\varphi - i\sin\varphi - 2 + \cos\varphi + i\sin\varphi) = \\ = 1 - 2r(1 - \cos\varphi) = 1 - 4r\sin^2\frac{\varphi}{2}.$$

Очевидно, что при всех φ выполняется неравенство $\lambda \leq 1$. Поэтому, для того, чтобы $|\lambda| \leq 1$ достаточно потребовать, чтобы при всех φ

$$-1 \leq 1 - 4r\sin^2\frac{\varphi}{2}.$$

Отсюда следует, что должно выполняться неравенство

$$r\sin^2\frac{\varphi}{2} \leq \frac{1}{2}.$$

Так как наибольшее значение, принимаемое функцией синус равно 1, получаем окончательно условие устойчивости разностной схемы (8.36)

$$\frac{a\tau}{h^2} = r \leq \frac{1}{2}. \tag{8.56}$$

Условие (8.56) означает, что если, например, $a = 1,\ h = 0.1$, то шаг τ нельзя выбрать таким, что $\tau > 0.005$. Так как схема получается устойчивой не при любых значениях τ и h, а только при тех значениях, которые удовлетворяют определенному условию, ее называют **условно устойчивой**

Рассмотрим для задачи (8.35) другую разностную схему

$$\frac{y_j^{n+1} - y_j^n}{\tau} = a\frac{y_{j-1}^{n+1} - 2y_j^{n+1} + y_{j+1}^{n+1}}{h^2} + f_j^n, \quad y_j^0 = g_j. \tag{8.57}$$

Это неявная разностная схема, так как ее шаблон содержит три точки на слое $n+1$. Исследуем ее на устойчивость. Имеем

$$\frac{\lambda v_0 e^{ij\varphi} - v_0 e^{ij\varphi}}{\tau} = a\lambda\frac{v_0 e^{i(j-1)\varphi} - 2v_0 e^{ij\varphi} + v_0 e^{i(j+1)\varphi}}{h^2}.$$

Выражая отсюда λ, получим

$$\lambda = \frac{1}{1 + 4r\sin^2\frac{\varphi}{2}}.$$

Следовательно, в отличии от явной схемы спектральный критерий выполнен при любых достаточно малых положительных шагах сетки. Такую схему называют **безусловно** или **абсолютно устойчивой**.

Следствие 2. Достоинство спектрального критерия состоит в том, что он легко обобщается на более сложные задачи, в частности на системы уравнений. Если считать, что $\mathbf{y}^n, \mathbf{f}^n, \mathbf{v}$ — сеточные вектор-функции, то есть в каждой точке сетки их значения есть вектор, то все приведенные выше рассуждения, связанные с доказательством необходимости выполнения спектрального критерия для устойчивости разностной схемы остаются в силе. Надо только учесть, что $\mathbf{v}_0$ — вектор, а $\mathbf{a}_p$ — матрицы. Тогда равенство (8.41) представляет собой систему линейных однородных алгебраических уравнений относительно вектора $\mathbf{v}_0$. Так как по определению собственная функция не может быть равной нулю, эта система должна иметь ненулевое решение, что возможно тогда и только тогда, когда определитель системы равен нулю. Следовательно, для нахождения λ получаем уравнение

$$\det\Big(\sum_p e^{ip\varphi}\mathbf{a}_p - \lambda\mathbf{E}\Big) = 0,$$

где $\mathbf{E}$ — единичная матрица.

Аналогично тому, как это было указано в следствии 1, для получения равенства (8.41) достаточно в однородную разностную схему формально подставить вместо решения в j-ой точке на n-ом слое выражение $\mathbf{v}_0 e^{ij\varphi}$, а на $(n+1)$-ом слое — выражение $\lambda\mathbf{v}_0 e^{ij\varphi}$.

Различного рода условия, при которых спектральный критерий является достаточным для устойчивости разностной задачи Коши для систем уравнений можно найти в [28].

В качестве примера рассмотрим разностную схему для системы, описывающей распространение одномерных звуковых волн

$$\frac{\partial u}{\partial t} = c\frac{\partial w}{\partial x}, \qquad \frac{\partial w}{\partial t} = c\frac{\partial u}{\partial x}.$$

Разностные уравнения запишем в виде

$$\frac{u_j^{n+1} - u_j^n}{\tau} = c\frac{w_{j+1}^n - w_{j-1}^n}{2h}, \qquad \frac{w_j^{n+1} - w_j^n}{\tau} = c\frac{u_{j+1}^n - u_{j-1}^n}{2h}.$$

Совершив формальную замену, имеем

$$\frac{\lambda v_{01}e^{ij\varphi} - v_{01}e^{ij\varphi}}{\tau} = c\frac{v_{02}e^{i(j+1)\varphi} - v_{02}e^{i(j-1)\varphi}}{2h},$$
$$\frac{\lambda v_{02}e^{ij\varphi} - v_{02}e^{ij\varphi}}{\tau} = c\frac{v_{01}e^{i(j+1)\varphi} - v_{01}e^{i(j-1)\varphi}}{2h}.$$

Введем обозначение $\kappa = c\tau/h$. Тогда, сократив оба уравнения на $e^{ij\varphi}$ и приведя подобные, получим

$$\begin{cases} (\lambda-1)v_{01} - i\kappa\sin\varphi v_{02} = 0, \\ -i\kappa\sin\varphi v_{01} + (\lambda-1)v_{02} = 0. \end{cases}$$

Приравнивая нулю определитель матрицы этой системы уравнений, имеем

$$\begin{vmatrix} \lambda-1 & -i\kappa\sin\varphi \\ -i\kappa\sin\varphi & \lambda-1 \end{vmatrix} = (\lambda-1)^2 + \kappa^2\sin^2\varphi = 0.$$

Следовательно, $\lambda = 1 \pm i\kappa \sin\varphi$, $|\lambda|^2 = 1 + \kappa^2 \sin^2\varphi > 1$ при $\varphi \neq m\pi$. Таким образом, условие устойчивости всегда не выполнено. Такую схему называют **абсолютно** или **безусловно неустойчивой.**

Следствие 3. Спектральный критерий легко обобщается и на случай многослойных схем. Рассмотрим для простоты трехслойную разностную схему

$$y^{n+1} = S_1 y^n + S_2 y^{n-1} + \tau f^n, \quad y^0 = g^0, \quad y^1 = g^1. \tag{8.58}$$

где $S_1,\ S_2$ — разностные операторы, действующие на слое. Если ввести вектор

$$\mathbf{Y}^n = (Y^n_{(1)}, Y^n_{(2)}) = (y^n, y^{n-1}),$$

то разностная схема (8.58) перепишется в виде двухслойной разностной схемы для системы

$$Y^{n+1}_{(1)} = S_1 Y^n_{(1)} + S_2 Y^n_{(2)} + \tau f^n, \quad Y^{n+1}_{(2)} = Y^n_{(1)}, \quad (Y^1_{(1)}, Y^1_{(2)}) = (g^1, g^0). \tag{8.59}$$

Так как схема (8.59) двухслойная, к ней применим спектральный критерий. Как отмечалось в следствии 2, для нахождения λ надо совершить формальную замену: положить $f^n = 0$ и подставить вместо $\mathbf{Y}^{n+1}_j$ вектор $\lambda \mathbf{v}_0 e^{ij\varphi}$, а вместо $\mathbf{Y}^n_j$ — вектор $\mathbf{v}_0 e^{ij\varphi}$. В результате получим

$$\lambda v_{01} e^{ij\varphi} = S_1 v_{01} e^{ij\varphi} + S_2 v_{02} e^{ij\varphi}, \quad \lambda v_{02} e^{ij\varphi} = v_{01} e^{ij\varphi}. \tag{8.60}$$

Из второго уравнения имеем $v_{02} e^{ij\varphi} = \frac{1}{\lambda} v_{01} e^{ij\varphi}$. Подставим это соотношение в первой из уравнений (8.60), в результате получим равенство, из которого можно найти λ:

$$\lambda v_{01} e^{ij\varphi} = S_1 v_{01} e^{ij\varphi} + \frac{1}{\lambda} S_2 v_{01} e^{ij\varphi}. \tag{8.61}$$

Заметим теперь, что для получения равенства (8.61) нет необходимости переходить к двухслойной схеме. Это равенство получается, если разностное уравнение (8.58) сделать однородным, положив $f^n = 0$, после чего произвести формальную замену, подставив $\lambda^k v_0 e^{ij\varphi}$ вместо $y^{n+k}_j,\ k = -1, 0, 1$[6].

В качестве примера рассмотрим разностную схему для задачи (8.35):

$$\frac{y^{n+1}_j - y^{n-1}_j}{2\tau} = a \frac{y^n_{j-1} - 2y^n_j + y^n_{j+1}}{h^2} + f^n_j, \quad y^0_j = g_j. \tag{8.62}$$

Совершая указанную выше замену, получим

$$\frac{\lambda v_0 e^{ij\varphi} - \lambda^{-1} v_0 e^{ij\varphi}_j}{2\tau} = a \frac{v_0 e^{i(j-1)\varphi} - 2 v_0 e^{ij\varphi} + v_0 e^{i(j+1)\varphi}}{h^2}.$$

Сократим обе части равенства на $v_0 e^{ij\varphi}$, умножим на $2\lambda\tau$ и введем обозначение $r = 2a\tau/h^2$. В результате имеем

$$\lambda^2 - 1 = r\lambda(e^{-i\varphi} - 2 + e^{i\varphi}).$$

Учитывая, что

$$e^{-i\varphi} - 2 + e^{i\varphi} = 2\cos\varphi - 2 = -4\sin^2(\varphi/2),$$

[6]Во избежании путаницы еще раз хотелось бы подчеркнуть, что индекс k у λ означает степень.

получаем квадратное уравнение для нахождения λ

$$\lambda^2 + 4r\lambda \sin^2(\varphi/2) - 1 = 0.$$

Дискриминант этого квадратного уравнения положителен, значит, уравнение имеет два действительных корня. Так как согласно теореме Виета произведение корней равно -1, один корень по модулю меньше 1, а другой больше. Отсюда следует, что схема неустойчива.

Следствие 4. Спектральный критерий применим для исследования разностной задачи Коши и в случае, когда пространственных переменных две и более. Если, например, имеется две пространственные переменные, собственную функцию надо выбирать в виде $v = v_0 e^{i(j\varphi + l\psi)}$.

Пусть для решения задачи Коши

$$\frac{\partial u}{\partial t} = a\left(\frac{\partial^2 u}{\partial {x^{(1)}}^2} + \frac{\partial^2 u}{\partial {x^{(2)}}^2}\right), \quad 0 < t < T, \quad |x^{(1)}|, |x^{(2)}| < \infty, \quad u(0, x^{(1)}, x^{(2)}) = g(x^{(1)}, x^{(2)}) \tag{8.63}$$

на сетке $(t_n, x_j^{(1)}, x_l^{(2)}) = (\tau n, h_1 j, h_2 l)$ построена явная разностная схема

$$\frac{y_{jl}^{n+1} - y_{jl}^n}{\tau} = a\left(\frac{y_{j-1\,l}^n - 2y_{jl}^n + y_{j+1\,l}^n}{h_1^2} + \frac{y_{jl-1}^n - 2y_{jl}^n + y_{jl+1}^n}{h_2^2}\right), \quad y_{jl}^0 = g(x_j^{(1)}, x_l^{(2)}). \tag{8.64}$$

Подставляя в разностное уравнение $v_0 e^{i(j\varphi + l\psi)}$ вместо y_{jl}^n и $\lambda v_0 e^{i(j\varphi + l\psi)}$ вместо y_{jl}^{n+1}, после сокращений и тождественных преобразований найдем

$$\lambda = 1 - 4r_1 \sin^2(\varphi/2) - 4r_2 \sin^2(\psi/2).$$

Здесь $r_s = a\tau/h_s^2$, $s = 1, 2$. При любых ψ, φ для λ выполняется неравенство

$$1 - 4r_1 - 4r_2 \le \lambda \le 1.$$

Отсюда следует, что условие устойчивости будет выполнено, если $-1 \le 1 - 4r_1 - 4r_2$, то есть при $r_1 + r_2 \le 1/2$. В частности, если $h_1 = h_2 = h$, то $r_1 = r_2 = a\tau/h^2$ и условие устойчивости выполнено при $a\tau/h^2 \le 1/4$. Сравнивая это условие с условием (8.56), полученным для случая одной пространственной переменной, видим, что для двух пространственных переменных при одном и том же шаге h по переменным $x^{(1)}, x^{(2)}$ приходится брать шаг τ в два раза меньше. Очевидно, что если бы пространственных переменных было три, то условие на шаг по временной переменной было бы в три раза более жестким и.т.д.

Рассмотрим теперь прием, позволяющий использовать спектральный критерий не только для задачи Коши для уравнений с постоянными коэффициентами, но и для уравнений с непрерывными переменными коэффициентами, для некоторых нелинейных уравнений и для нестационарных краевых задач.

Зафиксируем произвольную точку внутри области, в которой ищется решение, и вычислим в этой точке коэффициенты. Рассмотрим теперь задачу Коши, в которой в качестве коэффициентов уравнений взяты зафиксированные коэффициенты. Сформулируем **принцип замороженных коэффициентов**: *для устойчивости исходной задачи необходимо, чтобы задача Коши для разностного уравнения с постоянными коэффициентами удовлетворяла спектральному признаку устойчивости.*

Приведем не строгие рассуждения в обоснование принципа замороженных коэффициентов. Напомним, что одна из трактовок понятия устойчивости — малость

влияния на решение возмущений, внесенных во входные данные. Таким образом, исследование устойчивости есть ни что иное, как исследование поведения возмущения. При измельчении шагов сетки коэффициенты, в силу непрерывности, начинают меняться слабо в окрестности фиксированной точки, если окрестность определяется заданным числом шагов, то есть числом шагов, которое не зависит от их величины. Кроме того, расстояние до границ области, измеряемое числом шагов сетки, стремится к бесконечности. Поэтому, при мелкой сетке возмущения, наложенные на решение в фиксированный момент времени в окрестности фиксированной по пространственным переменным точки, развиваются за малый промежуток времени так же как и для задачи Коши с замороженными коэффициентами. Заметим, что это рассуждение носит общий характер и не зависит ни от числа переменных, ни от числа уравнений.

Если необходимое условие устойчивости, полученное путем рассмотрения задачи Коши с замороженными в произвольной точке области коэффициентами, окажется нарушенным, то устойчивость нельзя ожидать ни при каких граничных условиях. Подчеркнем, что принцип замороженных коэффициентов никак не учитывает влияния граничных условий. Поэтому в случае выполнения принципа замороженных коэффициентов, устойчивость может иметь место при одних граничных условиях и не иметь при других.

Рассмотрим примеры. Пусть в задаче (8.35) коэффициент a зависит от t, x. Тогда разностная явная схема будет иметь почти такой же вид как и схема (8.36). единственное отличие заключается в том, что вместо коэффициента a надо будет записать $a_j^n = a(t_n, x_j)$. Ранее уже было получено с помощью спектрального критерия для схемы (8.36) условие устойчивости $a\tau/h^2 \le 1/2$. В соответствии с принципом замороженных коэффициентов это условие должно выполняться для коэффициентов, вычисленных в произвольной точке области. Отсюда получаем достаточное условие устойчивости

$$\frac{\max\limits_{t,x} a(t,x)\tau}{h^2} \le \frac{1}{2}. \tag{8.65}$$

Рассмотрим теперь пример нелинейной задачи

$$\begin{gathered}\frac{\partial u(t,x)}{\partial t} = (1+u^2)\frac{\partial^2 u(t,x)}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t \le T, \\ u(t,0) = \mu_1(t), \quad u(t,1) = \mu_2(t), \quad u(0,x) = g(x).\end{gathered} \tag{8.66}$$

Введем сетку по пространственной переменной $x_j = jh,\ j = 0, 1, \ldots, M,\ Mh = 1$. Будем считать, что шаг по временной переменной может меняться от слоя к слою. Обозначим $t_n = \tau_0 + \cdots + \tau_{n-1}$. Тогда для задачи (8.66) можно предложить разностную схему

$$\frac{y_j^{n+1} - y_j^n}{\tau_n} = \left(1 + (y_j^n)^2\right)\frac{y_{j-1}^n - 2y_j^n + y_{j+1}^n}{h^2}, \quad y_0^n = \mu_1(t_n), \quad y_M^n = \mu_2(t_n), \quad y_j^0 = g(x_j). \tag{8.67}$$

Эта схема позволяет последовательно, слой за слоем, вычислить решение. Сначала из граничных условий находим $y_0^1,\ y_M^1$, затем, положив $n = 0$, из разностного уравнения находим $y_1^1, y_2^1, \ldots, y_{M-1}^1$, то есть определяем решение на первом слое. При этом необходимые для проведения вычислений значения y_j^0 берутся из начальных условий. После этого процесс продолжаем для второго слоя и так далее. Осталось описать процесс нахождения шагов по временной переменной.

Пусть на слое t_n решение уже найдено. Используем для определения τ_n принцип замороженных коэффициентов. В соответствии с этим принципом запишем разностную задачу Коши, которая для изучаемого случая совпадает с задачей (8.36).

Коэффициент $a = 1 + (y^n_{j_0})^2$, где j_0 — некоторое фиксированное положительное целое число меньшее M. Как уже было получено выше, для устойчивости разностной схемы (8.36) необходимо выполнение условия $a\tau/h^2 \le 1/2$. Так как это неравенство должно выполняться при любой фиксированной точке сетки, в которой вычисляется коэффициент a, получаем условие на шаг τ_n:

$$\tau_n \le \tau_n^{max} = \frac{h^2}{2 \cdot \max\limits_{0<j<M}\left(1 + (y^n_j)^2\right)}.$$

Обычно, при проведении расчетов для шага τ_n выбирается не максимально возможное значение, а несколько меньшее, например, берут $\tau_n = \kappa\tau_n^{max}$. При этом коэффициент κ выбирается из промежутка $[0.7, 0.9]$.

8.3.2 Принцип максимума

В отличие от спектрального критерия, метод исследования устойчивости, носящий название принцип максимума, позволяет исследовать разностные схемы для нестационарных задач Коши и краевых задач, коэффициенты которых переменные. Принцип максимума является достаточным условием устойчивости в C - норме, то есть норме на слое определенной по формуле $\|y^n\| = \max\limits_j |y^n_j|$.

Предположим, что разностная схема записана в виде

$$\sum_j a_j y^{n+1}_{k+j} = \sum_j b_j y^n_{k+j} + \tau f^n_k, \quad n = 0, 1, \ldots, N-1, \quad N\tau = T. \tag{8.68}$$

Здесь коэффициенты a_j, b_j могут зависеть от k, n хотя для сокращения записи схемы это явно не указано. Суммирование на каждом слое производится по узлам шаблона в окрестности точки k-го узла. При этом коэффициенты a_j перенумерованы так, что $|a_0| = \max\limits_j |a_j|$.

Теорема 8.3.2 (Принцип максимума) *Для того, чтобы разностная схема (8.68) была устойчива, достаточно, чтобы существовали такие, не зависящие от шагов сетки константы $c_1 \ge 0,\ c_2 > 0,\ c_3 > 0$, что выполняются неравенства*

$$(1 + c_1\tau)|a_0| \ge \sum_{j\ne 0} |a_j| + \sum_j |b_j|, \tag{8.69}$$

$$|a_0| - \sum_{j\ne 0} |a_j| \ge c_2, \tag{8.70}$$

$$|a_0| \le c_3. \tag{8.71}$$

Доказательство. Из уравнения (8.68) следует, что для всех допустимых значение числа k выполняется неравенство

$$|a_0||y^{n+1}_k| \le \sum_{j\ne 0} |a_j||y^{n+1}_{k+j}| + \sum_j |b_j||y^n_{k+j}| + \tau|f^n_k|.$$

Применим это неравенство к узлу с номером m, где номер m выбран таким образом, что

$$|y^{n+1}_m| = \max_j |y^{n+1}_j| = \|y^{n+1}\|.$$

В результате получим

$$|a_0|\|y^{n+1}\| = |a_0||y_m^{n+1}| \le \sum_{j\ne 0}|a_j||y_{m+j}^{n+1}| + \sum_j |b_j||y_{m+j}^n| + \tau|f_m^n| \le$$
$$\le \Big(\sum_{j\ne 0}|a_j|\Big)\|y^{n+1}\| + \Big(\sum_j |b_j|\Big)\|y^n\| + \tau\|f^n\|.$$

Отсюда следует, что

$$\Big(|a_0| - \sum_{j\ne 0}|a_j|\Big)\|y^{n+1}\| \le \Big(\sum_j |b_j|\Big)\|y^n\| + \tau\|f^n\|. \tag{8.72}$$

В силу условия (8.69)

$$\sum_j |b_j| \le (1+c_1\tau)|a_0| - \sum_{j\ne 0}|a_j| \le \Big(|a_0| - \sum_{j\ne 0}|a_j|\Big) + c_1\tau|a_0|.$$

Подставим это неравенство в (8.72) и разделим обе части полученного неравенства на коэффициент при $\|y^{n+1}\|$. В результате имеем

$$\|y^{n+1}\| \le \left(1 + \frac{c_1\tau|a_0|}{|a_0| - \sum\limits_{j\ne 0}|a_j|}\right)\|y^n\| + \tau\frac{1}{|a_0| - \sum\limits_{j\ne 0}|a_j|}\|f^n\|. \tag{8.73}$$

Из условий (8.70),(8.71) следует, что

$$\frac{c_1|a_0|}{|a_0| - \sum\limits_{j\ne 0}|a_j|} \le \frac{c_1c_3}{c_2} = c_4, \quad \frac{1}{|a_0| - \sum\limits_{j\ne 0}|a_j|} \le \frac{1}{c_2} = c_5.$$

Тогда, учитывая (8.73), имеем

$$\|y^{n+1}\| \le (1+c_4\tau)\|y^n\| + c_5\tau\|f^n\|. \tag{8.74}$$

Заметим, что $1 + c_4\tau \le e^{c_4\tau}$, $n\tau = t_n < T$. Применяя теперь неравенство (8.74) при различных значениях n, получим

$$\|y^{n+1}\| \le (1+c_4\tau)\|y^n\| + c_5\tau\|f^n\| \le$$
$$\le (1+c_4\tau)\big[(1+c_4\tau)\|y^{n-1}\| + c_5\tau\|f^{n-1}\|\big] + c_5\tau\|f^n\| \le \cdots \le$$
$$\le (1+c_4\tau)^{n+1}\|y^0\| + c_5\tau\big[(1+c_4\tau)^n\|f^0\| + (1+c_4\tau)^{n-1}\|f^1\| + \cdots + \|f^n\|\big] \le$$
$$\le e^{c_4T}\big(\|y^0\| + Tc_5\max_{k=0,\dots,n}\|f^k\|\big).$$

Это неравенство означает, что схема устойчива.

Замечание 1. Из доказательства теоремы следует, что если $c_1 = 0$, то условие (8.71) является лишним.

Для иллюстрации применения принципа максимума рассмотрим задачу

$$\frac{\partial u}{\partial t} = a^2(t,x)\frac{\partial^2 u}{\partial x^2} + f(t,x), \quad 0 < t \le T, \quad 0 < x < 1,$$
$$u(t,0) = 0, \quad u(t,1) = 0, \quad u(0,x) = g(x). \tag{8.75}$$

На сетке $x_k = kh,\ k = 0, \dots, K,\ Kh = 1,\ t_n = n\tau,\ n = 0, \dots, N,\ N\tau = T$ рассмотрим явную разностную схему

$$\frac{y_k^{n+1} - y_k^n}{\tau} = a^2(t_n, x_k)\frac{y_{k-1}^n - 2y_k^n + y_{k+1}^n}{h^2} + f(t_n, x_k), \tag{8.76}$$

$$k = 1, \dots, K-1,\ n = 0, \dots, N-1,$$

$$y_0^{n+1} = 0, \quad y_K^{n+1} = 0, \quad y_k^0 = g(x_k). \tag{8.77}$$

Умножая (8.76) на τ, получим разностное уравнение, записанное в форме (8.68). При этом $a_0 = 1$ во всех точках сетки,

$$b_{-1} = b_1 = \frac{a^2(t_n, x_k)\tau}{h^2},\ b_0 = 1 - 2\frac{a^2(t_n, x_k)\tau}{h^2},\ \text{при } k = 1, \dots, K-1.$$

Остальные коэффициенты равны нулю. Заметим, что в граничных точках, то есть при $x = 0,\ x = 1$, разностная схема также имеет вид (8.68), так как в этих точках ее можно записать в виде $a_0 y_0^{n+1} = \tau \cdot 0, \quad a_0 y_K^{n+1} = \tau \cdot 0$.

Если применить к разностному уравнению (8.76) принцип замороженных коэффициентов, рассмотренный в предыдущем пункте, получим необходимое условие устойчивости

$$\frac{\tau \max\limits_{t,x} a^2(t, x)}{h^2} \le \frac{1}{2}. \tag{8.78}$$

Заметим теперь, что если это условие выполнено, то коэффициенты $b_j,\ j = -1, 0, 1$ не отрицательны. Поэтому сумма их модулей равна их сумме и равна 1, если $k \ne 0,\ k \ne K$ и равна нулю, если $k = 0,\ k = K$. Следовательно, выполнены условия, которые накладываются в принципе максимума, при этом $c_1 = 0,\ c_2 = 1,\ c_3 = 1$. Таким образом, условие (8.78) является не только необходимым, но и достаточным для устойчивости разностной схемы (8.77).

Если разностное уравнение (8.76) заменить на

$$\frac{y_k^{n+1} - y_k^n}{\tau} = a^2(t_{n+1}, x_k)\frac{y_{k-1}^{n+1} - 2y_k^{n+1} + y_{k+1}^{n+1}}{h^2} + f(t_{n+1}, x_k), \tag{8.79}$$

получится неявная разностная схема. Переписав уравнение (8.79) в виде

$$-\frac{a^2(t_{n+1}, x_k)\tau}{h^2}y_{k-1}^{n+1} + \Big(1 + \frac{2a^2(t_{n+1}, x_k)\tau}{h^2}\Big)y_k^{n+1} - \frac{a^2(t_{n+1}, x_k)\tau}{h^2}y_{k-1}^{n+1} = y_k^n + \tau f(t_{n+1}, x_k),$$

получаем, что при $k = 1, \dots, K-1$

$$b_0 = 1, \quad a_1 = a_{-1} = -\frac{a^2(t_{n+1}, x_k)\tau}{h^2}, \quad a_0 = 1 + \frac{2a^2(t_{n+1}, x_k)\tau}{h^2}.$$

При $k = 0$ и при $k = K$ все коэффициенты кроме a_0 равны нулю, а $a_0 = 1$. Очевидно, что в этом случае все условия принципа максимума выполнены, и, следовательно, схема абсолютно устойчива, то есть устойчива при любом выборе шагов сетки.

Замечание 2. В рассмотренных примерах граничные условия были однородными. Достаточно просто обобщить утверждение, связанное с устойчивостью и для неоднородных граничных условий. Мы этого делать сейчас не будем, однако заметим, что устойчивость нам нужна была, прежде всего, для установления факта сходимости. Если же граничные условия неоднородные, но аппроксимируются точно, то из замечания 1 к теореме сходимости следует, что проверку устойчивости достаточно проверить для однородных граничных условий.

8.4 РАЗНОСТНЫЕ СХЕМЫ ДЛЯ ГИПЕРБОЛИЧЕСКИХ УРАВНЕНИЙ

В этом параграфе основное внимание будет уделено **уравнению переноса**

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + f(t,x). \tag{8.80}$$

Выбор этого уравнения не случаен. К решению уравнений такого типа сводится гиперболическая система уравнений в случае двух пространственных переменных, уравнение колебания струны. Докажем, например, последнее утверждение. Пусть $u,\ v$ — дважды непрерывно дифференцируемые решения системы

$$\frac{\partial u}{\partial t} = a\frac{\partial v}{\partial x}, \quad \frac{\partial v}{\partial t} = a\frac{\partial u}{\partial x}, \quad a = const. \tag{8.81}$$

Продифференцируем первое из этих уравнений по t, а второе по x, умножим второе уравнение на a и сложим полученные результаты. Тогда имеем

$$\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}, \tag{8.82}$$

то есть u есть решение уравнения колебания струны. Если теперь сложить и вычесть уравнения (8.81), получим

$$\frac{\partial(u+v)}{\partial t} = a\frac{\partial(u+v)}{\partial x}, \quad \frac{\partial(u-v)}{\partial t} = -a\frac{\partial(u-v)}{\partial x}.$$

Таким образом, каждая из функций $u+v$ и $u-v$ является решением однородного уравнения переноса[7].

Уравнение переноса является одним из простейших, поэтому на его примере легко исследовать разностные схемы, выявить те особенности, которые затем можно будет учесть при построении схем для более сложных задач. Такой прием исследования является на сегодняшний день типичным для задач математической физики, где провести изучение проблемы в общем виде пока зачастую не представляется возможным, либо является чрезвычайно сложным.

8.4.1 Разностные схемы для уравнения переноса (схемы бегущего счета)

Рассмотрим сначала некоторые свойства решения уравнения переноса, которые помогут объяснить особенности построения разностных схем. Ранее уже отмечалось, что в случае постоянного коэффициента a общее решение однородного уравнения имеет вид $u = F(x-at)$, где F — произвольная гладкая функция. Таким образом, решение принимает постоянное значение вдоль прямых $x-at = const$, которые называют **характеристиками**. Поэтому для определения решения в некоторой области вдоль характеристик, достаточно найти точку пересечения характеристики с той частью границы области, где задано решение, то есть задано начальное или граничное условие, и затем распространить значение решения в этой точке на всю характеристику. Поэтому говорят, что начальные и граничные условия переносятся вдоль характеристик.

[7]Для каждой из функций коэффициент при производной по x разумеется свой.

Для неоднородного уравнения и для уравнения с переменным коэффициентом ситуация качественно остается такой же. В случае неоднородного уравнения решение меняется вдоль характеристики, характер этого изменения зависит от правой части. Для переменного коэффициента характеристики не являются прямыми. Но в обоих случаях по значению решение в точке пересечения характеристики с границей области можно найти решение вдоль характеристики. Поэтому наклон характеристик играет важное значение для постановки краевых задач.

Возвращаясь для простоты к случаю однородного уравнения с постоянным коэффициентом и анализируя вид решения, заключаем следующее. Если в начальный момент времени график решения на некотором промежутке имел вид $u = F(x)$, то по прошествии времени t график станет $u = F(x - at)$. Это означает, что график сместился на величину $|a|t$ влево при отрицательном a и вправо при положительном[8]. Скорость перемещения при этом равна $|a|$. Если $u = F(x)$ означает, например, плотность вещества в начальный момент времени, то получается, что уравнение описывает перенос вещества в направлении определяемом знаком коэффициента a. При этом, если известно начальное распределение плотности на каком-то отрезке, например, в определенном участке канала, то для определения распределение вещества на этом участке в последующие моменты времени, надо знать сколько вещества втекает и не надо задавать условие на границе, где вещество вытекает. Таким образом, коэффициент a определяет правило постановки граничных условий. Сформулируем это правило.

Для определенности, если не оговорено противное, будем всюду в этом параграфе считать, что решение ищется в области $\{(t, x) : 0 \le t \le T,\ 0 \le x \le 1\}$. Тогда, если коэффициент a положительный, граничное условие задается при $x = 0$ и имеет вид

$$u(t, 0) = \mu(t). \tag{8.83}$$

Для отрицательного коэффициента граничное условие ставится при $x = 1$. Если коэффициент переменный, на левой границе принимает положительное, а на правой отрицательное значение, граничные условия задаются на двух границах. И, наконец, если на левой границе коэффициент отрицательный, а на правой положительный, граничное условие не задается вообще. Во всех случаях при $t = 0$ задается начальное условие

$$u(0, x) = g(x). \tag{8.84}$$

Прежде чем переходить к анализу разностных схем, сделаем еще одно замечание относительно свойств решения задачи Коши для однородного уравнения с постоянным коэффициентом. Если в какой-то фиксированный момент времени решение было монотонной функцией переменной x [9], то в любой последующий фиксированный момент времени свойство монотонности сохранится.

В параграфе 8.1 рассматривалась разностная схема (8.6) для задачи Коши (8.5) с положительным коэффициентом a. Было показано, что схема имеет первый порядок аппроксимации и абсолютно неустойчива, то есть неустойчива при любом соотношении шагов сетки. Из физических соображений нетрудно объяснить природу неустойчивости. Предложенная схема соответствует той ситуации, когда в какой-то точке реки хотят определить параметры течения, зная в предыдущие моменты времени характеристики течения не вверх по течению от этой точки, а вниз. Таким образом, пытаются учесть не тот поток, который втекает в точку, где проводится его изучение,

[8] Предполагается, что ось Ox направлена слева направо.

[9] В этом случае говорят, что решение имеет в данный момент времени монотонный профиль.

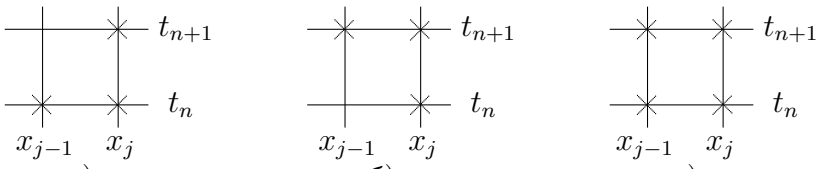


Рис. 8.4 Шаблоны разностных схем для уравнения переноса, $a > 0$.

а тот, который уже давно вытек или никогда в этой точке не был и не будет. Говорят в этом случае, что использовались данные по потоку, а не против потока.

Понятно теперь, что для получения устойчивой схемы, надо изменить ее шаблон таким образом, чтобы при вычислении использовались значения решения против потока. Следовательно, для коэффициента $a > 0$ шаблон должен быть таким, как он представлен на рисунке 8.4.

Итак, явная схема для задачи (8.80),(8.83), (8.84) имеет вид

$$\begin{gathered} \frac{y_j^{n+1} - y_j^n}{\tau} + a\frac{y_j^n - y_{j-1}^n}{h} = f(t_n, x_j), \\ x_j = jh, \quad j = 1, \ldots, J, \quad Jh = 1, \quad t_n = n\tau, \quad n = 0, 1, \ldots, N-1, \ N\tau = T, \\ y_j^0 = g(x_j), \quad j = 0, \ldots, J, \ y_0^{n+1} = \mu(t_{n+1}). \end{gathered} \tag{8.85}$$

Ее шаблон изображен на рисунке 8.4а). Анализируя шаблон, получаем, что решение на $n+1$-ом слое выражается через значения решения на n-ом слое. Отсюда следует алгоритм получения решения. Полагаем $n = 0$ и из граничного условия находим y_0^1. После этого из разностного уравнения последовательно находим y_j^{n+1} для $j = 1, \ldots, J$, используя значения решения, известные из начального условия. Таким образом, двигаясь слева направо, находится решение на всем первом слое. После этого увеличиваем n на единицу и повторяем процесс нахождения решения на втором слое и так далее.

Так же как и в параграфе 8.1 проверяется, что разностная схема аппроксимирует задачу (8.80),(8.83), (8.84) с первым порядком по τ и h на дважды непрерывно дифференцируемом решении. При этом легко заметить, что если в разностной схеме $f(t_n, x_j), g(x_j), \mu(t_{n+1})$ заменить на величины, отличающиеся от соответствующих значения на $O(\tau + h)$, то порядок аппроксимации не изменится. Поэтому с точки зрения аппроксимации ничего не изменится, если в разностной схеме, например, правую часть уравнения вычислять в точке (t_{n+1}, x_j).

Неявная схема, шаблон которой изображен на рисунке 8.4б) имеет вид

$$\begin{gathered} \frac{y_j^{n+1} - y_j^n}{\tau} + a\frac{y_j^{n+1} - y_{j-1}^{n+1}}{h} = f(t_n, x_j), \\ x_j = jh, \quad j = 1, \ldots, J, \quad Jh = 1, \quad t_n = n\tau, \quad n = 0, 1, \ldots, N-1, \ N\tau = T, \\ y_j^0 = g(x_j), \quad j = 0, \ldots, J, \ y_0^{n+1} = \mu(t_{n+1}). \end{gathered} \tag{8.86}$$

Она также имеет первый порядок аппроксимации и процесс организации вычислений такой же как и для явной схемы. Однако эта схема не применима, когда ставится задача Коши, то есть нет граничного условия.

Сравним теперь обе схемы с точки зрения устойчивости. Применим сначала к схеме (8.85) спектральный критерий. Имеем

$$\frac{\lambda e^{ij\varphi} - e^{ij\varphi}}{\tau} + a\frac{e^{ij\varphi} - e^{i(j-1)\varphi}}{h} = 0.$$

Выражая отсюда λ и вводя обозначение $r = a\tau/h > 0$, имеем

$$\lambda = 1 - r(1 - e^{-i\varphi}) = (1 - r + r\cos\varphi) - ir\sin\varphi.$$

Отсюда в соответствии со спектральным критерием

$$|\lambda|^2 = (1 - r + r\cos\varphi)^2 + r^2\sin^2\varphi = 1 - 2r + 2r^2 + 2r(1 - r)\cos\varphi \le 1$$

или $(1 - r)(\cos\varphi - 1) \le 0$. Это неравенство справедливо при любом φ, если $r \le 1$. Таким образом, в соответствии со спектральным критерием, для того, чтобы схема (8.85) была устойчива, необходимо выполнение условия

$$\frac{a\tau}{h} \le 1. \tag{8.87}$$

Это условие принято называть **условием Куранта**.

Для того, чтобы показать, что условие Куранта является достаточным для устойчивости, воспользуемся принципом максимума. Перепишем разностное уравнение (8.85) в виде

$$y_j^{n+1} = (1 - r)y_j^n + ry_{j-1}^n + \tau f(t_n, x_j).$$

Используя обозначения теоремы, имеем $a_0 = 1,\ b_0 = 1 - r,\ b_{-1} = r$, а остальные коэффициенты равны нулю при $j > 0$. Если же $j = 0$, то все коэффициенты равны нулю кроме коэффициента a_0, который равен 1. Заметим, что при выполнении условия Куранта все коэффициенты неотрицательны. Поэтому легко проверить, что выполнены условия принципа максимума.

Для неявной схемы (8.86) разностное уравнение записывается в виде

$$(1 + r)y_j^{n+1} - ry_{j-1}^{n+1} = y_j^n + \tau f(t_n, x_j).$$

Значит $a_0 = 1 + r,\ a_{-1} = -r,\ b_0 = 1$ при $j > 0$. Если же $j = 0$,то $a_0 = 1$. Остальные коэффициенты равны нуля. Очевидно, что требования принципа максимума выполняются всегда.

Подведем теперь итоги на основании результатов, полученных для схем (8.85), (8.86). Из теоремы сходимости следует, что если решение краевой задачи (8.80),(8.83), (8.84) имеет непрерывные вторые производные, то

- при выполнении условия Куранта решение явной схемы (8.85) сходится со скоростью $O(\tau+h)$ по норме пространства C, то есть $\max\limits_j |u(t_n, x_j) - y_j^n| = O(\tau+h)$;
- неявная схема (8.86) безусловно сходится со скоростью $O(\tau + h)$ по норме пространства C.

Неявная схема считается более предпочтительной, чем явная, так как трудоемкость расчетов по обеим схемам одинакова, но при использовании неявной схемы надо заботиться в выполнении условия Куранта. Обе схемы, однако, имеют невысокую точность. Постараемся построить схему более высокого порядка точности. Рассмотрим схему

$$\begin{gathered}
\frac{y_j^{n+1} - y_j^n + y_{j-1}^{n+1} - y_{j-1}^n}{2\tau} + a\frac{y_j^{n+1} - y_{j-1}^{n+1} + y_j^n - y_{j-1}^n}{2h} = f(t_{n+1/2}, x_{j-1/2}), \\
x_j = jh,\quad j = 1, \dots, J,\quad Jh = 1,\quad x_{j-1/2} = x_j - h/2, \\
t_n = n\tau,\quad n = 0, 1, \dots, N-1,\ N\tau = T,\quad t_{n+1/2} = t_n + \tau/2, \\
y_j^0 = g(x_j),\quad j = 0, \dots, J,\ y_0^{n+1} = \mu(t_{n+1}).
\end{gathered} \tag{8.88}$$

Это тоже неявная схема. Ее шаблон изображен на рисунке 8.4в) и является симметричным относительно точки $(t_{n+1/2}, x_{j-1/2})$. Покажем, что если решение дифференциальной задачи трижды непрерывно дифференцируемо, то разностная схема аппроксимирует со вторым порядком. Начальное и граничное условия в разностной схеме заданы точно. Следует проверить только аппроксимацию для разностного уравнения, то есть оценить величину $\psi_h^{(1)}$ (см. обозначения из параграфа 8.1). Для этого воспользуемся формулой Тейлора в окрестности точки $(t_{n+1/2}, x_{j-1/2})$. Для краткости записи у этой точки будем опускать индексы, а у функций, вычисленных в этой точке, будем опускать аргумент. Имеем

$$\begin{aligned}
\psi_h^{(1)} = \frac{1}{2\tau}\Big(u(t+\tau/2,x+h/2) - u(t-\tau/2,x+h/2) + \\
+u(t+\tau/2,x-h/2) - u(t-\tau/2,x-h/2)\Big) + \\
+\frac{a}{2h}\Big(u(t+\tau/2,x+h/2) - u(t+\tau/2,x-h/2) + \\
+u(t-\tau/2,x+h/2) - u(t-\tau/2,x-h/2)\Big) - f = \\
= \frac{1}{2\tau}\Bigg[\Big(u + \frac{\tau}{2}u_t + \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} + \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big) - \\
-\Big(u - \frac{\tau}{2}u_t + \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} - \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big) + \\
+\Big(u + \frac{\tau}{2}u_t - \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} - \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big) - \\
-\Big(u - \frac{\tau}{2}u_t - \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} + \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big)\Bigg] + \\
+\frac{a}{2h}\Bigg[\Big(u + \frac{\tau}{2}u_t + \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} + \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big) - \\
-\Big(u + \frac{\tau}{2}u_t - \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} - \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big) + \\
+\Big(u - \frac{\tau}{2}u_t + \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} - \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big) - \\
-\Big(u - \frac{\tau}{2}u_t - \frac{h}{2}u_x + \frac{\tau^2}{8}u_{tt} + \frac{\tau h}{4}u_{tx} + \frac{h^2}{8}u_{xx} + O(\tau^3+\tau^2 h+\tau h^2+h^3)\Big)\Bigg] - f = \\
= u_t + a u_x - f + O(\tau^2 + \tau h + h^2 + \frac{h^3}{\tau} + \frac{\tau^3}{h}) = O(\tau^2 + \tau h + h^2 + \frac{h^3}{\tau} + \frac{\tau^3}{h}).
\end{aligned} \tag{8.89}$$

Итак, если $\tau = O(h)$, то $\psi_h^{(1)} = O(h^2)$. Таким образом, схема имеет второй порядок аппроксимации.

Замечание 1. В выражениях, которые обозначены $O(\tau^3+\tau^2 h+\tau h^2+h^3)$ в формуле (8.89) коэффициентами при степенях τ и h стоит третьи производные решения u. Поэтому, если u является полиномом второй степени, погрешность аппроксимации $\psi_h^{(1)} = 0$. Значит, в этом случае решение u дифференциальной задачи совпадает с решением y разностной схемы.

Замечание 2. Второй порядок аппроксимации получился благодаря тому, что в силу симметричного расположения точек слагаемые со вторыми производными имели разные знаки и сократились. Такая ситуация является довольно типичной для симметричных разностных схем.

Легко заметить, что порядок расчетов по разностной схеме (8.88) такой же, как и для схем (8.85),(8.86).

Для исследования устойчивости не удается применить принцип максимума. Поэтому ограничимся спектральным критерием. Положим в разностном уравнении $f = 0,\ y_j^{n+1} = \lambda e^{ij\varphi},\ y_j^n = e^{ij\varphi}$. Тогда, после сокращения на $e^{ij\varphi}$, получим

$$\frac{(\lambda-1)+(\lambda-1)e^{-i\varphi}}{2\tau} + a\frac{\lambda(1-e^{-i\varphi})+(1-e^{-i\varphi})}{2h} = 0.$$

Пусть, как и ранее, $r = a\tau/h$. Тогда, выражая λ, имеем

$$\lambda = \frac{(1-r)+e^{-i\varphi}(1+r))}{(1+r)+e^{-i\varphi}(1-r))} = e^{-i\varphi}\frac{(1+r)+e^{i\varphi}(1-r))}{(1+r)+e^{-i\varphi}(1-r))}.$$

Отсюда следует, что $|\lambda| = 1$, так как числитель и знаменатель дроби являются комплексно сопряженными числами, значит, их модули совпадают, а $|e^{-i\varphi}| = 1$. Таким образом, условие спектрального критерия всегда выполняется.

При нахождении гладких решений разностная схема (8.88) имеет преимущества перед схемами (8.85), (8.86), так как позволяет получить более точное приближение к решению при незначительно большем объеме вычислений. Однако, как будет сейчас показано, при расчете разрывных решений или решений с большими градиентами, но на сетке с относительно крупными шагами, преимущества схемы исчезают.

Поясним суть дела. В начале параграфа отмечалось, что если в какой-то момент времени решение однородного уравнения переноса является монотонной функцией, то и во все другие моменты времени оно монотонно. Естественно потребовать, чтобы это свойство сохранялось и для решений разностных схем.

Найдем с помощью изученных выше разностных схем решение однородного уравнения (8.80), которое удовлетворяет граничному условию $u(t,0) = 1$ и монотонному начальному условию $u(0,x) = 1$ при $x < 0.5$, $u(0,x) = 0$ при $x \geq 0.5$. В момент времени $t = \tau$ точное решение этой задачи имеет вид

$$u(\tau,x) = \begin{cases} 1, & x < 0.5 + a\tau, \\ 0, & x \geq 0.5 + a\tau. \end{cases}$$

Вычисленные значения y_j^1 для различных значений $r = a\tau/h$ при $\tau = h = 0.1$ приведены в таблице

Схема	r	x_j									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(8.85)	0.25	1	1	1	1	0.250	0.000	0.000	0.000	0.000	0.000
	0.50	1	1	1	1	0.500	0.000	0.000	0.000	0.000	0.000
	1.00	1	1	1	1	1.000	0.000	0.000	0.000	0.000	0.000
(8.86)	0.25	1	1	1	1	0.200	0.040	0.008	0.002	0.000	0.000
	0.50	1	1	1	1	0.333	0.111	0.012	0.004	0.001	0.000
	1.00	1	1	1	1	0.500	0.250	0.125	0.065	0.031	0.016
	2.00	1	1	1	1	0.667	0.444	0.296	0.198	0.0132	0.088
	4.00	1	1	1	1	0.800	0.640	0.512	0.410	0.328	0.262
(8.88)	0.25	1	1	1	1	0.400	-0.240	0.144	-0.086	0.052	-0.031
	0.50	1	1	1	1	0.667	-0.212	0.074	-0.025	0.008	-0.003
	1.00	1	1	1	1	1.000	0.000	0.000	0.000	0.000	0.000
	2.00	1	1	1	1	1.333	0.444	0.148	0.049	0.016	0.005
	4.00	1	1	1	1	1.600	0.960	0.576	0.346	0.207	0.124

В таблице отсутствуют результаты расчетов по схеме (8.85) для r равного 2 и 4, так как при этих значениях схема неустойчива. Схемы (8.85) и (8.86) сохраняют решение монотонным, причем схема (8.86) "размазывает" скачек решения. В решении же полученном по схеме (8.88) возникают колебания, при этом решение принимает значения большие 1 или отрицательные значения, которые лишены физического смысла. Действительно, если $u(t,x)$ — концентрация вещества, движение которого описывается уравнением переноса, то $u(t,x) \in [0,1]$.

Для изучения возникающей ситуации рассмотрим некоторые свойства, касающиеся произвольных разностных схем, то есть разностных схем не обязательно связанных с уравнением переноса.

Определение 8.4.1 *Однородная разностная схема называется* **монотонной**, *если она монотонный профиль переводит в монотонный.*

Теорема 8.4.1 *Для того, чтобы двухслойная линейная разностная схема*

$$y_j^{n+1} = \sum_k \alpha_k y_{j+k}^n \tag{8.90}$$

была монотонной, необходимо и достаточно, чтобы при всех k выполнялись неравенства $\alpha_k \geq 0$.

Доказательство. Докажем сначала достаточность. Пусть $\alpha_k \geq 0$. Тогда

$$y_j^{n+1} - y_{j-1}^{n+1} = \sum_k \alpha_k y_{j+k}^n - \sum_k \alpha_k y_{j-1+k}^n = \sum_k \alpha_k (y_{j+k}^n - y_{j-1+k}^n). \tag{8.91}$$

Если профиль y_j^n монотонный, например, неубывающий, то выражение, стоящее в правой части равенства (8.91) неотрицательно. Значит, $y_j^{n+1} \geq y_{j-1}^{n+1}$, а это означает, что y_j^{n+1} неубывающая сеточная функция переменной j.

Перейдем теперь к доказательству необходимости. Предположим, что схема монотонна, однако среди ее коэффициентов есть хотя бы один отрицательный, например, α_s. Выберем монотонно возрастающий начальный профиль $y_j^0 = 0$ при $j < s$ и $y_j^0 = 1$ при $j \geq s$. В соответствии с формулой (8.91)

$$y_0^1 - y_{-1}^1 = \sum_k \alpha_k (y_k^0 - y_{k-1}^0) = \alpha_s < 0.$$

Таким образом, y_j^1 не является монотонно возрастающей сеточной функцией, что противоречит монотонности схемы. Теорема доказана.

При выполнении условия Куранта явная однородная разностная схема (8.85) монотонна. Для доказательства достаточно переписать ее в виде (8.90) и заметить, что при выполнении условия Куранта, получатся положительные коэффициенты. Согласно теореме это означает монотонность схемы.

Если двухслойная разностная схема неявная, то ее можно привести к виду (8.90) с бесконечными пределами суммирования, после чего воспользоваться теоремой для проверки монотонности. Рассмотрим однородную неявную схему (8.86). Перепишем ее в виде

$$y_j^{n+1} = \frac{1}{1+r}(r y_{j-1}^{n+1} + y_j^n). \tag{8.92}$$

Здесь, как и ранее $r = a\tau/h$. Уменьшим индекс j на единицу, выразим y_{j-1}^{n+1} через y_{j-2}^{n+1} и подставим это выражение в правую часть (8.92). Продолжая аналогичную процедуру, получим

$$y_j^{n+1} = \frac{1}{1+r} \sum_{k=0}^{\infty} \left(\frac{r}{1+r} \right)^k y_{j-k}^n.$$

Все коэффициенты здесь положительны, поэтому однородная схема (8.86) монотонна при любых τ и h.

Теорема 8.4.2 *Двухслойная линейная монотонная схема для однородного уравнения переноса (8.80) может иметь только первый порядок аппроксимации.*

Доказательство. Предположим противное, то есть что существует линейная монотонная схема, порядок аппроксимации которой равен двум или более. Запишем схему в виде (8.90). По предыдущей теореме все коэффициенты $\alpha_k \geq 0$.

Возьмем в качестве начальных данных функцию $u(0,x) = (x/h - 1/2)^2 - 1/4$. Решение задачи Коши имеет вид

$$u(t,x) = \Big(\frac{x - at}{h} - \frac{1}{2} \Big)^2 - \frac{1}{4},$$

то есть является полиномом второй степени. Как отмечалось в замечании 1 текущего параграфа, погрешность аппроксимации в этом случае $\psi_h^{(1)} = 0$. Значит, решение u дифференциальной задачи совпадает с решением y разностной схемы.

На равномерной сетке с шагом h по переменной x имеем тогда

$$y_j^0 = \Big(j - \frac{1}{2} \Big)^2 - \frac{1}{4} \geq 0, \quad y_j^1 = \Big(j - \frac{a\tau}{h} - \frac{1}{2} \Big)^2 - \frac{1}{4}.$$

Но эти сеточные функции должны удовлетворять уравнению (8.90). Подставляя эти функции в (8.90), получим

$$\Big(j - \frac{a\tau}{h} - \frac{1}{2} \Big)^2 - \frac{1}{4} = \sum_k \alpha_k \left(\Big(j + k - \frac{1}{2} \Big)^2 - \frac{1}{4} \right). \tag{8.93}$$

Выражение, стоящее справа, неотрицательно при любом j. Если же подобрать шаги сетки так, что $a\tau/h$ не является целым или полуцелым числом, то при значении j таком, что $|j - a\tau/h - 1/2| < 1/2$, выражение, стоящее в левой части равенства (8.93) будет отрицательным. Полученное противоречие доказывает теорему.

Из теоремы следует, что схема (8.88) не является монотонной. Это свойство и сказывается при нахождении разрывных или быстро меняющихся решений. Если же шаг сетки мал и решение дифференциальной задачи достаточно гладкое то расчет по немонотонным схемам не нарушает монотонности и тогда они являются предпочтительнее схем первого порядка точности.

До сих пор рассматривался случай, когда коэффициент a уравнения (8.80) был постоянным и положительным. В том случае, когда коэффициент $a < 0$, перенос вещества происходит справа налево. Поэтому во всех схемах шаблон следует зеркально преобразовать в другую сторону. Для этого в разностных схемах производятся очевидные изменения, связанные с заменой разностной производной $(y_j - y_{j-1})/h$ на производную $(y_{j+1} - y_j)/h$.

Перейдем теперь к рассмотрению случая переменного коэффициента a. Ограничимся рассмотрением неявной схемы вида (8.86). В том случае, когда коэффициент

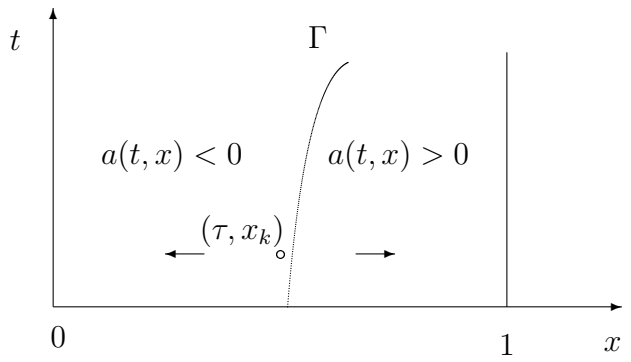


Рис. 8.5 Область, в которой ищется решение уравнения 8.80

не меняет знак, никаких проблем не возникает. Рассмотрим случай, когда коэффициент принимает значения разных знаков, причем $a(t,0) < 0,\ a(t,1) > 0$. Граничное условие тогда не ставится ни на одной из границ.

Пусть, например, выполняется случай, изображенный на рисунке 8.5, где Γ — кривая, вдоль которой коэффициент $a(t,x) = 0$. Для нахождения решения на первом слое находим точку сетки, ближайшую к кривой Γ. На рисунке это точка (τ, x_k). Находим в этой точке решение, используя явное разностное уравнение

$$\frac{y_k^{n+1} - y_k^n}{\tau} + a(t_n, x_k)\frac{y_{k+1}^n - y_k^n}{h} = f(t_n, x_k), \quad n = 0.$$

Так как точка x_k расположена вблизи кривой Γ, на которой коэффициент $a(t,x) = 0$, условие Куранта в этой точке выполнено. Далее используются неявные разностные схемы:

$$\frac{y_j^{n+1} - y_j^n}{\tau} + a\frac{y_j^{n+1} - y_{j-1}^{n+1}}{h} = f(t_n, x_j),$$

где $j = k+1, \ldots, J$;

$$\frac{y_j^{n+1} - y_j^n}{\tau} + a\frac{y_{j+1}^{n+1} - y_j^{n+1}}{h} = f(t_n, x_j)$$

для $j = k-1, \ldots, 0$. Таким образом, в той части области, где коэффициент положительный, вычисления проводятся от точки x_k слева направо, а там, где коэффициент отрицательный — справа налево. На рисунке 8.5 направления, в которых ведутся вычисления, указаны стрелками. Затем, по рассмотренному алгоритму вычисления проводятся для второго слоя и так далее.

Рассмотрим теперь вкратце многомерный случай. Для простоты будем исследовать задачу с двумя пространственными переменными и постоянными коэффициентами:

$$\frac{\partial u}{\partial t} + a_1\frac{\partial u}{\partial x^{(1)}} + a_2\frac{\partial u}{\partial x^{(2)}} = f(t, x^{(1)}, x^{(2)}), \quad x^{(i)} \in [0, A_i],\ i = 1,2, \quad t \in [0,T]. \tag{8.94}$$

Предположим для определенности, что $a_1 > 0,\ a_2 < 0$. Тогда граничное условие следует задать при $x^{(1)} = 0$ и при $x^{(2)} = A_2$:

$$u(t, 0, x^{(2)}) = \mu_1(t, x^{(2)}), \quad u(t, x^{(1)}, A_2) = \mu_2(t, x^{(1)}). \tag{8.95}$$

Кроме того, следует задать начальное условие:

$$u(0, x^{(1)}, x^{(2)}) = g(x^{(1)}, x^{(2)}). \tag{8.96}$$

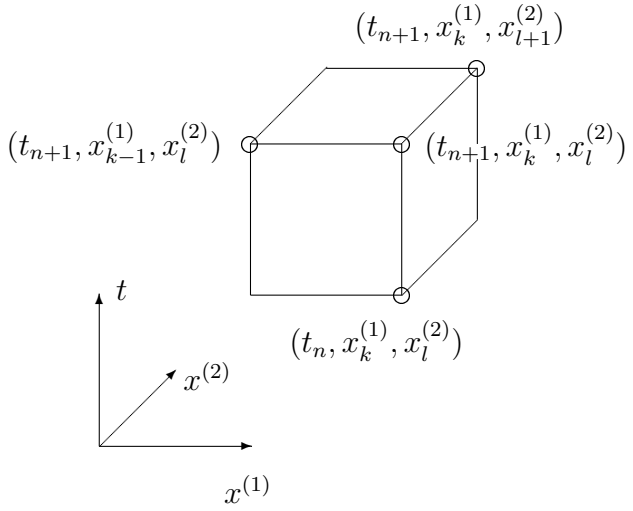


Рис. 8.6 Шаблон разностной схемы 8.97

Введем сетку по каждой переменной

$$x_k^{(1)} = kh_1,\ k = 0, \dots, K,\ Kh_1 = A_1, \qquad x_l^{(2)} = lh_2,\ l = 0, \dots, L,\ Lh_2 = A_2,$$
$$t_n = n\tau,\ n = 0, \dots, N,\ N\tau = T.$$

Значение решения разностной схемы в точке $(t_n, x_k^{(1)}, x_l^{(2)})$ обозначим y_{kl}^n. Построим аналог одной из схем, изученных для случая одной пространственной переменной, например, аналог схемы (8.86). Учитывая направление "потоков" по каждой из переменных, определяем, что шаблон разностной схемы должен выглядеть так, как показано на рисунке 8.6.

Составим по этому шаблону схему

$$\begin{gathered}\frac{y_{kl}^{n+1} - y_{kl}^n}{\tau} + a_1 \frac{y_{kl}^{n+1} - y_{k-1\,l}^{n+1}}{h_1} + a_2 \frac{y_{k\,l+1}^{n+1} - y_{kl}^{n+1}}{h_2} = f(t_{n+1}, x_k^{(1)}, x_l^{(2)}), \\ k = 1, \dots, K, \quad l = 0, \dots, L-1, \quad n = 0, \dots, N-1, \\ y_{0\,l}^{n+1} = \mu_1(t_{n+1}, x_l^{(2)}), \quad y_{k\,L}^{n+1} = \mu_2(t_{n+1}, x_k^{(1)}), \quad y_{k\,l}^0 = g(x_k^{(1)}, x_l^{(2)}).\end{gathered} \tag{8.97}$$

Погрешность аппроксимации разностного уравнения находится, как и в случае одной пространственной переменной с помощью формулы Тейлора и имеет порядок $O(\tau + h_1 + h_2)$ на дважды непрерывно дифференцируемом решении. Остальные условия аппроксимируются точно. Из принципа максимума следует безусловная устойчивость схемы.

Остановимся на процессе вычислений по схеме (8.97). Из разностного уравнения выражается y_{kl}^{n+1} через значения в других точках. Вычисляются значения решения при $t = 0$ и на границах. Если решение на слое t_n найдено, то на следующем слое вычисления производятся с помощью двух вложенных циклов, внутренний, например, по k от 1 до K, внешний по l от $L-1$ до 0.

Следует отметить, что для уравнения переноса многомерность не приводит к принципиальным усложнениям. Неявная схема оказывается такой же простой в реализации как и явная схема. Как увидим в дальнейшем этот факт не типичен и является скорее исключением из правил.

8.4.2 Разностные схемы для волнового уравнения

Другим типичным представителем гиперболических уравнений является уравнение колебания струны (8.82). Выше отмечалось, что его решение можно свести к нахождению решения уравнения переноса. Здесь будет проведено непосредственное исследование разностных схем для этого уравнения, не опирающееся на уравнение переноса. Рассмотрим уравнение колебания струны в области $x \in [0, l],\ t \in [0, T]$. Зададим граничные условия

$$u(t, 0) = \mu_1(t), \quad u(t, l) = \mu_2(t) \tag{8.98}$$

и начальные данные

$$u(0, x) = f(x), \quad \frac{\partial u(0, x)}{\partial t} = g(x). \tag{8.99}$$

Выберем равномерную сетку

$$(t_n, x_k),\ t_n = n\tau,\ n = 0, \ldots, N,\ N\tau = T, \quad x_k = kh,\ k = 0, \ldots, K,\ Kh = l.$$

Поскольку для аппроксимации второй производной требуется использовать значение функции как минимум в трех точках, минимальный шаблон для аппроксимации уравнений колебания струны содержит пять точек. Рассмотрим следующую аппроксимацию дифференциального уравнения:

$$\frac{y_k^{n+1} - 2y_k^n + y_k^{n-1}}{\tau^2} = a^2 \frac{y_{k+1}^n - 2y_k^n + y_{k-1}^n}{h^2}, \quad k = 1, \ldots, K,\ n = 1, \ldots, N-1. \tag{8.100}$$

Погрешность аппроксимации разностного уравнения

$$\begin{aligned}
\psi_h^{(1)} &= \frac{u(t+\tau, x) - 2u(t.x) + u(t-\tau, x)}{\tau^2} - a^2 \frac{u(t, x+h) - 2u(t.x) + u(t, x-h)}{h^2} = \\
&= \frac{u + u_t\tau + u_{tt}\frac{\tau^2}{2} + u_{ttt}\frac{\tau^3}{6} + O(\tau^4) - 2u + u - u_t\tau + u_{tt}\frac{\tau^2}{2} - u_{ttt}\frac{\tau^3}{6} + O(\tau^4)}{\tau^2} + \\
&+ a^2 \frac{u + u_x h + u_{xx}\frac{h^2}{2} + u_{xxx}\frac{h^3}{6} + O(h^4) - 2u + u - u_x h + u_{xx}\frac{h^2}{2} - u_{xxx}\frac{h^3}{6} + O(h^4)}{h^2} = \\
&= u_{tt} - a^2 u_{xx} + O(\tau^2 + h^2) = O(\tau^2 + h^2).
\end{aligned}$$

Таким образом, разностное уравнение аппроксимирует дифференциальное уравнение со вторым порядком по каждой из переменных.

Граничные условия аппроксимируем точно:

$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_K^{n+1} = \mu_2(t_{n+1}). \tag{8.101}$$

Не возникает проблем с аппроксимацией первого начального условия, которое также аппроксимируется точно:

$$y_k^0 = f(x_k). \tag{8.102}$$

Для аппроксимации второго начального условия воспользуемся одним из методов, описанных в пункте 8.2.4. По формуле Тейлора

$$\frac{u(\tau, x) - u(0, x)}{\tau} = \frac{\partial u(0, x)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(0, x)}{\partial t^2} + O(\tau^2). \tag{8.103}$$

В выражении, стоящем в правой части этого равенства значение первой производной известно из второго начального условия. Для нахождения значения второго слагаемого воспользуемся дифференциальным уравнением и первым начальным условием. В результате получим

$$\frac{\partial^2 u(0,x)}{\partial t^2} = a^2 \frac{\partial^2 u(0,x)}{\partial x^2} = a^2 f''(x).$$

Подставляя это выражение в (8.103), имеем

$$\frac{u(\tau,x) - u(0,x)}{\tau} = g(x) + \frac{a^2\tau}{2} f''(x) + O(\tau^2).$$

Отбрасывая члены порядка $O(\tau^2)$, получим аппроксимацию второго начального условия со вторым порядком.

$$\frac{y_k^1 - y_k^0}{\tau} = g(x_k) + \frac{a^2\tau}{2} f''(x_k). \tag{8.104}$$

Таким образом, разностная схема (8.100)-(8.102), (8.104) аппроксимирует задачу (8.82), (8.98), (8.99) со вторым порядком.

Для исследования устойчивости воспользуемся спектральным критерием. Подставляя в разностное уравнение $\lambda^p e^{ik\varphi}$ вместо y_k^{n+p}, после сокращения на $e^{ik\varphi}$ получим

$$\frac{\lambda - 2 + 1/\lambda}{\tau^2} = a^2 \frac{e^{i\varphi} - 2 + e^{-i\varphi}}{h^2}.$$

Отсюда, учитывая, что

$$e^{i\varphi} - 2 + e^{-i\varphi} = 2\cos\varphi - 2 = -4\sin^2\frac{\varphi}{2},$$

имеем квадратное уравнение для нахождения λ:

$$\lambda^2 - 2\Big(1 - 2r^2\sin^2\frac{\varphi}{2}\Big) + 1 = 0, \quad r = \frac{a\tau}{h}.$$

Согласно теореме Виета произведение корней этого уравнения равно 1. Тогда, если корни действительны и различны, один из корней меньше, а второй больше 1. Поэтому, в этом случае схема неустойчива. Если корни равны, то их модули равны 1. Если же корни комплексные, то они комплексно сопряженные и их модули равны и, следовательно, модули равны 1. То есть в случае комплексных корней спектральное условие выполняется. Для того, чтобы корни были комплексными или равными, дискриминант должен быть не положительным, то есть

$$\Big(1 - 2r^2\sin^2\frac{\varphi}{2}\Big)^2 - 1 \le 0.$$

Отсюда следует, что

$$-1 \le 1 - 2r^2\sin^2\frac{\varphi}{2} \le 1.$$

Правая часть этого неравенства справедлива всегда, для выполнения левой части, должно выполняться неравенство

$$r^2\sin^2\frac{\varphi}{2} \le 1.$$

Так как неравенство должно быть справедливым для всех значений φ, отсюда следует, что необходимо, чтобы $r = a\tau/h \leq 1$. Таким образом, при выборе шагов разностной схемы необходимо учитывать условие Куранта.

Организация процесса вычислений по этой схеме не вызывает труда. Сначала из соотношений (8.102),(8.104) находятся значения решения разностной схемы на первых двух временных слоях, после чего из разностных граничных условий и разностного уравнения последовательно определяется решение на втором, третьем слоях и так далее.

8.5 РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

В предыдущих параграфах уравнение теплопроводности встречалось несколько раз в качестве примеров, иллюстрирующих применение различных методов. В этом параграфе будут рассматривается обобщение ранее встречавшихся схем, в частности, так называемое, однопараметрическое семейство схем с весам, уравнения с переменными и нелинейными коэффициентами, задачи с граничными условиями третьего рода.

8.5.1 Схема с весами для уравнения теплопроводности

Рассматрим первую краевую задачу для уравнения теплопроводности

$$\begin{gathered}\frac{\partial u(t,x)}{\partial t} = a^2\frac{\partial^2 u(t,x)}{\partial x^2} + f(t,x), \quad 0 < x < l,\ 0 < t \leq T, \\ u(0,x) = u_0(x), \quad u(t,0) = \mu_1(t), \quad u(t,l) = \mu_2(t).\end{gathered} \tag{8.105}$$

Будем считать, что $a = const$ и что решение этой задачи имеет столько производных, сколько необходимо при последующих рассуждениях.

Рассмотрим следующее однопараметрическое семейство разностных схем:

$$\begin{gathered}\frac{y_k^{n+1} - y_k^n}{\tau} = a^2\left(\sigma\frac{y_{k+1}^{n+1} - 2y_k^{n+1} + y_{k-1}^{n+1}}{h^2} + (1-\sigma)\frac{y_{k+1}^n - 2y_k^n + y_{k-1}^n}{h^2}\right) + F_k^n, \\ x_k = kh, \quad k = 1,2,\ldots,K-1, \quad Kh = l, \quad t_n = n\tau, \quad n = 0,1,\ldots,N-1, \quad N\tau = T, \\ y_k^0 = u_0(x_k), \quad y_0^{n+1} = \mu_1(t_{n+1}), \quad y_K^{n+1} = \mu_2(t_{n+1}).\end{gathered} \tag{8.106}$$

Здесь σ — действительное число, называемое **весом**, а само семейство разностных схем (8.106) называется **схемой с весами.** При $\sigma = 0$ получается **явная разностная схема**, при $\sigma = 1$ — **чисто неявная схема**, при $\sigma = 0.5$ — **симметричная схема** или, как ее еще называют **схема Кранка-Николсон**. Шаблоны этих схем представлены не рисунке 8.7

Начнем с исследования порядка аппроксимации схемы при различных значениях веса σ. Так как начальные и граничные условия задаются для разностной схемы точно, необходимо исследовать только погрешность аппроксимации на решении разностного уравнения. Как обычно, это делается с помощью формулы Тейлора. В качестве точки, в окрестности которой будем производить разложение по формуле

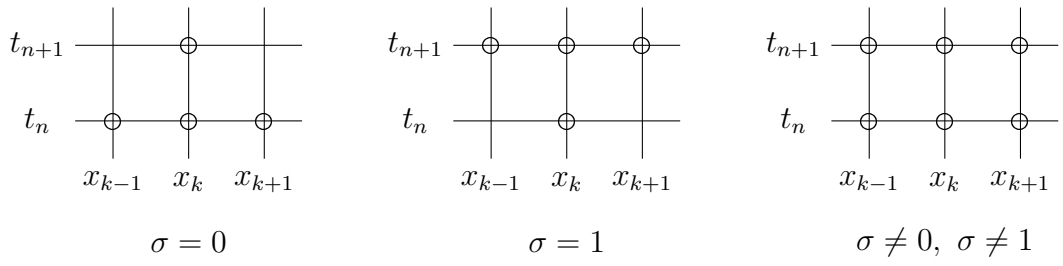


Рис. 8.7 Шаблон схемы с весами при различных значениях σ.

Тейлора, выберем точку $(t_n + 0.5\tau, x_k)$, так как при всех значениях σ кроме 0 и 1 шаблон симметричен относительно этой точки.

Воспользуемся формулой (8.3), в соответствии с которой

$$\frac{u(t, x+h) - 2u(t,x) + u(t, x-h)}{h^2} = \frac{\partial^2 u(t,x)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(t,x)}{\partial x^4} + O(h^4).$$

Кроме того, учтем, что

$$\begin{aligned}\frac{u(t+\tau, x) - u(t,x)}{\tau} = \\ = \frac{1}{\tau}\Bigg(u(t+0.5\tau, x) + \frac{\tau}{2}\frac{\partial u(t+0.5\tau, x)}{\partial t} + \frac{\tau^2}{8}\frac{\partial^2 u(t+0.5\tau, x)}{\partial t^2} + O(\tau^3) - \\ - u(t+0.5\tau, x) + \frac{\tau}{2}\frac{\partial u(t+0.5\tau, x)}{\partial t} - \frac{\tau^2}{8}\frac{\partial^2 u(t+0.5\tau, x)}{\partial t^2} + O(\tau^3)\Bigg) = \\ = \frac{\partial u(t+0.5\tau, x)}{\partial t} + O(\tau^2)\end{aligned}$$

и

$$\begin{aligned}\sigma v(t+\tau, x) + (1-\sigma)v(t,x) = \sigma\left(v(t+0.5\tau, x) + \frac{\tau}{2}\frac{\partial v(t+0.5\tau, x)}{\partial t} + O(\tau^2)\right) + \\ + (1-\sigma)\left(v(t+0.5\tau, x) - \frac{\tau}{2}\frac{\partial v(t+0.5\tau, x)}{\partial t} + O(\tau^2)\right) = \\ = v(t+0.5\tau, x) + \tau\left(\sigma - \frac{1}{2}\right)\frac{\partial v(t+0.5\tau, x)}{\partial t} + O(\tau^2), \quad (8.107)\end{aligned}$$

где $v(t,x)$ — произвольная гладкая функция. Тогда погрешность аппроксимации рав-

на (аргумент $(t_n + 0.5\tau, x_k)$ будем для краткости опускать)

$$\psi_h^{(1)} =$$
$$= \frac{u(t_n+\tau,x_k)-u(t_n,x_k)}{\tau} - a^2\sigma\frac{u(t_n+\tau,x_k+h)-2u(t_n+\tau,x_k)+u(t_n+\tau,x_k-h)}{h^2} -$$
$$-a^2(1-\sigma)\frac{u(t_n,x_k+h)-2u(t_n,x_k)+u(t_n,x_k-h)}{h^2} - F_k^n =$$
$$= \frac{\partial u}{\partial t} + O(\tau^2) - a^2\sigma\left(\frac{\partial^2 u(t_n+\tau,x_k)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(t_n+\tau,x_k)}{\partial x^4} + O(h^4)\right) -$$
$$-a^2(1-\sigma)\left(\frac{\partial^2 u(t_n,x_k)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(t_n,x_k)}{\partial x^4} + O(h^4)\right) - F_k^n =$$
$$= \frac{\partial u}{\partial t} - a^2\frac{\partial^2 u}{\partial x^2} - a^2\tau\Big(\sigma-\frac{1}{2}\Big)\frac{\partial}{\partial t}\frac{\partial^2 u}{\partial x^2} - \frac{a^2h^2}{12}\frac{\partial^4 u}{\partial x^4} - F_k^n + O(\tau^2+\tau h^2+h^4). \tag{8.108}$$

Из дифференциального уравнения (8.105) следует, что

$$\frac{\partial u}{\partial t} - a^2\frac{\partial^2 u}{\partial x^2} = f, \quad a^2\frac{\partial^4 u}{\partial x^4} = \frac{\partial^2}{\partial x^2}\frac{\partial u}{\partial t} - \frac{\partial^2 f}{\partial x^2}.$$

Подставляя эти соотношения в (8.108), получаем

$$\psi_h^{(1)} = \Big(f + \frac{h^2}{12}\frac{\partial^2 f}{\partial x^2} - F_k^n\Big) - a^2\tau\Big(\sigma - \frac{1}{2} + \frac{h^2}{12a^2\tau}\Big)\frac{\partial^3 u}{\partial x^2\partial t} + O(\tau^2+\tau h^2+h^4). \tag{8.109}$$

Формула (8.109) позволяет сделать следующие выводы.

- Если $F_k^n = f(t_n+0.5\tau, x_k)+O(\tau+h^2)$, то разностная схема (8.106) имеет первый порядок аппроксимации по τ и второй по h, то есть погрешность аппроксимации равна $O(\tau + h^2)$.

- Если $\sigma = \dfrac{1}{2}$ и $F_k^n = f(t_n + 0.5\tau, x_k) + O(\tau^2 + h^2)$, то разностная схема (8.106) имеет второй порядок аппроксимации по τ и h.

- Если $\sigma = \sigma^* = \dfrac{1}{2} - \dfrac{h^2}{12a^2\tau}$ и

$$F_k^n = f(t_n + 0.5\tau, x_k) + \frac{h^2}{12}\frac{\partial^2 f(t_n+0.5\tau,x_k)}{\partial x^2} + O(\tau^2+\tau h^2+h^4), \tag{8.110}$$

  то погрешность аппроксимации равна $O(\tau^2 + \tau h^2 + h^4)$. В этом случае схема называется **схемой повышенного порядка аппроксимации**.

Замечание. Для того чтобы не задавать вторую производную функции f в случае расчетов по схеме повышенного порядка аппроксимации, достаточно взять

$$F_k^n = f(t_n+0.5\tau,x_k) + \frac{h^2}{12}\frac{f(t_n+0.5\tau,x_{k+1})-2f(t_n+0.5\tau,x_k)+f(t_n+0.5\tau,x_{k-1})}{h^2} =$$
$$= \frac{f(t_n+0.5\tau,x_{k+1})+10f(t_n+0.5\tau,x_k)+f(t_n+0.5\tau,x_{k-1})}{12}.$$

Из формулы (8.3) следует, что при таком выборе F_k^n условие (8.110) будет выполнено.

Перейдем к исследованию устойчивости. В параграфе 8.3 было показано, что явная схема условно устойчива, причем условием устойчивости служит неравенство $a^2\tau/h^2 \le 1/2$. Чисто неявная схема безусловно устойчива. Поэтому здесь проведем исследование для других значений σ. Постараемся найти те значения σ, при которых схема безусловно устойчива.

Ограничимся нахождением тех значений σ, при которых выполняются условия спектрального критерия. Для этого, как обычно, возьмем в разностном уравнении (8.106) $F_k^n = 0$ и заменим формально y_k^{n+1} на $\lambda e^{ik\varphi}$, а y_k^n на $e^{ik\varphi}$. Вводя обозначение $r = a^2\tau/h^2$ и учитывая, что

$$e^{i(k+1)\varphi} - 2e^{ik\varphi} + e^{i(k-1)\varphi} = -4e^{ik\varphi}\sin^2\varphi/2,$$

получим

$$\lambda - 1 = r\sigma\lambda(-4\sin^2\varphi/2) + r(1-\sigma)(-4\sin^2\varphi/2).$$

Отсюда следует, что

$$\lambda = \frac{1 - 4r(1-\sigma)\sin^2\varphi/2}{1 + 4r\sigma\sin^2\varphi/2}.$$

Необходимо, чтобы при всех значениях φ выполнялось неравенство $|\lambda| \le 1$. Таким образом, имеем

$$-1 - 4r\sigma\sin^2\varphi/2 \le 1 - 4r(1-\sigma)\sin^2\varphi/2 \le 1 + 4r\sigma\sin^2\varphi/2.$$

Правая часть этого неравенства, очевидно, справедлива всегда. Левая часть эквивалентна неравенству $2r\sin^2\varphi/2 - 1 \le 4r\sigma\sin^2\varphi/2$, которое выполняется при любом φ, если

$$\sigma \ge \frac{1}{2} - \frac{1}{4r} = \frac{1}{2} - \frac{h^2}{4a^2\tau}. \tag{8.111}$$

Итак, для выполнения требований спектрального критерия должно выполняться условие (8.111). Легко заметить, что схема повышенного порядка точности, схема Кранка-Николсон удовлетворяют условию (8.111).

Остановимся теперь на вопросе нахождения решения по схеме (8.106). Если $\sigma = 0$, то есть схема явная, процесс вычислений на представляет трудностей. Разностное уравнение переписывается в виде

$$y_k^{n+1} = ry_{k+1}^n + (1-2r)y_k^n + ry_{k-1}^n + \tau F_k^n, \quad k = 1, \ldots, K-1. \tag{8.112}$$

После этого послойно, начиная с первого слоя вычисляется решение. При этом y_k^0 вычисляется из начального условия, а значения решения на каждом слое при $k = 0$ и $k = K$ — из граничных условий. Следует еще раз подчеркнуть, что так как схема условно устойчива, если выбран шаг h, то шаг τ должен удовлетворять условию $\tau \le h^2/(2a^2)$. Например, если $h = 0.1,\ a^2 = 10$, то $\tau \le 0.0005$, то есть очень мал. Поэтому для доведения расчетов до заданного момента времени T может потребоваться сделать много шагов по времени, а, значит, выполнить большой объем вычислений. По этой причине, несмотря на простоту вычислений, явные схемы для нахождения решений параболических уравнений используются редко.

Заметим еще, что при выполнении условия устойчивости для однородного уравнения (8.112) выполняется условие монотонности, подробно рассмотренное в предыдущем параграфе.

Если $\sigma \neq 0$, то есть схема неявная. Перепишем ее в виде

$$\begin{aligned} y_0^{n+1} &= \mu_1(t_{n+1}), \\ -r\sigma y_{k+1}^{n+1} + (1+2r\sigma)y_k^{n+1} + r\sigma y_{k-1}^{n+1}) &= \\ = r(1-\sigma)(y_{k-1}^n + y_{k+1}^n) + (1-2r(1-\sigma))y_k^n + \tau F_k^n, &\quad k = 1, \ldots, K-1, \\ y_K^{n+1} &= \mu_2(t_{n+1}). \end{aligned} \tag{8.113}$$

При каждом фиксированном n (8.113) представляет собой систему линейных алгебраических уравнений относительно неизвестных $y_k^{n+1},\ k = 0, \ldots, K$. Если на n-ом слое решение уже найдено, то правая часть всех уравнений известна. При $n = 0$ решение разностной схемы находится из начального условия. Матрица системы (8.113) трехдиагональная, поэтому система решается методом прогонки. Заметим, что при $\sigma > 0$ выполняется условие диагонального преобладания, поэтому система имеет единственное решение и, кроме того, прогонка устойчива. Итак, начиная с первого слоя, решая на каждом слое систему уравнений (8.113) можно найти решение разностной схемы.

Следует отметить, что процесс решения системы методом прогонки требует $O(K)$ арифметических операций, то есть только некоторым коэффициентом, не зависящим от шагов сетки, трудоемкость нахождения решения на слое по неявной схеме отличается от явной. Обычно для расчетов выбирают неявные безусловно устойчивые схемы. Чаще всего используют симметричную схему ($\sigma = 0.5$) или схему повышенного порядка точности, которые обеспечивают хорошую точность вычислений при не слишком малых шагах τ и h.

Чисто неявная схема в случае постоянного коэффициента a используется не часто из-за невысокого порядка аппроксимации. Однако, можно показать, что как и явная схема, чисто неявная схема является монотонной (при решении однородного уравнения на бесконечной прямой).

Доказывается, что симметричная схема монотонна тогда и только тогда, когда $\tau \leq 1.5h^2/a^2$. Таким образом, монотонность выполняется только при очень малом шаге по времени $\tau = O(h^2)$, в то время как для симметричной схемы есть смысл выбирать $\tau = O(h)$. Достаточно гладкое решение на подобных сетках можно находить и по немонотонным схемам. Однако на грубых сетках, при разрывных начальных данных, например, симметричная схема может привести к "разболтке" счета. Чисто неявная схема даже в этих условиях дает плавно меняющееся разностное решение, хотя его точность и невелика. В том случае, когда коэффициент $a = a(u)$ часто применяется чисто неявная схема благодаря своей монотонности.

8.5.2 Граничные условия третьего рода

Рассмотрим следующую задачу

$$\begin{gathered} \frac{\partial u(t,x)}{\partial t} = a^2 \frac{\partial^2 u(t,x)}{\partial x^2} + f(t,x), \quad 0 < x < l,\ 0 < t \leq T, \\ u(0,x) = u_0(x), \quad \frac{\partial u(t,0)}{\partial x} = \alpha_1 u(t,0) - \mu_1(t), \quad -\frac{\partial u(t,l)}{\partial x} = \alpha_2 u(t,l) - \mu_2(t). \end{gathered} \tag{8.114}$$

Здесь $a,\ \alpha_1\ \alpha_2$ константы.

Покажем, что выбирая разностные условия в виде

$$\sigma\left(\frac{y_1^{n+1}-y_0^{n+1}}{h}-\alpha_1 y_0^{n+1}\right)+(1-\sigma)\left(\frac{y_1^n-y_0^n}{h}-\alpha_1 y_0^n\right)=$$
$$=\frac{h}{2a^2}\frac{y_0^{n+1}-y_0^n}{\tau}-\mu_1(t_n+0.5\tau)-\frac{h}{2a^2}f(t_n+0.5\tau,0), \quad (8.115)$$

$$-\sigma\left(\frac{y_K^{n+1}-y_{K-1}^{n+1}}{h}+\alpha_2 y_K^{n+1}\right)-(1-\sigma)\left(\frac{y_K^n-y_{K-1}^n}{h}+\alpha_2 y_K^n\right)=$$
$$=\frac{h}{2a^2}\frac{y_K^{n+1}-y_K^n}{\tau}-\mu_2(t_n+0.5\tau)-\frac{h}{2a^2}f(t_n+0.5\tau,l), \quad (8.116)$$

получим аппроксимацию граничных условий порядка $O(\tau+h^2)$ при $\sigma\neq 0.5$ и $O(\tau^2+h^2)$ при $\sigma=0.5$ [10].

Для доказательства достаточно выписать погрешность аппроксимации граничного условия $\psi_h^{(2)}$ (см. параграф 8.1) и с помощью формулы Тейлора определить порядок малости этой величины. Поступим по другому — приведем рассуждения, которые приводят к формулам (8.115),(8.116). Так как для граничных условий на левой и правой границах рассуждения аналогичны, рассмотрим только одну из границ, например, левую.

Легко проверить, что для любой гладкой функции w справедливо равенство:

$$\frac{w(h)-w(0)}{h}=w'(0)+\frac{h}{2}w''(0)+O(h^2).$$

Отсюда и из (8.107) имеем

$$\sigma\left(\frac{u(t_n+\tau,h)-u(t_n+\tau,0)}{h}-\alpha_1 u(t_n+\tau,0)\right)+(1-\sigma)\left(\frac{u(t_n,h)-u(t_n,0)}{h}-\alpha_1 u(t_n,0)\right)=$$
$$=\sigma\left(\frac{\partial u(t_n+\tau,0)}{\partial x}-\alpha_1 u(t_n+\tau,0)+\frac{h}{2}\frac{\partial^2 u(t_n+\tau,0)}{\partial x^2}\right)+$$
$$+(1-\sigma)\left(\frac{\partial u(t_n,0)}{\partial x}-\alpha_1 u(t_n,0)+\frac{h}{2}\frac{\partial^2 u(t_n,0)}{\partial x^2}\right)+O(h^2)=$$
$$=\frac{\partial u(t_n+0.5\tau,0)}{\partial x}-\alpha_1 u(t_n+0.5\tau,0)+\frac{h}{2}\frac{\partial^2 u(t_n+0.5\tau,0)}{\partial x^2}+$$
$$+\left(\sigma-\frac{1}{2}\right)\tau\frac{\partial}{\partial t}\left(\frac{\partial u(t_n+0.5\tau,0)}{\partial x}-\alpha_1 u(t_n+0.5\tau,0)+\frac{h}{2}\frac{\partial^2 u(t_n+0.5\tau,0)}{\partial x^2}\right)+O(\tau^2+h^2). \quad (8.117)$$

Из дифференциального уравнения следует, что

$$\frac{\partial^2 u(t_n+0.5\tau,0)}{\partial x^2}=\frac{1}{a^2}\frac{\partial u(t_n+0.5\tau,0)}{\partial t}-\frac{1}{a^2}f(t_n+0.5\tau,0).$$

В свою очередь

$$\frac{\partial u(t_n+0.5\tau,0)}{\partial t}=\frac{u(t_n+\tau,0)-u(t_n,0)}{\tau}+O(\tau^2).$$

[10] Аппроксимация граничных условий для схемы повышенного порядка точности приведена в [31]

Поэтому

$$\frac{\partial^2 u(t_n+0.5\tau,0)}{\partial x^2}=\frac{1}{a^2}\frac{u(t_n+\tau,0)-u(t_n,0)}{\tau}-\frac{1}{a^2}f(t_n+0.5\tau,0).$$

Подставляя это соотношение в (8.117) и учитывая граничное условие, имеем

$$\sigma\left(\frac{u(t_n+\tau,h)-u(t_n+\tau,0)}{h}-\alpha_1 u(t_n+\tau,0)\right)+(1-\sigma)\left(\frac{u(t_n,h)-u(t_n,0)}{h}-\alpha_1 u(t_n,0)\right)=$$
$$=\frac{h}{2a^2}\frac{u(t_n+\tau,0)-u(t_n,0)}{\tau}-\mu_1(t_n+0.5\tau)-\frac{h}{2a^2}f(t_n+0.5\tau,0)+$$
$$+\Big(\sigma-\frac{1}{2}\Big)\tau\frac{\partial}{\partial t}\left(\frac{\partial u(t_n+0.5\tau,0)}{\partial x}-\alpha_1 u(t_n+0.5\tau,0)+\frac{h}{2}\frac{\partial^2 u(t_n+0.5\tau,0)}{\partial x^2}\right)+O(\tau^2+h^2). \tag{8.118}$$

Из равенства (8.118) следует требуемое утверждение.

Как и в случае граничного условия первого рода, при $\sigma\neq 0$ решение разностной схемы находится с помощью прогонки послойно.

Если разностная схема явная, то сначала с помощью разностного уравнения вычисляются $y_1^{n+1},\ y_{K-1}^{n+1}$, а затем, из граничных условий находятся $y_0^{n+1},\ y_K^{n+1}$.

8.5.3 Уравнения с переменными коэффициентами и квазилинейные уравнения

В этом пункте очень кратко остановимся на методах нахождения решения в случае, когда коэффициент $a\neq const$.

Чаще всего, уравнение с переменным коэффициентом встречается записанным в виде

$$\rho(t,x)\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\Big(p(t,x)\frac{\partial u}{\partial x}\Big)+f(t,x),\quad 0<x<l,\quad 0<t\leq T. \tag{8.119}$$

Здесь предполагается, что все заданные функции достаточно гладкие и $0<\rho_0\leq\rho(t,x),\ 0<p_0\leq p(t,x)$. Для простоты рассмотрим первую краевую задачу, то есть будем считать, что дополнительные условия (начальные и граничные) заданы в виде

$$u(0,x)=u_0(x),\quad u(t,0)=\mu_1(t),\quad u(t,l)=\mu_2(t). \tag{8.120}$$

Введем сначала дифференциальный оператор

$$Lu=\frac{\partial}{\partial x}\Big(p(t,x)\frac{\partial u}{\partial x}\Big)$$

и разностный оператор

$$\Lambda(t)z_k=\frac{1}{h}\left(A_{k+1}(t)\frac{z_{k+1}-z_k}{h}-A_k(t)\frac{z_k-z_{k-1}}{h}\right). \tag{8.121}$$

Здесь для $A_k(t)$ используется одно из следующих выражений

$$A_k(t)=p(t,x_k-h/2),\quad A_k(t)=\frac{p(t,x_{k-1})+p(t,x_k)}{2},\quad A_k(t)=\frac{2p(t,x_{k-1})p(t,x_k)}{p(t,x_{k-1})+p(t,x_k)}.$$

Легко проверить, что при таком выборе коэффициента $A_k(t)$ оператор Λ аппроксимирует оператор L со вторым порядком по h на гладких функциях u.

Заметим, что такой выбор оператора Λ и его коэффициентов не случайны. Соответствующие выражения уже встречались при аппроксимации стационарного уравнения теплопроводности в пункте 8.2.3.

Схема с весами для задачи (8.119) (8.120) записывается теперь в виде

$$\rho(t,x_k)\frac{y_k^{n+1}-y_k^n}{\tau} = \sigma\Lambda(t)y_k^{n+1} + (1-\sigma)\Lambda(t)y_k^n + f(t,x_k),$$
$$x_k = kh, \quad k = 1,2,\ldots,K-1, \quad Kh = l, \quad t_n = n\tau, \quad n = 0,1,\ldots,N-1, \quad N\tau = T,$$
$$y_k^0 = u_0(x_k), \quad y_0^{n+1} = \mu_1(t_{n+1}), \quad y_K^{n+1} = \mu_2(t_{n+1}). \tag{8.122}$$

В качестве t можно взять любое значение из промежутка $[t_n, t_{n+1}]$. Если выбрать $t = t_n + \tau/2$, а $\sigma = 1/2$, получится схема, имеющая второй порядок аппроксимации по τ и h. При любых других значениях t и σ порядок аппроксимации первый по τ и второй по h.

Для исследования устойчивости применим принцип замороженных коэффициентов. Считая коэффициенты ρ и p постоянными и полагая $f = 0$ получим уравнение

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2}, \quad a = \frac{p}{\rho}.$$

В параграфе 8.2 уже было проведено исследование устойчивости явной схемы $\sigma = 0$ для этого уравнения. Было получено, что для устойчивости необходимо выполнение условия (8.65). С учетом определения коэффициента a имеем необходимое условие устойчивости явной схемы:

$$\max_{t,x}\frac{p(t,x)}{\rho(t,x)}\frac{\tau}{h^2} \le \frac{1}{2}.$$

Заметим, что если использовать явную разностную схему и брать в ней $t = t_n$, то она допускает счет с переменным шагом по времени τ, который подбирается следующим образом, при условии, что на слое t_n решение уже найдено. Величину следующего шага по времени выбирают исходя из условия

$$\tau_n \le \frac{h^2}{2\max\limits_{t_n,x}\frac{p(t_n,x)}{\rho(t_n,x)}}.$$

При $\sigma \ge 1/2$ из принцип замороженных коэффициентов следует безусловная устойчивость схемы (8.122).

Рассмотрим теперь квазилинейное уравнение

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(p(u)\frac{\partial u}{\partial x}\Big) + f(t,x,u), \quad 0 < x < l, \quad 0 < t \le T, \tag{8.123}$$

для которого зададим еще условия (8.120).

Так как пределы изменения коэффициента $p(u)$ не известны, явную схему стараются не применять. Неявные схемы можно строить по разному. Первая схема, которую рассмотрим, так называемая, **чисто неявная линеаризованная схема**

$$\frac{y_k^{n+1}-y_k^n}{\tau} = \frac{1}{h}\left(A_{k+1}\frac{y_{k+1}^{n+1}-y_k^{n+1}}{h} - A_k\frac{y_k^{n+1}-y_{k-1}^{n+1}}{h}\right) + f(t_n,x_k,y_k^n), \tag{8.124}$$

где $A_k = \dfrac{p(y_k^n) + p(y_{k-1}^n)}{2}$. Термин "линеаризованная" здесь означает, что относительно y_k^{n+1} уравнения являются линейными. Схема безусловно устойчива и обладает первым порядком аппроксимации по τ и вторым по h. Для нахождения решения по этой схеме используется метод прогонки.

Использование **чисто неявной схемы без линеаризации**

$$\frac{y_k^{n+1} - y_k^n}{\tau} = \frac{1}{h}\left(A_{k+1}\frac{y_{k+1}^{n+1} - y_k^{n+1}}{h} - A_k\frac{y_k^{n+1} - y_{k-1}^{n+1}}{h}\right) + f(t_{n+1}, x_k, y_k^{n+1}),$$
$$A_k = \frac{p(y_k^{n+1}) + p(y_{k-1}^{n+1})}{2} \tag{8.125}$$

приводит к нелинейной системе уравнений, которая дополняется граничными условиями. Решать эту систему на каждом слое приходится с помощью итераций, которые можно организовать, например, следующим образом:

$$y_k^{(0)} = y_k^n, \quad k = 0, \ldots, K,$$
$$\frac{y_k^{(j+1)} - y_k^n}{\tau} = \frac{1}{h}\left(A_{k+1}\frac{y_{k+1}^{(j+1)} - y_k^{(j+1)}}{h} - A_k\frac{y_k^{(j+1)} - y_{k-1}^{(j+1)}}{h}\right) + f(t_{n+1}, x_k, y_k^{(j)}),$$
$$k = 1, \ldots, K-1,$$
$$A_k = \frac{p(y_k^{(j)}) + p(y_{k-1}^{(j)})}{2},\ k = 1, \ldots, K, \qquad y_0^{j+1} = \mu_1(t_{n+1}),\ y_K^{j+1} = \mu_2(t_{n+1}),$$
$$j = 0, 1, \ldots, J-1,$$
$$y_k^{n+1} = y_k^{(J)}, \quad k = 0, \ldots, K.$$

Здесь j — номер итераций. Число итераций J определяют из соображений точности. Как правило, число итераций невелико — две - три. Это связано с тем, что начальное приближение y_k^n при малом шаге по времени является достаточно хорошим приближением к y_k^{n+1}. Заметим, что при $J = 1$ получается фактически схема (8.125). Значение $y_k^{(j+1)}$ на каждой итерации находится методом прогонки.

8.6 РАЗНОСТНЫЕ СХЕМЫ ДЛЯ МНОГОМЕРНЫХ НЕСТАЦИОНАРНЫХ УРАВНЕНИЙ

Рассмотрим теперь методы решения многомерных нестационарных задач. В качестве примера, на котором будут изучаться возникающие проблемы и подходы к нахождению численного решения, выберем уравнение теплопроводности

$$\frac{\partial u(t, \mathbf{x})}{\partial t} = a^2 \Delta u(t, \mathbf{x}) + f(t, \mathbf{x}),$$
$$\mathbf{x} = (x^{(1)}, \ldots, x^{(m)}),\ 0 < x^{(i)} < l_i,\ i = 1, \ldots, m, \quad 0 < t \le T, \tag{8.126}$$
$$\Delta u(t, \mathbf{x}) = \sum_{i=1}^{m} \frac{\partial^2 u(t, \mathbf{x})}{\partial (x^{(i)})^2}.$$

Как правило, если не оговорено противное, будет рассматриваться двумерное уравнение, то есть случай, когда $m = 2$.

Для этого уравнения поставим первую краевую задачу

$$\begin{aligned} u(0,\mathbf{x}) &= g(\mathbf{x}), \\ u(t,0,x^{(2)}) = \mu_1(t,x^{(2)}), \quad u(t,l_1,x^{(2)}) &= \mu_2(t,x^{(2)}), \\ u(t,x^{(1)},0) = \mu_3(t,x^{(1)}), \quad u(t,x^{(1)},l_2) &= \mu_4(t,x^{(1)}). \end{aligned} \tag{8.127}$$

Для простоты будем считать, что по каждой переменной выбрана равномерная сетка с шагами h_i по пространственным переменным и τ по времени. Таким образом, узлами сетки будут точки

$$\begin{aligned} &(t_n, x_k^{(1)}, x_j^{(2)}), \quad t_n = n\tau,\ n = 0,1,\dots,N,\ N\tau = T, \\ &x_k^{(1)} = kh_1,\ k = 0,\dots,K,\ Kh_1 = l_1, \quad x_j^{(2)} = jh_2,\ j = 0,\dots,J,\ Jh_2 = l_2. \end{aligned}$$

Решение разностного уравнения в точке $(t_n, x_k^{(1)}, x_j^{(2)})$ будем обозначать $y_{k\,j}^n$.

Введем разностные операторы

$$\Lambda_1 y_{k\,j} = \frac{y_{k+1\,j} - 2y_{k\,j} + y_{k-1\,j}}{h_1^2}, \qquad \Lambda_2 y_{k\,j} = \frac{y_{k\,j+1} - 2y_{k\,j} + y_{k\,j-1}}{h_2^2}, \tag{8.128}$$

которые аппроксимируют со вторым порядком производные $\dfrac{\partial^2}{\partial(x^{(1)})^2}$, $\dfrac{\partial^2}{\partial(x^{(2)})^2}$ соответственно. В том случае, когда в разностном уравнении все слагаемые будут иметь одни и те же нижние индексы k, j, будем их опускать.

Не представляет труда записать явную разностную схему для задачи (8.126), (8.127). Начальные и граничные условия аппроксимируем точно, а разностное уравнение запишем в виде

$$\frac{y^{n+1} - y^n}{\tau} = a^2(\Lambda_1 y^n + \Lambda_2 y^n) + f^n,$$

где $f^n = f(t_n, x_k^{(1)}, x_j^{(2)})$. Эта разностная схема аппроксимирует с первым порядком по τ и вторым по $h_1,\ h_2$.

Очевидна и организация расчетов по этой схеме. Из разностного уравнения имеем

$$y_{k\,j}^{n+1} = y_{k\,j}^n + a^2\tau(\Lambda_1 y_{k\,j}^n + \Lambda_2 y_{k\,j}^n) + \tau f_{k\,j}^n, \quad k = 1,\dots,K-1, \quad j = 1,\dots,J-1.$$

В остальных точках $n{+}1$-го слоя значение решения находится из граничного условия. На нулевом слое для определения решения используется начальное условие.

В следствие 4 пункта 8.3.1 с помощью спектрального критерия было получено необходимое условие устойчивости этой схемы, которое, легко обобщается на случай произвольного числа пространственных переменных и имеет вид

$$a^2\tau\sum_{i=1}^{m}\frac{1}{h_i^2} \le \frac{1}{2}.$$

Это означает, что в многомерном случае на шаг по времени накладывается еще более жесткое ограничение, чем в одномерном. Уже в одномерном случае от явных схем пришлось отказаться из-за необходимости выбирать слишком малый шаг τ. Для многомерного уравнения, следовательно, придется выбирать еще более малый шаг. Например, если $m = 2$, $l_1 = l_2 = T = 1$, $a^2 = 10$ и $h_1 = h_2 = 10^{-2}$, то $\tau \le 2.5\cdot 10^{-6}$. Значит, придется совершить не менее 400 000 шагов по времени. Из-за необходимости

вычислять с очень большим числом шагов по времени, явные схемы применяются очень редко.

Чисто неявная схема имеет вид

$$\frac{y^{n+1}-y^n}{\tau} = a^2(\Lambda_1 y^{n+1} + \Lambda_2 y^{n+1}) + f^{n+1}, \tag{8.129}$$

У нее такой же порядок аппроксимации как и у явной схемы, однако, в отличии от явной, неявная схема безусловно устойчива и, следовательно, шаг τ можно брать существенно большим, чем для явной схемы. Проблема здесь в другом — как на каждом слое находить решение. В одномерном случае относительно y^{n+1} решалась система линейных алгебраических уравнений. Благодаря специфике матрицы системы, решение находилось методом прогонки, который требовал $O(K)$ арифметических операций, то есть число арифметических операций имеет тот же порядок, что и явный метод. Уравнение (8.129) также является системой линейных алгебраических уравнений относительно y^{n+1}. Однако, теперь неизвестных $O(KJ)$ и, главное, теперь нет столь экономичного метода решения этой системы. Общие методы, например, метод Гаусса не подходит, из-за большого числа уравнений, которые, кроме того придется решать на каждом шаге по времени, то есть многократно. Есть, конечно, методы, которые учитывают специфику матрицы системы, и дают некоторый выигрыш во времени. Однако, по сравнению с явной схемой число арифметических операций в них для совершения одного шага по времени велико. В связи с этим встает вопрос о построении таких разностных схем, которые сочетали в себе простоту вычисления явных схем и безусловную устойчивость неявных схем. Такие схемы принято называть экономичными. Более точно, **экономичной** называется безусловно устойчивая разностная схема, у которой количество арифметических операций, требующихся для определении решения на $n+1$-ом слое при известных значениях решения на предыдущих слоях, пропорциональное числу точек сетки на слое.

Рассмотрению таких методов и будет посвящен этот параграф.

8.6.1 Схема переменных направлений

Исторически первой схемой, которая удовлетворяла условию экономичности была схема предложенная Писманом и Рекфордом. В литературе ее называют по-разному: схемой переменных направлений, продольно-поперечной схемой, схемой Писмана-Рекфорда.

Основная идея авторов заключается во введении дополнительного временного слоя $t_{n+1/2} = t_n + \tau/2$ и промежуточного значения сеточной функции $y^{n+1/2}$. Переход от слоя n к слою $n+1$ осуществляется в два этапа. Сначала от слоя n переходят к промежуточному слою $n+1/2$ — первый дробный шаг, а затем уже от слоя $n+1/2$ переходят к слою $n+1$ — второй дробный шаг. Записывается это следующим образом для всех внутренних точек сетки, то есть для $k \neq 0, K,\ j \neq 0, J$:

$$\frac{y^{n+1/2}-y^n}{0.5\tau} = a^2(\Lambda_1 y^{n+1/2} + \Lambda_2 y^n) + f^{n+1/2}, \tag{8.130}$$

$$\frac{y^{n+1}-y^{n+1/2}}{0.5\tau} = a^2(\Lambda_1 y^{n+1/2} + \Lambda_2 y^{n+1}) + f^{n+1/2}. \tag{8.131}$$

Первое из этих разностных уравнений является явным по направлению $x^{(2)}$ и неявным по $x^{(1)}$. Второе наоборот — явное по $x^{(1)}$ и неявное по $x^{(2)}$. К уравнениям добавляется начальное условие

$$y^0_{k\,j} = g(x^{(1)}_k, x^{(2)}_j), \quad k = 0, \ldots, K, \quad j = 0, \ldots, J, \tag{8.132}$$

и граничные условия

$$y_{0\,j}^{n+1/2} = (\widetilde{\mu}_1)_j^n, \quad y_{K\,j}^{n+1/2} = (\widetilde{\mu}_2)_j^n, \quad j = 1, \ldots, J-1, \tag{8.133}$$

$$y_{k\,0}^{n+1} = \mu_3(t_{n+1}, x_k^{(1)}), \quad y_{k\,J}^{n+1} = \mu_4(t_{n+1}, x_k^{(1)}), \quad k = 0, \ldots, K. \tag{8.134}$$

Здесь

$$(\widetilde{\mu}_i)_j^n = \frac{\mu_i(t_{n+1}, x_j^{(2)}) + \mu_i(t_n, x_j^{(2)})}{2} - \frac{a^2\tau}{4}\Lambda_2\Big(\mu_i(t_{n+1}, x_j^{(2)}) - \mu_i(t_n, x_j^{(2)})\Big), \ \ i = 1, 2. \tag{8.135}$$

Целесообразность такого выбора функций $\widetilde{\mu}_1,\ \widetilde{\mu}_2$ прояснится позднее при исследовании аппроксимации. Сейчас же отметим, что $(\widetilde{\mu}_i)_j^n$ только на величину порядка $O(\tau^2)$ отличается от $\mu_i(t_{n+1/2}, x_j^{(2)})$, которую на первый взгляд естественнее было бы определить в качестве граничного условия. В этом легко убедиться, переписав (8.135) в виде

$$(\widetilde{\mu}_i)_j^n = \frac{\mu_i(t_{n+1}, x_j^{(2)}) + \mu_i(t_n, x_j^{(2)})}{2} - \frac{a^2\tau^2}{4}\Lambda_2\left(\frac{\mu_i(t_{n+1}, x_j^{(2)}) - \mu_i(t_n, x_j^{(2)})}{\tau}\right) =$$

$$= \mu_i(t_{n+1/2}, x_j^{(2)}) - \frac{a^2\tau^2}{4}\frac{\partial^2}{\partial(x^{(2)})^2}\Big(\frac{\partial\mu_i(t_{n+1/2}, x_j^{(2)})}{\partial t}\Big) + O(\tau^2).$$

Прежде чем приступать к исследованию аппроксимации и устойчивости, рассмотрим как можно вычислить решение на $(n+1)$-ом слое, зная его на n-ом слое. Для этого перепишем (8.130), перенеся все слагаемые содержащие $y^{n+1/2}$ в левую часть, а y^n в правую. Умножим, кроме того, уравнение на 0.5τ и введем обозначение $r_i = a^2\tau/(2h_i^2)$. К полученным уравнениям добавим еще граничные условия для $y^{n+1/2}$. В результате получим:

$$\begin{aligned}
&y_{0\,j}^{n+1/2} = (\widetilde{\mu}_1)_j^n, \\
&-r_1 y_{k-1\,j}^{n+1/2} + (1+2r_1)y_{k\,j}^{n+1/2} - r_1 y_{k+1\,j}^{n+1/2} = \\
&\qquad = r_2 y_{k\,j-1}^n - (1+2r_2)y_{k\,j}^n + r_2 y_{k\,j+1}^n + \frac{1}{2}\tau f_{k\,j}^{n+1/2}, \quad k = 1, \ldots, K-1, \\
&y_{K\,j}^{n+1/2} = (\widetilde{\mu}_2)_j^n.
\end{aligned} \tag{8.136}$$

Аналогично преобразуем (8.131)

$$\begin{aligned}
&y_{k,0}^{n+1} = \mu_3(t_{n+1}, x_k^{(1)}) \\
&-r_2 y_{k\,j-1}^{n+1} + (1+2r_2)y_{k\,j}^{n+1} - r_2 y_{k\,j+1}^{n+1} = \\
&\qquad = r_1 y_{k-1\,j}^{n+1/2} - (1+2r_1)y_{k\,j}^{n+1/2} + r_1 y_{k+1\,j}^{n+1/2} + \frac{1}{2}\tau f_{k\,j}^{n+1/2}, \quad j = 1, \ldots, J-1, \\
&y_{k\,J}^{n+1} = \mu_4(t_{n+1} x_k^{(1)}).
\end{aligned} \tag{8.137}$$

При каждом фиксированном значении j (8.136) представляет собой систему линейных алгебраических уравнений относительно $y_{k\,j}^{n+1/2},\ k = 0, \ldots, K$. Матрица системы имеет трехдиагональный вид. Поэтому вдоль строк $j = 1, \ldots, J-1$ методом прогонки определяем $y^{n+1/2}$. Заметим, что условие диагонального преобладания матрицы системы выполнено, поэтому решение системы существует и единственно. Затем вдоль

столбцов $k = 1, \ldots, K-1$ прогонкой решаем (8.137) и находим y^{n+1}. При переходе к следующему слою процедура счета повторяется, то есть все время происходит чередование направлений, откуда и происходит одно из названий схемы.

Так как прогонка требует количество арифметических операций пропорциональное числу уравнений, общее число арифметических операций при переходе от слоя n к слою $n+1$ пропорционально KJ. Следовательно, для проверки того, что схема экономичная осталось доказать, что она безусловно устойчива.

Ограничимся проверкой выполнения спектрального критерия.

Если обозначить через S оператор перехода от слоя n к слою $n+1$, а через S_1 и S_2 операторы перехода от слоя n к слою $n+1/2$ и от слоя $n+1/2$ к слою $n+1$ соответственно, то $S = S_2 S_1$. Поскольку у всех трех операторов собственные функции одинаковы, $\lambda = \lambda_1 \cdot \lambda_2$, где $\lambda,\ \lambda_1,\ \lambda_2$ — собственные числа операторов $S,\ S_1,\ S_2$ соответственно.

Так как рассматриваются функции двух пространственных переменных, для нахождения собственных чисел вместо $y_{k\,j}$ надо подставлять $e^{i(k\varphi+j\psi)}$. Заметим, что

$$\Lambda_1 e^{i(k\varphi+j\psi)} = \left(-\frac{4}{h_1^2}\sin^2\frac{\varphi}{2}\right)e^{i(k\varphi+j\psi)},\quad \Lambda_2 e^{i(k\varphi+j\psi)} = \left(-\frac{4}{h_2^2}\sin^2\frac{\psi}{2}\right)e^{i(k\varphi+j\psi)}. \quad (8.138)$$

Подставляя теперь в однородное уравнение (8.130) $e^{i(k\varphi+j\psi)}$ вместо $y^n_{k\,j}$ и $\lambda_1 e^{i(k\varphi+j\psi)}$ вместо $y^{n+1/2}_{k\,j}$, получим

$$\lambda_1 - 1 = -4r_1\lambda_1\sin^2\frac{\varphi}{2} - 4r_2\sin^2\frac{\psi}{2},$$

откуда

$$\lambda_1 = \frac{1 - 4r_2\sin^2\frac{\psi}{2}}{1 + 4r_1\sin^2\frac{\varphi}{2}}.$$

Рассуждая аналогично, в однородном уравнении (8.131) заменим $y^{n+1/2}_{k\,j}$ на $e^{i(k\varphi+j\psi)}$, а $y^{n+1}_{k\,j}$ на $\lambda_2 e^{i(k\varphi+j\psi)}$. Тогда получим, что

$$\lambda_2 = \frac{1 - 4r_1\sin^2\frac{\varphi}{2}}{1 + 4r_2\sin^2\frac{\psi}{2}}.$$

Окончательно имеем

$$\lambda = \frac{1 - 4r_2\sin^2\frac{\psi}{2}}{1 + 4r_1\sin^2\frac{\varphi}{2}} \cdot \frac{1 - 4r_1\sin^2\frac{\varphi}{2}}{1 + 4r_2\sin^2\frac{\psi}{2}}. \quad (8.139)$$

Очевидно, что при любых $\varphi,\ \psi$ выполняется неравенство $|\lambda| \le 1$, тем самым требование спектрального критерия выполнено.

Осталось определить порядок аппроксимации этой схемы. С этой цедью исключим промежуточный дробный шаг. Для этого вычтем (8.131) из (8.130) и выразим $y^{n+1/2}_{k\,j}$. В результате получим

$$y^{n+1/2}_{k\,j} = \frac{y^{n+1}_{k\,j} + y^n_{k\,j}}{2} - \frac{a^2\tau}{4}\Lambda_2(y^{n+1}_{k\,j} - y^n_{k\,j}). \quad (8.140)$$

Заметим, что поскольку (8.130), (8.131) были определены для $k = 1, \ldots, K-1$, равенство (8.140) тоже справедливо только для этого диапазона индексов. Естественно потребовать, чтобы (8.140) выполнялось и для $k = 0, K$. Полагая $y^{n+1}_{0\,j} =$

$\mu_1(t_{n+1}, x_j^{(2)})$, $y_{0j}^n = \mu_1(t_n, x_j^{(2)})$, получаем первое граничное условие для решения на полуцелом шаге (8.133). Аналогично выводится второе условие (8.133).

Подставляя теперь (8.140) в (8.130), получим после простых преобразований

$$\frac{y^{n+1} - y^n}{\tau} = a^2(\Lambda_1 + \Lambda_2)\frac{y^{n+1} + y^n}{2} - \frac{a^4\tau^2}{4}\Lambda_1\Lambda_2\frac{y^{n+1} - y^n}{\tau} + f^{n+1/2}. \tag{8.141}$$

Схема (8.141) называется **схемой в целых шагах**.

Следует отметить, что если бы соотношение (8.140) не доопределили для $k = 0, K$ так как это сделано выше, равенство (8.141) было бы справедливо не для любых точек сетки. Его нельзя было бы записать для точек, в которых $k = 1, K-1$. Это связано с тем, что, например, при $k = 1$ в уравнение (8.130) входит $y_{0j}^{n+1/2}$, которое нельзя было бы тогда заменить по формуле (8.140).

Для определения погрешности аппроксимации достаточно исследовать схему в целых шагах. В параграфе 8.5 уже отмечалось, что

$$\frac{u(t_{n+1}, \mathbf{x}) - u(t_n, \mathbf{x})}{\tau} = \frac{\partial u(t_{n+1/2}, \mathbf{x})}{\partial t} + O(\tau^2), \quad \frac{u(t_{n+1}, \mathbf{x}) + u(t_n, \mathbf{x})}{2} = u(t_{n+1/2}, \mathbf{x}) + O(\tau^2).$$

Поэтому, учитывая что функция u является решением уравнения (8.126), а Λ_1, Λ_2 аппроксимируют вторые производные по пространственным переменным со вторым порядком, имеем

$$\begin{aligned}
\psi_h^{(1)} &= \frac{u(t_{n+1}, \mathbf{x}) - u(t_n, \mathbf{x})}{\tau} - a^2(\Lambda_1 + \Lambda_2)\frac{u(t_{n+1}, \mathbf{x}) + u(t_n, \mathbf{x})}{2} + \\
&+ \frac{a^4\tau^2}{4}\Lambda_1\Lambda_2\frac{u(t_{n+1}, \mathbf{x}) - u(t_n, \mathbf{x})}{\tau} - f(t_{n+1/2}, \mathbf{x}) = \frac{\partial u(t_{n+1/2}, \mathbf{x})}{\partial t} - \\
&- a^2(\Lambda_1 + \Lambda_2)u(t_{n+1/2}, \mathbf{x}) - f(t_{n+1/2}, \mathbf{x}) + \frac{a^4\tau^2}{4}\Lambda_1\Lambda_2\frac{\partial u(t_{n+1/2}, \mathbf{x})}{\partial t} + O(\tau^2) = \\
&= \frac{\partial u(t_{n+1/2}, \mathbf{x})}{\partial t} - a^2\Delta u(t_{n+1/2}, \mathbf{x}) - f(t_{n+1/2}, \mathbf{x}) + O(\tau^2 + h_1^2 + h_2^2) = \\
&= O(\tau^2 + h_1^2 + h_2^2).
\end{aligned} \tag{8.142}$$

Так как для y^n граничное и начальное условия выполняются точно, разностная схема аппроксимирует задачу (8.126), (8.127) на ее гладким решении со вторым порядком. Отсюда по теореме сходимости заключаем, что порядок скорости сходимости решения разностной y схемы к решению дифференциальной задачи u равен двум.

Основным недостатком изучаемой схемы переменных направлений является тот факт, что ее нельзя обобщить на случай трех и более переменных. Например, для трех переменных естественным обобщением схемы (8.130) будет следующая схема (для краткости будем считать $f = 0$, $a = 1$):

$$\begin{aligned}
\frac{y^{n+1/3} - y^n}{1/3\tau} &= \Lambda_1 y^{n+1/3} + \Lambda_2 y^n + \Lambda_3 y^n, \\
\frac{y^{n+2/3} - y^{n+1/3}}{1/3\tau} &= \Lambda_1 y^{n+1/3} + \Lambda_2 y^{n+2/3} + \Lambda_3 y^{n+1/3}, \\
\frac{y^{n+1} - y^{n+2/3}}{1/3\tau} &= \Lambda_1 y^{n+2/3} + \Lambda_2 y^{n+2/3} + \Lambda_3 y^{n+1}.
\end{aligned}$$

Теперь в каждом разностном уравнении только по одной переменной аппроксимация неявная, а по двум явная. По аналогии с (8.139) получим

$$\lambda = \frac{1 - 4r_2\sin^2\frac{\psi}{2} - 4r_3\sin^2\frac{\omega}{2}}{1 + 4r_1\sin^2\frac{\varphi}{2}} \cdot \frac{1 - 4r_1\sin^2\frac{\varphi}{2} - 4r_3\sin^2\frac{\omega}{2}}{1 + 4r_2\sin^2\frac{\psi}{2}} \cdot \frac{1 - 4r_1\sin^2\frac{\varphi}{2} - 4r_2\sin^2\frac{\psi}{2}}{1 + 4r_3\sin^2\frac{\omega}{2}},$$

где $r_i = \tau/(3h_i^2)$. Ясно, что $|\lambda|$ может принимать значения большие 1. Для того, чтобы в этом убедиться достаточно взять $\varphi = \psi = \omega = \pi$, $r_1 = r_2 = r_3 = 1$. Следовательно, схема не является безусловно устойчивой. Неустойчивость возникла из-за того, что на каждом промежуточном шаге по двум переменным применялась явная аппроксимация, которая "ухудшала" устойчивость схемы. Следовательно желательно избавляться от слагаемых, содержащих явную аппроксимацию. Эта идея и реализована в схеме, рассматриваемой в следующем пункте.

8.6.2 Метод расщепления (дробных шагов)

Для решения задачи (8.126), (8.127) предлагается следующая схема (**схема дробных шагов**). Разностные уравнения имеют вид

$$\frac{y^{n+1/2} - y^n}{\tau} = a^2 \Lambda_1 y^{n+1/2} + f^{n+1/2}, \tag{8.143}$$

$$\frac{y^{n+1} - y^{n+1/2}}{\tau} = a^2 \Lambda_2 y^{n+1}. \tag{8.144}$$

Заметим, что разностное уравнение (8.143) аппроксимирует дифференциальное уравнение

$$\frac{1}{2}\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial (x^{(1)})^2} + f, \tag{8.145}$$

а разностное уравнение (8.144) — дифференциальное уравнение

$$\frac{1}{2}\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial (x^{(2)})^2}. \tag{8.146}$$

Таким образом, на каждом шаге разностные уравнения не приближают исходное дифференциальное уравнение, однако, сумма уравнений (8.145), (8.146) дает исходное уравнение (8.126).

Используя спектральный критерий, нетрудно получить, что

$$0 < \lambda = \frac{1}{1 + 4r_1 \sin^2 \frac{\varphi}{2}} \cdot \frac{1}{1 + 4r_2 \sin^2 \frac{\psi}{2}} \le 1, \quad r_i = \frac{a^2 \tau}{h_i^2}, \quad i = 1, 2.$$

Следовательно, при любом соотношении шагов условие устойчивости выполнено.

Для определения порядка аппроксимации снова запишем схему в целых шагах. Для этого из (8.144) выразим $y^{n+1/2}$

$$y^{n+1/2} = y^{n+1} - a^2 \tau \Lambda_2 y^{n+1}. \tag{8.147}$$

Подставляя это значение в (8.143), получим схему в целых шагах, которая после простых преобразований примет вид

$$\frac{y^{n+1} - y^n}{\tau} = a^2 \Lambda_1 y^{n+1} + a^2 \Lambda_2 y^{n+1} - a^4 \tau \Lambda_1 \Lambda_2 y^{n+1} + f^{n+1/2}. \tag{8.148}$$

Нетрудно получить теперь, что погрешность аппроксимации равна $O(\tau + h_1^2 + h_2^2)$.

Замечание Правую часть f не обязательно вычислять в точке $t_{n+1/2}$. Она может быть без ущерба для аппроксимации подсчитана в любой точке отрезка $[t_n, t_{n+1}]$ и, даже отличаться от вычисленной величины не $O(\tau)$.

До сих пор ничего не говорилось о постановке граничных условий для разностных уравнений (8.143), (8.144). Для того, чтобы схема в целых шагах имела вид (8.148), равенство (8.148) должно выполняться и при $k = 0, K$. Поэтому граничное условие должно иметь вид

$$y_{0\,j}^{n+1/2} = (\widetilde{\mu}_1)_j^n, \quad y_{K\,j}^{n+1/2} = (\widetilde{\mu}_2)_j^n, \quad j = 1, \ldots, J-1,$$
$$(\widetilde{\mu}_i)_j^n = (\mu_i)_j^{n+1} - a^2\tau\Lambda_2(\mu_i)_j^{n+1}, \quad i = 1, 2,$$
$$y_{k\,0}^{n+1} = \mu_3(t_{n+1}, x_k^{(1)}), \quad y_{k\,J}^{n+1} = \mu_4(t_{n+1}, x_k^{(1)}), \quad k = 0, \ldots, K.$$

Так же как и для продольно-поперечной схемы на каждом дробном шаге решение находится прогонкой, прочем на первом дробном шаге прогонка осуществляется вдоль переменной $x^{(1)}$, а на втором — вдоль $x^{(2)}$.

К вопросу аппроксимации можно подойти иначе. Рассмотрим погрешность аппроксимации каждого из разностных уравнений (8.143), (8.144)

$$\psi_{h,1}^{(1)} = \frac{u(t_{n+1/2}, \mathbf{x}) - u(t_n, \mathbf{x})}{\tau} - a^2\Lambda_1 u(t_{n+1/2}, \mathbf{x}) - f(t_{n+1/2}, \mathbf{x}) =$$
$$= \frac{1}{2}\frac{\partial u(t_n, \mathbf{x})}{\partial t} - a^2\frac{\partial^2 u(t_n, \mathbf{x})}{\partial (x^{(1)})^2} - f(t_n, \mathbf{x}) + O(\tau + h_1^2) = O(1),$$
$$\psi_{h,2}^{(1)} = \frac{u(t_{n+1}, \mathbf{x}) - u(t_{n+1/2}, \mathbf{x})}{\tau} - a^2\Lambda_2 u(t_{n+1}, \mathbf{x}) =$$
$$= \frac{1}{2}\frac{\partial u(t_n, \mathbf{x})}{\partial t} - a^2\frac{\partial^2 u(t_n, \mathbf{x})}{\partial (x^{(2)})^2} + O(\tau + h_2^2) = O(1).$$

Как видим, каждая из погрешностей $\psi_{h,i}^{(1)} = O(1),\ i = 1, 2$, то есть является конечной величиной и, значит, не стремится к нулю при стремлении к нулю шагов сетки. В то же время $\psi_h^{(1)} = \psi_{h,1}^{(1)} + \psi_{h,2}^{(1)} = O(\tau + h_1^2 + h_2^2)$. В этом случае говорят об **аппроксимации в суммарном смысле**, а методы получения схем для которых выполняется аппроксимация в суммарном смысле называют **методами суммарной аппроксимации**, **методами дробных шагов** или **методами расщепления**.

В заключение этого параграфа рассмотрим общий подход к построению схем, обладающих суммарной аппроксимацией. Для простоты будем говорить о задаче Коши. Пусть имеется задача Коши

$$\frac{\partial u}{\partial t} = \mathcal{L}u + f, \quad t \in [0, T], \quad u\Big|_{t=0} = u_0. \tag{8.149}$$

Пусть $\mathcal{L} = \mathcal{L}_1 + \cdots + \mathcal{L}_p,\ f = f_1 + \cdots + f_p$. Введем по временной переменной сетку $t_n = n\tau,\ n = 0, \ldots, N,\ N\tau = T$. Кроме того, на каждом отрезке $[t_n, t_{n+1}]$ введем промежуточные точки

$$t_{n+i/p} = (n + i/p)\tau = t_n + (i/p)\tau, \quad i = 1, \ldots, p-1.$$

Рассмотрим теперь последовательность задач Коши:

$$\begin{aligned}
&\frac{\partial u_1}{\partial t} = p\mathcal{L}_1 + pf_1, \quad t_n < t \le t_{n+1/p}, \quad u_1\Big|_{t=t_n} = u_p\Big|_{t=t_n}, \\
&\frac{\partial u_2}{\partial t} = p\mathcal{L}_2 + pf_2, \quad t_{n+1/p} < t \le t_{n+2/p}, \quad u_2\Big|_{t=t_{n+1/p}} = u_1\Big|_{t=t_{n+1/p}}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\frac{\partial u_p}{\partial t} = p\mathcal{L}_p + pf_p, \quad t_{n+(p-1)/p} < t \le t_{n+1}, \quad u_p\Big|_{t=t_{n+(p-1)/p}} = u_{p-1}\Big|_{t=t_{n+(p-1)/p}}, \\
&u_1\Big|_{t=0} = u_0, \quad n = 0, 1, \ldots, N-1.
\end{aligned} \tag{8.150}$$

Если ввести функцию $u_\tau = u_i$ при $t \in (t_{n+i/p}, t_{n+(i+1)/p}]$, то можно показать, что при определенных условиях $u - u_\tau = O(\tau)$.

Если каждое из уравнений (8.150) заменить на разностное уравнение, получится схема обладающая аппроксимацией в суммарном смысле. Примером может послужить схема (8.143), (8.144). При ее построении в качестве уравнения (8.149) выступало уравнение (8.126), в котором $m = 2$. Оператор

$$\mathcal{L}u = a^2\frac{\partial^2 u(t,\mathbf{x})}{\partial(x^{(1)})^2} + a^2\frac{\partial^2 u(t,\mathbf{x})}{\partial(x^{(2)})^2}, \qquad \mathcal{L}_i u = a^2\frac{\partial^2 u(t,\mathbf{x})}{\partial(x^{(i)})^2},\ i = 1,2, \quad f_1 = f,\ f_2 = 0.$$

Уравнениям (8.150) соответствуют уравнения (8.145), (8.146).

Разбиение оператора $\mathcal{L}$ на сумму операторов может осуществляться по разному. В том случае, когда каждый из операторов $\mathcal{L}_i$ содержит дифференцирование только по одной переменной, их называют одномерными и говорят, что многомерная задача свелась к решению последовательности одномерных задач. Полученную в этом случае разностную схему называют **локально-одномерной**. Разбиение (расщепление) оператора на одномерные, это не единственный способ, который применяют при построении схем, обладающих суммарной аппроксимацией. Иногда применяют подход, при котором операторы $\mathcal{L}_i$ соответствуют различным физическим процессам, в этом случае говорят о **расщеплении по физическим процессам**. Очевидно, что в любом случае при расщеплении стараются получить операторы более простой структуры.

8.7 РАЗНОСТНЫЕ СХЕМЫ ДЛЯ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ

Типичным представителем уравнений эллиптического типа, разностные схемы для которого будет изучаться в этом параграфе, является уравнение Пуассона

$$\Delta u(\mathbf{x}) = -f(\mathbf{x}), \quad \mathbf{x} = (x^{(1)}, \ldots, x^{(m)}), \quad \mathbf{x} \in \Omega \subset \mathbb{R}^m, \quad \Delta u(\mathbf{x}) = \sum_{i=1}^{m}\frac{\partial^2 u(\mathbf{x})}{\partial(x^{(i)})^2}. \quad (8.151)$$

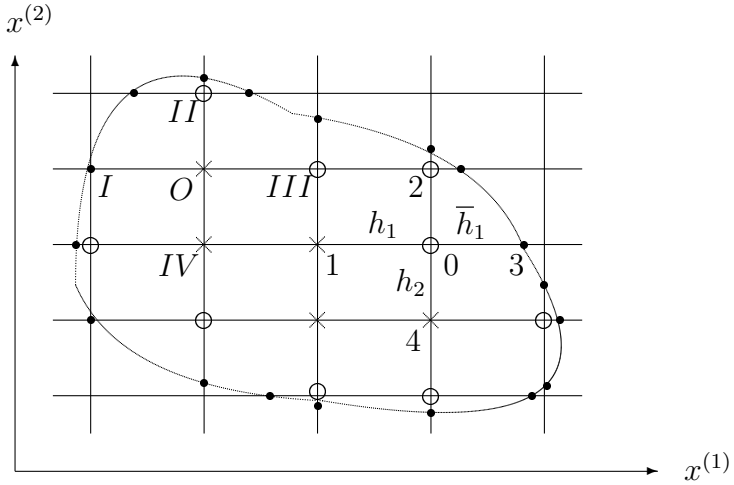
Будем считать, что Ω — ограниченная область из m-мерного евклидова пространства и граница Γ области Ω является кусочно-гладкой. Для уравнения (8.151) могут задаваться различного рода граничные условия: Дирихле, Неймана, третьего рода. В основном, в этом параграфе речь будет идти об условии Дирихле

$$u(\mathbf{x})\Big|_{\mathbf{x}\in\Gamma} = g(\mathbf{x}). \quad (8.152)$$

Для простоты в дальнейшем будем рассматривать случай $m = 2$.

8.7.1 Построение разностной схемы

Как обычно, построение разностной схемы начнем с задания сетки. Для этого проведем прямые $x_k^{(1)} = kh_1,\ x_j^{(2)} = jh_2\ \ k, j = 0, \pm1, \pm2, \ldots$. Точки пересечения этих прямых, которые попадут внутрь области Ω назовем **внутренними узлами сетки** и их совокупность обозначим ω. Точки пересечения прямых с границей Γ назовем **граничными узлами**, их множество обозначим γ.



$\times$ — регулярные, $\circ$ — нерегулярные, $\bullet$ — граничные узлы

Рис. 8.8 Сетка для решения задачи (8.151),(8.152).

Пусть $\overline{\omega} = \omega \bigcup \gamma$ — совокупность внутренних и граничных узлов.

Не составляет труда записать разностное уравнение

$$\frac{y_{k+1\,j} - 2y_{k\,j} + y_{k-1\,j}}{h_1^2} + \frac{y_{k\,j+1} - 2y_{k\,j} + y_{k\,j-1}}{h_2^2} = -f_{k\,j}, \tag{8.153}$$

где $f_{k\,j} = f(x_k^{(1)}, x_j^{(2)})$. Это же уравнение можно переписать в операторном виде:

$$\Lambda y = \Lambda_1 y + \Lambda_2 y = -f. \tag{8.154}$$

Здесь $\Lambda_1,\ \Lambda_2$ — операторы, аппроксимирующие со вторым порядком вторые производные по $x^{(1)}, x^{(2)}$ соответственно и определенные по формулам (8.128). У сеточных функций опущены индексы k, j.

Необходимо выделить те точки сетки, для которых можно записать разностные уравнения (8.153). С этой целью введем понятия соседних, регулярных и нерегулярных узлов. Сразу оговоримся, что приведенное ниже определение соседнего узла связано с шаблоном разностной схемы.

Два узла сетки $\mathbf{x_1}$ и $\mathbf{x_2}$ назовем **соседними**, если у них одна из координат совпадает, а вторая отличается на величину шага, соответствующего этой координате. На рисунке 8.8 для точки O римскими цифрами I,II,III,IV отмечены соседние точки.

Внутренний узел сетки назовем **регулярным**, если все его соседние узлы принадлежат $\overline{\omega}_h$ и **нерегулярным** в противном случае.

Совокупность регулярных узлов обозначим ω_r , а нерегулярных — ω_u. Из определения следует, что $\omega = \omega_r \bigcup \omega_u$.

В некоторых случаях одно из множеств ω_r или ω_u может быть пустым. Если пусто множество ω_r, это означает, что шаги сетки очень большие, поэтому этот случай не интересен. Множество ω_u пусто, например, в том случае, когда область прямоугольник

$$\Omega = \{(x^{(1)}, x^{(2)}) \ : \ 0 < x^{(i)} < l_i,\ i = 1, 2\}$$

и шаги подобраны таким образом, что отношения l_i/h_i — целые числа.

Очевидно теперь, что разностные уравнения (8.153) могут быть записаны для всех регулярных узлов сетки, то есть для тех точек, для которых $(x_k^{(1)}, x_j^{(2)}) \in \omega_r$. Порядок аппроксимации разностных уравнений в регулярных точках равен двум.

Для граничных узлов достаточно положить

$$y_{k\,j} = g(x_k^{(1)}, x_j^{(2)}), \quad (x_k^{(1)}, x_j^{(2)}) \in \gamma. \tag{8.155}$$

Для того случая, когда множество нерегулярных узлов пусто, разностная схема построена. Поэтому при возможности стараются выбирать сетку так, чтобы нерегулярные узлы отсутствовали. Такую сетку называют **согласованной**.

Если же нерегулярные узлы существуют, надо записать аппроксимацию для нерегулярных узлов. Существуют различные варианты аппроксимации. Рассмотрим некоторые из них.

- В простейшем случае для нерегулярного узла $\mathbf{x}$ находят ближайший граничный $\mathbf{x}_\gamma(\mathbf{x})$ и полагают

  $$y\Big|_{\mathbf{x}} = g(\mathbf{x}_\gamma(\mathbf{x})), \quad \mathbf{x} \in \omega_u. \tag{8.156}$$

  В нерегулярном узле функции просто приписывается значение, взятое из ближайшей граничной точки. Погрешность аппроксимации граничного условия при таком подходе равна $O(h_1 + h_2)$.

  По существу в этом подходе происходит изменение понятий внутренних и граничных узлов. Граничными становятся все нерегулярные узлы, а внутренними — регулярные.

- Для повышения порядка аппроксимации значение функции в нерегулярный узел не переносится из ближайшей граничной точки, а получается путем интерполяции многочленом первого порядка. Так, например, для ситуации, изображенной на рисунке 8.8 значение решения в узле "0" получается интерполяцией вдоль оси $x^{(1)}$ по узлам "1" и "3".

  $$y_0 = \frac{\overline{h}_1 y_1 + h_1 y_3}{\overline{h}_1 + h_1}, \tag{8.157}$$

  где $\overline{h}_1$ — расстояние между точками "0" и "3 y_i — значение искомого разностного решения в точке с номером i. Погрешность аппроксимации в этом случае равна $O(h_1^2 + h_2^2)$. Как и прежде, нерегулярные узлы здесь трактуются как граничные.

- Нерегулярные узлы считаются внутренними и в них записывается разностное уравнение, но с учетом неравномерности сетки. Так, например, для узла под номером "0" (см. рисунок 8.8) разностное уравнение имеет вид

  $$\frac{2}{\overline{h}_1 + h_1}\left(\frac{y_3 - y_0}{\overline{h}_1} - \frac{y_0 - y_1}{h_1}\right) + \frac{y_2 - 2y_0 + y_4}{h_2} = -f_0. \tag{8.158}$$

  По аналогии можно записать уравнения для других вариантов расположения узлов. Погрешность аппроксимации в нерегулярных имеет первый порядок. Так, например, для уравнения (8.158) разложением по формуле Тейлора получаем, что погрешность аппроксимации равна $O(h_1 + h_2^2)$.

  В граничных узлах разностное решение определяется по формуле (8.155).

Итак, разностная схема построена. Она представляет собой систему линейных алгебраических уравнений относительно значений сеточной функции y. Возникают вопросы: существует ли решение этой системы и если да, то единственно ли оно; как

найти это решение, если оно существует; сходится ли решение разностной схемы к решению исходной задачи (8.151),(8.152).

Для ответа на последний вопрос в соответствии с теоремой сходимости достаточно проверить устойчивость разностной схемы. Как будет видно из дальнейшего, из устойчивости будет следовать существование и единственность решения системы алгебраических уравнений.

8.7.2 Устойчивость разностной схемы

Для того, чтобы не усложнять рассуждения деталями, будем считать, что при построении схемы выбран простейший способ аппроксимации граничных условий, заключающийся в том, что в нерегулярные точки перенесены значения разностного решения из ближайшего граничного узла. Когда далее речь будет идти о внутренних точках, будем иметь в виду регулярные точки, а нерегулярные относить к граничным.

Для доказательства устойчивости потребуется следующая лемма

Лемма 8.7.1 (Принцип максимума) *Пусть v — произвольная сеточная функция, определенная на $\overline{\omega}$, причем $v \le 0$ на γ и $\Lambda v \ge 0$ на ω. Тогда $v \le 0$ на $\overline{\omega}$.*

Доказательство. Предположим противное, то есть, что существует один или несколько внутренних узлов сетки, в которых значение функции v положительно. Тогда среди этих узлов найдется такой, что значение функции v в нем не меньше, чем в соседних узлах, причем хотя бы в одном из соседних узлов значение функции строго меньше. Действительно, достаточно выбрать узел, в котором значение функции v наибольшее и, если во всех соседних узлах значения функции совпадают, начать двигаться от него вдоль одной из осей, например, $x^{(1)}$. Либо в процессе перемещения обязательно встретится нужная ситуация, либо, так как область ограничена, дойдем до приграничного узла и в нем значение больше, чем в граничном узле, поскольку на границе значения $v \le 0$. Пусть $(x_k^{(1)}, x_j^{(2)})$ такой узел и, например, $v_{k+1\,j} < v_{k\,j}$. Тогда

$$\Lambda v_{k\,j} = \frac{(v_{k+1\,j} - v_{k\,j}) + (v_{k-1\,j} - v_{k\,j})}{h_1^2} + \frac{(v_{k\,j+1} - v_{k\,j}) + (v_{k\,j-1} - v_{k\,j})}{h_2^2}.$$

В числителе каждой дроби все скобки не положительны, причем первая отрицательна. Поэтому $\Lambda v_{k\,j} < 0$, что противоречит условию леммы. Полученное противоречие означает, что предположение не верно и, следовательно, утверждение леммы доказано.

Введем следующие нормы:

$$\|v\|^{(1)} = \max_{(x_k^{(1)}, x_j^{(2)}) \in \omega} |v_{k\,j}|, \qquad \|g\|^{(2)} = \max_{(x_k^{(1)}, x_j^{(2)}) \in \gamma} |g_{k\,j}|, \qquad \|v\|^{(3)} = \max_{(x_k^{(1)}, x_j^{(2)}) \in \overline{\omega}} |v_{k\,j}|.$$

Теорема 8.7.1 *Если решение разностной схемы (8.153),(8.155) существует, то для него справедлива оценка*

$$\|y_h\|^{(3)} \le C\|f\|^{(1)} + \|g_h\|^{(2)}, \tag{8.159}$$

где C — константа, значение которой зависит от области Ω.

Заметим, что неравенство (8.159) означает устойчивость разностной схемы.

Доказательство. Введем две вспомогательные функции v_+ и v_-, определив их

$$v_\pm = \pm y + a\big((x^{(1)})^2 + (x^{(2)})^2\big) + b,$$

где a, b — числа, которые будут определены позднее, а y — решение разностной схемы.

Заметим, что

$$\Lambda_1(x^{(1)})^2 = \Lambda_2(x^{(2)})^2 = 2, \quad \Lambda_1(x^{(2)})^2 = \Lambda_2(x^{(1)})^2 = 0.$$

Эти соотношения проверяются непосредственным вычислением, например,

$$\Lambda_1(x^{(1)})^2 = \frac{(x^{(1)} + h_1)^2 - 2(x^{(1)})^2 + (x^{(1)} - h_1)^2}{h_1^2} = 2.$$

Тогда $\Lambda\big((x^{(1)})^2 + (x^{(2)})^2\big) = 4$. Отсюда и из того, что y — решение разностной схемы, имеем для внутренних узлов сетки

$$\Lambda v_\pm = \mp f + 4a.$$

Подберем a так, чтобы всюду на ω выполнялось неравенство $\Lambda v_\pm \geq 0$. Для этого достаточно положить

$$a = \frac{1}{4} \max_{(x_k^{(1)}, x_j^{(2)}) \in \omega} |f_{kj}| = \frac{1}{4}\|f\|^{(1)}.$$

Определим теперь b так, чтобы всюду на γ выполнялось неравенство $v_\pm \leq 0$. Неравенство будет выполнено, если взять

$$\begin{aligned} b &= -\|g\|^{(2)} - a \max_{(x^{(1)}, x^{(2)}) \in \Omega \bigcup \Gamma} \big((x^{(1)})^2 + (x^{(2)})^2\big) = \\ &= -\|g\|^{(2)} - \frac{1}{4} \max_{(x^{(1)}, x^{(2)}) \in \Omega \bigcup \Gamma} \big((x^{(1)})^2 + (x^{(2)})^2\big)\|f\|^{(1)} = -\|g\|^{(2)} - C\|f\|^{(1)}. \end{aligned}$$

Здесь

$$C = \frac{1}{4} \max_{(x^{(1)}, x^{(2)}) \in \Omega \bigcup \Gamma} \big((x^{(1)})^2 + (x^{(2)})^2\big).$$

Из леммы следует, что $v_\pm \leq 0$ на ω. Учитывая определение функций $v_\pm$, имеем

$$\pm y + a\big((x^{(1)})^2 + (x^{(2)})^2\big) + b \leq 0. \tag{8.160}$$

Следовательно, $\pm y + b \leq 0$ или $|y| \leq -b$. Подставляя в это неравенство выражение для b, получим, что для всех внутренних точек сетки справедливо неравенство

$$|y| \leq \|g\|^{(2)} + C\|f\|^{(1)}.$$

Так как для граничных узлов $|y| \leq \|g\|^{(2)}$, можно заключить, что неравенство (8.160) выполняется всюду на $\overline{\omega}$, то есть справедливо (8.159). Теорема доказана.

Теорема 8.7.2 *Разностная схемы (8.153), (8.155) имеет решение, причем только одно.*

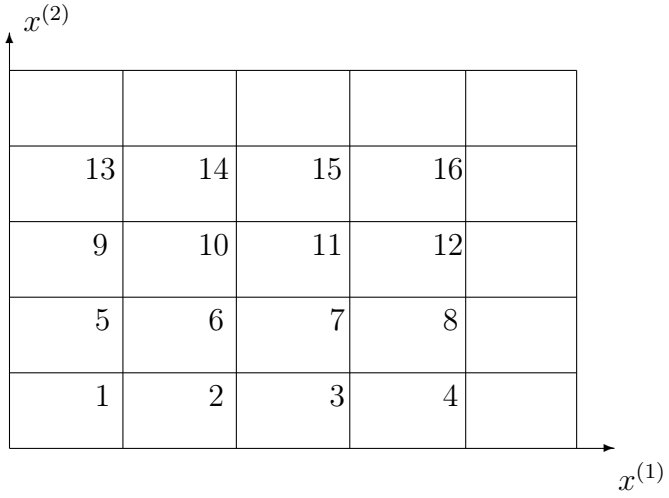


Рис. 8.9 Одномерная нумерация двумерного массива

Доказательство. Разностная схема представляет собой систему линейных алгебраических уравнений. Известно, что такая система имеет решение и при том единственное тогда и только тогда, когда ее определитель не равен нулю. В этом случае однородная система уравнений имеет единственное нулевое решение. Если же определитель равен нулю, однородная система имеет ненулевое решение. Поэтому, если доказать что у однородной системы уравнений нет никакого другого решения кроме нулевого, это будет означать, что определитель не равен нулю и, значит для любых f, g решение разностной схемы существует.

Для получения однородной системы достаточно взять $f = 0,\ g = 0$. Предположим, что у однородной системы существует какое-нибудь решение кроме нулевого. Для него справедлива оценка (8.159). Но в правой части этого неравенства стоит ноль, значит $\|y\|^{(3)} = 0$, откуда следует, что $y \equiv 0$. Теорема доказана.

Осталась не решенной одна проблема — каким образом находить решение разностной схемы. Решение этой задачи будет рассмотрено в следующем параграфе.

8.8 МЕТОДЫ РЕШЕНИЯ РАЗНОСТНЫХ СХЕМ ДЛЯ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ

Полученная в предыдущем параграфе система линейных алгебраических уравнений (8.153),(8.155) обладает рядом особенностей. Прежде всего система характеризуется большой размерностью, то есть большим числом неизвестных в системе. Это связано с желанием получить решение исходной дифференциальной задачи с большой точностью, что требует мелкого шага сетки.

Если произвести нумерацию компонент вектора неизвестных каким угодно способом (смотри, например, рисунок 8.9), систему можно будет записать в традиционном виде. Матрица системы при этом будет иметь в каждой строке большое число нулей, так как в каждой строке не более пяти ненулевых элементов (столько неизвестных входим в каждое разностное уравнение). Кроме того, как будет показано ниже, матрица системы является плохо обусловленной.

Перечисленные особенности системы определяет и подход к нахождению ее решения. Применение для решения системы, например, метода Гаусса нецелесообразно. Для того, чтобы пояснить причину этого предположим, что область Ω квадрат

со стороной равной 1 и число шагов по каждой переменной одинаково и равно K. Тогда исключая граничные значение сеточной функции, получим $(K-1)^2$ уравнений. Метод Гаусса потребует $O(2(K-1)^6/3)$ арифметических операций. При этом большинство операций будут малосодержательными, так как будут производиться с нулевыми элементами. Кроме того, потребуется хранить в памяти $(K-1)^2$ элементов матрицы, большинство из которых нули.

Таким образом, возникает необходимость разработки специальных методов, учитывающих особенность системы. Наибольшее распространение, особенно для областей отличных от прямоугольника, получили итерационные методы, которые и будут рассмотрены в этом параграфе.

8.8.1 Вспомогательные утверждения

Систему линейных алгебраических уравнений будем записывать в виде

$$\mathcal{A}y = b, \tag{8.161}$$

рассматривая $\mathcal{A}$ как линейный оператор, заданный на линейном нормированном пространстве $\mathcal{H}$ сеточных функций.

Проводить детальный анализ методов удобно для случая прямоугольной области $\Omega = \{(x^{(1)}, x^{(2)}) \ : \ 0 \le x^{(i)} \le l_i, \ i = 1, 2\}$. Будем считать, что сетка выбрана таким образом, что узлами сетки ω являются точки $(x_k^{(1)}, x_j^{(2)}) = (kh_1, jh_2)$, $k = 1, \ldots, K-1$, $j = 1, \ldots, J-1$, $Kh_1 = l_1, Jh_2 = l_2$. Граница γ состоит из точек $(0, jh_2), (K, jh_2), (kh_1, 0), (kh_1, J)$.

Тогда для разностных уравнений (8.153) оператор $\mathcal{A} = -\Lambda = -\Lambda_1 - \Lambda_2$, а в качестве пространства $\mathcal{H}$ выбираются сеточные функции определенные на ω, и доопределенные нулем на γ.

Следует сделать пояснения, связанные с требованием равенства функций нулю на γ, которое не согласовано с условием (8.155). Выберем, например, узел $(x_1^{(1)}, x_2^{(2)})$ и запишем в нем разностное уравнение

$$\frac{y_{0\,2} - 2y_{1\,2} + y_{2\,2}}{h_1^2} + \frac{y_{1\,1} - 2y_{1\,2} + y_{1\,3}}{h_2^2} = -f_{1\,2}. \tag{8.162}$$

Сюда следует подставить значение

$$y_{0\,2} = g(0, x_2^{(2)}), \tag{8.163}$$

полученное из граничного условия. Перенося это значение в правую часть, имели бы

$$\frac{y_{0\,2} - 2y_{1\,2} + y_{2\,2}}{h_1^2} + \frac{y_{1\,1} - 2y_{1\,2} + y_{1\,3}}{h_2^2} = -f_{1\,2} - \frac{g(x_0^{(1)}, x_2^{(2)})}{h_1^2}. \tag{8.164}$$

Значит, $b_{1\,2} = f_{1\,2} + \frac{g(x_0^{(1)}, x_2^{(2)})}{h_1^2}$, а

$$\mathcal{A}y_{1\,2} = \frac{-2y_{1\,2} + y_{2\,2}}{h_1^2} + \frac{y_{1\,1} - 2y_{1\,2} + y_{1\,3}}{h_2^2}.$$

Таким образом, у оператора $\mathcal{A}$ в приграничном узле коэффициенты не такие, как в узле, все соседи которого внутренние узлы сетки. Для единообразия записи удобно считать, что во всех внутренних узлах, включая приграничные, оператор выглядит

одинаково, просто граничное значение функции равно нулю. Фактически равенства $\Lambda y_{1\,2} = -f_{1\,2}$ и (8.163) заменили на равенства $\Lambda y_{1\,2} = b_{1\,2},\ y_{0\,2} = 0$. На значение решения разностной схемы (8.153), (8.155) такая трактовка уравнений не повлияет. Она удобнее для теоретических рассуждений.

Для дальнейшего понадобятся некоторые свойства оператора $\mathcal{A} = -\Lambda$ и пространства $\mathcal{H}$ сеточных функций, на котором он определен.

Прежде всего отметим, что поскольку сеточную функцию можно трактовать как одномерный вектор, число компонент которого совпадает с числом узлов сетки ω, размерность пространства сеточных функций $\mathcal{H}$ равна $(K-1)(J-1)$.

Скалярное произведение векторов определяется как сумма произведений соответствующих координат, поэтому в пространстве $\mathcal{H}$ можно ввести скалярное произведение по формуле

$$(v, w) = h_1 h_2 \sum_{k=1}^{K-1} \sum_{j=1}^{J-1} v_{k\,j} w_{k\,j}. \tag{8.165}$$

Здесь перед суммами добавлен множитель $h_1 h_2$ для того, чтобы в пределе при $h_i \to 0,\ i = 1, 2$ получить скалярное произведение пространства L_2, то есть интеграл от произведения функций.

Лемма 8.8.1 *Оператор $\mathcal{A}$ является положительно определенным, то есть для любых $v, w \in \mathcal{H}$ справедливы утверждения:*

- $(\mathcal{A}v, w) = (v, \mathcal{A}w)$ *— свойство самосопряженности оператора;*
- $(\mathcal{A}v, v) > 0$, *если* $v \neq 0$.

Доказательство. Так как $\mathcal{A} = -\Lambda_1 - \Lambda_2$, достаточно доказать утверждения для каждого из операторов $-\Lambda_i$. Проведем доказательство для оператора $-\Lambda_2$, так как для второго оператора доказательство аналогично.

$$\begin{aligned}
-(\Lambda_2 v, w) &= -\frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J-1} (v_{k\,j+1} - 2v_{k\,j} + v_{k\,j-1}) w_{k\,j} = \\
&= \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J-1} (v_{k\,j} - v_{k\,j-1}) w_{k\,j} - \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J-1} (v_{k\,j+1} - v_{k\,j}) w_{k\,j} = \\
&= \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J-1} (v_{k\,j} - v_{k\,j-1}) w_{k\,j} - \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{s=2}^{J} (v_{k\,s} - v_{k\,s-1}) w_{k\,s-1} = \\
&= \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J} (v_{k\,j} - v_{k\,j-1}) w_{k\,j} - \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J} (v_{k\,j} - v_{k\,j-1}) w_{k\,j-1} = \\
&= \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J} (v_{k\,j} - v_{k\,j-1})(w_{k\,j} - w_{k\,j-1}).
\end{aligned} \tag{8.166}$$

Последнее равенство получено с учетом того, что $w_{k\,0} = w_{k\,J} = 0$. В правую часть равенства (8.166) функции v и w входят симметрично, поэтому если их поменять местами, правая часть не изменится. Но тогда не изменится левая часть, что означает выполнение равенства $(\mathcal{A}v, w) = (v, \mathcal{A}w)$. Взяв в равенстве (8.166) $w = v \neq 0$, получим

$$(\mathcal{A}v, v) = \frac{h_1}{h_2} \sum_{k=1}^{K-1} \sum_{j=1}^{J} (v_{k\,j} - v_{k\,j-1})^2 \geq 0.$$

Среди слагаемых, входящих в сумму есть хотя бы одно ненулевое. Действительно, если бы все слагаемые были нулевыми, то имели бы равенства $0 = v_{k\,0} = v_{k\,1} = \ldots$. А это означало бы, что $v = 0$, что противоречит выбору v. Значит, $(\mathcal{A}v, v) > 0$. Лемма доказана.

Следствие. Из самосопряженности оператора $\mathcal{A}$ следует, что его собственные функции, соответствующие различным собственным числам ортогональны. Из неравенства $(\mathcal{A}v, v) > 0$ следует, что собственные числа положительны. Доказательства этих утверждений см. в [20].

Лемма 8.8.2 *Сеточные функции*

$$v^{(p\,q)} = \sin\frac{p\pi x^{(1)}}{l_1}\sin\frac{q\pi x^{(2)}}{l_2}, \quad (x^{(1)}, x^{(2)}) \in \omega, \;\; p = 1, \ldots, K-1, \; q = 1, \ldots, J-1 \quad (8.167)$$

являются собственными функциями оператора $\mathcal{A}$, соответствующими числам

$$\lambda^{(p\,q)} = \frac{4}{h_1^2}\sin^2\frac{p\pi}{2K} + \frac{4}{h_2^2}\sin^2\frac{q\pi}{2J}. \quad (8.168)$$

Доказательство. С учетом того, что $h_1 K = l_1$, имеем

$$\Lambda_1\Big(\sin\frac{p\pi x^{(1)}}{l_1}\Big) = \\ = \frac{1}{h_1^2}\Big(\sin\frac{p\pi(x^{(1)}+h_1)}{l_1} - 2\sin\frac{p\pi x^{(1)}}{l_1} + \sin\frac{p\pi(x^{(1)}-h_1)}{l_1}\Big) = \\ = \frac{2}{h_1^2}\Big(\cos\frac{p\pi h_1}{l_1} - 1\Big)\sin\frac{p\pi x^{(1)}}{l_1} = -\frac{4}{h_1^2}\sin^2\frac{p\pi}{2K}\cdot\sin\frac{p\pi x^{(1)}}{l_1}.$$

Тогда

$$-\Lambda_1 v^{(p\,q)} = -\Lambda_1\Big(\sin\frac{p\pi x^{(1)}}{l_1}\Big)\sin\frac{q\pi x^{(2)}}{l_2} = \\ = \frac{4}{h_1^2}\sin^2\frac{p\pi}{2K}\cdot\sin\frac{p\pi x^{(1)}}{l_1}\sin\frac{q\pi x^{(2)}}{l_2} = \frac{4}{h_1^2}\sin^2\frac{p\pi}{2K}v^{(p\,q)}.$$

Аналогичное равенство получается для Λ_2. Из двух равенств следует утверждение леммы.

Лемма 8.8.3 *Сеточные функции (8.167) образуют ортогональный базис в пространстве $\mathcal{H}$.*

Доказательство. Количество функций $v^{(k\,j)}$ совпадает с размерностью пространства. Поэтому для доказательства того, что эти функции образуют базис, достаточно показать, что они линейно независимы. Из линейной алгебры известен факт — если имеется некоторое множество попарно ортогональных векторов, то они линейно независимы. Ортогональность функций $v^{(k\,j)}$ следует из леммы 8.8.2 и следствия из леммы 8.8.1.

Лемма 8.8.4 *Число обусловленности $M_{\mathcal{A}}$ матрицы, соответствующей оператору $\mathcal{A}$ равно*

$$M_{\mathcal{A}} = \frac{\frac{4}{h_1^2}\cos^2\frac{\pi}{2K} + \frac{4}{h_2^2}\cos^2\frac{\pi}{2J}}{\frac{4}{h_1^2}\sin^2\frac{\pi}{2K} + \frac{4}{h_2^2}\sin^2\frac{\pi}{2J}}, \quad (8.169)$$

если в качестве нормы в пространстве сеточных функций выбрать $\|v\| = \sqrt{(v, v)}$.

Доказательство. В параграфе 2.1.4 было доказано, что если в пространстве существует базис из собственных векторов матрицы и норма вычисляется как корень из скалярного произведения векторов, то число обусловленности матрицы равно отношению наибольшего по модулю собственного числа $\lambda_{\max}$ к собственному числу, модуль которого минимален $\lambda_{\min}$. Из условия данной леммы, а также лемм 8.8.2, 8.8.3 следует, что все эти требования выполнены. Поэтому доказываемое утверждение справедливо, так как согласно формуле (8.168)

$$\begin{aligned}\lambda_{\min} =&\lambda_{1\,1} = \frac{4}{h_1^2}\sin^2\frac{\pi}{2K} + \frac{4}{h_2^2}\sin^2\frac{\pi}{2J},\\ \lambda_{\max} =&\lambda_{K-1\,J-1} = \frac{4}{h_1^2}\sin^2\frac{(K-1)\pi}{2K} + \frac{4}{h_2^2}\sin^2\frac{(J-1)\pi}{2J} = \frac{4}{h_1^2}\cos^2\frac{\pi}{2K} + \frac{4}{h_2^2}\cos^2\frac{\pi}{2J}.\end{aligned} \tag{8.170}$$

Лемма доказана.

Если, например, $K = J$ и $h_1 = h_2$, то $M_{\mathcal{A}} = \operatorname{ctg}^2\dfrac{\pi}{2K} \approx \dfrac{4K^2}{\pi^2}$. Отсюда видно, что с уменьшением шага сетки число обусловленности растет, причем уменьшение шага, например, в 10 раз приводит к увеличению числа обусловленности в 100 раз.

8.8.2 Метод простой итерации

Перейдем теперь непосредственно к рассмотрению методов нахождения решения системы (8.161). Начнем с метода простой итерации. В пункте 2.2.2 рассматривался итерационный метод, который записывается в виде

$$y^{n+1} = y^n - \tau(\mathcal{A}y^n - b), \qquad y^0 \text{ произвольно.} \tag{8.171}$$

Здесь n — номер итерации, τ — итерационный параметр. было доказано, что если собственные числа матрицы, соответствующей оператору $\mathcal{A}$ положительны, то метод сходится при условии, что $0 < \tau\lambda_{max} < 2$, а при $\tau_0 = 2/(\lambda_{max} + \lambda_{min})$ скорость сходимости максимальная. В формулах (8.170) записаны выражения для λ_{max}, λ_{min}, учитывая их, имеем

$$\tau_0 = \frac{h_1^2h_2^2}{2(h_1^2 + h_2^2)}.$$

В частности, при $h_1 = h_2 = h$ выражение для τ_0 упрощается $\tau_0 = h^2/4$.

Учитывая определение оператора $\mathcal{A}$, итерационный процесс (8.171) с оптимальным значение параметра τ для решения системы уравнений (8.153) имеет вид

$$y_{k\,j}^{n+1} = y_{k\,j}^n + \frac{h_1^2h_2^2}{2(h_1^2 + h_2^2)}\left(\frac{y_{k+1\,j}^n - 2y_{k\,j}^n + y_{k-1\,j}^n}{h_1^2} + \frac{y_{k\,j+1}^n - 2y_{k\,j}^n + y_{k\,j-1}^n}{h_2^2} + f_{k\,j}\right).$$

Приводя подобные, имеем

$$\begin{aligned} y_{k\,j}^{n+1} = \frac{h_1^2h_2^2}{2(h_1^2 + h_2^2)}&\left(\frac{y_{k+1\,j}^n + y_{k-1\,j}^n}{h_1^2} + \frac{y_{k\,j+1}^n + y_{k\,j-1}^n}{h_2^2} + f_{k\,j}\right),\\ &k = 1,\ldots,K-1, \quad j = 1,\ldots,J-1,\\ y_{0\,j}^{n+1} = g(0, x_j^{(2)}), \quad y_{K\,j}^{n+1} = g(l_1, x_j^{(2)}), \quad &y_{k\,0}^{n+1} = g(x_k^{(1)}, 0), \quad y_{k\,J}^{n+1} = g(x_k^{(1)}, l_2),\\ &y_{k\,j}^0 \text{ произвольно.}\end{aligned} \tag{8.172}$$

В том случае, когда $h_1 = h_2 = h$ (8.172) упрощается

$$y_{k\,j}^{n+1} = \frac{1}{4}(y_{k+1\,j}^{n} + y_{k-1\,j}^{n} + y_{k\,j+1}^{n} + y_{k\,j-1}^{n}) + \frac{h^2}{4} f_{k\,j}. \tag{8.173}$$

Итерационный метод (8.173) иногда называют **методом Либмана**.

Определим число итераций n, которые необходимо совершить для того, чтобы

$$\|y^n - y\| \le \varepsilon \|y^0 - y\|. \tag{8.174}$$

Разложим $y^n - y$ по базису, состоящему из собственных функций $v^{(p\,q)}$ оператора $\mathcal{A}$. Пусть

$$y^n - y = \sum_{p=1}^{K-1} \sum_{q=1}^{J-1} \alpha_{p\,q}^{(n)} v^{(p\,q)}.$$

Здесь $\alpha_{p\,q}^{(n)}$ — коэффициенты разложения. Так как $y^n - y = (E - \tau_0 \mathcal{A})(y^{n-1} - y)$, где E — единичный оператор, имеем

$$\|y^n - y\|^2 = \|(E - \tau_0 \mathcal{A})(y^{n-1} - y)\|^2 = \left\| \sum_{p=1}^{K-1} \sum_{q=1}^{J-1} \alpha_{p\,q}^{(n-1)} (E - \tau_0 \mathcal{A}) v^{(p\,q)} \right\|^2 =$$

$$= \left\| \sum_{p=1}^{K-1} \sum_{q=1}^{J-1} \alpha_{p\,q}^{(n-1)} (1 - \tau_0 \lambda^{(p\,q)}) v^{(p\,q)} \right\|^2 = \sum_{p=1}^{K-1} \sum_{q=1}^{J-1} \left(\alpha_{p\,q}^{(n-1)}\right)^2 (1 - \tau_0 \lambda^{(p\,q)})^2 \|v^{(p\,q)}\|^2 \le$$

$$\le (1 - \tau_0 \lambda_{\min})^2 \sum_{p=1}^{K-1} \sum_{q=1}^{J-1} \left(\alpha_{p\,q}^{(n-1)}\right)^2 \|v^{(p\,q)}\|^2 = (1 - \tau_0 \lambda_{\min})^2 \|y^{n-1} - y\|^2.$$

Здесь при выводе неравенства использовалась теорема Пифагора [20], согласно которой квадрат нормы суммы ортогональных функций равен сумме квадратов норм этих функций. Заметим, что число $\rho = 1 - \tau_0 \lambda_{\min}$ является наибольшим по модулю собственным числом оператора $E - \tau_0 \mathcal{A}$. Легко теперь получить, что $\|y^n - y\| \le \rho^n \|y^0 - y\|$.

Для выполнения неравенства (8.174) достаточно потребовать, чтобы $\rho^n \le \varepsilon$ или

$$n \ge \frac{\ln \varepsilon}{\ln \rho}.$$

Заметим, что

$$\rho = 1 - \frac{2}{\lambda_{\min} + \lambda_{\max}} \lambda_{\min} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\min} + \lambda_{\max}} = \frac{1 - M_{\mathcal{A}}^{-1}}{1 + M_{\mathcal{A}}^{-1}},$$

где $M_{\mathcal{A}} = \lambda_{\max}/\lambda_{\min}$ — число обусловленности матрицы, соответствующей оператору $\mathcal{A}$, которое, как отмечалось, велико. Поэтому

$$\ln \rho = \ln (1 - M_{\mathcal{A}}^{-1}) - \ln (1 + M_{\mathcal{A}}^{-1}) \approx -2 M_{\mathcal{A}}^{-1}.$$

Таким образом,

$$n \ge \frac{\ln \varepsilon}{-2 M_{\mathcal{A}}^{-1}}. \tag{8.175}$$

Если, например, $K = J$ и $h_1 = h_2$, то уже отмечалось, что $M_{\mathcal{A}} \approx \dfrac{4K^2}{\pi^2}$ и

$$n \ge \frac{-2K^2 \ln \varepsilon}{\pi^2}. \tag{8.176}$$

Следует сделать одно замечание, касающееся точности ε, с которой решается уравнение (8.161). Это уравнение аппроксимирует дифференциальную задачу, решение u которой ищется. Погрешность аппроксимации, например, для прямоугольной области равна $O(h^2)$, поэтому $u-y^n=(u-y)+(y-y^n)=O(h^2)+O(\varepsilon)$. Следовательно, целесообразно выполнение равенства $\varepsilon=O(h^2)$.

Заметим, что в том случае, когда область, в которой ищется решение не прямоугольная, расчет значений приближения на $(n+1)$-ой итерации во внутренних узлах сетки по-прежнему можно проводить по формуле (8.173) для случая равных шагов или (8.172) для различных шагов сетки. В граничных же узлах значение решения известно. Следует, однако отметить, что для непрямоугольной области, собственные числа оператора $\mathcal{A}$ другие, поэтому оптимальное значение итерационного параметра τ не равно тому, которым пользовались при получении (8.173). Поэтому сходиться метод будет несколько медленнее, чем для прямоугольной области.

Итерационный метод (8.172) может быть получен основываясь на методе Якоби. Для этого, согласно идее, заложенной при получении расчетных формул метода, надо переписать уравнение (8.153), выразив из него $y_{k\,j}$. После этого необходимо $y_{k\,j}$ брать на $n+1$-ой итерации, а значения y в остальных узлах — на n-ой итерации. В итоге получится в точности формула (8.172). Поэтому иногда метод (8.172) называют методом Якоби.

Известно, что без увеличения объема вычислений на одной итерации можно добиться увеличения скорости сходимости, применяя к системе (8.153) метод Зейделя. Соответствующая формула будет иметь вид

$$y_{k\,j}^{n+1}=\frac{h_1^2h_2^2}{2(h_1^2+h_2^2)}\left(\frac{y_{k+1\,j}^{n}+y_{k-1\,j}^{n+1}}{h_1^2}+\frac{y_{k\,j+1}^{n}+y_{k\,j-1}^{n+1}}{h_2^2}+f_{k\,j}\right),$$
$$k=1,\ldots,K-1,\quad j=1,\ldots,J-1, \tag{8.177}$$
$$y_{0\,j}^{n+1}=g(0,x_j^{(2)}),\quad y_{K\,j}^{n+1}=g(l_1,x_j^{(2)}),\quad y_{k\,0}^{n+1}=g(x_k^{(1)},0),\quad y_{k\,J}^{n+1}=g(x_k^{(1)},l_2),$$
$$y_{k\,j}^{0}\ \text{произвольно.}$$

Один из вариантов нахождения неизвестных значений $y_{k\,j}^{n+1}$ заключается в следующем: вычисляются значения $y_{k\,1}^{n+1}$ в порядке возрастания индекса k от 1 до $K-1$, после чего значение индекса j увеличивается на 1, процесс повторяется и так далее.

Доказывается, что число итераций, требуемое для получения неравенства(8.174) (см. [32]), примерно в два раза меньше, чем в методе Якоби. Считается, что такая скорость сходимости мала, вследствие чего методы Якоби и Зейделя применяются при решении серьезных задач не часто.

Можно ускорить процесс сходимости метода Зейделя, вводя в него **параметр релаксации** ω. Для этого обозначим результат вычислений по формуле (8.177) через $\widetilde{y}_{k\,j}^{n+1}$. Значение приближения на $(n+1)$-ой итерации будем вычислять по формуле

$$y_{k\,j}^{n+1}=(1-\omega)y_{k\,j}^{n}+\omega\widetilde{y}_{k\,j}^{n+1}. \tag{8.178}$$

Полученный метод называется **методом верхней релаксации**. Он сходится при условии $\omega\in(0,2)$[11]. Параметр релаксации выбирается таким образом, чтобы ускорить сходимость. Например см.[32], при $l_1=l_2=l$, $K=J$ оптимальное значение $\omega_0\approx\dfrac{2}{1+\pi/K}$, а число итераций для выполнения неравенства (8.174)

[11]Иногда, говоря о методе верхней релаксации, считают, что $\omega\in(1,2)$

$n \approx \dfrac{-2K\ln(\varepsilon)}{\pi}$. В отличии от методов Якоби и Зейделя, где $n = O(K^2)$, у метода релаксации $n = O(K)$.

Если область отлична от прямоугольника, вычисление оптимального значения параметра релаксации, как правило, невозможно. Тогда величину ω, которая обеспечивает ускорение итерационного процесса подбирают из промежутка $(1, 2)$ путем проведения вычислительных экспериментов.

8.8.3 Методы установления

Вернемся к итерационному методу (8.171) и перепишем его в виде

$$\frac{y^{n+1} - y^n}{\tau} = \mathcal{A}y^n - b.$$

Если же вспомнить определение оператора $\mathcal{A}$ и сеточной функции b, то этот же итерационный метод запишется следующим образом

$$\frac{y^{n+1} - y^n}{\tau} = \Lambda y^n + f, \quad y^{n+1}\big|_\gamma = g, \quad y^0 \text{ задано произвольным образом.}$$

В этой записи итерационный метод напоминает явную разностную схему для решения первой краевой задачи для нестационарного уравнения теплопроводности, которое изучалось в параграфе 8.6. Такое совпадение не случайно, оно имеет как физическое, так и математической толкование. Начнем с физической интерпретации. Поместим в комнату два одинаковых тела, одно из которых извлечем из холодильника, а другое из нагревательного шкафа. Тогда через некоторый промежуток времени в обеих телах установится одинаковое распределение температуры, не зависящее от начального состояния тел и определяемое только температурой в комнате. Таким образом можно говорить о том, что процесс нагревания (остывания) тел установился (сошелся), температура перестает меняться и принимает стационарное значение.

Перейдем к математическому описанию явления. Процесс нагревания (остывания) тела описывается нестационарным уравнением теплопроводности, распределение температуры, которое установилось в теле — уравнением Пуассона. Таким образом имеем две задачи (для простоты будем считать тело плоской пластиной квадратной формы со стороной 1)

$$\frac{\partial \widetilde{u}}{\partial t} = \frac{\partial^2 \widetilde{u}}{\partial (x^{(1)})^2} + \frac{\partial^2 \widetilde{u}}{\partial (x^{(2)})^2} + f, \quad 0 < x^{(i)} < 1, \ i = 1, 2, \quad \widetilde{u}\big|_\Gamma = g, \quad \widetilde{u}\big|_{t=0} = \widetilde{u}_0 \tag{8.179}$$

и

$$\frac{\partial^2 \overline{u}}{\partial (x^{(1)})^2} + \frac{\partial^2 \overline{u}}{\partial (x^{(2)})^2} = -f, \quad 0 < x^{(i)} < 1, \ i = 1, 2, \quad \overline{u}\big|_\Gamma = g, \tag{8.180}$$

где Γ — граница квадрата.

Покажем, что $\lim\limits_{t\to\infty} \widetilde{u}(t, \mathbf{x}) = \overline{u}(\mathbf{x})$, $\mathbf{x} = (x^{(1)}, x^{(2)})$. Пусть $u = \widetilde{u} - \overline{u}$. Легко заметить, что u является решением задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial (x^{(1)})^2} + \frac{\partial^2 u}{\partial (x^{(2)})^2}, \quad 0 < x^{(i)} < 1, \ i = 1, 2, \quad u\big|_\Gamma = 0, \quad u\big|_{t=0} = \widetilde{u}_0 - \overline{u} = u_0. \tag{8.181}$$

Задачу (8.181) можно решить методом разделения переменных (Фурье) [20]. В результате имеем

$$u = \sum_{m_1=1}^{\infty} \sum_{m_2=1}^{\infty} a_{m_1 m_2} e^{-(m_1+m_2)t} \sin m_1 \pi x^{(1)} \sin m_2 \pi x^{(2)}, \tag{8.182}$$

где $a_{m_1\,m_2}$ — коэффициенты разложения функции u_0 в ряд Фурье по синусам, то есть

$$u_0 = \sum_{m_1=1}^{\infty}\sum_{m_2=1}^{\infty} a_{m_1\,m_2}\sin m_1\pi x^{(1)}\sin m_2\pi x^{(2)}.$$

Из формулы (8.182) следует, что $u \to 0$ при $t \to \infty$. Полученное утверждение означает, что для того чтобы решить стационарную задачу (8.180), можно решать нестационарную задаче до того момента времени t, пока ее решение не перестанет меняться в пределах заданной точности. В этом суть метода **установления**. Итерационный метод (8.171) это как раз и есть пример счета на установление.

В более общем виде одношаговый итерационный метод для решения системы уравнений (8.161) можно записать следующим образом

$$\mathcal{B}_n\frac{y^{n+1}-y^n}{\tau_n} = b - \mathcal{A}y^n. \tag{8.183}$$

Легко заметить, что решение y системы (8.161) удовлетворяет (8.183) при любых $\mathcal{B}$ и τ_n. Вместо сходимости y^n к y удобнее говорить о сходимости $v^n = y^n - y$ к нулю. При этом для v^n можно записать следующий итерацоионный метод

$$\mathcal{B}_n\frac{v^{n+1}-v^n}{\tau_n} = -\mathcal{A}v^n, \quad v^0 \ \text{ произвольная сеточная функция.} \tag{8.184}$$

Как уже отмечалось, (8.183) внешне напоминает разностную схему для уравнения

$$\mathcal{B}\frac{\partial u}{\partial t} = -\mathcal{A}u + b.$$

Параметр τ_n может рассматриваться как шаг по фиктивному времени $t_n = \tau_0 + \cdots + \tau_n$. Однако между разностной схемой для нестационарного уравнения и итерационным методом (8.183) есть и определенные различия. Если в разностной схеме шаг τ выбирается из требований аппроксимации и устойчивости, то в итерационной схеме выбор оператора $\mathcal{B}_n$ и итерационного параметра τ_n подчинены условию сходимости итераций и минимума арифметических операций для нахождения приближенного решения с заданной точностью.

Сравнение итерационных методов происходит на основании сравнения числа арифметических операций $Q(\varepsilon)$, требуемых для получения заданной точности ε при произвольном начальном приближении y^0. Если обозначить через q_n число арифметических операций на n-ой итерации, то $Q(\varepsilon) = q_1 + \cdots + q_{n(\varepsilon)}$, где $n(\varepsilon)$ — минимальное число итераций, требуемых для достижения точности ε. В случае, когда все итерации одинаковы по трудоемкости, $q_n = q$ и $Q(\varepsilon) = qn(\varepsilon)$. В рассмотренном в предыдущем пункте методе простой итерации оператор $\mathcal{B}$ был единичным, итерационный параметр τ_n не зависел от номера итерации, число q было пропорционально числу внутренних узлов сетки, а $n(\varepsilon) = O(1/h^2\ln(1/\varepsilon))$.

Если переписать (8.183) в виде

$$\mathcal{B}_n y^{n+1} = \mathcal{B}_n y^n + \tau_n(b - \mathcal{A}y^n),$$

то становится ясным, что величина q_n во многом зависит от оператора $\mathcal{B}_n$. Его выбирают либо единичным и тогда метод называют явным, либо легко обратимым.

Для числовой характеристики скорости сходимости методов часто используется величина S, называемая **асимптотической скоростью сходимости**. Она определяется следующим образом:

$$S = -\lim_{n\to\infty}\frac{1}{n}\ln\Big(\sup_{v^0\neq 0}\frac{\|v^n\|}{\|v^0\|}\Big).$$

где $v^n = y^n - y$.

Для того, чтобы пояснить смысл числа S будем считать что $\mathcal{B}$ и τ не зависят от номера итерации. Такие итерационные методы называют **стационарными**. Для разности v^n из (8.184) следует $v^{n+1} = \mathcal{U}v^n$, где $\mathcal{U} = E - \tau\mathcal{B}^{-1}\mathcal{A}$, а E — единичный оператор. Тогда

$$\sup_{v^0 \neq 0} \frac{\|v^n\|}{\|v^0\|} = \sup_{v^0 \neq 0} \frac{\|\mathcal{U}^n v^0\|}{\|v^0\|} = \|\mathcal{U}^n\|.$$

Справедливо следующее утверждение, которое примем без доказательства: для любого линейного оператора $\mathcal{M}$, действующего в конечномерном линейном нормированном пространстве, выполняется равенство

$$\lim_{n\to\infty} \|\mathcal{M}^n\|^{1/n} = \rho(\mathcal{M}),$$

где $\rho(\mathcal{M})$ — максимальное по модулю собственное число оператора $\mathcal{M}$.

Следовательно,

$$S = -\ln \rho(\mathcal{U}).$$

Для стационарного итерационного метода это равенство выбирают в качестве определения асимптотической скорости сходимости.

Для уменьшения ошибки при задании начального приближения в $1/\varepsilon$ раз имеем

$$\|v^n\| \leq \|\mathcal{U}^n\|\|v^0\| \leq \varepsilon\|v^0\|.$$

Следовательно, число

$$n \geq \frac{\ln \varepsilon}{\ln \|\mathcal{U}^n\|^{1/n}}.$$

Так как при больших значениях n выполняется приблизительное равенство

$$\|\mathcal{U}^n\|^{1/n} \approx \rho(\mathcal{U}),$$

имеем

$$n \geq \frac{\ln(1/\varepsilon)}{-\ln \rho(\mathcal{U})} = S^{-1}\ln(1/\varepsilon). \tag{8.185}$$

Заметим, наконец, что если норму ввести по формуле $\|v\| = \sqrt{(v,v)}$, то можно показать, что $\|\mathcal{U}^n\| = \rho^n(\mathcal{U})$. Тогда неравенство (8.185) справедливо без предположения о том, что n велико. Если взять теперь $1/\varepsilon = e$, то $n \geq S^{-1}$. Следовательно, число S^{-1} показывает сколько достаточно сделать итераций, чтобы уменьшить норму ошибки в e раз. Чем S^{-1} меньше, то есть S больше, тем метод лучше.

Из (8.175) следует, что для метода простой итерации $S \approx \dfrac{2}{M_{\mathcal{A}}}$, где $M_{\mathcal{A}}$ — число обусловленности матрицы, соответствующей оператору $\mathcal{A}$. Отмечалось, что для метода Зейделя это число в два раза больше. Для метода верхней релаксации в [32] показано, что $S \approx \dfrac{1}{\sqrt{M_{\mathcal{A}}}}$.

В следующих пунктах будут приведены примеры методов вида (8.183). Для одного из них $\mathcal{B}$ — единичный оператор, а итерационный параметр меняется от шага к шагу, для другого итерационный параметр постоянный, а оператор $\mathcal{B}$ не единичный.

8.8.4 Метод минимальных невязок

Для оптимального выбора итерационного параметра в рассмотренных выше методах необходима информация о границах промежутка, на котором расположены собственные числа оператора. В рассматриваемом в этом пункта методе такая информация не нужна. Будем считать, что оператор $\mathcal{A}$ положительно определенный. Напомним, что согласно лемме 8.8.1 для прямоугольной области это свойство выполнено.

Запишем итерационный метод в виде

$$y^{n+1} = y^n - \tau_n(\mathcal{A}y^n - b). \tag{8.186}$$

Обозначим через ξ^n невязку, то есть $\xi^n = \mathcal{A}y^n - b$.

Тогда, подействовав на обе части равенства (8.186) оператором $\mathcal{A}$ и вычитая затем из обеих частей равенства b, получим соотношение, которому удовлетворяет невязка

$$\xi^{n+1} = (E - \tau_n\mathcal{A})\xi^n. \tag{8.187}$$

Для проведения итераций необходимо выбрать значение параметра τ_n. Так как при стремлении y^n к y невязка стремится к нулю, будем выбирать τ_n таким образом, чтобы на каждом шаге невязка была как можно меньше. Для этого запишем

$$\|\xi^{n+1}\|^2 = \big((E-\tau_n\mathcal{A})\xi^n, (E-\tau_n\mathcal{A})\xi^n\big) = (\xi^n,\xi^n) - 2\tau_n(\mathcal{A}\xi^n,\xi^n) + \tau_n^2(\mathcal{A}\xi^n,\mathcal{A}\xi^n). \tag{8.188}$$

Продифференцировав выражение, стоящее в правой части равенства, по τ_n и приравняв производную нулю, найдем интересующее нас значение

$$\tau_n = \frac{(\mathcal{A}\xi^n,\xi^n)}{(\mathcal{A}\xi^n,\mathcal{A}\xi^n)}.$$

При таком выборе τ_n имеем из (8.188)

$$\|\xi^{n+1}\| = \sqrt{1 - \frac{(\mathcal{A}\xi^n,\xi^n)^2}{(\mathcal{A}\xi^n,\mathcal{A}\xi^n)(\xi^n,\xi^n)}}\,\|\xi^n\| = \rho_n\|\xi^n\|.$$

Очевидно, что $\rho_n < 1$. Можно показать, что асимптотическая скорость сходимости метода $S \approx \dfrac{2}{M_{\mathcal{A}}}$. Формально асимптотическая скорость сходимости этого метода такая же как и для метода простой итерации. Однако, численные эксперименты показывают, что на нескольких первых итерациях стремление приближенного решения к точному происходит быстрее, чем у многих других методов. Это означает, что асимптотическая оценка слишком груба.

В заключении этого пункта выпишем расчетные формулы метода для случая решения задачи Дирихле для уравнения Пуассона в прямоугольнике с помощью разностной схемы (8.153),(8.155). Начальное приближение — сеточная функция, определенная произвольным образом на ω и принимающая на границе γ заданное значение

g. На остальных итерациях вычисления производятся по следующим формулам:

$$\xi^n = -\frac{y^n_{k+1\,j} - 2y^n_{k\,j} + y^n_{k-1\,j}}{h_1^2} - \frac{y^n_{k\,j+1} - 2y^n_{k\,j} + y^n_{k\,j-1}}{h_2^2} - f_{k\,j},$$

$$k = 1, \ldots, K-1, \quad j = 1, \ldots, J-1,$$

$$y^n\big|_\gamma = g, \quad \xi^n\big|_\gamma = 0,$$

$$\mathcal{A}\xi^n_{k\,j} = -\frac{\xi^n_{k+1\,j} - 2\xi^n_{k\,j} + \xi^n_{k-1\,j}}{h_1^2} - \frac{\xi^n_{k\,j+1} - 2\xi^n_{k\,j} + \xi^n_{k\,j-1}}{h_2^2},$$

$$k = 1, \ldots, K-1, \quad j = 1, \ldots, J-1,$$

$$\tau_n^u = h_1 h_2 \sum_{k=1}^{K-1}\sum_{j=1}^{J-1} \mathcal{A}\xi^n_{k\,j} \cdot \xi_{k\,j}, \quad \tau_n^d = h_1 h_2 \sum_{k=1}^{K-1}\sum_{j=1}^{J-1} \mathcal{A}\xi^n_{k\,j} \cdot \mathcal{A}\xi_{k\,j}, \quad \tau_n = \frac{\tau_n^u}{\tau_n^d},$$

$$y^{n+1}_{k\,j} = y^n_{k\,j} - \tau_n \xi^n, \quad k = 1, \ldots, K-1, \quad j = 1, \ldots, J-1.$$

8.8.5 Итерационная схема переменных направлений

В пункте 8.8.3 отмечалось, что решение стационарной задачи может быть сведено к решению нестационарной задачи. Для нестационарных задач хорошо зарекомендовали себя схемы, когда переход от одного шага к другому осуществляется в несколько приемов. Рассмотрим итерационный метод, аналогичный методу переменных направлений.

Пусть оператор $\mathcal{A}$ представим в виде $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$ и начальное приближение y^0 задано произвольно. Тогда переход от итерации n к итерации $n+1$ осуществляется в два этапа. На первом находится промежуточное значение $y^{n+1/2}$ из уравнения

$$\frac{y^{n+1/2} - y^n}{\tau} + \mathcal{A}_1 y^{n+1/2} + \mathcal{A}_2 y^n = b. \tag{8.189}$$

На втором этапе находится y^{n+1}

$$\frac{y^{n+1} - y^{n+1/2}}{\tau} + \mathcal{A}_1 y^{n+1/2} + \mathcal{A}_2 y^{n+1} = b. \tag{8.190}$$

Выбор итерационного параметра τ обсудим позднее.

Итерационная схема (8.189),(8.190) может быть представлена в виде одношагового итерационного метода (8.183). Для этого надо исключить из (8.189),(8.190) промежуточное значение $y^{n+1/2}$. Достаточно вычесть из равенства (8.189) равенство (8.190), выразить $y^{n+1/2}$ и подставить его в (8.189). Тогда после приведения подобных получится равенство

$$\frac{1}{2}(E + \tau\mathcal{A}_1)(E + \tau\mathcal{A}_2)\frac{y^{n+1} - y^n}{\tau} = b - \mathcal{A}y^n. \tag{8.191}$$

Следовательно, (8.191) имеет вид (8.183) с оператором $\mathcal{B} = \frac{1}{2}(E + \tau\mathcal{A}_1)(E + \tau\mathcal{A}_2)$.

Займемся вопросом сходимости и выбора параметра τ. Будем считать, что операторы $\mathcal{A}_1, \mathcal{A}_2$ положительно определены и перестановочны.

Перепишем (8.189),(8.190) в виде

$$(E + \tau\mathcal{A}_1)y^{n+1/2} = (E - \tau\mathcal{A}_2)y^n + \tau b, \quad (E + \tau\mathcal{A}_2)y^{n+1} = (E - \tau\mathcal{A}_1)y^{n+1/2} + \tau b. \tag{8.192}$$

Из этих уравнений можно найти сначала $y^{n+1/2}$, а затем y^{n+1}. Заметим, что

$$(E + \tau\mathcal{A}_1)y = (E - \tau\mathcal{A}_2)y + \tau b, \quad (E + \tau\mathcal{A}_2)y = (E - \tau\mathcal{A}_1)y + \tau b, \tag{8.193}$$

где y — решение уравнения (8.161). Поэтому, вычитая (8.193) из (8.192) и обозначая $v^{n+1/2} = y^{n+1/2} - y,\ v^n = y^n - y$, получим.

$$(E + \tau\mathcal{A}_1)v^{n+1/2} = (E - \tau\mathcal{A}_2)v^n,\quad (E + \tau\mathcal{A}_2)v^{n+1} = (E - \tau\mathcal{A}_1)v^{n+1/2}. \tag{8.194}$$

Если подействовать на первое из этих уравнений оператором $E - \tau\mathcal{A}_1$, а на второе оператором $E + \tau\mathcal{A}_1$ и сложить результаты, то слагаемые, содержащие $v^{n+1/2}$ исчезнут. В результате получим

$$(E + \tau\mathcal{A}_2)(E + \tau\mathcal{A}_1)v^{n+1} = (E - \tau\mathcal{A}_1)(E - \tau\mathcal{A}_2)v^n. \tag{8.195}$$

Из равенства (8.195) следует, что

$$v^{n+1} = (E + \tau\mathcal{A}_1)^{-1}(E + \tau\mathcal{A}_2)^{-1}(E - \tau\mathcal{A}_1)(E - \tau\mathcal{A}_2)z^n = \mathcal{E}v^n. \tag{8.196}$$

В параграфе 2.2 было доказано, что для сходимости необходимо, чтобы все собственные числа оператора $\mathcal{E}$ были по модулю меньше 1. Из линейной алгебры известно, что если операторы самосопряженные и перестановочны, то они имеют общий базис, состоящий из собственных векторов. Отсюда следует, что

$$\lambda_s(\mathcal{E}) = \frac{(1 - \tau\lambda_s(\mathcal{A}_1))(1 - \tau\lambda_s(\mathcal{A}_2))}{(1 + \tau\lambda_s(\mathcal{A}_1))(1 + \tau\lambda_s(\mathcal{A}_2))}, \tag{8.197}$$

где $\lambda_s(\mathcal{E}), \lambda_s(\mathcal{A}_1), \lambda_s(\mathcal{A}_2)$ — собственные числа операторов $\mathcal{E}, \mathcal{A}_1, \mathcal{A}_2$ соответствующих одному и тому же собственному вектору с номером s. Поскольку операторы $\mathcal{A}_1, \mathcal{A}_2$ положительно определены, числа $\lambda_s(\mathcal{A}_1), \lambda_s(\mathcal{A}_2)$ положительны. Тогда из (8.197) следует, что при любом $\tau > 0$ выполняется неравенство $|\lambda_s(\mathcal{E})| < 1$, и, значит, итерации сходятся. Таким образом, доказана теорема:

Теорема 8.8.1 *Если $\mathcal{A}_1, \mathcal{A}_2$ — положительно определенные перестановочные операторы, то при любом $\tau > 0$ итерационная схема переменных направлений (8.189), (8.190) сходится.*

Осталось подобрать значение τ, обеспечивающее быструю сходимость метода.

Как уже отмечалось ранее, скорость сходимости определяется наибольшим по модулю собственным числом оператора $\mathcal{E}$. Из выражения (8.197) для собственных чисел оператора $\mathcal{E}$ следует, что оно представимо в виде произведения $F(x)F(z)$, где $F(x) = (1 - x)/(1 + x)$. Так как $F'(x) = -2/(1 + x)^2 < 0$, функция $F(x)$ является монотонно убывающей. Поэтому свое экстремальное значение эта функция принимает на концах промежутка изменения аргумента. Если предположить, что известны границы $\lambda_{\min}(\mathcal{A}_i),\ \lambda_{\max}(\mathcal{A}_i)$ изменения собственных чисел операторов $\mathcal{A}_i, i = 1, 2$, то

$$\alpha_i = \frac{1 - \tau\lambda_{\max}(\mathcal{A}_i)}{1 + \tau\lambda_{\max}(\mathcal{A}_i)} \le \frac{1 - \tau\lambda_s(\mathcal{A}_i)}{1 + \tau\lambda_s(\mathcal{A}_i)} \le \frac{1 - \tau\lambda_{\min}(\mathcal{A}_i)}{1 + \tau\lambda_{\min}(\mathcal{A}_i)} = \beta_i.$$

Увеличение τ вызывает сдвиг точек α_i, β_i влево вдоль числовой прямой[12], при уменьшении τ точки α_i, β_i сдвигаются вправо. Поэтому величина $\max\limits_s |\lambda_s(\mathcal{A}_i)|$ будет наименьшей, когда $-\alpha_i = \beta_i$. Отсюда следует, что

$$\tau_i = \frac{1}{\sqrt{\lambda_{\min}(\mathcal{A}_i)\lambda_{\max}(\mathcal{A}_i)}}. \tag{8.198}$$

[12]Предполагается, что на числовой прямой числа возрастают слева направо.

Мы имеем в своем распоряжении только один параметр τ. Поэтому придется сделать еще одно дополнительное предположение[13]: будем считать, что

$$\lambda_{\min}(\mathcal{A}_1) = \lambda_{\min}(\mathcal{A}_2) = \lambda_{\min}, \quad \lambda_{\max}(\mathcal{A}_1) = \lambda_{\max}(\mathcal{A}_2) = \lambda_{\max}. \tag{8.199}$$

В этом случае при

$$\tau = \frac{1}{\sqrt{\lambda_{\min}\lambda_{\max}}} \tag{8.200}$$

имеем

$$|\lambda_s(\mathcal{E})| \le \left(\frac{1 - \sqrt{\lambda_{\min}/\lambda_{\max}}}{1 + \sqrt{\lambda_{\min}/\lambda_{\max}}}\right)^2 = \rho. \tag{8.201}$$

Ранее уже доказывалось, что если ввести норму $\|y\|^2 = (y, y)$, то $\|\mathcal{E}\| = \max\limits_s |\lambda_s(\mathcal{E})|$. Значит, $\|\mathcal{E}\| \le \rho$. Таким образом, доказана теорема:

Теорема 8.8.2 *Пусть $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$ и $\mathcal{A}_1, \mathcal{A}_2$ — положительно определенные, перестановочные операторы. Тогда итерационная схема (8.189),(8.190) сходится при любом положительном значении итерационного параметра τ. Если выполнено условие (8.199), то при выборе τ в соответствии с равенством (8.200) справедливо неравенство*

$$\|y^{n+1} - y\| \le \left(\frac{1 - \sqrt{\lambda_{\min}/\lambda_{\max}}}{1 + \sqrt{\lambda_{\min}/\lambda_{\max}}}\right)^2 \|y^n - y\|.$$

Замечание. В том случае, когда выполнены равенства (8.199), причем у каждой из пар собственных чисел $\lambda_{\min}(\mathcal{A}_1), \lambda_{\min}(\mathcal{A}_2)$ и $\lambda_{\max}(\mathcal{A}_1), \lambda_{\max}(\mathcal{A}_2)$ одинаковые собственные вектора (для каждой из пар, разумеется свой собственный вектор) выполняется равенство $\max\limits_s |\lambda_s(\mathcal{E})| = \rho$, поэтому $\|\mathcal{E}\| = \rho$. Кроме того, в этом случае $\lambda_{\min}(\mathcal{A}) = 2\lambda_{\min}$, $\lambda_{\max}(\mathcal{A}) = 2\lambda_{\max}$ и число обусловленности $M_{\mathcal{A}} = \lambda_{\max}/\lambda_{\min}$. При большом числе обусловленности имеем

$$\rho = \left(\frac{1 - \sqrt{1/M_{\mathcal{A}}}}{1 + \sqrt{1/M_{\mathcal{A}}}}\right)^2 \approx 1 - 4\sqrt{1/M_{\mathcal{A}}}.$$

Асимптотическая скорость сходимости итерационной схемы переменных направлений $S = -\ln\rho \approx 4/\sqrt{M_{\mathcal{A}}}$, что намного больше, чем у рассмотренных ранее методов.

Рассмотрим теперь итерационную схему применительно к задаче Дирихле для уравнения Пуассона решаемой с помощью разностной схемы (8.153),(8.155) при условии, что область Ω — квадрат, то есть $l_1 = l_2 = l$ и $K = J$. Последнее предположение означает, что шаги равны. Тогда $\mathcal{A}_i = -\Lambda_i$, выполняются условия, о которых говорится в замечании,

$$\lambda_{\min} = \frac{4}{h^2}\sin^2\frac{\pi}{2K}, \quad \lambda_{\max} = \frac{4}{h^2}\cos^2\frac{\pi}{2K}.$$

Тогда

$$\tau = \frac{h^2}{2\sin\pi/K} \approx \frac{hl}{2\pi}.$$

[13]В том случае, когда дополнительное предположение не делается, итерационный метод переменных направлений рассмотрен в [32].

Первое из уравнений (8.193) принимает вид

$$
\begin{aligned}
-\frac{\tau}{h^2}y^{n+1/2}_{k-1\,j} + \Big(1+2\frac{\tau}{h^2}\Big)y^{n+1/2}_{k\,j} - \frac{\tau}{h^2}y^{n+1/2}_{k+1\,j} = \\
= \frac{\tau}{h^2}y^{n}_{k\,j-1} + \Big(1-2\frac{\tau}{h^2}\Big)y^{n}_{k\,j} + \frac{\tau}{h^2}y^{n}_{k\,j+1} + \tau f(x^{(1)}_k, x^{(2)}_j), \qquad (8.202)\\
y^{n+1/2}_{0\,j} = g(0, x^{(2)}_j), \quad y^{n+1/2}_{K\,j} = g(l, x^{(2)}_j), \quad k,j = 1,\dots,K-1.
\end{aligned}
$$

Второе уравнение (8.193) запишется в виде

$$
\begin{aligned}
-\frac{\tau}{h^2}y^{n+1}_{k\,j-1} + \Big(1+2\frac{\tau}{h^2}\Big)y^{n+1}_{k\,j} - \frac{\tau}{h^2}y^{n+1}_{k\,j+1} = \\
= \frac{\tau}{h^2}y^{n+1/2}_{k-1\,j} + \Big(1-2\frac{\tau}{h^2}\Big)y^{n+1/2}_{k\,j} + \frac{\tau}{h^2}y^{n+1/2}_{k+1\,j} + \tau f(x^{(1)}_k, x^{(2)}_j), \qquad (8.203)\\
y^{n+1}_{k\,0} = g(x^{(1)}_k, 0), \quad y^{n+1}_{k\,K} = g(x^{(1)}_k, l), \quad k,j = 1,\dots,K-1.
\end{aligned}
$$

Каждая из систем (8.202), (8.203) решается методом прогонки.

Рассмотренными в этом параграфе итерационными схемами не ограничиваются методы решения сеточных уравнений. Весьма эффективным является попеременно-треугольный метод. Многие методы допускают модификацию, связанную с специальным набором параметров итерации. Однако все эти вопросы выходят за пределы данного учебного пособия. Подробное изложение большого числа методов решения разностных схем можно найти в книге [33].

8.9 ЗАДАЧИ К ГЛАВЕ 8

8.9.1 Примеры решения задач

1. Исследовать с помощью спектрального критерия устойчивости разностное уравнение

$$\frac{y^{n+1}_j - y^n_j}{\tau} + a\frac{y^{n+1}_j - y^{n+1}_{j-1}}{h} = 0, \quad a = const.$$

Решение. Заменим формально y^n_j на $e^{ij\varphi}$, а y^{n+1}_j на $\lambda e^{ij\varphi}$, подставим их в разностное уравнение и выразим λ. В результате получим

$$\lambda = \lambda(\varphi) = \frac{1}{1+r-re^{-i\varphi}} = \frac{1}{1+r(1-\cos\varphi)+ir\sin\varphi},$$

где $r = \dfrac{a\tau}{h}$. Отсюда следует, что

$$|\lambda|^2 = \big(1+r(1-\cos\varphi)^2 + r^2\sin^2\varphi\big)^{-1} = \big(1+4(r+r^2)\sin^2(\varphi/2)\big)^{-1}.$$

Следовательно, при $r \ge 0$ или $r \le -1$ имеем $|\lambda| \le 1$, значит необходимое условие устойчивости выполнено. Если же $-1 < r < 0$, то $r + r^2 < 0$ и

$$\max_{0\le\varphi\le 2\pi}|\lambda(\varphi)| = \frac{1}{|1+2r|} > 1.$$

В этом случае необходимое условие устойчивости не выполнено.

3. Для уравнения

$$\frac{\partial u}{\partial t} = 2\frac{\partial^2 u}{\partial x^2} + 5\frac{\partial^2 u}{\partial y^2} - \frac{\partial u}{\partial x} + xtu + x^2$$

построить экономичную разностную схему.

Решение Для построения экономичной разностной схемы воспользуемся методом, описанным в конце параграфа 8.6. Оператор $\mathcal{L}u = 2\dfrac{\partial^2 u}{\partial x^2} + 5\dfrac{\partial^2 u}{\partial y^2} - \dfrac{\partial u}{\partial x} + xtu$ представим в виде суммы двух операторов $\mathcal{L}_1$ и $\mathcal{L}_2$, где

$$\mathcal{L}_1 u = 2\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}, \qquad \mathcal{L}_2 u = 5\frac{\partial^2 u}{\partial y^2} + xtu.$$

Заменим теперь исходное уравнение на два уравнения

$$\frac{1}{2}\frac{\partial u}{\partial t} = 2\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x} + x^2, \quad \frac{1}{2}\frac{\partial u}{\partial t} = 5\frac{\partial^2 u}{\partial y^2} + xtu. \tag{8.204}$$

Заметим, что если сложить эти два уравнения, то получится исходное уравнение, а каждое из уравнений обладает той особенностью, что помимо производной по переменной t содержит еще производную только по одной переменной.

Пусть $t_n = n\tau$. Для того, чтобы найти решение на промежутке (t_n, t_{n+1}) введем $t_{n+1/2} = t_n + \tau/2$ и на промежутке $(t_n, t_{n+1/2})$ выберем первое уравнение (8.204), а на $(t_{n+1/2}, t_{n+1})$ — второе уравнение (8.204). Для каждого из этих уравнений запишем теперь неявную разностную схему

$$\frac{z_{kl}^{n+1/2} - z_{kl}^{n}}{\tau} = 2\frac{z_{k+1\,l}^{n+1/2} - 2z_{kl}^{n+1/2} + z_{k-1\,l}^{n+1/2}}{h_1^2} - \frac{z_{k+1\,l}^{n+1/2} - z_{k-1\,l}^{n+1/2}}{2h_1} + x_k^2,$$
$$\frac{z_{kl}^{n+1} - z_{kl}^{n+1/2}}{\tau} = 5\frac{z_{k\,l+1}^{n+1} - 2z_{kl}^{n+1} + z_{k\,l-1}^{n+1}}{h_2^2} + x_k t_{n+1} z_{kl}^{n+1}.$$

Построенная разностная схема является искомой схемой расщепления.

Заметим, что разбиение оператора $\mathcal{L}$ на сумму двух операторов можно было произвести и другим способом. Важно только, что один из операторов содержит дифференцирование по переменной x, а другой — по y. Таким образом, каждое из уравнений можно будет решать как одномерное. В связи с этим фактом, построенная разностная схема называется локально-одномерной.

4. Дана дифференциальная задача

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad (x,y) \in (0,1)\times(0,1), \quad t \in (0,T],$$
$$u(0,x,y) = f(x,y),$$
$$u(t,0,y) = u(t,1,y) = \mu_1(t,y), \qquad u(t,x,0) = u(t,x,1) = \mu_2(t,x)$$

и разностная схема

$$\Big(E - \frac{\tau}{2}\Lambda_1\Big) z^{n+1/4} = (\Lambda_1 + \Lambda_2) z^n, \quad \Big(E - \frac{\tau}{2}\Lambda_2\Big) z^{n+1/2} = z^{n+1/4}, \quad z^{n+1} = z^n + \tau z^{n+1/2},$$
$$(\Lambda_1 z)_{kl} = \frac{z_{k+1\,l} - 2z_{kl} + z_{k-1\,l}}{h^2}, \quad (\Lambda_2 z)_{kl} = \frac{z_{k\,l+1} - 2z_{kl} + z_{k\,l-1}}{h^2},$$
$$k,l = 1,2\ldots,M-1, \quad Mh = 1, \quad n = 0,1,\ldots,N-1, \quad N\tau = T,$$
$$z_{kl}^0 = f(kh,lh), \quad z_{0l}^n = z_{Ml}^n = \mu_1(\tau n, lh), \quad z_{k0}^n = z_{kM}^n = \mu_2(\tau n, kh).$$

Определить порядок аппроксимации разностного уравнения и сформулировать граничные условия на слоях $n+1/4$, $n+1/2$ таким образом, чтобы сохранить порядок аппроксимации по τ и h.

Решение. Для определения порядка аппроксимации получим сначала схему в целых шагах. Исключим сначала слой $n+1/4$. Для этого подействуем на второе разностное уравнение оператором $E-\frac{\tau}{2}\Lambda_1$ и полученное уравнение сложим с первым. В результате имеем

$$\Big(E-\frac{\tau}{2}\Lambda_1\Big)\Big(E-\frac{\tau}{2}\Lambda_2\Big)z^{n+1/2}=(\Lambda_1+\Lambda_2)z^n.$$

Теперь, для исключения слоя $n+1/2$, к полученному уравнению прибавим уравнение $z^{n+1}=z^n+\tau z^{n+1/2}$, на которое предварительно подействуем оператором

$$\frac{1}{\tau}\Big(E-\frac{\tau}{2}\Lambda_1\Big)\Big(E-\frac{\tau}{2}\Lambda_2\Big)$$

Имеем:

$$\frac{1}{\tau}\Big(E-\frac{\tau}{2}\Lambda_1\Big)\Big(E-\frac{\tau}{2}\Lambda_2\Big)z^{n+1}=\frac{1}{\tau}\Big(E-\frac{\tau}{2}\Lambda_1\Big)\Big(E-\frac{\tau}{2}\Lambda_2\Big)z^n+(\Lambda_1+\Lambda_2)z^n.$$

Перепишем теперь это уравнение в удобном для анализа виде для чего следует раскрыть скобки и привести подобные

$$\frac{z^{n+1}-z^n}{\tau}=(\Lambda_1+\Lambda_2)\frac{z^{n+1}+z^n}{2}-\frac{\tau^2}{4}\Lambda_1\Lambda_2\frac{z^{n+1}-z^n}{\tau}.$$

Это и есть искомая схема в целых шагах. Если теперь выбрать точку $\big((n+1/2)\tau,x_k,y_l\big)$, в окрестности которой проводить разложение по формуле Тейлора и учесть, что в окрестности этой точки

$$\Lambda_1u=\frac{\partial^2u}{\partial x^2}+O(h^2),\ \ \Lambda_2u=\frac{\partial^2u}{\partial y^2}+O(h^2),\ \ \frac{u^{n+1}-u^n}{\tau}=\frac{\partial u}{\partial t}+O(\tau^2),\ \ \frac{u^{n+1}+u^n}{2}=u+O(\tau^2),$$

то легко заметить, что погрешность аппроксимации равна $O(\tau^2+h^2)$.

Из разностного уравнения $z^{n+1}=z^n+\tau z^{n+1/2}$ взятого в точках границы, и граничного условия на целых шагах получаем:

$$z_{0l}^{n+1/2}=z_{Ml}^{n+1/2}=\frac{\mu_1((n+1)\tau,lh)-\mu_1((n)\tau,lh)}{\tau},$$
$$z_{k0}^{n+1/2}=z_{kM}^{n+1/2}=\frac{\mu_2((n+1)\tau,kh)-\mu_2((n)\tau,kh)}{\tau}.$$

Граничное условие на слое $n+1/4$ требуется только в точках $(0,y),\ (1,y)$. Оно вычисляется из разностного уравнения $z^{n+1/4}=\Big(E-\frac{\tau}{2}\Lambda_2\Big)z^{n+1/2}$ с учетом того, что $z^{n+1/2}$ уже найдено в этих точках.

8.9.2 Задачи

1. Доказать, что $L_h^0=0.5(L_h^++L_h^-),\ \Lambda_h=L_h^+L_h^-=L_h^-L_h^+$. Здесь использованы обозначения примера из параграфа 8.1.2

2. Показать, что разностная схема (8.26) с величинами a_i,d_i,f_i, выбранными в соответствии с формулами (8.27) или (8.28) аппроксимирует задачу (8.22) со вторым порядком, в случае гладких функций $k(x),q(x),f(x),u(x)$.

3. Используя метод неопределенных коэффициентов, построить на шаблоне

$$(t_{n+1}, x_k), (t_n, x_k), (t_n, x_{k-1}), (t_n, x_{k+1})$$

разностную схему максимально возможного порядка аппроксимации для уравнения

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}.$$

4. Используя спектральный критерий устойчивости, показать, что разностная схема (8.6) при положительном коэффициенте a абсолютно неустойчива.

5. Исследовать с помощью спектрального критерия устойчивости разностное уравнение

$$\frac{y_j^{n+1} - y_j^n}{\tau} + a\frac{y_{j+1}^n - y_{j-1}^n}{2h} = 0, \quad a = const.$$

Ответ. Условие спектрального критерия выполнено, если $\tau = Ah^2$, где $A = const$.

6. Для задачи (8.94)-(8.96) построить и исследовать явную разностную схему.

7. При каком соотношении между τ и h разностное уравнение

$$\frac{y_j^{n+1} - y_j^n}{\tau} = \frac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2}$$

аппроксимирует уравнение теплопроводности с порядком $O(\tau^2 + h^4)$?

8. Во второй половине 20-го века было предложено много различных схем для решения уравнений газовой динамики. Ниже приведены некоторые из них, записанные для модельного уравнения

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 :$$

– схема "чехарда"

$$\frac{y_k^{n+1} - y_k^{n-1}}{2\tau} + a\frac{y_{k+1}^n - y_{k-1}^n}{2h} = 0;$$

– схема Лакса-Вендроффа

$$\frac{y_k^{n+1} - y_k^n}{\tau} + a\frac{y_{k+1}^n - y_{k-1}^n}{2h} = a^2\tau\frac{y_{k+1}^n - 2y_k^n + y_{k-1}^n}{2h^2};$$

– схема Мак-Кормака

$$\frac{\widetilde{y}_k^{n+1} - y_k^n}{\tau} + a\frac{y_{k+1}^n - y_k^n}{h} = 0, \quad \frac{y_k^{n+1} - (\widetilde{y}_k^{n+1} + y_k^n)/2}{\tau} + a\frac{\widetilde{y}_k^{n+1} - \widetilde{y}_{k-1}^{n+1}}{2h} = 0.$$

Определить порядок аппроксимации этих схем и исследовать их устойчивость с помощью спектрального критерия.

9. Используя принцип максимума, исследовать на устойчивость разностную схему

$$\frac{y_k^{n+1} - y_k^n}{\tau} = \frac{y_{k+1}^n - 2y_k^n + y_{k-1}^n}{h^2} - \frac{y_{k+1}^n - y_{k-1}^n}{2h} - \sin n\tau, \tag{8.205}$$

$$k = 1, 2, \ldots, K-1, \quad Kh = 1, \quad n = 0, 1, \ldots, N-1, \quad N\tau = T, \tag{8.206}$$

$$y_k^0 = \sin \pi kh, \quad y_0^{n+1} = y_K^{n+1} = 0, \tag{8.207}$$

предложенную для решения задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x} + \sin t,\ 0 < x < 1,\ 0 < t \le T,\quad u(0,x) = \sin \pi x,\quad u(t,0) = u(t,1) = 0.$$

10. При каком условии разностная схема

$$\frac{y_k^{n+1} - y_k^n}{\tau} - \frac{y_{k+1}^n - y_{k-1}^n}{2h} - \frac{\tau}{2h^2}(y_{k+1}^n - 2y_k^n + y_{k-1}^n) = 0$$

монотонна?

11. Предложить разностную схему для решения задачи

$$\frac{\partial u}{\partial t} + 3x(t^2+1)\frac{\partial u}{\partial x} = e^x \cos t,\quad |x| \le 1,\ 0 \le t \le 2,\quad u(0,x) = (x - 1^2).$$

12. Исследовать с каким порядком разностная схема

$$\frac{z^{n+1/2} - z^n}{\tau} = \Lambda_1 \frac{z^{n+1/2} + z^n}{2}, \qquad \frac{z^{n+1} - z^{n+1/2}}{\tau} = \Lambda_2 \frac{z^{n+1} + z^{n+1/2}}{2}$$
$$(\Lambda_1 z)_{kl} = \frac{z_{k+1\,l} - 2z_{kl} + z_{k-1\,l}}{h^2},\quad (\Lambda_2 z)_{kl} = \frac{z_{k\,l+1} - 2z_{kl} + z_{k\,l-1}}{h^2},$$
$$k,l = 1,2\ldots,M-1,\quad Mh = 1,\quad n = 0,1,\ldots,N-1,\quad N\tau = T,$$
$$z_{kl}^0 = f(kh,lh),\quad z_{0l}^{n+1} = z_{Ml}^{n+1} = z_{k0}^{n+1} = z_{kM}^{n+1} = 0$$

аппроксимирует дифференциальную задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},\quad (x,y) \in [0,1] \times [0,1],\quad t \in [0,T],$$
$$u(0,x,y) = f(x,y),$$
$$u(t,0,y) = u(t,1,y) = u(t,x,0) = u(t,x,1) = 0.$$

13. Для задачи

$$\frac{\partial u}{\partial t} = 3\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial^2 u}{\partial y^2} + y\frac{\partial u}{\partial x} + x\frac{\partial u}{\partial y} - 5\sin xy,\quad (x,y) \in (0,2) \times (0,2),\quad t \in (0,T],$$
$$u(0,x,y) = x - y,$$
$$u(t,0,y) = u(t,1,y) = u(t,x,0) = u(t,x,1) = 0$$

построить экономичную разностную схему.

8.9.3 Примеры тестовых вопросов к главе 8

1. Для разностной схемы

$$\begin{cases} L_h y_h = f_h & \text{в области } \omega_h, \\ l_h y_h = g_h & \text{на границе } \gamma_h \text{ области } \omega_h, \end{cases}$$

где $L_h,\ l_h$ — линейные разностные операторы, устойчивость означает

а) существование такой константы C, что $\|y_h\|_h^{(3)} \le C\|f_h\|_h^{(1)} \cdot \|g_h\|_h^{(2)}$;

б) существование такой константы C, что $\|y_h\|_h^{(3)} \le C\|f_h\|_h^{(1)} + \|g_h\|_h^{(2)}$;

в) существование такой константы C, что $\|y_h\|_h^{(3)} \le C\big(\|f_h\|_h^{(1)} + \|g_h\|_h^{(2)}\big)$;

г) существование такой константы C, что $\|f_h\|_h^{(1)} + \|g_h\|_h^{(2)} \le C\|y_h\|_h^{(3)}$;

д) при достаточно малом h для любого $\varepsilon > 0$ найдется такое $\delta(\varepsilon)$, не зависящее от шага, что $\|y_h - \widetilde{y}_h\|_h^{(3)} \le \varepsilon$, если $\|f_h - \widetilde{f}_h\|_h^{(1)} \le \delta$ и $\|g_h - \widetilde{g}_h\|_h^{(2)} \le \delta$, где $\widetilde{y}_h$ — решение разностной схемы $L\widetilde{y}_h = \widetilde{f}_h$, $l\widetilde{y}_h = \widetilde{g}_h$.

2. Какие из схем не могут применяться для решения задачи

$$\frac{\partial u}{\partial t} + 5\frac{\partial u}{\partial x} = \frac{\sin x}{x},\ x \ge 0,\ t \ge 0; \qquad u(0,x) = 0,\ u(t,0) = t?$$

Во всех схемах $x_k = kh,\ k = 0, 1, \ldots,\ y_k^0 = 0,\ y_0^n = n\tau$.

а) $\dfrac{y_k^{n+1} - y_k^n}{\tau} + 5\dfrac{y_k^n - y_{k-1}^n}{h} = \dfrac{\sin x_k}{x_k}$;

б) $\dfrac{y_k^{n+1} - y_k^n}{\tau} + 5\dfrac{y_{k+1}^n - y_k^n}{h} = \dfrac{\sin x_k}{x_k}$;

в) $\dfrac{y_k^{n+1} - y_k^n}{\tau} + 5\dfrac{y_{k+1}^n - y_{k-1}^n}{h} = \dfrac{\sin x_k}{x_k}$;

г) $\dfrac{y_k^{n+1} - y_k^n}{\tau} + 5\dfrac{y_k^{n+1} - y_{k-1}^{n+1}}{h} = \dfrac{\sin x_k}{x_k}$;

д) $\dfrac{y_k^{n+1} - y_k^n}{\tau} + 5\dfrac{y_{k+1}^{n+1} - y_k^{n+1}}{h} = \dfrac{\sin x_k}{x_k}$.

3. Согласно условию Куранта для схемы $\dfrac{y_k^{n+1} - y_k^n}{\tau} + a\dfrac{y_k^n - y_{k-1}^n}{h} = 0$, должно выполняться условие

а) $\dfrac{a\tau}{h} = 1$;

б) $\dfrac{a\tau}{h^2} = \dfrac{1}{2}$;

в) $\tau \le ah$;

г) $\tau \le \dfrac{ah^2}{2}$;

д) $0 \le \dfrac{a\tau}{h} \le 1$.

4. Разностная схема $y_k^{n+1} = \sum\limits_k a_k y_k^n$ монотонна тогда и только тогда, когда

а) $a_k > 0$;

б) $a_k > 0$, если $k > 0$, $a_k < 0$, если $k < 0$ и $a_0 = 0$;

в) $a_k < 0$;

г) $a_k > 0$, если $k < 0$, $a_0 = 0$ и $a_k < 0$, если $k > 0$;

д) $a_k \le 0$;

е) $a_k \ge 0$.

5. Для уравнения теплопроводности записывается разностная схема с весами

$$\frac{y^{n+1} - y^n}{\tau} = \Lambda\big(\sigma y^{n+1} + (1-\sigma)y^n\big),$$

где $(\Lambda y)_k = \dfrac{y_{k+1} - 2y_k + y_{k-1}}{h^2}$. Какие из утверждений справедливы?

а) Если $\sigma = 1$, то схема абсолютно устойчива.

б) Если $\sigma = 1/2$, то схема абсолютно устойчива.

в) Если $\sigma = 0$, то схема абсолютно устойчива.

г) Если $\sigma \ne 0$, то схема абсолютно устойчива.

д) При любом σ схема условно устойчива.

6. Для краевой задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(t,x), \quad 0 < x < 1,\ t > 0, \quad u(0,x) = g(x), \quad u(t,0) = u(t,1) = 0$$

записана разностная схема

$$\frac{y_k^{n+1} - y_k^n}{\tau} = \frac{y_{k+1}^{n+1} - 2y_k^{n+1} + y_{k-1}^{n+1}}{h^2} + f_k^{n+1},$$
$$k = 1, \ldots, K-1, \quad Kh = 1, \quad n = 0, \ldots, N-1, \quad y_k^0 = g(kh), \quad y_0^n = y_K^n = 0.$$

Вычислительная сложность алгоритма нахождения решения равна

а) $O(K + N)$;

б) $O(K \cdot N)$;

в) $O(K + N^3)$;

г) $O(1/K + 1/N)$;

д) $O(K^3 + N)$;

е) $O(K^2 N)$.

9 ПЕРЕЧЕНЬ ЗАДАНИЙ К ЛАБОРАТОРНЫМ РАБОТАМ

Полноценное изучение дисциплины "Вычислительная математика" невозможно без проведения вычислений. Поэтому лабораторный практикум является неотъемлемой частью этой дисциплины. Цель лабораторных работ состоит в закреплении теоретического материала, приобретении практического опыта проведения вычислений, получении навыков тестирования численных методов и анализа полученных результатов.

Наилучшее понимание алгоритмов, по которым производятся расчеты, достигается не при использовании готовых программных продуктов, а при самостоятельном написании и отладке программ. Учебное пособие предназначено в первую очередь для лиц, изучающих программирование в повышенном объеме. Поэтому для студентов, понявших алгоритм, процесс получения исходного кода и его отладки не займет много времени и не является основным в работе.

При выполнении лабораторных работ выбор языка программирования предоставляется студентам. Не следует уделять большое внимание интерфейсу. К нему предъявляются минимальные требования, связанные с удобством ввода данных, просмотра и анализа результатов расчетов.

Зачастую, при составлении тестовых примеров для отладки программ можно воспользоваться следующим приемом, который поясним на примере решения уравнения вида $\mathcal{A}x = b$. Задается значение x_0 и вычисляется $b = \mathcal{A}x_0$. После чего решается уравнение с полученным значением b. Если оператор $\mathcal{A}$ выбран так, что решение единственно, например, для систем линейных алгебраических уравнений матрица системы не вырождена, то найденное решение должно быть равно x_0 с точностью до погрешности метода и ошибок округления.

Для некоторых лабораторных работ приведены варианты заданий. В этом случае номер варианта совпадает с номером студента в списке группы.

Отчет по лабораторной работе должен включать в себя:

- задание;
- краткое описание метода, расчетные формулы;
- текст программы с комментариями;
- тестовые примеры;
- результаты выполнения других пунктов задания.

9.1 ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ ЛИНЕЙНОЙ АЛГЕБРЫ

9.1.1 Решение систем линейных уравнений методом Гаусса

Задание к лабораторной работе

- Составить программу для решения системы линейных алгебраических уравнений методом Гаусса с выбором главного элемента, нахождения определителя матрицы системы и обратной матрицы. Исходные данные — матрица системы уравнений и столбец свободных членов должны читаться из файла, а результаты расчетов помещаться в файл. В случае, когда матрица системы вырождена, выдать об этом сообщение. В противном случае вывести решение системы, невязки, величину определителя, обратную матрицу. Подобрать тестовые примеры, предусматривающие различные ситуации (матрица вырожденная, невырожденная) и провести вычисления.

- Найти число обусловленности матрицы системы в некоторой, выбранной Вами норме для Ваших тестовых примеров.

- Отключить в программе процедуру выбора главного элемента и найти решение системы (предполагается, что вещественные переменные имеют тип float)

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2.1 & 6 \\ 5 & -1 & 5 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 7 \\ 3.9 \\ 6 \end{pmatrix}.$$

  Объяснить результат.

- Решить систему

$$\begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{pmatrix} \mathbf{x} = \mathbf{b},$$

  выбирая $\mathbf{b}$ равным

$$\begin{pmatrix} 23 \\ 32 \\ 33 \\ 31 \end{pmatrix}, \quad \begin{pmatrix} 23.01 \\ 31.99 \\ 32.99 \\ 31.01 \end{pmatrix}, \quad \begin{pmatrix} 23.1 \\ 31.9 \\ 32.9 \\ 31.1 \end{pmatrix}$$

  Пояснить полученные результаты.

9.1.2 Решение систем линейных уравнений методом квадратного корня

Задание к лабораторной работе

- Составить программу для решения системы линейных алгебраических уравнений с симметричной матрицей методом квадратного корня. Предусмотреть в программе возможность вычисления обратной матрицы. Исходные данные —

матрица системы уравнений и столбец свободных членов должны читаться из файла, а результаты расчетов помещаться в файл. Для Ваших тестовых примеров сравнить результаты расчетов, полученных по методу Гаусса и методу квадратного корня.

- Исследовать зависимость числа обусловленности матрицы Гильберта

$$\mathbf{H} = \Big(\frac{1}{i+j-1}\Big)_{i\,j=1}^{n}$$

от числа n для $n = 2, 3, \ldots, 7$.

9.1.3 Решение систем линейных уравнений методами Якоби и Зейделя

Задание к лабораторной работе

- Составить программу для решения системы линейных алгебраических уравнений методами Якоби и Зейделя. Исходные данные — матрица системы уравнений и столбец свободных членов, точность ε должны читаться из файла, а результаты расчетов помещаться в файл. Предусмотреть вывод числа итераций, необходимых для получения решения с заданной точностью ε.

- Исследовать зависимость числа итераций от начального приближения, точности, выбора метода решения.

- Изучить влияние на сходимость величины диагонального преобладания матрицы, то есть величины отношения суммы модулей недиагональных элементов строки к модулю диагонального элемента.

- Используя оценку (2.27), найти число итераций для определения решения с заданной точностью ε. Для тестового примера, решение которого известно, сравнить полученную оценку с тем значением числа итераций, начиная с которого заданная точность достигается. Объяснить результат сравнения.

- Подобрать примеры, показывающие, что диагональное преобладание не является необходимым условием сходимости.

9.1.4 Частичная проблема собственных чисел

Задание к лабораторной работе

- Составить программу, которая позволяет: а) найти методом итераций два наибольших по модулю собственных числа матрицы и соответствующих им собственные вектора; б) методом обратных итераций найти собственное число, ближайшее к заданному числу λ_0. Исходные данные — матрица, точность, число λ_0 и начальный вектор должны читаться из файла, а результаты расчетов помещаться в файл. Предусмотрите вывод числа итераций, которые пришлось совершить для нахождения заданной точности.

- Используя Вашу программу найти наибольшее и наименьшее собственные числа матрицы.

- Исследовать зависимость результатов расчетов от начального вектора, например, при нахождении собственных чисел матрицы

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix}$$

выбирая начальные вектора $(1,1,1,1)$, $(1,2,3,4)$, объяснить полученные результаты расчетов.

- Попытайтесь с помощью Вашей программы найти собственные числа матрицы

$$\begin{pmatrix} -5.7 & -61.1 & -32.9 \\ 0.8 & 11.9 & 7.1 \\ -1.1 & -11.8 & -7.2 \end{pmatrix}.$$

Объясните почему не сходится метод. Как с помощью Вашей программы подтвердить Ваши предположения.

- Сделайте попытку найти с помощью Вашей программы собственные числа матрицы

$$\begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}.$$

Объясните результат.

- Найдите наименьшее по модулю собственное число заданной матрицы.

9.1.5 Полная проблема собственных чисел

Задание к лабораторной работе

- Составить программу для нахождения собственных чисел симметричной матрицы и соответствующих им собственных векторов методом вращения. Входные данные программы — матрица и точность вычислений, выходные — собственные числа, собственные вектора и невязки. В качестве тестового примера можете взять матрицу из предыдущей лабораторной работы.

9.2 ПРИБЛИЖЕНИЕ ФУНКЦИЙ

9.2.1 Интерполирование многочленами

Задание к лабораторной работе

- Составить программу для построения интерполяционного многочлена Лагранжа (Ньютона). Программа должна работать в двух режимах:

а) по заданной таблице значений функции определять приближенное значение функции в некоторой точке, вводимой пользователем;

б) по заданной аналитически функции $y = f(x)$ и массиву значений аргумента (массив читается из файла) вычислить таблицу значений функции. Используя полученную таблицу, построить интерполяционный многочлен после чего нарисовать графики функции $y = f(x)$ и интерполяционного многочлена.

- Исследовать путем проведения вычислительных экспериментов влияние количества и расположения узлов интерполирования, участков интерполирования на величину погрешности интерполирования. В качестве функций, для которых проводится анализ, помимо придуманных Вами функций рекомендуется рассмотреть $y = |x|$ при $|x| \le 1$, $y = e^{-x^2}$ при $|x| \le 4$, $y = \sin x$ при $|x| \le \pi$.

9.2.2 Интерполирование сплайнами

Задание к лабораторной работе

- Составить программу для интерполяции функции кубическим сплайном дефекта 1. Выполнить пункт 2 предыдущей работы в случае интерполяции сплайном. Сравнить результаты, полученные в данной и предыдущей работе.

- Экспериментально исследовать чувствительность сплайна к изменению таблицы значения функции. Для этого задать таблицу значений функции, построить по этой таблице сплайн, нарисовать его график, после чего изменить значения в одной из точек таблицы и посмотреть как изменится график сплайна для новой таблицы.

- Пусть задан набор точек на плоскости $(x_i, y_i),\ i = 0, 1 \ldots, n$, которые лежат на некоторой плоской кривой. Предполагается, что точки занумерованы в порядке следования вдоль кривой. Кривая может иметь произвольную форму, например, быть замкнутой, с самопересечениями, поэтому ее удобнее задать параметрически, то есть в виде $x = x(t),\ y = y(t)$. Постройте по заданному набору точек кривую, используя сплайн-интерполяцию для нахождения функций $x = x(t),\ y = y(t)$. Проанализируйте влияние метода ввода параметра на форму интерполирующей кривой. Приведем два примера ввода параметра.

  1) $(i, x_i),\ (i, y_i)$, то есть в i-ой точке значение параметра t_i равно i.

  2) Пусть

$$\Delta t_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}, \quad t_0 = 0, \quad t_i = \sum_{i=0}^{i-1} \Delta t_i.$$

  Придумайте еще один-два способа задания параметра и сравните кривые, полученные при различном способе параметризации.

9.3 ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ И ИНТЕГРИРОВАНИЕ

9.3.1 Численное дифференцирование

Задание к лабораторной работе

- Для заданной гладкой функции вычислить таблицу ее значений с некоторым заданным шагом. Используя эту таблицу в качестве исходных данных, найти таблицу значений первой, второй и третьей производной. Сравнить полученные значения с точными значениями производных. Как зависит точность вычисления производных от шага таблицы?

- Используя датчик случайных чисел, внести в исходную таблицу возмущения, не превосходящие по модулю некоторого заданного уровня δ. Исследовать путем проведения вычислительных экспериментов влияние шага таблицы и уровня возмущений на значения производных.

9.3.2 Вычисление определенных интегралов методами прямоугольников трапеций и Симпсона

Задание к лабораторной работе

- Составить программу для вычисления интеграла методами прямоугольников, трапеций и Симпсона. Предусмотреть режимы проведения расчетов с постоянным и автоматическим выбором шага интегрирования.

- Путем проведения вычислительных экспериментов при постоянном шаге интегрирования исследовать зависимость точности вычисления интеграла от выбора метода и шага.

- В случае автоматического выбора шага, исследовать зависимость количества шагов от метода и заданной точности.

9.3.3 Вычисление интегралов методом Монте-Карло

Задание к лабораторной работе

- Имеется множество точек, лежащих на границе некоторой области, принадлежащей плоскости. Точки заданы в порядке, соответствующем некоторому обходу границы. Используя интерполяцию кубическими сплайнами нарисовать область (границу области задать параметрически применяя метод, описанный в лабораторной работе 9.2.2).

- Вычислить площадь области, применяя методы Симпсона и Монте-Карло. При использовании метода Монте-Карло генерируемые точки отображать на экране.

Замечание. Если функции $x = x(t),\ y = y(t),\ t_b \le t \le t_e$ параметрически задают кривую без самопересечения, которая является границей области, то величина площади S, ограниченной этой кривой вычисляется по одной из формул:

$$S = \left| \int_{t_b}^{t_e} y(t)x'(t)\,dt \right| = \left| \int_{t_b}^{t_e} x(t)y'(t)\,dt \right|.$$

9.4 РЕШЕНИЕ НЕЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ И СИСТЕМ

9.4.1 Решение одного нелинейного алгебраического уравнения

Задание к лабораторной работе

Составить программу решения нелинейного алгебраического уравнения комбинированным методом хорд и касательных. Исходными данными являются: отрезок, на котором ищутся корни; точность с которой требуется найти корни; шаг, с которым делится отрезок для отделения корней. Функцию, ее первую и вторую производные задать непосредственно в программе.

9.4.2 Нахождение корней многочленов методом парабол

Задание к лабораторной работе

Найти корни заданного многочлена $P(x)$, используя метод парабол. Предусмотреть возможность нахождения комплексных корней. Многочлен задавать набором его коэффициентов. Для каждого полученного корня вычислить значение невязки.

9.4.3 Метод Ньютона для решения систем нелинейных уравнений

Задание к лабораторной работе

Для системы нелинейных алгебраических уравнений найти решение методом Ньютона. Функции, определяющие систему и матрицу Якоби задать непосредственно в программе. Точность вычислений и начальное приближение задавать в диалоге.

9.5 РЕШЕНИЕ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

9.5.1 Нахождение методами Рунге-Кутта и Адамса решения задачи Коши для системы обыкновенных дифференциальных уравнений

Задание к лабораторной работе

- Составить программу решения задачи Коши для системы n обыкновенных дифференциальных уравнений методом Рунге-Кутта четвертого порядка точности. В программе предусмотреть два режима работы — с постоянным шагом и автоматическим выбором шага. В основу автоматического выбора шага для студентов, чей номер в списке группы четный, положить метод Рунге, а для тех, у кого номер нечетный, — вложенные методы.

Правые части системы уравнений задать в программе, а остальные параметры: начальные данные, промежуток интегрирования, шаг, точность ε — в файле или в диалоге в процессе выполнения программы.

- Вывести в файл таблицу значений функций, получить графики решений.
- Для тестовых примеров, решение которых известно, наряду с графиками приближенного решения **y** построить графики точного решения **u** и получить в сопоставимых с приближенным решением точках x_j таблицу значений функции. Вычислить норму глобальной погрешности $\max\limits_{i=1,\dots,n} \max\limits_{j} |y_i(x_j) - u_i(x_j)|$.
- Путем проведения численных экспериментов исследовать зависимость нормы глобальной погрешности от величины шага (для вычислений с постоянным шагом) и точности ε (для вычислений с переменным шагом).
- Найти, используя Вашу программу с постоянным шагом, решение задачи

$$u' + 30u = 0, \quad u(0) = 1, \quad 0 \le x \le 1.$$

Расчеты провести дважды, первый раз разбить промежуток интегрирования на 10 частей, а второй — на 11. Объяснить почему результаты расчетов качественно различаются.

- Составить программу решения задачи Коши для системы обыкновенных дифференциальных уравнений явно-неявным (предиктор-корректор) методом Адамса четвертого порядка. Провести расчеты и сравнить результаты с решением, полученным по методу Рунге-Кутта.

Указание Для тестирования программы можно рассмотреть следующую задачу Коши

$$u' = f'(x) + p(u - f(x)), \quad 0 \le x \le 1, \quad u(0) = f(0).$$

Здесь $p \neq 0$ — заданная константа, а $f(x)$ — заданная функция. Очевидно, что решением этой задачи Коши является функция $u = f(x)$.

9.5.2 Задача Коши для жестких систем

Задание к лабораторной работе

- Заданы два числа p_1, p_2. Пусть кроме того, $b_{11}, b_{12}, b_{21}, b_{22}$ заданные числа такие, что $\Delta = b_{11}b_{22} - b_{12}b_{21} \neq 0$. Положим

$$a_{11} = \frac{b_{11}b_{22}p_1 - b_{12}b_{21}p_2}{\Delta},$$
$$a_{12} = \frac{b_{11}b_{12}(p_2 - p_1)}{\Delta},$$
$$a_{21} = \frac{b_{21}b_{22}(p_1 - p_1)}{\Delta},$$
$$a_{22} = \frac{b_{11}b_{22}p_2 - b_{12}b_{21}p_1}{\Delta}$$

Доказать, что числа p_1, p_2 являются собственными числами матрицы **A**, элементами которой являются числа $a_{ij},\ i, j = 1, 2$, а соответствующими этим числам собственные вектора являются столбцами матрицы **B**.

- Пусть $f_1(x), f_2(x)$ две заданные непрерывно дифференцируемые функции. Показать, что общее решение системы

$$\begin{cases} u'_1 = a_{11}u_1 + a_{12}u_2 + f'_1 - a_{11}f_1 - a_{12}f_2, \\ u'_2 = a_{21}u_1 + a_{22}u_2 + f'_2 - a_{21}f_1 - a_{22}f_2, \end{cases} \tag{9.1}$$

имеет вид

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = C_1 \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix} e^{p_1 x} + C_2 \begin{pmatrix} b_{12} \\ b_{22} \end{pmatrix} e^{p_2 x} + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

Таким образом, если числа p_1, p_2 отрицательные, причем $p_2/p_1 \gg 1$, например, $p_1 = -1, p_2 = -1000$, то системе (9.1) является жесткой.

- Составить программу, решения системы линейных дифференциальных уравнений явным и неявным методами Эйлера.

- Для жесткой системы уравнений убедиться в преимуществе неявного метода перед явным. В качестве тестового примера достаточно рассмотреть систему (9.1) на отрезке $[0, 1]$ и задать начальные условия $u_1 = f_1(0), u_2 = f_2(0)$. Тогда функции $f_1(x), f_2(x)$ будут решением полученной задачи Коши.

9.6 РЕШЕНИЕ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ И ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

9.6.1 Решение краевых задач методом стрельбы

Задание к лабораторной работе

- Составить программу решения нелинейной краевой задачи для дифференциального уравнения второго порядка методом стрельбы.

- Проанализировать полученный результат, исходя из физического смысла задачи, и при необходимости провести дополнительные вычисления.

Задачи для вариантов заданий взяты из книги [26]. Варианты разбиты на блоки по 5 задач. Внутри каждого блока различные значения параметра соответствуют различным номерам вариантов. Например, вариант 3 это задача об изгибе консольной балки, в которой параметр $\beta = 0.3$, вариант 7 это задача о распределении температуры в пластине, причем параметр $\beta = 0.4$.

Варианты 1–5

При исследовании изгиба консольной балки была получена нелинейная граничная задача, определяемая дифференциальным уравнением

$$\frac{d^2\vartheta}{ds^2} + \beta \cos\vartheta = 0$$

и граничными условиями

$$\vartheta(0) = 0, \quad \frac{d\vartheta(1)}{ds} = 0,$$

где ϑ — угол поворота сечения относительно оси балки, ось Os направлена вдоль оси балки. Найти решение задачи для $\beta = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

Варианты 6–10

Увеличение поверхности тела для выделения тепла в виде излучения имеет большое значение при проектировании сильноточных проводников, так как оно является единственным способом отвода излишков тепла. Чтобы масса теплоизлучателя была по возможности наименьшей, его делают в виде набора тонких кольцевых пластин с малым углом между боковыми гранями. Распределение температуры в такой пластине является решением уравнения

$$\frac{d^2U}{dR^2} + \left(\frac{1}{R+\rho} - \frac{\operatorname{tg}\alpha}{(1-R)\operatorname{tg}\alpha + \theta} \right) \frac{dU}{dR} - \frac{\beta U^4}{(1-R)\operatorname{tg}\alpha + \theta} = 0$$

с граничными условиями

$$U(0) = 1, \quad \frac{dU(1)}{dR} = 0,$$

где

$$R = \frac{r - r_B}{r_T - r_B}, \quad U = \frac{T}{T_B}, \quad \theta = \frac{z_T}{r_T - r_B}, \quad \rho = \frac{z_B}{r_T - r_B}.$$

Здесь α — угол между гранями пластины; r, r_B, r_T и z_T — текущий радиус, радиус основания, радиус вершины и толщина пластины в вершине соответственно; T и T_B — текущая температура и температура основания.

Найти относительную температуру U, если $\rho = 0.5$, $\theta = 0.05$, $\alpha = 6^o$, а $\beta = 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1$.

Варианты 11–15

Трубчатый реактор — это труба, через которую протекает смесь химических реагентов. Такие реакторы являются важнейшими составляющими при проектировании химических заводов. Пусть в реакторе протекает изотермическая реакция n-го порядка вида $A+A \to B$. Тогда относительная концентрация y компонента A находится из граничной задачи

$$\frac{d^2y}{dz^2} - N\frac{dy}{dz} - NRy^n = 0, \quad y(0) - \frac{1}{N}\frac{dy(0)}{dz} = 1, \quad \frac{dy(1)}{dz} = 0.$$

Здесь

$$y = \frac{C_A}{C_{A0}}, \quad z = \frac{x}{L}, \quad N = \frac{vL}{E_a}, \quad R = \frac{kL}{v},$$

где C_A — концентрация компонента A; C_{A0} — концентрация компонента A в жидкости, поступающей в трубу; L — длина трубы (предполагается, что длина трубы достаточна для завершения реакции); v — осевая скорость потока компонента A; E_a — эффективный коэффициент диффузии; k — константа скорости химической реакции. Найти относительную концентрацию y компонента A, если $n = 2$, а

- $N = 1, \quad R = 1.94,$
- $N = 1, \quad R = 3.24,$
- $N = 2, \quad R = 0.4,$
- $N = 2, \quad R = 0.815,$
- $N = 2, \quad R = 2.16.$

Варианты 16– 20

При рассмотрении переноса тепла и массы в пористом катализаторе получается следующая граничная задача:

$$\frac{d^2y}{dx^2} = \delta y \exp\left(\frac{\gamma\beta(1-y)}{1+\beta(1-y)}\right), \quad \frac{dy(0)}{dx} = 0, \quad y(1) = 1.$$

Найти решение этой задачи при

- $\beta = 0.4, \quad \gamma = 10, \quad \delta = 0.14;$
- $\beta = 0.3, \quad \gamma = 15, \quad \delta = 0.10;$
- $\beta = 0.4, \quad \gamma = 10, \quad \delta = 0.12;$
- $\beta = 0.3, \quad \gamma = 15, \quad \delta = 0.08;$
- $\beta = 0.2, \quad \gamma = 20, \quad \delta = 0.10.$

Варианты 21– 25

Поведение неньютоновских жидкостей представляет интерес для специалистов в области химических технологий. При исследовании естественной конвекции таких жидкостей между параллельными пластинами установлено, что скорость жидкости представляется решением граничной задачи

$$\frac{d^2v}{ds^2} = \frac{s-1}{1+1/[\beta\sqrt{(\alpha dv/ds)^2+1}]}, \quad v(0) = 0, \quad v(1) = 0.$$

Найти распределение скорости при $\alpha = 0.01$ и $\beta = 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0.$

9.6.2 Решение краевых задач для линейных дифференциальных уравнений второго порядка методом прогонки

Постановка задачи

Требуется найти приближенное значение решения следующей краевой задачи:

$$\frac{d^2u(x)}{dx^2} + A(x)\frac{du(x)}{dx} + B(x)u(x) = C(x), \quad x \in [a,b], \tag{9.2}$$

$$F_1u(a) + D_1\frac{du(a)}{dx} = E_1, \tag{9.3}$$

$$F_2u(b) + D_2\frac{du(b)}{dx} = E_2. \tag{9.4}$$

Подробное описание метода построения разностной схемы, аппроксимирующей со вторым порядком, и получения расчетных формул для численной реализации дано в параграфе 7.2.

Задание к лабораторной работе

- По аналогии с описанным в параграфе 7.2 методом, вывести формулы для коэффициентов системы линейных алгебраических уравнений с трехдиагональной матрицей, которая получается для метода первого порядка аппроксимации.

- Составить программу для решения краевой задачи для линейного дифференциального уравнения второго порядка методом прогонки в случае первого и второго порядка аппроксимации.
- Провести вычисления, разбивая отрезок интегрирования на различное число частей, например, 25, 50, 100, 200.
- Проанализировать зависимость нормы разности между точным и приближенным решением от шага сетки для методов первого и второго порядка аппроксимации. Для этого составить таблицу или построить графики нормы разносить для методов различного порядка аппроксимации. В качестве нормы можно взять C-норму. В этом случае норма функции равна максимальному значению модуля этой функции.
- Исследовать для Вашего варианта вопрос об устойчивости прогонки.

Варианты задач к лабораторной работе представлены в таблице 9.6.2.

9.6.3 Решение интегральных уравнений

Задание к лабораторной работе

- Составить программу для решения интегрального уравнения Фредгольма второго рода.
- Проводя вычислительные эксперименты, исследовать на тестовых примерах влияние выбора квадратурной формулы и шага интегрирования на норму погрешности решения.

9.7 РЕШЕНИЕ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

9.7.1 Краевые задачи для уравнения теплопроводности

Постановка краевой задачи

Рассматривается краевая задача для уравнения теплопроводности:

$$\frac{\partial u(t,x)}{\partial t} = a^2 \frac{\partial^2 u(t,x)}{\partial x^2} + f(t,x), \quad 0 < x < l, \quad 0 < t \le T, \tag{9.5}$$

При $t = 0$ для функции $u = u(t,x)$ задается начальное условие

$$u(0,x) = u_0(x), \tag{9.6}$$

а на границах $x = 0$ и $x = l$ — граничные условия, вообще говоря, третьего рода

$$\alpha_1 \frac{\partial u(t,0)}{\partial x} = \beta_1 u(t,0) - \mu_1(t), \tag{9.7}$$

$$-\alpha_2 \frac{\partial u(t,l)}{\partial x} = \beta_2 u(t,l) - \mu_2(t). \tag{9.8}$$

Таблица 9.6.2

№	$A(x)$	$B(x)$	$C(x)$	a	b	F_1	D_1	E_1	F_2	D_2	E_2	Решение $u(x)$
1	2	-1	$-3e^{-x}\sin x$	0	π	1	-1	-1	1	1	$-e^{-\pi}$	$e^{-x}\sin x$
2	-2	-1	$-3e^{x}\sin x$	0	π	1	-1	-1	1	1	$-e^{\pi}$	$e^{x}\sin x$
3	0	$\dfrac{-2}{x^2+1}$	0	0	1	1	0	0	-1	1	1	$x+(x^2+1)\operatorname{arctg} x$
4	$\dfrac{-2}{e^x+1}$	$\dfrac{-e^x}{e^x+1}$	0	0	1	1	-1	-1	0	1	e	e^x-1
5	$\dfrac{2e^x}{e^x+1}$	$\dfrac{e^x}{e^x+1}$	0	0	1	1	0	1	e	$e+1$	0	$\dfrac{2}{e^x+1}$
6	0	$\dfrac{-2}{x^3+x^2}$	0	1	2	0	1	-1	1	2	1	$1+\dfrac{1}{x}$
7	$\dfrac{4x}{2x+1}$	$\dfrac{-4}{2x+1}$	0	0	1	1	1	0	2	1	3	$x+e^{-2x}$
8	$\dfrac{4x}{2x+1}$	$\dfrac{-4}{2x+1}$	0	0	1	3	1	0	2	1	3	$x-e^{-2x}$
9	$\dfrac{4x}{2x+1}$	$\dfrac{-4}{2x+1}$	0	0	1	2	-1	-3	2	1	3	$x-\dfrac{1}{2}e^{-2x}$
10	$\dfrac{2}{x}$	-1	0	1	2	1	0	e	$\dfrac{1}{2}$	-1	0	$\dfrac{e^x}{x}$
11	$\dfrac{2}{x}$	-1	0	1	2	1	1	$\dfrac{-1}{e}$	$\dfrac{3}{2}$	1	0	$\dfrac{e^{-x}}{x}$
12	0	$\dfrac{-2}{\cos^2 x}$	0	0	$\dfrac{\pi}{4}$	0	1	1	2	1	4	$\operatorname{tg} x$
13	1	-2	$3xe^x$	0	2	1	-3	1	-3	3	$5e^2$	$\left(\dfrac{x^2}{2}-\dfrac{x}{3}\right)e^x$
14	-2	0	$2e^x$	0	1	2	-1	$2e-4$	0	1	0	$\dfrac{e^{2x}}{e}-2e^x+e-1$
15	0	$\dfrac{-2}{x^2}$	$\dfrac{6\ln x}{x^3}$	1	e	1	0	1	0	1	$-2e^{-2}$	$\dfrac{3-2\ln x-3\ln^2 x}{3x}$
16	$\dfrac{1}{x}$	$\dfrac{-1}{x^2}$	1	$\dfrac{1}{2}$	2	4	-1	11	0	1	$\dfrac{1}{12}$	$\dfrac{x^2}{3}-x+\dfrac{1}{x}$
17	1	0	$\sin^2 x$	0	$\dfrac{\pi}{4}$	4	-1	0	0	1	$\dfrac{3}{5}$	$\dfrac{x}{2}+\dfrac{2\cos 2x-\sin 2x}{20}$
18	0	-1	$2x\sin x$	0	π	1	-1	-1	0	1	π	$-x\sin x-\cos x$
19	-2	0	$e^x(x^2+x-3)$	0	1	0	1	2	-1	1	$e(e-3)$	$e^x(e^x-x^2-x+1)$
20	-1	0	$2(1-x)$	0	2	0	1	1	1	-1	0	x^2+e^x
21	2	-2	$e^x-3\sin x+2\cos x$	0	π	0	1	2	1	-1	1	$e^x+\sin x$
22	0	-1	e^x-e^{-x}	0	1	0	1	1	0	1	e	$x\dfrac{e^x+e^{-x}}{2}$
23	-1	0	$2x-1-3e^x$	0	1	0	1	-4	1	1	$-9e-3$	$2-3xe^x-x-x^2$
24	0	-1	xe^x	0	1	1	4	-1	1	0	0	$\dfrac{(x^2-x)e^x}{4}$
25	$\dfrac{2}{x}$	0	$\dfrac{-\sin x}{x}$	$\dfrac{\pi}{2}$	π	0	1	$\dfrac{-4}{\pi^2}$	2	1	$\dfrac{-1}{\pi}$	$\dfrac{\sin x}{x}$

Здесь $\alpha_i, \beta_i,\ i = 1, 2$ — неотрицательные константы, причем $\alpha_i + \beta_i \neq 0, i = 1, 2$. В том случае, когда $\alpha_i = 0,\ i = 1, 2$ получаются граничные условия первого рода, если же $\beta_i = 0,\ i = 1, 2$ — второго рода.

Описание разностной схемы

Для построения разностной схемы с весами выберем в качестве сетки множество точек

$$\overline{\omega}_{\tau h} = \{(t_n, x_k) : t_n = n\tau,\ n = 0, \ldots, N,\ N\tau = T;\ x_k = kh,\ k = 0, \ldots, K,\ Kh = l\}.$$

Уравнение (9.5) аппроксимируем разностным уравнением

$$\frac{y_k^{n+1} - y_k^n}{\tau} = a^2 \left(\sigma \frac{y_{k+1}^{n+1} - 2y_k^{n+1} + y_{k-1}^{n+1}}{h^2} + (1-\sigma) \frac{y_{k+1}^n - 2y_k^n + y_{k-1}^n}{h^2} \right) + \varphi_k^n, \tag{9.9}$$
$$k = 1, \ldots, K-1,\ n = 0, \ldots, N-1, \quad \varphi_k^n = f(t_n + \sigma\tau, x_k),$$

Здесь σ — вещественный параметр, называемый весом, $0 \leq \sigma \leq 1$. При $\sigma = 0$ получается явная схема, $\sigma = 1/2$ — схема Кранка-Николсон, а при $\sigma = 1$ — чисто неявная схема.

Начальное условие (9.6) аппроксимируем точно, то есть положим

$$y_k^0 = u_0(x_k). \tag{9.10}$$

В том случае, когда $\alpha_i \neq 0,\ i = 1, 2$, граничные условия (9.7),(9.8) заменим следующими разностными граничными условиями:

при $\sigma \neq 0,\ \sigma \neq 1$

$$\begin{gathered}
\sigma\Big(\alpha_1 \frac{y_1^{n+1} - y_0^{n+1}}{h} - \beta_1 y_0^{n+1}\Big) + (1-\sigma)\Big(\alpha_1 \frac{y_1^n - y_0^n}{h} - \beta_1 y_0^n\Big) = \frac{\alpha_1 h}{2a^2} \frac{y_0^{n+1} - y_0^n}{\tau} - \widetilde{\mu}_1^n, \\
-\sigma\Big(\alpha_2 \frac{y_K^{n+1} - y_{K-1}^{n+1}}{h} + \beta_2 y_K^{n+1}\Big) - (1-\sigma)\Big(\alpha_2 \frac{y_K^n - y_{K-1}^n}{h} + \beta_2 y_K^n\Big) = \frac{\alpha_2 h}{2a^2} \frac{y_K^{n+1} - y_K^n}{\tau} - \widetilde{\mu}_2^n, \\
\widetilde{\mu}_1^n = \mu_1(t_n + \tau/2) + \frac{\alpha_1 h}{2a^2} f(t_n + \tau/2, 0), \quad \widetilde{\mu}_2^n = \mu_2(t_n + \tau/2) + \frac{\alpha_2 h}{2a^2} f(t_n + \tau/2, l);
\end{gathered} \tag{9.11}$$

при $\sigma = 0$ или $\sigma = 1$

$$\begin{gathered}
\alpha_1 \frac{y_1^{n+1} - y_0^{n+1}}{h} - \beta_1 y_0^{n+1} = \frac{\alpha_1 h}{2a^2} \frac{y_0^{n+1} - y_0^n}{\tau} - \widetilde{\mu}_1^n, \\
-\Big(\alpha_2 \frac{y_K^{n+1} - y_{K-1}^{n+1}}{h} + \beta_2 y_K^{n+1}\Big) = \frac{\alpha_2 h}{2a^2} \frac{y_K^{n+1} - y_K^n}{\tau} - \widetilde{\mu}_2^n, \\
\widetilde{\mu}_1^n = \mu_1(t_{n+1}) + \frac{\alpha_1 h}{2a^2} f(t_{n+1}, 0), \quad \widetilde{\mu}_2^n = \mu_2(t_{n+1}) + \frac{\alpha_2 h}{2a^2} f(t_{n+1}, l).
\end{gathered} \tag{9.12}$$

Если же, например, $\alpha_1 = 0$, то независимо от значения σ левое граничное условие аппроксимируется точно, то есть в (9.11) или (9.12) оно заменяется на

$$\beta_1 y_0^{n+1} = \mu_1(t_{n+1}). \tag{9.13}$$

Аналогично, следует поступить при $\alpha_2 = 0$.

На гладких решениях полученная разностная схема аппроксимирует исходную задачу с порядком $\tau + h^2$ при $\sigma \neq 1/2$ и $\tau^2 + h^2$ при $\sigma = 1/2$. При $\sigma \geq 1/2$ схема

абсолютно устойчива. Если же $\sigma = 0$, то схема устойчива если $r = \dfrac{a^2\tau}{h^2} \le \dfrac{1}{2}$, то есть условно устойчива.

Рассмотрим теперь вопрос об организации вычислений по разностной схеме. Явная схема переписывается в виде

$$y_k^{n+1} = (1-2r)y_k^n + r(y_{k-1}^n + y_{k+1}^n) + \tau\varphi_k^n, \quad k = 1, \dots, K-1. \tag{9.14}$$

Таким образом, зная решение на слое с номером n, легко найти его в точках с номерами $k = 1, \dots, K-1$ на слое $n+1$. Разностные граничные условия позволяют найти после этого решение в точках x_0 и x_K.

Для нахождение решения по неявной схеме $\sigma > 0$ перепишем ее в виде

$$\begin{cases} y_0^{n+1} = \zeta_1 y_1^{n+1} + \nu_1, \\ Ay_{k-1}^{n+1} - By_k^{n+1} + Cy_{k+1}^{n+1} = D_k^n, \quad k = 1, \dots, K-1, \\ y_K^{n+1} = \zeta_2 y_{K-1}^{n+1} + \nu_2. \end{cases} \tag{9.15}$$

При $\alpha_1 \neq 0,\ \alpha_2 \neq 0$ коэффициенты здесь равны:

$$\zeta_1 = \frac{\alpha_1\sigma}{\Delta_1}, \quad \Delta_1 = \sigma(\alpha_1 + h\beta_1) + \frac{\alpha_1 h^2}{2a^2\tau},$$

$$\nu_1 = \frac{1}{\Delta_1}\left(h\widetilde{\mu}_1^n + (1-\sigma)\alpha_1 y_1^n + y_0^n\Big(\frac{\alpha_1 h^2}{2a^2\tau} - (1-\sigma)(\alpha_1 + \beta_1 h)\Big)\right),$$

$$\zeta_2 = \frac{\alpha_2\sigma}{\Delta_2}, \quad \Delta_2 = \sigma(\alpha_2 + h\beta_2) + \frac{\alpha_2 h^2}{2a^2\tau},$$

$$\nu_2 = \frac{1}{\Delta_2}\left(h\widetilde{\mu}_2^n + (1-\sigma)\alpha_2 y_{K-1}^n + y_K^n\Big(\frac{\alpha_2 h^2}{2a^2\tau} - (1-\sigma)(\alpha_2 + \beta_2 h)\Big)\right),$$

$$A = C = a^2\sigma, \quad B = 2a^2\sigma + \frac{h^2}{\tau},$$

$$D_k^n = \Big(2a^2(1-\sigma) - \frac{h^2}{\tau}\Big)y_k^n - a^2(1-\sigma)(y_{k-1}^n + y_{k+1}^n) - \varphi_k^n h^2.$$

Если же $\alpha_1 = 0$, то $\zeta_1 = 0, \quad \nu_1 = \mu_1(t_{n+1})/\beta_1$. При $\alpha_2 = 0$ имеем $\zeta_2 = 0, \quad \nu_2 = \mu_2(t_{n+1})/\beta_2$.

Система (9.15) решается методом прогонки. Заметим, что условие устойчивости прогонки выполнено.

Задание к лабораторной работе

- Показать расчетами сходимость решения, полученного по явной разностной схеме, к точному решению. Для этого зафиксировав значение $r < 1/2$ и уменьшая h, провести серию расчетов. Результаты сравнить на выбранном контрольном временном слое, в сопоставимых точках x_k. Вычислить на этом слое норму относительной погрешности $\|u - y\|/\|y\|$ и исследовать ее зависимость от шага h. В качестве контрольного временного слоя можно взять множество точек сетки при $t = 1$. Шаг h удобно уменьшать, удваивая каждый раз число интервалов разбиения отрезка $[0, l]$.

- Проведя вычислительный эксперимент, убедиться, что при $r > 1/2$ явная схема расходится.

- Провести вычисления для неявных схем ($\sigma = 1/2,\ \sigma = 1$) при $r > 1/2$. На выбранном контрольном временном слое сравнить результаты с точным решением, сравнить $\|u-y\|/\|y\|$ для разностных решений, полученных при различных значениях шагов h и весов σ.

- Рассмотреть задачу (9.5)-(9.7)

$$\alpha_1 = \alpha_2 = 0,\ \beta_1 = \beta_2 = 1,\ \mu_1 = 0,\ \mu_2 = 1,\ f = 0,\ u_0 = \begin{cases} 0, & \text{при } x \in [0, l/2), \\ 1, & \text{при } x \in [l/2, l] \end{cases}$$

  и дать ее физическую интерпретацию. Сделав 1-2 шага при времени, убедиться, что при соблюдении условия устойчивости и $\sigma = 0$, а также при любом r и $\sigma = 1$ разностная схема обладает свойством монотонности, в то время как при $\sigma = 1/2$ и $r > 2$ свойство монотонности схемы не выполнено.

- Какие изменения произойдут в разностной схеме, если $a = a(t, x)$? Какую можно предложить разностную схему и как произвести по ней расчеты, если $a = a(u)$?

Варианты задач к лабораторной работе представлены в таблице 9.7.1. В ней параметр p равен номеру варианта.

9.7.2 Решение задачи Дирихле для уравнения Пуассона в прямоугольнике

Постановка краевой задачи

Пусть $\Omega = \{(x, y) :\ 0 < x < a,\ 0 < y < b\}$ и $f = f(x, y)$ заданная в области Ω непрерывная функция. Требуется найти функцию $u = u(x, y)$, которая непрерывна в замыкании области Ω, имеет в области Ω непрерывные вторые производные, удовлетворяет уравнению Пуассона

$$\Delta u = \frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = -f(x, y), \quad (x, y) \in \Omega \tag{9.16}$$

и принимает на границе $\partial\Omega$ области Ω заданные значения:

$$\begin{aligned} u(0, y) = g_1(y), \quad u(a, y) = g_2(y), \quad y \in [0, b], \\ u(x, 0) = g_3(x), \quad u(x, b) = g_4(x), \quad x \in [0, a]. \end{aligned} \tag{9.17}$$

Непрерывные функции $g_1, \ldots, g_4$ должны удовлетворять условиям

$$g_1(0) = g_3(0),\ g_1(b) = g_4(0),\ g_2(0) = g_3(a),\ g_2(b) = g_4(a),$$

обеспечивающим непрерывность функции u на границе области.

Разностная схема

В области Ω построим равномерную разностную сетку и шагами h_1 по направлению оси x и h_2 по направлению оси y:

$$\omega = \{(x_i, y_j) :\ x_i = ih_1,\ y_j = jh_2,\ i = 1, \ldots, I-1,\ j = 1, \ldots, J-1,\ Ih_1 = a,\ Jh_2 = b\}.$$

Таблица 9.7.1

№	a	l	$u_0(x)$	$f(t,x)$	α_1	β_1	$\mu_1(t)$	α_2	β_2	μ_2	Решение $u(t,x)$
1 2 3 4 5	$\sqrt{\frac{p}{2}}$	$\frac{\pi}{2}$	$\cos x$	$\frac{p}{2}e^{-z}\cos x$, $z=\frac{pt}{2}$	3	1	$(1+z)e^{-z}$	0	1	0	$(1+z)e^{-z}\cos x$
6 7 8 9 10	1	1	$\sin\sqrt{p}x$	0	1	1	$-\sqrt{p}e^{-pt}$	0	1	$e^{-pt}\sin\sqrt{p}$	$e^{-pt}\sin\sqrt{p}x$
11 12 13 14 15	1	1	$xe^{\sqrt{s}x}$, $s=\frac{p+1}{p}$	$-2\sqrt{s}e^{\sqrt{s}x+st}$	0	1	0	-1	$1+\sqrt{s}$	0	$xe^{\sqrt{s}x+st}$
16 17 18 19 20	$\frac{\sqrt{p}}{4}$	$\frac{\pi}{2}$	$\sin x$	$\frac{p}{16}e^{z}\sin x$, $z=\frac{pt}{16}$	0	1	0	1	0	0	$\frac{e^{z}+e^{-z}}{2}\sin x$
21 22 23 24 25	1	1	$x\cos qx$, $q=\sqrt[6]{p}$	$2qe^{-q^2t}\sin qx$	-1	0	e^{-q^2t}	0	1	$e^{-q^2t}\cos q$	$xe^{-q^2t}\cos qx$

Пусть

$$\gamma = \{(0, y_j) : j = 0, \ldots, J\} \cup \{(a, y_j) : j = 0, \ldots, J\} \cup \\ \cup \{(x_i, 0) : i = 0, \ldots, I\} \cup \{(x_i, b) : i = 0, \ldots, I\}$$

множество граничных узлов. Обозначим через v_{ij} значение сеточной функции в узле (x_i, y_j) и введем разностные операторы

$$\Lambda_1 v_{i\,j} = \frac{v_{i-1\,j} - 2v_{ij} + v_{i+1\,j}}{h_1^2}, \quad \Lambda_2 v_{i\,j} = \frac{v_{i\,j-1} - 2v_{ij} + v_{i\,j+1}}{h_2^2},$$

которые аппроксимируют на гладких функциях операторы $\dfrac{\partial^2}{\partial x^2}, \dfrac{\partial^2}{\partial y^2}$ соответственно со вторым порядком. Тогда в каждом узле сетки ω уравнение (9.16) можно заменить разностным уравнением

$$\Lambda_1 v_{ij} + \Lambda_2 v_{ij} = -f_{ij}, \tag{9.18}$$

где $f_{ij} = f(x_i, y_j)$. В граничных узлах условия (9.17) аппроксимируются точно

$$v_{0,j} = g_1(y_j), \ v_{I,j} = g_2(y_j), \ v_{i,0} = g_3(x_i), \ v_{i,J} = g_4(x_i). \tag{9.19}$$

Разностная схема (9.18), (9.19) устойчива и аппроксимирует задачу (9.16), (9.17) со вторым порядком на решении, обладающем четвертыми непрерывными производными. Из теоремы о сходимости следует, что $|u(x_i, y_j) - v_{ij}| = O(h_1^2 + h_2^2)$.

Разностная схема представляет собой систему $(I+1)(J+1)$ линейных алгебраических уравнений относительно значений сеточной функции в точках из $\omega \cup \gamma$.

Методы решения системы линейных алгебраических уравнений

Рассмотрим два метода решения получившейся системы линейных алгебраических уравнений (9.18), (9.19). Первый — **метод простой итерации**. В соответствии с этим методом запишем следующий итерационный процесс

$$\frac{v_{ij}^{n+1} - v_{ij}^n}{\tau} = \Lambda_1 v_{ij}^n + \Lambda_2 v_{ij}^n + f_{ij}, \ i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1, \quad n = 0, 1, \ldots \\ v_{0,j}^n = g_1(y_j), \ v_{I,j}^n = g_2(y_j), \ v_{i,0}^n = g_3(x_i), \ v_{i,J}^n = g_4(x_i). \tag{9.20}$$

Здесь верхний индекс n соответствует номеру итерации. Начальное приближение $v_{i,j}^0$ задается произвольным образом, например, можно положить $v_{i,j}^0 = 0$.

Если существует предел $\lim\limits_{n\to\infty} v_{ij}^n$, то, очевидно, этот предел и есть решение системы (9.18), (9.19). Параметр τ выбирается так, чтобы обеспечить максимальную скорость сходимости метода. В рассматриваемом случае, когда область, в которой ищется решение прямоугольник, оптимальное значение τ равно

$$\tau = \Big(\frac{2}{h_1^2} + \frac{2}{h_2^2}\Big)^{-1}.$$

Тогда, подставляя это значение τ в (9.20) и учитывая определение $\Lambda_1, \ \Lambda_2$, получаем окончательно следующие расчетные формулы для внутренних точек сетки:

$$v_{ij}^{n+1} = \frac{h_2^2}{2(h_1^2 + h_2^2)}\big(v_{i-1\,j}^n + v_{i+1\,j}^n\big) + \frac{h_1^2}{2(h_1^2 + h_2^2)}\big(v_{i\,j-1}^n + v_{i\,j+1}^n\big) + \frac{h_1^2 h_2^2}{2(h_1^2 + h_2^2)} f_{ij}. \tag{9.21}$$

В том случае, когда $h_1 = h_2 = h$ формула (9.21) упрощается:

$$v_{ij}^{n+1} = \frac{1}{4}\big(v_{i-1\,j}^{n} + v_{i+1\,j}^{n} + v_{i\,j-1}^{n} + v_{i\,j+1}^{n}\big) + \frac{h^2}{4} f_{ij}. \tag{9.22}$$

Таким образом, получается следующий алгоритм нахождения последовательных приближений:

- вычисляются значения v_{ij} в точках границы γ, которые остаются неизменными на всех итерациях;
- произвольным образом задаются значения v_{ij}^0 в точках из ω;
- по формулам (9.21) или (9.22) в ω вычисляются $v^1, v^2, \ldots$

Расчеты проводятся до тех пор, пока приближения не перестанут меняться в пределах заданной точности. При выборе точности следует руководствоваться следующими соображениями. v_{ij} отличается от $u(x_i, y_j)$ на величину порядка $O(h_1^2 + h_2^2)$, так как разностная схема устойчива и имеет второй порядок аппроксимации. В связи с этим нет смысла искать решение (9.18), (9.19) с большей точностью. Критерием остановки итерационного процесса может служить условие

$$\frac{\|v^{n+1} - v^n\|}{\|v^n\|} < \varepsilon, \tag{9.23}$$

где $\varepsilon = ves\big(\frac{h_1+h_2}{2}\big)^2$, а коэффициент ves подбирается, например, $ves = 0.1$ или $ves = 0.01$.

Второй метод решения системы алгебраических уравнений (9.18), (9.19) — **метод переменных направлений или метод продольно поперечной прогонки**. В этом методе итерационная схема записывается в виде

$$\begin{aligned} \frac{v_{ij}^{n+1/2} - v_{ij}^{n}}{0.5\tau} &= \Lambda_1 v_{ij}^{n+1/2} + \Lambda_2 v_{ij}^{n} + f_{ij}, \quad v_{ij}^{n+1/2}|_\gamma = g, \\ \frac{v_{ij}^{n+1} - v_{ij}^{n+1/2}}{0.5\tau} &= \Lambda_1 v_{ij}^{n+1/2} + \Lambda_2 v_{ij}^{n+1} + f_{ij}, \quad v_{ij}^{n+1}|_\gamma = g. \end{aligned} \tag{9.24}$$

Здесь n — номер итерации, $n + 1/2$ — номер промежуточной итерации, $\tau > 0$ — итерационный параметр. Подставляя в (9.24) выражения для Λ_1, Λ_2 получаем, что переход от n-ой к $(n+1)$-ой итерации получается путем применения метода прогонки вдоль строк и столбцов для трехточечных уравнений:

$$-\frac{\tau}{2h_1^2} v_{i-1\,j}^{n+1/2} + \Big(1 + \frac{\tau}{h_1^2}\Big) v_{i\,j}^{n+1/2} - \frac{\tau}{2h_1^2} v_{i+1\,j}^{n+1/2} = \frac{\tau}{2} f_{ij} + \frac{\tau}{2h_2^2} v_{i\,j-1}^{n} + \Big(1 - \frac{\tau}{h_2^2}\Big) v_{i\,j}^{n} + \frac{\tau}{2h_2^2} v_{i\,j+1}^{n}, \tag{9.25}$$

$$-\frac{\tau}{2h_2^2} v_{i\,j-1}^{n+1} + \Big(1 + \frac{\tau}{h_2^2}\Big) v_{i\,j}^{n+1} - \frac{\tau}{2h_2^2} v_{i\,j+1}^{n+1} = \frac{\tau}{2} f_{ij} + \frac{\tau}{2h_1^2} v_{i-1\,j}^{n+1/2} + \Big(1 - \frac{\tau}{h_1^2}\Big) v_{i\,j}^{n+1/2} + \frac{\tau}{2h_1^2} v_{i+1\,j}^{n+1/2}. \tag{9.26}$$

Сначала находится $v_{ij}^{n+1/2}$, для чего выбирается $j = 1$ и прогонкой по i решается система (9.25). Затем полагают $j = 2$ и снова применяют прогонку и так далее до $j = J - 1$ (прогонка вдоль строк). Затем из (9.26) находят v_{ij}^{n+1}, придавая i последовательно значения $1, 2, \ldots, I - 1$ и применяя прогонку по переменной j (прогонка вдоль столбцов).

Таблица 9.7.2

№	a	b	$g_1(y)$	$g_2(y)$	$g_3(x)$	$g_4(x)$	$f(x,y)$	Решение $u(x,y)$
1 2 3 4 5	1	$\frac{\pi}{2}$	$\sin py$	$e^p \sin py$	0	$e^{px} \sin \frac{p\pi}{2}$	0	$e^{px} \sin py$
6 7 8 9 10	2	1	0	$\frac{py}{4} + 4\cos y$	$2x$	$\frac{px}{8} + 2x \cos 1$	$2x \cos y$	$\frac{pxy}{8} + 2x \cos y$
11 12 13 14 15	1	1	$-\sqrt{p}y^2$	$(\sqrt{p}+y)(1-y^2)$	$\sqrt{p}x^2$	$(\sqrt{p}+x)(x^2-1)$	0	$(\sqrt{p}+xy)(x^2-y^2)$
16 17 18 19 20	1	1	$-2y-4y^2$	$4-12y-4y^2$	$x+3x^2$	$3x^2-9x-6$	$2+$ $+2\pi^2(p-15)*$ $*\sin \pi x \sin \pi y$	$x+3x^2-10xy-$ $-2y-4y^2+$ $+(p-15)\sin \pi x \sin \pi y$
21 22 23 24 25	1	1	e^{-y^2}	$\frac{y}{p} + e^{1-y^2}$	e^{x^2}	$\frac{x}{p} + e^{x^2-1}$	$-4x^2e^{x^2-y^2}+$ $-4y^2e^{x^2-y^2}$	$\frac{xy}{p} + e^{x^2-y^2}$

Оптимальное значение τ вычисляется по формуле

$$\tau = \frac{ab}{\pi}\frac{\sqrt{h_1^2 + h_2^2}}{\sqrt{a^2 + b^2}}.$$

Задание к лабораторной работе

- Найти решение задачи Дирихле для уравнения Пуассона, применяя методы простой итерации и продольно поперечной прогонки.
- Провести вычисления для различных чисел I, J. Сравнить число итераций для различных методов и различного числа разбиений.
- Оценить число арифметических операций, которые требуется совершить для получения решения для каждого из методов при условии, что значения функции f и граничных функций уже найдены.
- Найти норму разности между точным и приближенным решением при различных значениях шагов сетки.

Варианты задач к лабораторной работе представлены в таблице 9.7.2. В ней параметр p равен номеру варианта.

ОТВЕТЫ К ТЕСТАМ

	Номера вариантов					
Номера тестов	1	2	3	4	5	6
Тест к главе 1	6	0.0625	0.004	б в г д	*a*	а
Тест к главе 2	б в	в	б	б	в	в
Тест к главе 3	г	3	1	0.17	2.25	1 -2 2
Тест к главе 4	1 4 2 3	а	6	0	0.4	f
Тест к главе 5	б	в	19	д	а	г
Тест к главе 6	а	е	в	в	А234 РК15	г
Тест к главе 7	г	С12 П456	в	д	д	в д е
Тест к главе 8	в д	б в д	д	е	а б	б

Литература

1. *Амосов, А.А. Вычислительные методы для инженеров. Учебное пособие / А.А. Амосов, Ю.А. Дубинский, Н.В. Копченова. — М.: Высшая школа, 1994. — 544 с.*

2. *Арушанян, О.Б. Численные методы решения обыкновенных дифференциальных уравнений (задача Коши) / О.Б. Арушанян, С.Ф. Залеткин. — М.: МГУ, (http://www.srcc.msu.su/num_anal/list_wrk/sb3_doc/part6.htm)*

3. *Арушанян, О.Б. Практикум на ЭВМ по вычислительным методам. Численное решение нелинейных уравнений / О.Б. Арушанян. — М.: МГУ, 2003. — 37 с. (http://www.srcc.msu.su/num_anal/meth_mat/page_2.htm)*

4. *Арушанян, О.Б. Практикум на ЭВМ по вычислительным методам. Решение задачи Коши для обыкновенных дифференциальных уравнений одношаговыми методами / О.Б. Арушанян, С.Ф. Залеткин. — М.: МГУ, 2002. — 51 с. (http://www.srcc.msu.su/num_anal/meth_mat/page_2.htm)*

5. *Бахвалов, Н.С. Численные методы: учеб. пособие / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. — М.: Бином, 2008. — 636 с.*

6. *Бахвалов, Н.С. Численные методы в задачах и упражнениях: учеб. пособие / Н.С. Бахвалов, А.В. Лапин, Е.В. Чижонков.— М.: Высшая школа, 2000.— 190 с.*

7. *Бахвалов, Н.С. Численные методы. Решение задач и упражнения: учеб. пособие / Н.С. Бахвалов, А.А. Корнеев, Е.В. Чижонков. — М.: Дрофа., 2009. — 393 с.*

8. *Березин, И.С. Методы вычислений: т1. / И.С.Березин, Н.П. Жидков. — М.: Наука, 1966. — 632 с.*

9. *Вержбицкий, В.М. Основы численных методов / В.М. Вержбицкий. — М.: Высшая школа, 2002. — 848 с.*

10. *Воеводин, В.В. Матрицы и вычисления / В.В. Воеводин, Ю.А. Кузнецов. — М.: Наука, 1984. — 320 с.*

11. *Волков, Е.А. Численные методы: учеб. пособие / Е.А. Волков. — СПб.: Лань, 2008. — 256 с.*

12. *Годунов, С.К. Разностные схемы / С.К. Годунов, В.С. Рябенький. — М.: Наука, 1973. — 400 с.*

13. *Двайт, Г.Б. Таблицы интегралов и другие математические формулы / Г.Б. Двайт. — М.: Наука, 1973. — 228 с.*

14. *Демидович, Б.П. Основы вычислительной математики / Б.П. Демидович, И.А. Марон. — СПб.: Лань, 2009. — 672 с.*

15. *Деккер, К. Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений / К. Деккер, Я. Вервер. — М.: Мир, 1988. — 332 с.*

16. *Дробышевич, В.И. Задачи по вычислительной математике / В.И. Дробышевич, В.П. Дымников, Г.С. Ривин. — М.: Наука, 1980. — 144 с.*

17. *Дьяченко, В.Ф. Основные понятия вычислительной математики / В.Ф. Дьяченко. — М.: Наука, 1977. — 256 с.*

18. *Калиткин, Н.Н. Численные методы / Н.Н. Калиткин. — М.: Наука, 1978. — 512 с.*

19. *Каханер, Д. Численные методы и программное обеспечение / Д. Каханер, К. Моулер, С. Нэш. — М.: Мир, 1998. — 575 с.*

20. *Кантор, С.А. Специальные главы высшей математики: учебное пособие / С.А. Кантор. — Барнаул: АлтГТУ им. И.И. Ползунова, 2003. — 183 с.*

21. *Коллатц, Л. Задачи по прикладной математике / Л. Коллатц, Ю. Альбрехт. — М.: Мир, 1978. — 168 с.*

22. *Копченова, Н.В. Вычислительная математика в примерах и задачах Н.В. Копченова, И.А. Марон. — СПб.: Лань, 2009. — 357 с.*

23. *Крылов, В.И. Вычислительные методы: т1 / В.И. Крылов, В.В. Бобков, П.И. Монастырский. — М.: Наука, 1976. — 303 с.*

24. *Марчук, Г.И. Методы вычислительной математики / Г.И. Марчук. — СПб: Лань, 2009. — 608 с.*

25. *Михайлов, А.П. Задания вычислительного практикума на ЭВМ: метод. пособие / А.П. Михайлов. — Новосибирск.: НГУ, 1998. — 59 с.*

26. *На, Ц. Вычислительные методы решения прикладных граничных задач / Ц. На. — М.: Мир, 1982. — 296 с.*

27. *Ортега, Дж. Введение в численные методы решения дифференциальных уравнений / Дж. Ортега, У. Пул. — М.: Наука, 1986. — 288 с.*

28. *Рихтмайер, Р. Разностные методы решения краевых задач / Р. Рихтмайер, К. Мортон.— М.: Мир, 1972. — 418 с.*

29. *Рябенький, В.С. Введение в вычислительную математику: учеб. пособие / В.С. Рябенький. — М.: ФИЗМАТЛИТ, 2008. — 288 с.*

30. *Самарский, А.А. Введение в численные методы / А.А. Самарский. — СПб.: Лань, 2009. — 272 с.*

31. *Самарский, А.А. Теория разностных схем / А.А. Самарский. — М.: Наука, 1977. — 656 с.*

32. *Самарский, А.А. Численные методы / А.А. Самарский, А.В. Гулин. — М.: Наука, 1989. — 432 с.*

33. *Самарский, А.А. Методы решения сеточных уравнений / А.А. Самарский, Е.С. Николаев. — М.: Наука, 1978. — 591 с.*

34. *Соболь И.М. Численные методы Монте-Карло / И.М. Соболь. — М.: Наука, 1973. — 311 с.*

35. *Турчак, Л.И. Основы численных методов / Л.И. Турчак. — М.: Наука, 1987. — 320 с.*

36. *Фаддеев Д.К. Вычислительные методы линейной алгебры / Д.К. Фаддеев, В.Н. Фаддеева. — СПб.: Лань, 2009. — 736 с.*

37. *Хайрер, Э. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи / Э. Хайрер, С. Нерсетт, Г. Ваннер. — М.: Мир, 1990. — 512 с.*

38. *Холл, Дж. Современные численные методы решения обыкновенных дифференциальных уравнений / Дж. Холл, Дж. Уатт. — М.: Мир, 1979. — 312 с.*

Предметный указатель

Оглавление

Подписано в печать - 03.11.2010.
Печать – цифровая. Усл.п.л. 41,50.
Тираж 50 экз. Заказ 2010 - 566

Отпечатано в типографии АлтГТУ,
656038, г. Барнаул, пр-т Ленина, 46
тел.: (8–3852) 36–84–61

Лицензия на полиграфическую деятельность
ПЛД №28–35 от 15.07.97 г.